Imperial College London

Imperial College London
Department of Electrical and Electronic Engineering

Balancing Privacy and Data Access:

An Interdisciplinary Approach to Markets for Differentially-Private Smart Meter Data

---

Saurab Chhachhi

July 24, 2024

Supervised by Dr. Fei Teng

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Electrical and Electronic Engineering of Imperial College London and the Diploma of Imperial College London

Abstract

Access to high-resolution smart meter data has many operational benefits for energy suppliers, network operators, and the energy system as a whole. However, access also raises privacy concerns, hindering the adoption of smart meters and the sharing of high-resolution smart meter data. This thesis addresses this dilemma by employing an interdisciplinary approach to designing a privacy-preserving data market mechanism for smart meter data as a means of balancing privacy and data access.

The first research direction determines the design criteria of a data market for smart meter data. Specifically, we map the data dependence of benefits as well as the potential privacy infringements and risks associated with smart meter data. Data resolution, both spatial and temporal play a significant role in determining both benefits and privacy risks. We investigate consumers' privacy concerns and their willingness-to-pay/accept for anonymisation through a novel survey and discrete choice experiment. Significant heterogeneity and endowment effects are observed with information asymmetries leading to depressed valuations for privacy protection. Finally, we assess the suitability of different privacy-preserving techniques for smart meter data, finding differential privacy to be a flexible, transparent, and easily integrated mechanism for ensuring privacy while allowing access to data.

The second research direction develops a novel data market framework, using the design criteria determined in the first. A novel data valuation mechanism is developed based on the Wasserstein distance, which embodies the drivers of smart meter data value, including the privacy-utility trade-off induced by differential privacy. This is integrated into a novel procurement mechanism, developed using incentive mechanism design theory, which can model data buyers' and consumers' preferences, while preserving privacy. A joint energy and market is developed, which through case studies is shown to be a viable proposition to balance privacy and access to smart meter data, given our estimations of consumers' willingness-to-accept.

Statement of Originality & Copyright Statement

I hereby declare that the material contained within this thesis is my own work, except where other work is appropriately referenced.

Saurab Chhachhi, July 24, 2024
Department of Electrical and Electronic Engineering
Imperial College London

Acknowledgements

First, I would like to thank, my supervisor, Dr. Fei Teng. His guidance and the generous time he has given me, have been invaluable. His insights pushed me to explore new aspects and angles, and his encouragement and belief gave me the motivation to continue.

I am thankful for my funding from the ESRC through the London Interdisciplinary Social Science Doctoral Training Partnership studentship (ES/P000703/1:2113082) and UKRI Research England's Strategic Priorities Fund(2021-2022).

I would like to thank members of the Control and Power research group at Imperial College London for the discussions and feedback provided throughout my study. Specifically, Pudong Ge, Alicia Blatiak, and Wangkun Xu.

In addition, I would like to thank Dr Chien Fei-Chen at the University of Tennessee, Dr Aruna Sivakumar at Imperial College London, Dr Paul J. Metcalfe of PJM Economics, Debbie Shuttlewood of Accent Market Research, and Nicole Watson and Anna Gorbatcheva at the UCL Energy Institute for their feedback on the design and analysis of the discrete choice experiment and survey.

I would like to thank my examiners Dr Phil Grunewald and Prof. Pierre Pinson for review this thesis and the stimulating viva. Your time, comments and suggestions are very much appreciated.

A special thanks goes to my friends, especially, Kishan, Aashish, John, Oualid, Dharmanshu and Bennet for putting up with me during this time and at various points listening to me talk about obscure concepts and ideas, and for giving up their time to proofread my dissertation.

I would like to thank my family for their continued support, my father, my aunt, and sisters. A special mention to Laila, my niece, the human Pomodoro timer.

I thank Georgia, for supporting me through, what can only be described as a journey, for both of us. You brought a much needed light into my life, and for that I am grateful.

Finally, I would like to thank my mother for everything, from discussing and providing

feedback on my research from the first day to the last, to supporting and encouraging me to continue through the hardest moments. This would simply not have been possible without you.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANN Artificial Neural Networks.

BEIS Department for Business, Energy & Industrial Strategy.

BF Budget Feasibility.

CAD Consumer Access Device.

CDF Cumulative Distribution Function.

CSP Communication Service Provider.

CVaR Conditional Value-at-Risk.

DAPF Data Access and Protection Framework.

DCC Data Communications Company.

DCE Discrete Choice Experiment.

DDP Discounted Differential Privacy.

DLC Daily Load Coefficient.

DNO Distribution Network Operator.

DP Differential Privacy.

DRO Distributionally-Robust Optimisation.

EMD Earth Mover's Distance.

EV Electric Vehicles.

FL Federated Learning.

GB Great Britain.

GDPR General Data Protection Regulation.

GSP Grid Supply Point.

HAN Home Area Network.

HH Half-Hourly.

HHS Half-Hourly Settlement.

I.I.D. Independent and Identically Distributed.

IC Incentive Compatibility.

ICO Information Commissioner's Office.

IHD In-Home Display.

IIA Independence from Irrelevant Alternatives.

IR Individual Rationality.

JSD Jensen-Shannon Divergence.

KLD Kullback-Leibler Divergence.

KS Kolmogorov-Smirnov Metric.

LASSO Least Absolute Shrinkage and Selection Operator.

LSE Load-Serving Entity.

MAE Mean Absolute Error.

MHHS Market-Wide Half-Hourly Settlement.

MILP Mixed-Integer Linear Program.

MIQP Mixed-Integer Quadratic Program.

MISOCP Mixed Integer Second Order Conic Program.

MNL Multinomial Logit Model.

MPC Multi-Party Computation.

MPL Mean Pinball Loss.

MSE Mean Squared Error.

MXL Mixed Logit Model.

NHHS Non-Half-Hourly Settlement.

NILM Non-Intrusive Load Monitoring.

OFGEM Office of Gas and Electricity Markets.

ONS Office for National Statistics.

P2P Peer-to-Peer.

PDF Probability Density Function.

PPT Privacy-Preserving Technique.

PSI-CA Private Set Intersection-Cardinality.

PV Photovoltaics.

RCT Randomised Control Trial.

RMSE Root Mean Squared Error.

SEG Socio-Economic Group.

SMETS Smart Metering Equipment Technical Specifications.

SMIP Smart Meter Implementation Program.

TTP Trusted-Third Party.

TVD Total Variation Distance.

UK United Kingdom.

WAN Wide Area Network.

WAPE Weighted Absolute Percentage Error.

WD Wasserstein Distance.

WTA Willingness-to-Accept.

WTP Willingness-to-Pay.

WTP/A Willingness-to-Pay/Accept.

WTS Willingness-to-Share.

CHAPTER 1

Introduction

1.1 Motivation and Background

Great Britain (GB) is currently upgrading its national electricity and gas metering infrastructure by introducing digital smart meters, through the Smart Meter Implementation Program (SMIP). This is seen as a key component of the digitalisation of the energy sector and facilitates the transition to a more dynamic, cost-effective, cost-reflective, and decarbonised electricity network. Smart meters allow for logging and accessing high resolution consumption data, enable the integration of smart devices and automated load control for the domestic sector, and open up possibilities for innovative business models and pricing schemes. The Department for Business, Energy & Industrial Strategy (BEIS) has stated that digitalisation of the energy sector is a vital component of the strategy to achieving net zero [1]. In order for many of the expected benefits to materialise high levels of adoption are required [2]. However, the programme has been significantly delayed, with only 58% of homes and small businesses having installed smart meters, as of June 2023 [3].

One major hurdle is the concern raised around privacy and data misuse[4], as smart meters can collect personal information on individuals at high resolution[5], [6]. Concurrent to the introduction of domestic smart metering, awareness of issues surrounding data privacy and misuse is growing. The introduction of the General Data Protection Regulation (GDPR) is forcing companies and regulators to rethink the way in which they engage with customers and how data is handled [7]. Infringements of GDPR can result in large fines (€20 million, or 4% of the firm's worldwide annual revenue) and in the three years since the introduction of GDPR over £250 million have been issued in fines [8]. The reasons for the fines have ranged from lack of valid consent and transparency to inadequate security mechanisms and data misuse. Increasingly high profile data breaches [9]

and exposés of how, seemingly harmless, data can be used to influence individuals' daily lives are putting these issues into the spotlight [10]. However, the energy sector remains an area of high apathy among consumers[4]. The potential privacy risks associated with sharing smart meter data are complex and evolving over time, as the availability of such data increases and an understanding of the personal information embedded within it grows [7]. Communicating these risks and obtaining informed consent for data sharing, a cornerstone of GDPR, is the key requirement and challenge.

The SMIP has taken the privacy issues around smart meter data into consideration, with the development of the Data Access and Protection Framework (DAPF)[11] and multiple privacy impact assessments [12], as part of a Privacy by Design approach. Privacy by Design places emphasis on protecting individuals' privacy and data security, through strong privacy defaults, a user-centric approach and the use of Privacy-Preserving Techniques (PPT). However, the latest changes to data sharing regulations, as part of the shift to Market-Wide Half-Hourly Settlement (MHHS), move to an opt-out, rather than opt-in data sharing model w.r.t. energy suppliers/retailers[1] and has ruled-out the use of PPTs[13]. In addition, the DAPF is built around consumers consenting to data sharing, with no regulations on communicating the privacy risks involved. PPTs provide a method of ensuring consumers are protected from potential privacy infringements, especially in the presence of complex privacy risks, while still providing access to high-resolution data and the associated benefits of data sharing. At the same time, implementing PPTs incurs costs and thus introduces a trade-off between the level of privacy protection offered and the value of data[14].

Additionally, questions around the value of data and how benefits of data sharing are distributed are increasingly being discussed[14]. The existing data sharing framework, encoded in the DAPF, provides consumers with options for how (e.g. monthly, daily or Half-Hourly (HH) data), with whom (e.g. suppliers or third parties), and for what purpose to share their smart meter data[11]. However, it does not provide a means for consumers to monetise their data or ensure that the resulting benefits are shared. The increasing use of data, coupled with a raised awareness amongst consumers as to the value of their data has led to an increased interest in the concept of data markets[15]. Indeed, the energy sector is no exception with a growing body of work on data sharing

[1]Given the interdisciplinary nature of this thesis, we use the terms supplier, retailer and load serving entity in different chapters/sections. Specifically, supplier is used in the context of discussion of the SMIP, retailer is used to maintain consistency with power systems literature, and load serving entity is more general term used to refer entities which include a retailer.

and market frameworks for a range of energy data, including smart meter data[16].

Concurrently, we see a push towards widening access to electricity data and the development of Open Data platforms[17], [18]. Wider access would allow the full potential and value of smart meter data to be realised, but will also result in an increasing number of entities, who may or may not have direct contractual relationships with consumers, accessing the data. Providing access to these entities increases the risks of privacy infringements as well as the potential for data misuse. Given that many of the benefits of smart metering are contingent on widespread adoption, and the uses and privacy vulnerabilities of smart meter data are evolving, PPTs can play an important role in future-proofing the privacy-preservation of smart meter data. However, a lack of understanding as to the costs and benefits of such mechanisms has meant that these have not been adopted as part of the SMIP/DAPF thus far[13]. As such, a privacy-preserving market mechanism which resolves the induced privacy-utility trade-off could provide a means to balancing of privacy and access to smart meter data. The appropriate structure and properties for a market for smart meter data depend on; the regulatory environment (e.g. the choices available to data buyers/users and data sellers/owners), the data sharing technologies employed (e.g. privacy utility trade-off, distributed or centralised data processing and market clearing), and the definition of data value (e.g. quality or quantity-based, task-specific, or a combination).

1.2 Research Directions

Overall, we identify four key challenges in developing a privacy-preserving market mechanism for smart meter data:

1. The ability to obtain truly informed consent is hindered by the complexity of potential privacy risks and the wide range of potential uses for smart meter data.

2. Given the variety of PPTs, identifying the most suitable PPT for smart meter data and its associated costs and benefits.

3. The valuations of smart meter data need to be assessed, both from the perspective of data users (e.g. retailers) and data owners (e.g. consumers).

4. A market design approach with the desired properties to be identified and tested for its viability.

This thesis employs an interdisciplinary approach consisting of two distinct, but interdependent research directions to address the challenges identified above. The first focuses on determining the design criteria of a data market for smart meter data. This involves: developing an understanding of the drivers of value for smart meter data and the potential privacy risks associated with sharing smart meter data, investigating consumers' privacy concerns and valuations of smart meter data, and establishing the suitability of different PPTs for smart meter data. The second research direction focuses on developing a novel data market framework, based on the design criteria determined in the first. This consists of a data valuation mechanism, which embodies the drivers of smart meter data value, including the effect of PPTs, and a procurement mechanism which can model data buyers' and consumers' (data owners) preferences, while adhering to Privacy by Design principles.

1.2.1 Determining the Design Criteria

Data Dependence of Benefits and Privacy Risks

Smart metering provides a range of potential benefits for consumers, energy suppliers and network operators. These include operational benefits for suppliers, through the automation of meter readings and customer switching, consumer benefits, in terms of time savings and bill savings from energy usage reductions, and improved system management through greater network visibility and demand flexibility. These benefits, and therefore the drivers of value of smart meter data, have been extensively explored, quantified and researched in the context of the SMIP[19], and more broadly in the academic literature (e.g. [20]). However, a detailed understanding of the dependence of these benefits on smart meter data, including the specific form of the data (e.g. the frequency, level of aggregation) is lacking. Such a mapping is necessary, in order to accurately assess whether data access is proportionate, and to adequately inform consumers as to how and for what purpose their data may be used.

Similarly, the types and severity of potential privacy infringements depend on how, in what form, and with whom smart meter data is shared. Smart meter data has a vast amount of personal information embedded within it. This includes energy usage related information (e.g.the use of certain appliances[21]), derived information(e.g. daily routines and home occupancy[6], [22]), as well as socio-demographic information (e.g. age and income[5]). The ability to extract such information and the accuracy with which it can be done depends on the technical tools used, but notably, also on the form of the smart

meter data[23]. Again, in order to appropriately inform consumers as to the implications of sharing their smart meter data a comprehensive assessment of the dependencies is required.

Consumer Privacy Concerns

Privacy is not necessarily achieved by hiding personal data but rather by giving consumers control over their data[14]. The consent-based DAPF does this, however, it does not provide sufficient information as to the implications of sharing, to allow for truly informed consent. Existing literature, mainly using surveys and focus groups, on understanding consumers' awareness, perceptions and concerns around smart metering suggest that consumers are generally willing to share their smart meter data[24]. However, they do not account for the information asymmetries that may exist and resulting bounded rationality in decision making[25]. A deeper understanding of consumers' privacy preferences is required, with explicit consideration of the complexity of privacy risks associated with smart meter data.

A key component of any market-mechanism, to deal with the privacy-utility trade-off, is consumers' valuations or Willingness-to-Pay/Accept (WTP/A). Existing literature on valuing data privacy show a clear distinction between smart meter data and other forms of personal data[26], [27], again highlighting the information asymmetries that exist in relation to the privacy implications of sharing smart meter data. These surveys focus on the WTP, framing the choice as a payment consumers would have to make to maintain data privacy, as opposed to the WTA which presents the choice as compensation for sharing data.The endowment effect, where the WTA is larger than WTP, is a well-studied phenomenon, generally assumed to reflect a sense of ownership. It appears to be particularly pronounced in relation to privacy, potentially suggesting a sense of moral outrage at the framing[28]. From a regulatory perspective, this has significant implications. Indeed, under a Privacy by Design approach, establishing strong privacy defaults in the form of opt-ins rather than opt-outs would make the WTA the more appropriate valuation.

Suitability of Privacy-Preserving Techniques

A multitude of PPTs have been proposed for smart meter data, ranging from general techniques, such as, Differential Privacy (DP)[29] to domain specific methods such as user demand shaping[30]. PPTs vary in their definitions of privacy (e.g. anonymity,

limiting inference), which may be continuous (e.g. DP) or discrete (e.g. homomorphic encryption), and hence, so do the privacy infringements (e.g. membership inference, non-intrusive load monitoring, linking attacks or data breaches) they protect against. They also differ in terms of architecture, with techniques which require a trusted centralised entity and decentralised techniques which do not require a Trusted-Third Party (TTP).

The form of and extent to which data utility is degraded depends on whether the technique restricts data access (e.g. aggregation) or obfuscation (e.g. noise addition) and if it allows for heterogeneous consumer preferences or provides all consumers with the same protections. This is a well-studied area with many comprehensive reviews covering the technical considerations of applying PPTs in the context of smart metering (e.g. for aggregation or billing)[16], [31]–[36]. However, an assessment of how these techniques could be integrated into the existing SMIP, as well as facilitating a Privacy by Design approach, and accommodating consumer preferences remains an open question.

1.2.2 Developing a Privacy-Preserving Data Market

A data market can broadly be defined by a valuation mechanism and procurement mechanism. A valuation metric may be as simple as the data quantity or a more complex mechanism such as the value of a loss function for a specific task (e.g. mean squared error of a forecast). A procurement/pricing mechanism translates the valuation into a procurement decision, whether to buy the data or not, and how much to pay for it. Aside from ensuring privacy protection of procured data, it is also desirable to ensure privacy during the procurement process, motivating the need for a privacy-preserving valuation metric and procurement mechanism.

Data Valuation Metrics

Existing approaches to data valuation in the energy domain focus on task specific value, usually the improvement in forecast error [37]–[41]. However, given that data, in general, is an infinitely durable commodity and that smart meter data has many applications, a more generic notion of value, applicable across multiple potential uses may be more appropriate. In addition, calculating task-specific value requires data sharing, model sharing or both, with the buyer or a TTP (e.g. market platform). While this can be overcome by adapting the framework to provide performance information in a privacy-preserving or distributed manner, the results still correspond only to the output space and are therefore task specific. As a result, such techniques are also vulnerable to manipulation

through model mis-specification.

An alternative is to shift to the input space, and consider the dataset under valuation to be a distribution[42]. Consequently, a different, task-agnostic, notion of value can be developed. Statistical distances such as, the Jensen-Shannon Divergence (JSD) or the Wasserstein Distance (WD), provide a measure of the difference between the dataset and a reference or target distribution[43]–[49]. As such, the distance provides an indication of how close a given dataset is to the desired data, with the assumption being that statistically similar data will lead to similar model performance. By pursuing a task-agnostic metric there is an inevitable loss of performance guarantees and valuation accuracy for a specific task in the output space. In both cases, task-specific and task-agnostic, a method to incorporate the effect of PPTs on value as well as a privacy-preserving calculation technique is required.

Data Procurement Mechanism

The cooperative game formulation provides the current state-of-the-art for data procurement mechanisms, in the energy domain. Examples include general markets for regression features [38], historical wind production data [37] and consumer demand forecasts[39]. Payments are determined using the Shapley value, the average marginal contribution to a valuation metric, which is a computationally intensive process, growing exponentially with the number of data sellers. This allows the platform to capture the decision-dependent structure inherent to data procurement problems. Specifically, it implies that the budget available to pay for data is dependent on the value created (e.g. by reducing uncertainty) of the data. The mechanism assumes the buyer or a TTP has access to the data and is able to calculate every combination of data sources. Again, distributed calculation is possible, however, this assumes sufficient computational power at the nodes and also requires sharing of the model. In addition, the cooperative game framework assumes data sellers do not have reserve prices, that there are no barriers to their participation in the market, i.e. they do not have any privacy concerns or otherwise have their own valuations of their data.

An alternative approach is incentive mechanism design, which to date has mainly been applied to data procurement for Federated Learning (FL)[43], [48], [50]–[52]. Here, it is assumed that sellers have reserve prices and some value, with the market platform determining which data to buy and how much to pay each seller. Although, reserve prices are private, the value of each data sellers' data is assumed to be accessible to

the platform, and therefore, like the cooperative game, must be calculated in a privacy-preserving manner. The computational complexity of the mechanism depends on the objective of the platform. The platform either optimises value or a combination of value and payments. For example, if value is additive there is an efficient polynomial time approximation scheme[43]. However, an additive model of value may not adequately capture the decision-dependent structure of the problem.

1.3 Original Contributions

The main contributions of this thesis are:

- A critical assessment of the existing SMIP against Privacy by Design principles, and a mapping the data dependence of the benefits as well as the privacy risks of sharing smart meter data, providing a basis for obtaining informed consent.
- A comprehensive review of the suitability of different privacy-preserving techniques for smart meter data, with specific focus on the requirements of consumers and the existing SMIP infrastructure, adding to the evidence base for policymakers.
- A novel survey and discrete choice experiment of a sample of nationally-representative bill payers (n=686) quantifying their WTA and WTP for smart meter data anonymisation and investigating the effect of information asymmetries by means of a Randomised Control Trial (RCT). This broadens empirical evidence on data valuations, substantiates and quantifies behavioural economics effects, specifically, the endowment effect and bounded rationality, in the context of smart meter data privacy valuations.
- A novel data valuation mechanism and framework based on the WD. This includes a method to endogenously incorporate the effect of DP on data utility, a theoretically grounded approach to ensuring performance guarantees in the output space using Lipschitz bounds and a linear-time approximation scheme for calculating the combinatorial value using the Hoeffding bound.
- A novel set of procurement mechanisms based on incentive mechanism design theory to accommodate task-agnostic budget feasible procurement, task-specific value maximising procurement and task-specific profit maximising procurement. The latter two are able to capture decision-dependent structure. The task-specific

profit maximising procurement mechanism is used to develop a joint energy and data market.

We note that parts of Chapter 2 and 4, specifically, Sections 2.4, 4.2.1, 4.2.4, and 4.2.7 were written by me, based on a joint review of literature with Pudong Ge and Jemima Graham.

1.4 Thesis Structure

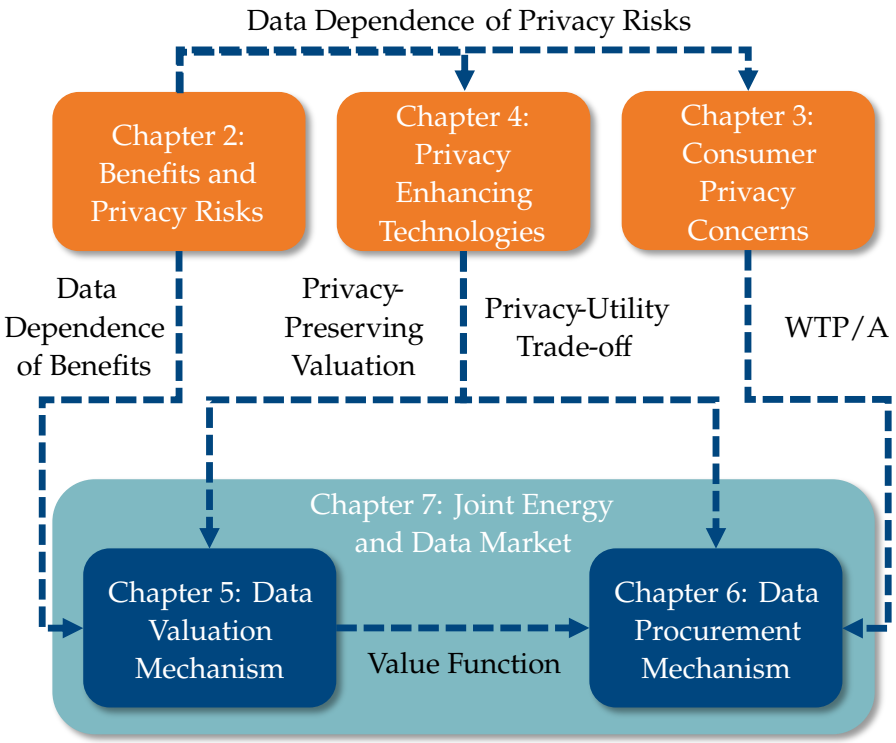


Figure 1.1: Outline of Thesis

An outline of the thesis and the connections across chapters is shown in Figure 1.1. Chapter 2 provides an assessment of the United Kingdom (UK)'s current smart meter data sharing regulations guided by the Privacy by Design framework. It lays out the potential benefits of wider access to smart meter data, as well as, the associated privacy infringements and risk, with a focus on their dependence on data.

Chapter 3 explores consumers' understanding of smart meter data and the associated privacy implications, their valuations of privacy, and the possibility for informed consent. It presents the results of the survey and discrete choice experiment on a representative sample of GB bill payers to quantify WTP/A for anonymising smart meter data.

Chapter 4 considers the use of privacy-preserving technologies as a means to balance

privacy and access to smart meter data. A review of existing techniques is conducted focusing on their suitability given the specific needs of the SMIP, the privacy-utility trade-offs induced, and consumers' concerns identified in Chapter 2 and 3.

Chapters 5 to 7 presents the novel data valuation and procurement framework. Chapter 5 translates the key findings of Chapters 2 to 4 into requirements for the framework and presents the Wasserstein distance as a data valuation metric. Chapter 6 develops the data market or procurement mechanism, built upon incentive mechanism design theory. Chapter 7 then applies the valuation and procurement mechanisms to the energy procurement problem, a core use of smart meter data, to develop a joint energy and data market.

Finally, Chapter 8 summarises the results of the project, discusses the limitations of the work and proposes future research paths to address these.

1.5 List of Publications

This project has led to a number of publications and working papers:

[Paper A] S. Chhachhi and F. Teng, "Market value of differentially-private smart meter data," in *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, IEEE, Feb. 2021, pp. 1–5, ISBN: 978-1-7281-8897-3, DOI: 10.1109/ISGT49243.2021.9372228.

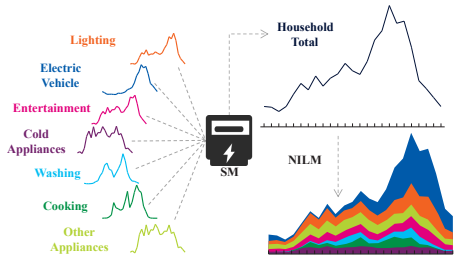
[Paper B] S. Chhachhi, F. Teng, P. Ge, *et al.*, "Balancing privacy and access to smart meter data: An Energy Futures Lab briefing paper," *Imperial College London*, pp. 1–64, May 2022, DOI: 10.25561/96974.

[Paper C] S. Chhachhi and F. Teng, "On the 1-Wasserstein distance between location-scale distributions and the effect of differential privacy," *arXiv preprint*, Apr. 2023, DOI: 10.48550/arXiv.2304.14869.

[Paper D] S. Chhachhi and F. Teng, "Balancing privacy and access to smart meter data: A willingness-to-pay/accept study of GB consumers," *Working Paper to be submitted to Nature Energy*, Dec. 2023.

[Paper E] S. Chhachhi and F. Teng, "Wasserstein distance based market for trading differentially-private data - Part I: Valuation and procurement mechanism," *Working Paper to be submitted to IEEE Transactions on Energy Markets, Policy and Regulation*, Dec. 2023.

[Paper F] S. Chhachhi and F. Teng, "Wasserstein distance based market for differentially-private data - Part II: A joint energy and smart meter data market," *Working Paper to be submitted to IEEE Transactions on Energy Markets, Policy and Regulation*, Dec. 2023.



CHAPTER 2
Smart Meter Data Privacy in the UK

The SMIP requires the roll-out of 53 million digital smart gas and electricity meters for the 30 million domestic and smaller non-domestic properties in GB [53]. Smart metering provides a range of potential benefits for consumers, energy suppliers and network operators[19]. However, the programme has been significantly delayed, with only 58% of homes and small businesses having installed a smart meter, as of June 2023 [3]. One major hurdle is the concern raised around privacy and data misuse[4], as smart meters can collect personal information on individuals at high resolution[5], [6].

In order to tackle the question of balancing access and privacy for smart meter data, we need to determine whether access is proportionate given the use case, and whether consumers are adequately informed as to how and for what purpose their data may be used, as well as the privacy risks associated with data sharing. As such, we need to develop an understanding of the dependence, of these benefits and potential privacy infringements, on the form of smart meter data. This chapter starts by outlining Privacy by Design, the framework used throughout this project to assess the adherence to privacy standards[54]. We then detail the core components of SMIP, as well as, how and with whom smart meter data can be shared. Finally, we present a detailed mapping of the dependence of the benefits of smart metering, as well as the privacy risk, on smart meter data. The majority of this chapter forms part of [Paper B].

2.1 Privacy by Design

Privacy by Design is a framework and certification approach that sets out seven key principles of a system to ensure that data breaches and privacy infringements are avoided to the extent possible and that in the event of such infringements the implications are limited. This framework for designing data collection and access systems places an emphasis on protecting individual privacy and data security thereby providing organisations a way to ensure compliance with privacy regulations such as GDPR.

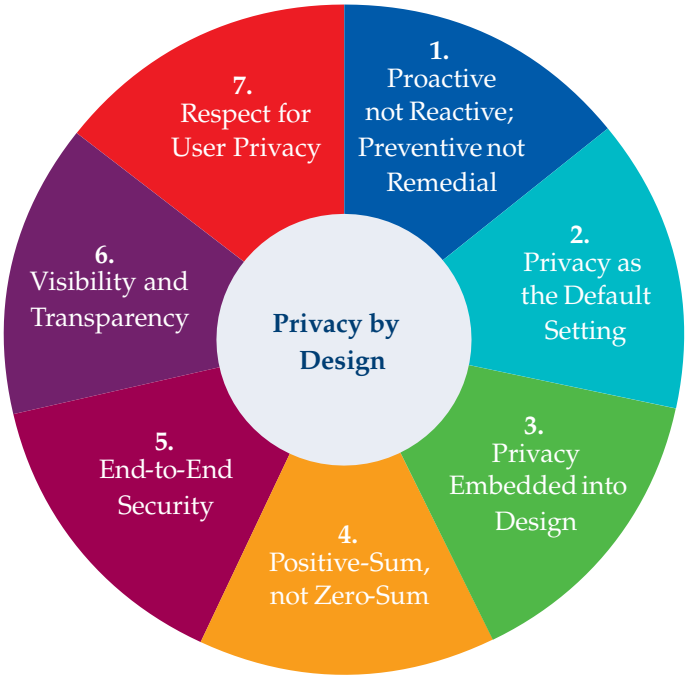


Figure 2.1: Foundational Principles of Privacy by Design. Reproduced from [55, p. 2].

The seven principles, shown in Figure 2.1, are [55]:

1. Anticipate, identify, and prevent invasive events before they happen; this means taking action before the fact, not afterwards.

2. Ensure personal data is automatically protected in all IT systems or business practices, with no added action required by any individual.

3. Privacy measures should not be add-ons, but fully integrated components of the system.

4. Employ a "win-win" approach to all legitimate system design goals; that is, privacy and security are important, and no unnecessary trade-offs need to be made to achieve both.

5. Data life-cycle security means all data should be securely retained as needed and destroyed when no longer needed.

6. Assure stakeholders that business practices and technologies are operating according to objectives and subject to independent verification.

7. Keep things user-centric; individual privacy interests must be supported by strong privacy defaults, appropriate notice, and user-friendly options.

The SMIP has taken the privacy issues around smart meter data into consideration, with the development of the DAPF[11] and multiple privacy impact assessments [12], as part of a Privacy by Design approach. However, the extent to which the principles are followed has been contested [56], with recent changes as part of a move to MHHS resulting in concerns being raised by the Information Commissioner's Office (ICO)[57]. This section provides an overview of the data sharing framework under the SMIP and the potential benefits of smart metering and their dependence on data. It also details the personal information embedded within smart meter data, and hence, the privacy risk associated with data sharing. We assess the UK's existing data sharing regulations against the principles of Privacy by Design. In particular, we focus on the technical aspects of privacy.

2.2 The Smart Meter Implementation Programme

The SMIP consists of several components which we summarise below. The dataflow between these is shown in Figure 2.2.

- Smart meters: digital electricity and gas meters capable of recording consumption as well as storing tariff and credit information.

- Communications hubs: devices that create a Home Area Network (HAN) to which consumers can connect an In-Home Display (IHD), to see near-real time usage and cost information, and other smart devices. The communications hub connects to the dedicated Wide Area Network (WAN), run by a Communication Service Provider (CSP), through which it is possible to share the data logged by smart meters with other users. Additionally, Consumer Access Device (CAD) can connect to the HAN providing an alternative means of data access.

- Data Communications Company (DCC): a centralised regulated entity that is responsible for gathering data from smart meters across the country, verifying and

processing the data, and providing a gateway for authorised users (e.g. suppliers, Distribution Network Operator (DNO) and others) to access smart meter data.

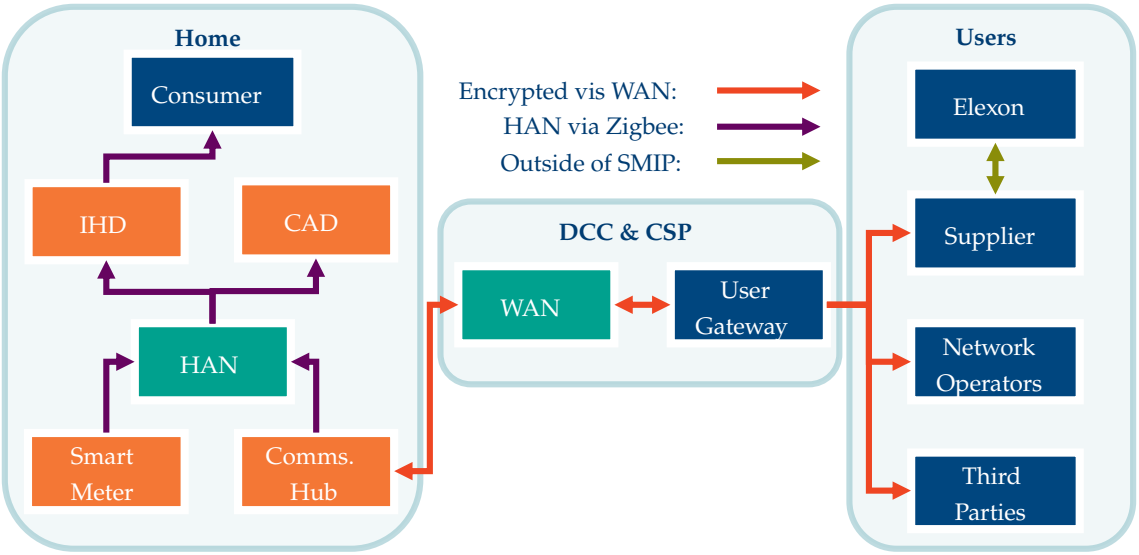


Figure 2.2: Dataflow for Smart Metering in GB. Adapted from [58].

Several features of the SMIP distinguish it from smart meter roll-outs elsewhere [2]:

- Supplier-led: Most national smart meter roll-outs have been led by DNOs, as they oversee maintaining network infrastructure and their remit is usually defined by geography. In GB it was decided that the roll-out would be supplier led as they already own existing traditional meters and have existing relationships with consumers.
- Voluntary/Opt-in: Although the SMIP aims to achieve 95% coverage, getting a smart meter is entirely voluntary. This is mainly due to privacy concerns and successful legal challenges to a mandatory roll-out in other countries [59].
- Mandatory IHDs: One of the major expected benefits of roll-out is energy usage reductions from feedback on energy usage. IHDs play a key role in this as they provide near real-time information on energy usage and costs. The SMIP mandates that energy suppliers must offer consumers a free IHD with their smart meter.
- DAPF: Through multiple consultations and reviews, privacy and security have been considered throughout the planning and implementation process. This has resulted in a framework to safeguard consumers' privacy interests whilst enabling proportionate access to data [11].

The process of installing a smart meter under the SMIP can be summarised into the following four steps [60]:

1. Opt-in: Customers must request a smart meter from their energy supplier. Energy suppliers are therefore required to actively encourage their customers to sign up. In addition, Smart Energy GB, a government entity, provides information on the benefits of smart meters, runs advertising campaigns and research into customer experience.

2. Meter Infrastructure: The smart meter roll-out in the UK is supplier-led. Suppliers are free to choose which smart meters to install as long as they conform to minimum technical specifications and ensure interoperability in the event a customer wishes to switch supplier.

3. Installation: Suppliers book appointments directly with their customers. Upon installation suppliers are required to provide information on data sharing options. In addition, they offer an IHD which allows customers to see their energy consumption and bills in near real-time [61].

4. Post-installation: Suppliers are required to remind customers of their current data sharing options on regular basis. Customers can change their options at any time.

2.2.1 Technical capabilities – SMETS2

The SMIP allows suppliers to choose which smart meters they want to install and what functionality it should provide. To ensure minimum standards for smart metering infrastructure and interoperability, BEIS set out the Smart Metering Equipment Technical Specifications (SMETS)[62]. It specifies requirements relating to data collection, data transmission and other functionality, as summarised in Table 2.1. The initial specifications, SMETS 1, were set out in 2014 with an updated version, SMETS 2, being published in 2018. Although the core data logging and other functionality are the same, SMETS 2 ensures interoperability across suppliers and enhanced security measures through end-to-end encryption and a dedicated closed communications network. SMETS 2 certified smart meters can log key parameters including active and reactive energy consumption as well as export (where households have an electricity generation source such as solar panels or storage with grid export capabilities such as Electric Vehicles (EV) or batteries) and voltage readings at high resolution (up to 10 second resolution). The meters are also able to store historical data going back 13 months at a HH resolution. In addition, the smart meter stores tariff, credit and debt information, sent by the supplier, to provide near real-time consumption costs and billing information.

Table 2.1: SMETS 2 Minimum Functionality. Summarised from [62].

Type	Description
Data Logging	Active & reactive energy imports and exports at 10 second resolution Time of use pricing rates Meter balance and debt registers Historical consumption and cost data at different resolution going back to the previous thirteen months
Data Transmission	Two-way communications Encrypting and decrypting data using the following cryptographic algorithms: Elliptic Curve DSA, Elliptic Curve DH and SHA-256. Joining a ZigBee SEP v1.2 Smart Metering Home Area Network
Other	Payment Mode (Credit or prepayment) Auxiliary load control switches, including randomised offset capabilities. Load Limiting, thresholding and remote disconnection Maximum meter power consumption of 4 Watts

SMETS 2 also incorporates smart control facilities that allow for remote load management and protection as well as remote disconnection of supply. Auxiliary load control switches, which can either be connected physically to the meter or wirelessly through the HAN, can be programmed to turn on/off based on schedules set by the supplier and stored in switching table or on an ad-hoc basis through a 'Boost' function. Although this functionality is currently limited to on/off events, future versions of SMETS will look to implement Proportional Load Control, which will allow for more fine-grained control actions to take place [63].

A key component of the smart meters is their ability to run encryption algorithms to securely verify commands coming from suppliers and other authorised parties and send data over the WAN network to these parties via the DCC. Many of the SMETS 2 certified smart meters also include tamper proofing and detection mechanisms which automatically report back to suppliers.

2.2.2 Data sharing options

Consumers can access the data stored on their smart meter either through the smart meter itself[1] or using an IHD, which must be offered to consumers and provides a more user-friendly interface. The functionality of the IHD varies between suppliers but the SMETS regulations set out minimum standards. The IHD must be able to display near real-time consumption information both in terms of usage (kWh) and cost. It must also facilitate access to billing and debt information and allow those on pre-payment plans to top-up their accounts. The physical infrastructure, smart meters, and communications devices, are owned by suppliers but the consumers have control over how and with whom the consumption data, logged by these devices, is shared. Although SMETS specifications allow for data to be recorded at very high resolutions, the DAPF sets out three data sharing frequencies from which consumers can choose: monthly, daily (default) and HH. These options have been summarised into a Data Guide developed by Citizen's Advice [61].

Additionally, suppliers require explicit consent to use the data for marketing purposes or to pass it on to third parties. Suppliers are therefore forced to incentivise consumers to provide higher resolution data by explaining how the data will be used and how it may benefit them. For example, one supplier's information leaflet states that at a monthly resolution suppliers can provide accurate bills, at daily resolution they can also provide useful energy savings and efficiency advice and improve their forecasting [61]. At half-hourly resolution they state that, in addition to the benefits offered at daily resolution, they can provide greater visibility of energy consumption across the day. We note that of the 58% of GB homes with smart meters, 92% of these are currently in smart mode [3]. However, a survey by Citizens Advice showed that many consumers were not providing HH data (49%) or simply did not know (37%) what their chosen data sharing options were [64]. This suggests that the availability of HH data is significantly less than the proportion of smart meters.

The DAPF defines which entities may access smart meter data and how it can be used. Parties and entities are split by those who undertake regulated activities relating to electricity supply and operation and therefore do not require explicit consent to access low resolution, monthly, smart meter data. These are summarised in Table 2.2. There are

[1]This is theoretically possible as the communications protocols and specifications are laid out in the SMETS. However, most available solutions to access the data, in the UK, rely on third party CADs which can then only be accessed through their dedicated apps or web portals. An exception is the CAD offered by GlowMarkt which offers local access, but requires significant technical expertise to install and operate.

Table 2.2: Authorised parties and activities under DAPF. Summarised from [11], [65], [66]

Authorised Parties	Regulatory Duties
Energy suppliers	Billing
Distribution network operators	Settlement and forecasting
Law enforcement	Investigating suspected theft
Government	Business readiness
Authorised third parties registered with DCC	

currently two exceptions:

- Following a consultation by the Office of Gas and Electricity Markets (OFGEM) on MHHS, suppliers will now have access to consumers HH consumption data for settlement and forecasting purposes by default with an option to opt-out to daily data sharing [65]. This exception has recently been expanded to allow supplier access for business readiness purposes [66]. OFGEM provides a broad definition of business readiness purposes encompassing forecasting and trading functions as well as the development of new products and services. Additionally, OFGEM states that data used for purposes beyond settlement must be anonymised and aggregated where practicable.

- DNOs can access HH data subject to privacy plans approved by OFGEM [67].

Any other parties such as tariff comparison websites, energy switching services as well as entities outside the electricity sector are required to first register with the DCC and will still require explicit consent from consumers before they can access consumers' data. In addition to the protections laid out in the DAPF, companies are also bound by GDPR regulations as set out in the Data Protection Act 2018, which is currently under review[68], as individual smart meter data is deemed personally identifiable information [69]. The DAPF provides consumers with options as to how their data is shared but does not explicitly provide the ability to share anonymous data as is offered in other countries such as the US and Canada [70]. Smart meter data accessed through CADs, such as a suppliers' dedicated app would be subject to GDPR but not the DAPF. Although the DAPF is meant to cover data access for smart meter data, the introduction of CADs allows companies to access smart meter data directly with their own terms and conditions. Importantly, CADs also offer suppliers and other entities a means to access the highest

resolution data SMETS2 offers, namely 10 second resolution. As such, arguably the most privacy sensitive data, as we will discuss in more detail in Section 2.4, is not governed by the DAPF.

2.2.3 Beyond the SMIP

The DAPF was developed in 2012 during the early stages of the SMIP and currently remains the main framework for which smart meter data access is governed. Access to smart meter data has been limited mainly to energy suppliers and DNOs. However, there have a been a number of recent developments which could fundamentally change how and by whom smart meter and associated consumption data can be accessed. Smart appliances and demand response are expected to play a significant role in the energy transition. This will generate complementary data streams with additional, more granular data on energy consumption. BEIS has recently set out codes to standardise the operation of such devices and schemes which include specific requirements around data privacy and cybersecurity based on explicit consumer consent, data minimisation, and encryption, however these are not encompassed by the DAPF [71], [72]. As a result, there may be multiple entities (e.g. demand response aggregators, local energy system operators, etc.), other than energy suppliers, who will have access to smart meter data, each with their own data privacy policies. This creates a complex and confusing landscape for consumers to navigate. The concept of a Data Dashboard has been proposed by Citizens Advice, which would provide consumers with a centralised platform to manage access and permissions as well as see how the data is being used [73]. This consent-based dashboard has also been put forward as a key recommendation by the Energy Digitalisation Taskforce[74].

In addition, there are calls for widening access to smart meter data for uses beyond the day-to-day operation of the electricity network as part of a move to digitalise the energy sector. A key component of this is to develop Open Data platforms which will provide access to public interest actors such as government, regulators, local authorities, and other stakeholders to inform and shape policy. How such a platform might incorporate smart meter data is still under discussion but could involve the use of a trusted processor to provide appropriate privacy protections [18].

2.3 Benefits and Uses of Smart Meter Data

Smart meters, specifically their data logging and sharing capabilities, are seen as key enablers for a more efficient and cost-effective low carbon electricity network. Access to

granular data on consumption and the introduction of time-varying tariff structures will enable a myriad of potential benefits, for customers, energy suppliers and the system overall. As part of the UK's smart meter roll-out BEIS conducts regular cost-benefit analyses. According to the latest edition, completed in 2019, the SMIP is expected to cost £13.5bn with projected benefits of £19.5bn [19]. This constitutes an average net benefit per household of £250 over the appraisal period (2013 to 2034). However, many of these benefits are contingent on high levels of smart meter adoption and data sharing[2]. This section summarises some of the key benefits enabled by smart meters with a focus on what level of data sharing is required to achieve them. For details on the innovations enabled by smart metering, as well as the limitations, we refer to the Energy Futures Lab briefing paper series on residential demand response, digitalisation of energy and smart electric heating [75]–[77].

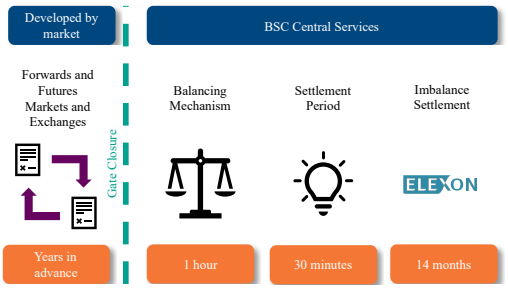
2.3.1 Projected benefits

Automated Meter Readings

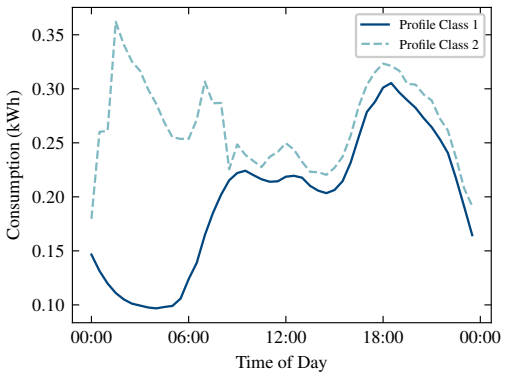
A major benefit of smart metering infrastructure is the avoided operational costs associated with meter readings. Automated meter readings will eliminate the need for customers to take meter readings and avoid estimated bills. Energy suppliers will avoid the costs of sending meter reading operatives to properties in order to obtain a meter reading. Similarly, during the customer switching process, smart meters provide automated meter readings upon change of supplier. It is also expected that accurate, automated billing and streamlined switching process will improve customer experience resulting in fewer inbound customer calls. In addition, debt handling is projected to improve with the possibility for more frequent billing, earlier identification of debt buildup and providing faster follow-up action to help consumers. BEIS estimates the benefits from automated meter readings to be almost £7.4bn consisting of £1.38bn in direct time savings for consumers and £5.12bn in reduced operational costs for suppliers [19]. Although these benefits are enabled by smart metering, they do not require customers to share their high-resolution consumption data (e.g. HH data) with suppliers or other entities. SMETS meters include tariff, credit and consumption registers allowing customers to see their usage and bills in real-time [62]. As a result, billing and verification can be performed without sharing high-resolution consumption data with suppliers.

Forecasting and Energy Procurement

In an electricity network, demand and supply must be matched at all times to ensure the stability of the electricity grid. Suppliers are responsible for procuring electricity on behalf of their customers. Therefore, they must forecast the expected demand of their customers for each half-hour of the day, known as a settlement period. If there is a mismatch between the amount procured and the amount consumed, the imbalance is settled through the Balancing Mechanism (see Figure 2.3). This is a near real-time market where National Grid, the transmission network operator, purchases changes in generation and consumption to correct imbalances.



(a) Settlement Stages and Timeframes



(b) Average Daily Consumption of Domestic Profile Classes for 2018.

Figure 2.3: Overview of the GB Balancing Mechanism. (a) Reproduced from Elexon[78]. (b) Default Period Profile Class Coefficients Provided by Elexon. Average Annual Consumption for 2018 (Profile Class 1 - 3,642 kWh, Profile Class 2 - 5,143 kWh) Obtained from [79, Table 3a-b].

In the absence of smart metering the actual usage of the electricity for each customer in a given settlement period is not known and customers are assigned to one of eight Profile Classes. Domestic consumers fall into either Profile Class 1, domestic unrestricted customers with a single rate tariff, or Profile Class 2, domestic Economy 7 customers with a two-rate tariff. The consumption for each settlement period is then estimated based on the average expected consumption for each Profile Class using 'Default Period Profile Class Coefficients' generated by monitoring a sample of houses across the country[80]. The average annual consumption per settlement period for the domestic Profile Classes in shown Figure 2.3b.

As a result, costs are not reflective of actual consumption and suppliers are not exposed

to the risks HH changes in consumption. Smart metering will enable domestic consumers to be settled on a HH basis, based on actual consumption. Although 58% of homes now have a smart meter installed, very few are being settled on a half-hourly basis [81]. A move towards mandatory MHHS could reduce overall system costs and improve efficiency. It would also allow suppliers and domestic consumers to fully harness the benefits of shifting demand from peak to off-peak periods. OFGEM has estimated this to accrue a net benefit to consumers of between £1.56bn to £4.51bn till 2045, with average annual savings per household between £2 and £9 [82].

MHHS will legally oblige suppliers to settle their electricity volumes using actual consumption data instead of the Profile Classes where half-hourly data is available. As smart meters are an opt-in process and consumers will still be allowed to opt-out of HH data sharing when they have a smart meter, such consumers will still be settled based on estimated load profiles. However, these profiles will be generated on an ongoing basis using the actual HH data available from consumers who do share their data. It is expected that the full transition to MHHS will be completed by October 2025, however OFGEM aims to introduce obligations on the Half-Hourly Settlement (HHS) on actual consumption sooner than this [65].

Energy Savings

The smart meter roll-out expects to deliver significant benefits through energy reductions driven by changes in consumers' energy consumption behaviour. Energy usage reductions also reduce the amount of energy that needs to be produced, bringing additional benefits of carbon emissions reductions (estimated at 34.4mln tonnes) and air quality improvements. BEIS has estimated the total benefits from energy and associated carbon reductions due to informational feedback to be £8.273bn [19]. The average household is expected to have bill savings of £2,903 due to energy reductions and reductions in suppliers' operational costs. Trials have shown that providing real-time information feedback on energy usage can substantially reduce overall energy consumption. A 2019 meta-analysis of trials, across 130 electricity and gas pilots including around 5.5 million residential customers, found an average reduction of 5.4% in electricity consumption and 3.9% gas consumption [83]. There are four main components identified by BEIS that contribute to the realisation of these benefits [19]:

1. Direct feedback – real-time consumption data through IHD (that are offered to all domestic smart metered households), smartphones, online services, or other

platforms. Trials indicate this to be most effective, with average reduction of 7.9% in electricity and 9.6%in gas consumption.

2. Indirect feedback – aggregated or non real-time feedback, e.g. accurate bills and historical or comparative information on bills. Pilots results for this type of informational feedback average reduction of 5% in electricity and 1.8% in gas consumption indicating direct feedback results in higher engagement.

3. Advice and guidance – on energy and energy reduction, e.g. advice that installers are required to offer during installations or applications and services that can help interpret data and point towards better choices. Trials providing general tips and advice on ways to reduce energy consumption showed an average reduction of 5.0%. More personalised advice based on disaggregation of consumption exhibited average savings of 7.7%.

4. Motivational campaigns – designed to raise energy literacy and motivation to reduce energy consumption. Smart Energy GB, the national communications campaign supporting the roll-out, has an objective to this effect.

None of the components above necessarily require consumption data to be shared with suppliers or third parties as information on consumption can be relayed directly by consumers through their IHD. Although BEIS used conservative estimates of annual reductions of 3% for electricity and 2.2% for gas, recent evidence suggests that energy savings have been lower than expected, and questions remain as to whether these will be sustained in the long term [2], [75]. It has been proposed that more personalised and targeted information could help stimulate greater reductions in consumption for which sharing of high-resolution consumption data would be required. This includes detailed bill breakdowns, comparisons with neighbours, personalised advice on energy efficiency tips and appliance level usage information. Such additional functionality would require sharing of high-resolution data with suppliers and/or third parties.

Demand Shifting and Smart Grids

Smart meters are a key enabler of large-scale domestic demand shifting and smart grids. By facilitating intra-day pricing and automated load control they provide the technical infrastructure to unlock potential flexibility in the domestic energy sector. The introduction of more intermittent renewables and the electrification of the heat and transport sectors will present significant challenges for the power system. Demand-side flexibility can

reduce system costs by reducing peak demand and consuming renewable energy when it is available, reducing energy procurement costs and carbon emissions. Research has highlighted the high potential value of up to £8bn per year of flexibility [84]. Within the domestic sector various models have been proposed:

- Time-of-use pricing such as the Octopus Agile tariff, which follows day-ahead market prices for each half-hour of the day [85]. With prices being linked to HH costs consumers are incentivised to shift consumption to cheaper times of day, lowering costs for consumers with flexibility as well as reducing system costs and emissions by reducing peak consumption and the need for reserve capacity. A 2019 study valued the potential average household savings between £5 per year (assuming current trends) and £90 per year (assuming the electrification of heat and transport and automation)[86].

- Energy-as-a-service models where customers pay a flat monthly fee and sign a performance contract with their energy supplier to, for example, supply 'warm hours' ensuring a minimum temperature in the home rather than paying for each kWh of fuel consumed [87]. This incentivises energy efficiency rather than consumption while giving energy suppliers control to optimise the heating systems' energy usage.

- Local energy systems or peer-to-peer networks, allow customers to trade electricity amongst each other and manage their consumption while avoiding imports from the rest of the electricity network. Many variations of these schemes have been proposed and are being trialled with centralised and decentralised structures. One of the main components of these schemes is the sharing of high volumes of consumption data, in near real-time with peers and operators [88], [89].

Existing demand response programs and trials have shown mixed results, raising questions around engagement, consumers' responsiveness, and persistence [75]. BEIS estimates demand shifting to bring in benefits of £1.363bn based on conservative assumptions on engagement (19% of households). However personalised tariffs, gamification and innovative incentive structures enabled by sharing high resolution consumption data could significantly increase engagement and usage flexibility [90]. It is likely that more innovative models such as local energy systems will bring in multiple parties beyond regulated energy suppliers and DNOs which would require wider data sharing with third parties [82], [91]. Although, on average, the demand shifting schemes discussed above are expected to have a net benefit, individual consumer savings are highly dependent on

consumers' flexibility and how operational savings are passed on to consumers. For example, a 2016 study on the distributional impacts of different incentive structures found that some can result in increased electricity costs, especially for vulnerable consumers [92].

Network Management

Smart meters offer benefits to the DNOs who manage the infrastructure used for electricity distribution. These benefits come from the increased data that network operators will have available. This would allow them to identify faults in the network and restore electricity supply more quickly when outages occur, and take better informed investment decisions. Historical smart meter data allows DNOs to identify areas in the existing network which are at risk and might require reinforcement more easily. This will result in investment for network reinforcement being better directed. BEIS has estimated this to be £380mln [19]. The expected energy savings and demand shifting could reduce overall network losses by reducing the total amount of electricity transported on the distribution network and reduce peak consumption. Access to granular data would allow suppliers to identify patterns of behaviour that may indicate theft allowing them to reduce energy theft more efficiently. This would require access to high-resolution HH data and BEIS estimates this capability to bring benefits of £260mln.

2.3.2 Dependence on Data Sharing

The projected benefits in BEIS's cost-benefit analysis of the SMIP are all enabled by the introduction of smart metering for the domestic sector. However, we find that few are wholly dependent on sharing of high-resolution consumption data. To map the data dependence of the benefits considered by BEIS, we consider the following three categories; (1) benefits which are dependent on, (2) those which may be enhanced with, and (3) those which are independent of access to high resolution data. Our mapping is summarised in Figure 2.4.

The direct consumer benefits; the time saving from not taking and submitting meter readings can be achieved by smart meters sending low-resolution data for accurate billing (e.g. monthly which is similar to what data suppliers receive with traditional metering systems) and the expected energy reductions (as well as the associated carbon and air quality benefits) can be realised through informational feedback on the IHD, although personalised advice and recommendations, for which high-resolution data may be re-

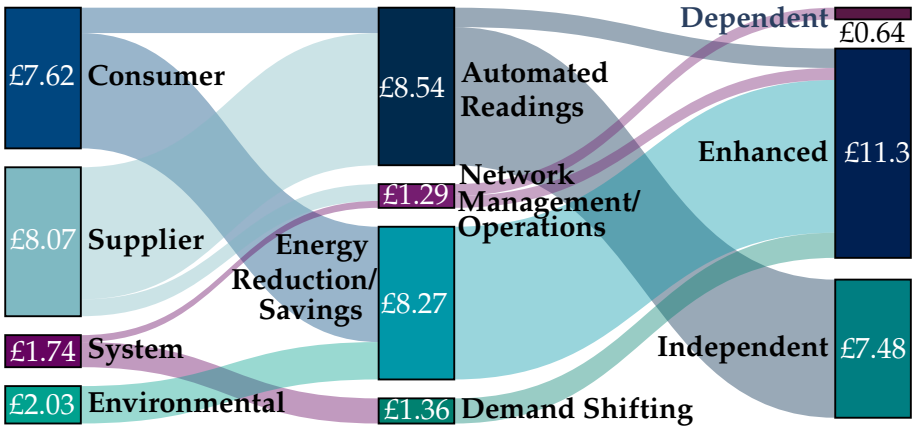


Figure 2.4: Dependence of Benefits on Sharing of High-Resolution Data. All values in £ billions. First Two Columns taken from 2019 BEIS Cost Benefit Analysis[19]. Final Column Presents our Mapping based on BEIS's Descriptions. Underlying Data can be Found in Table A.2 in Appendix A.

quired, could increase and sustain energy reductions. Similarly, demand shifting can be achieved without consumers having to share their high-resolution consumption data as smart meters are able to store tariff details locally and produce billing information on the IHD. Most of the operational benefits for suppliers are based on the ability for automated readings to avoid site visits but do not require access to high-resolution data. Suppliers' ability to identify theft and DNOs' improvement of network management do require access to HH data to provide meaningful benefits. Although many of the benefits discussed above are not wholly dependent on access to high-resolution data, the majority can be enhanced with access which could lead to improved accuracy and ability to personalise recommendations and actions. A full breakdown of the attributions can be found in Table A.2 in Appendix A.

A summary of the effect of data resolution on the realisable benefits is shown in Table 2.3. In addition to the temporal resolution of the smart meter data, the spatial resolution or level of aggregation also affects the benefits of data sharing. We note that similar to the temporal resolution many of the benefits quantified in BEIS's cost-benefit analysis can be realised using aggregate data. Only household specific functions require access to individual data. Specifically, automated meter readings and demand response do need data to be collected on an individual basis for verification purposes. Even targeted functions, such as direct informational feedback can be based on aggregate data, although trials suggest savings to be significantly less than individualised recommendations.

Table 2.3: Realisable Benefits at Different Data Resolutions

Benefit/Use	Temporal Resolution				Spatial Resolution	
	≤ 1 min	HH	Daily	Monthly	1 House	> 1 House
Avoided Meter Readings and Site Visits	✓	✓	✓	✓	✓	
HH Load Forecasting	✓	✓			✓	✓
Energy Savings	✓	✓	⋆	⋆	✓	⋆
Demand Shifting	✓	✓	⋆	⋆	✓	⋆
Network Visibility	✓	✓			✓	✓
Public Interest	✓	✓	✓		✓	✓
Research	✓	✓	✓		✓	⋆

Note: A ⋆ indicates that benefits can be accrued but are significantly diminished compared to a higher temporal resolution/lower spatial resolution.

2.3.3 Other Potential Benefits and Uses

Wider access to high resolution smart meter data creates opportunities for expanding energy research and exploring new information and inferences. Currently, most entities with access to smart meter data are directly involved in the operation of the electricity network (suppliers, network operators, switching websites etc.), however, many other entities, such as government, regulators, public interest groups, academic researchers and other stakeholders can benefit from access.

Public-Interest Uses

Smart meter data could play an important part in assessing the impact of policy interventions and reforms in the energy sector. It would, for example:

- Improve BEIS's national energy statistics, which are currently based on annualised consumption estimates, by providing accurate and temporally granular information on consumption patterns [18].
- Allow OFGEM to understand distributional impacts of policy choices such as ToU pricing schemes which could adversely affect vulnerable consumers [18], [92] or lead to discriminatory pricing [7].

- Enable projects such as the Virtual Energy System, led by National Grid, which aim to build a real-time replica of the GB energy system, could be enhanced with real-world smart meter data [93].
- Provide a means to improve citizen engagement around climate and environmental issues and the energy transition by providing information on the connections between our individual actions/choices and their energy use and environmental impacts. For example, information portals such as GridWatch, which provides near real-time information on the current generation mix in the UK, could be adapted to provide such information based on citizens' individual consumption.

Facilitating Research

The availability of high-resolution smart meter data opens possibilities for new research and insights. For example, a recent publication showed the immediate impacts of the lockdowns instituted in the UK during Covid-19 and their effect of energy consumption patterns, flexibility, and changes to daily routines [94]. The recently launched Smart Energy Research Lab will provide access to over 10,000 customers granular smart meter data and accompanying metadata to authorised researchers with aim to provide new insights[95].

Innovation and Third-Party Access

Smart meter data also has the potential to spur innovation in other sectors beyond energy. BEIS is actively encouraging businesses to sign up to the DCC platform [96]. Recent work has looked at the potential benefits of using smart meter data to monitor assisted living facilities and dementia patients [97]–[99]. Ongoing monitoring can provide relatives, carers and health practitioners with early warnings in the event someone has been incapacitated by a fall. Similarly, energy use patterns can be used to infer living conditions and behaviour changes which may be connected to heath issues. There may also be interest from other sectors and commercial entities such as, insurance providers and retailers, who could use smart meter data to better understand customers habits and tailor their services. As with other data streams (e.g. internet usage data or purchase history) such uses raises issues around privacy as well as ethical issues around discriminatory pricing.

2.4 Potential Privacy Infringements and Risks

The types of personal information that can be extracted from smart meter data are highly dependent on the data sharing options selected by customers. SMETS, the minimum technical standards for smart meters in the UK, are capable of recording and displaying consumption data at 0.1Hz (every 10 seconds)[62]. High resolution data, such as half-hourly data, can be used to infer a wide range of information about a household. The following section presents the potential privacy infringements and risks.

2.4.1 Load Disaggregation

Smart meters record the total electricity consumption of a particular house. This is effectively an aggregated representation of all the different electrical appliances in the home, as shown in Figure 2.5. Specialised techniques known as Non-Intrusive Load Monitoring (NILM) can disaggregate smart meter data to identify and estimate the consumption of different appliances with high accuracy. Appliances have characteristic profiles of energy use which make it possible to classify certain changes in consumption seen at the aggregate level and extract estimates of individual appliance usage. Supervised learning algorithms require a large amount of training data to calibrate. They provide very high accuracy, especially with high resolution data. This includes labelled data for individual appliances and aggregate consumption. Some examples of the techniques used are Hidden Markov Models [100], K-Nearest Neighbour [101], and deep learning [102]. Unsupervised learning algorithms can be run without access to such high-resolution labelled data. They perform particularly well for low-resolution datasets. These include adaptive Hidden Markov Models [103] and K-means clustering [104]. Depending on the data resolution it is possible to detect and estimate the power consumption of different appliances with varying degrees of accuracy. Appliances can be broadly categorised in the following:

- Small appliances such as TVs, laptops, lighting, and other consumer electronics devices.
- Cooking appliances such as electric stoves, kettles, and ovens.
- Heating and cooling appliances such as electric space heaters, electric water heaters, refrigerators, and air conditioning units.
- EVs, batteries, and distributed generation resources such as rooftop solar panels.

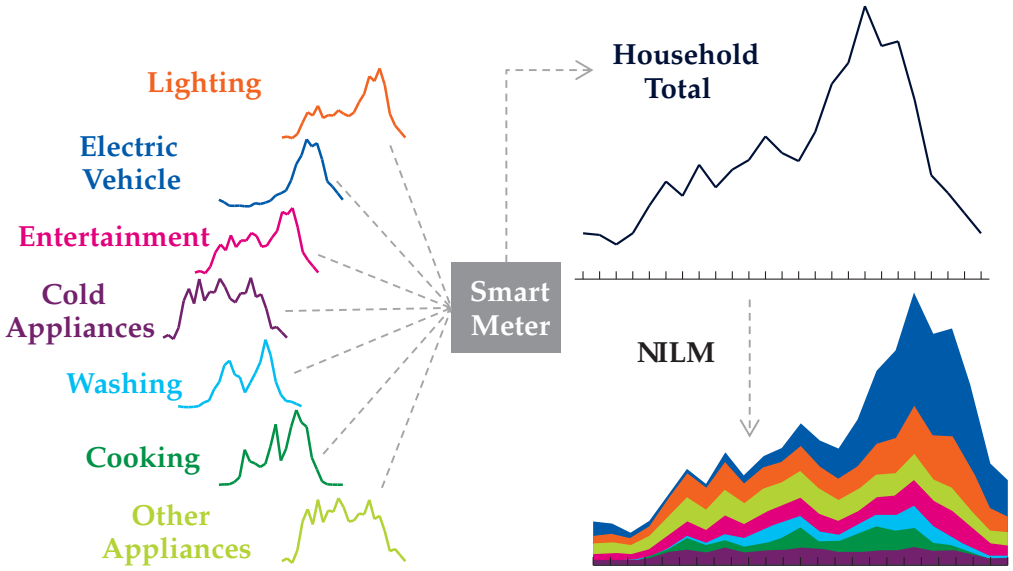


Figure 2.5: Illustrative Example of Appliance Signatures and Load Disaggregation. Synthetic Data for a Household in the City of Westminster. Sourced Directly from [105] which Employs a Markov-Chain Model to Generate the Data.

We conduct a rapid review of academic and grey literature of existing NILM algorithms to understand and map their dependence on data resolution, both temporal and spatial. The accuracy with which the state-of-the-art NILM algorithms can identify and estimate consumption of these appliances with respect to temporal resolution is summarised in Figure 2.6. A detailed breakdown of accuracy levels can be found in Table A.3 in Appendix A. Heating and cooling typically make up the bulk of a houses' consumption profile and are highly correlated with ambient temperature. As a result, it is possible to infer, for example, whether a house uses an electric or gas heating system even with daily or monthly data. EVs, distributed generation and electric cooking appliances also have a large impact on a houses' consumption profile. They have characteristic consumption profiles (e.g. solar panels produce electricity during the day reducing the net consumption observed during this period) and can therefore be identified even with hourly data. Smaller electronic appliances and lighting consume less electricity and have more irregular usage patterns making them difficult to identify without granular (sub-minute) data. The detection of appliances and their consumption patterns also provide other useful, derived, information about households. For example, it gives an indication of occupancy and activities. This is discussed in more detail in the following section.

When it comes to aggregated smart meter data we see that at 15 minute resolution most appliances can be detected, and their consumption estimated, when the level of ag-

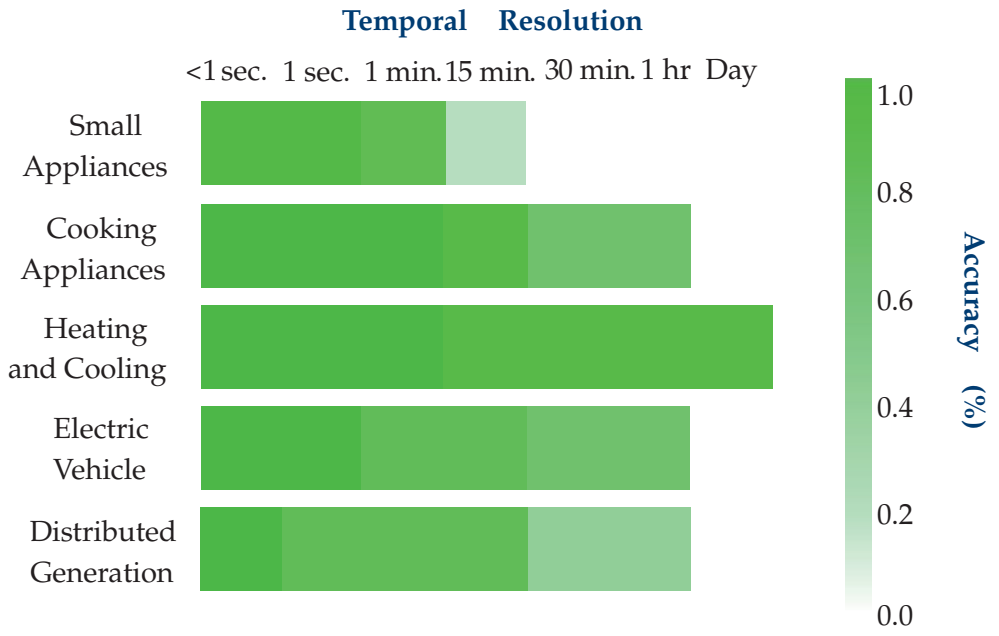


Figure 2.6: State-of-the-art NILM Accuracy. Maximum reported accuracy for identification and/or disaggregation of appliances summarised from [21], [100], [101], [104], [106]–[116]. A detailed breakdown can be found in Table A.3.

gregation, or spatial resolution is less than 10 houses[117]. At higher spatial resolutions, namely 100 households, only large loads such as electric vehicles and distributed generation can be [118]. For feeder-level data, with over 1000 households, large loads can still be inferred at 1 minute resolution[119]–[121], but not at lower levels of temporal granularity. Similarly, we see that increasing the level of spatial aggregation decreases the accuracy of occupancy inference. For example, [117] shows that as we increase the aggregation from 10 to 600 households, the accuracy decreases from 50% to 3%. A breakdown of the maximum reported accuracy, either in terms of consumption estimation or appliance identification, can be found in Table A.3 in Appendix A.

2.4.2 Beyond Energy Use

Demographics and Household Characteristics

Even at low resolutions, smart meter data can also be used to determine socio-economic and demographic information about a household. However, the potential privacy infringements and the consequences of data sharing are not limited to identifying specific household characteristics. Smart meters also store individuals' debt and payment records. In combination, access to such information can lead to unintended or unauthorised prac-

tices such as discriminative pricing and unsolicited targeted marketing [7]. Table 2.4 summarises the demographic information that can be extracted from smart meter data at different resolutions. Smart meter data is considered to be less sensitive than, for example, financial or medical data [26]. However, we see that financial and medical data such as income, employment status, or whether someone is regularly using electrical medical equipment is embedded within smart meter data.

Table 2.4: Identifiable Demographic Information by Temporal Resolutions

Type	≤ 1hr.	Daily	Monthly
Socio-Demographics	No. Residents		
	Residents Age		
	Martial Status		
	Employment Status		
	Long-term Illness		
	Household Income		
	Children and Pets		
	House Type		
Dwelling Characteristics	No. of Rooms		
	Size of House		
	House Location		
	House Ownership		

Note: Shaded columns indicate the characteristic is identifiable at corresponding temporal resolution. Considered identifiable if accuracy is greater 50%. Summarised from [5], [116], [122]–[128]. A detailed breakdown of sources can be found in Table A.1 in Appendix A.

Residential Activity Patterns

Another line of research that is drawing particular attention in the academic community is the linking of smart meter data with time use surveys to understand linkages between activities such as cooking, cleaning etc. and energy usage. A prime example of this is Meter.org, a research project at Oxford University aiming to understand what we use electricity for [129]. The researchers record consumption data for a household for a day and ask members of the household to fill out their activities as well as emotions on an app.

The project is ongoing but has already built a database of over 10,000 participants. This provides several insights into the drivers of electricity demand and identifying flexibility in energy usage as well as the differences in energy use across gender and household composition [22], [130], [131]. This research has shown strong correlations between smart meter data and peoples' activities and the occupancy of the house. This allows one to develop techniques to infer a much more comprehensive range of information about peoples' day-to-day lives from their smart meter data, beyond which appliances are running [6]. For example, high-frequency smart meter data has been used to determine what TV channel an individual is watching [132] or even to determine internet usage information[128]. As such, smart meter data has embedded within it significant amounts of personal and sensitive information. As larger high-quality datasets become available the accuracy with which such inferences can be made will increase.

Linking Datasets

Smart meter data can be linked with other data sources such as social media and other smart devices such as thermostats for further aggregation of multiple data streams. This will enable data analysts to build more detailed profiles of consumers and generate deep insights. Furthermore, the increasing use of Internet-of-Things devices, smart energy appliances and the electrification of the heat and transport sector will mean that smart meter data could play a key role in relating these various data sources. In particular, correlations between different activities and data streams could allow data users to infer seemingly unrelated information using smart meter data. Even before the availability of large datasets and the widespread use of machine learning techniques surprising correlations have led to serious privacy breaches. In the 1990s a pizza franchise owner in Washington D.C. was able to predict the occurrence of major global events based on the sudden increase in orders in the follow up to these events[133][2]. Additionally, the 'black-box' nature of machine learning makes it difficult even for data analysts to predict, in advance, what their machine learning algorithms might infer from these combined data sources. This unpredictability makes it practically impossible to inform consumers about potential future insights and uses of their data making it difficult to obtain truly informed consent [7].

[2]Thank you to Josh for bringing this example to our attention in Khao Sok.

2.5 Discussion

This chapter detailed the SMIP, focusing on the data sharing and access protocols set out in the DAPF. Although the framework provides consumers with some degree of control over how their smart meter data is used, it relies heavily on consumer consent and permissions controls. In addition, the introduction of smart appliances, demand response and CADs will further increase the complexity of data flows and introduce numerous new entities that have access to smart meter data. At the same time, there has been a push to widen data access, for example, through broadening the definition of regulated activities and moving to an opt-out rather than opt-in model for some use cases as part of the move to MHHS.

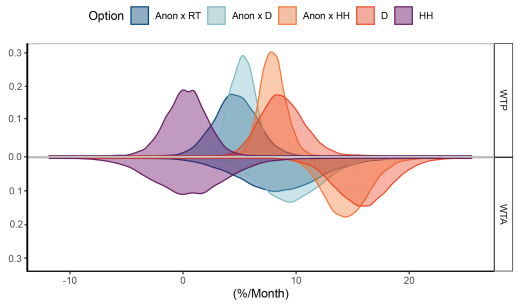
The use cases and resulting benefits of smart metering include operational benefits for suppliers, through the automation of meter readings and customer switching, consumer benefits, in terms of time savings and bill savings from energy usage reductions, and improved system management through greater network visibility and demand flexibility. We find that, of the benefits that have been quantified under BEIS's cost-benefit analyses of the SMIP, many are independent of sharing high-resolution (temporal) smart meter data. However, a majority of the benefits would be enhanced if such data were available to energy sector actors. Similarly, most benefits can be realised with spatially aggregated smart meter data and not require access to individual level data. Again, access to individual level data would enhance the benefits. We also find that many innovative uses for smart meter data lie beyond the energy sector, for example, in healthcare for monitoring assisted living facilities. As such, wider access would allow the full potential and value of smart meter data to be realised but will also require an increasing number of entities, who may or may not have direct contractual relationships with consumers, to access the data.

This widening of access is creating a difficult landscape for consumers to navigate. This is exacerbated by the complexity of potential privacy infringements and risks associated with sharing smart meter data. Smart meter data has a vast amount of personal information embedded within it. This includes energy usage related information (e.g.the use of certain appliances), derived information(e.g. daily routines and home occupancy), as well as socio-demographic information (e.g. age and income). The ability to extract such information and the accuracy with which it can be done depends on the technical tools used, but notably, also on the form of the smart meter data. Specifically, whether

the temporal resolution of the smart meter data and whether data is aggregated have a significant impact on inferrable information. Smaller appliances and activities can only be identified at sub-HH resolutions, where as larger appliances and demographic information can be extracted even at daily or monthly resolutions. In addition, given the quickly evolving nature of machine learning and the increasing availability of rich linked datasets it is difficult to properly define and communicate the potential implications of data sharing.

Overall, we see that obtaining informed consent in the case of high-resolution smart meter data, and specifically for uses outside of regulated activities outlined in the DAPF, is difficult, as the extent of potential privacy infringements is not known. As such, we see that the current data access framework is lacking in several key areas relating to Privacy by Design. The recent move to an opt-out for MHHS, moves away from the strong privacy defaults instituted by the DAPF. The current framework does not provide proactive and preventive protection of consumers privacy as it relies on consumer consent with no additional protections or regulations in place in event of, for example, data misuse or data breaches. As recommended by a multitude of stakeholders and organisations, Privacy-Preserving Techniques could play a significant role ensuring ongoing compliance of the SMIP with Privacy by Design principles. However, due to a lack of understanding as to the costs and complexity of implementation and the potential benefits, PPTs, thus far, have been ruled out.

To elucidate these trade-offs, the next chapter explores consumers' privacy concerns: their understanding of the privacy risks, their valuations of privacy in relation to smart meter data, and the demand for PPTs.



CHAPTER 3

Consumer Privacy Concerns and Valuations

The wealth of personal information embedded within smart meter data raises concerns around privacy and data misuse. These have been highlighted and their dependence on the form in which smart meter data is shared were mapped in Chapter 2. These issues around privacy have led to several surveys and focus groups to explore consumers awareness, understanding and concerns around smart metering. This includes academic literature as well as, assessments carried out in relation to the SMIP by OFGEM, DNOs, and Citizens Advice. Although most existing surveys have found that consumers are generally willing to share their smart meter datathere area a number of key contextual conditions that affect this.

This chapter first summarises the findings of these existing surveys, reviewing them in relation to four key concerns of our study: data sensitivity, anonymisation, trust and transparency, and monetary valuations or Willingness-to-Pay/Accept (WTP/A). It then present the results of a novel survey and discrete choice experiment to quantify consumers' WTP for anonymisation and privacy. We pay particular attention to the role of information asymmetries and informed consent, the distribution of privacy concerns and the impact of framing effects, i.e. whether privacy protection is the default option as advocated by Privacy by Design. This chapter forms the majority of [Paper D].

3.1 Existing Survey Results

3.1.1 Data Sensitivity – What Data?

Type of Data

A study conducted in 2020 by Ipsos Mori indicates that 55% of people consider electricity consumption non-sensitive, although younger respondents (under 35), were more likely to consider it sensitive 49% [134]. Other studies comparing the perceived sensitivity have shown that compared to other types of personal data such as financial details, location data, medical records, social media and contact details, smart meter data is considered less sensitive [24], [26].

However, the majority of consumers are unaware of the personal information that is embedded within smart meter data. A 2019 study by Citizens Advice found that only 30% knew that a smart meter could record the time when a person was in or out of the house [64]. This was even lower (18%) amongst people from lower socio-economic groups. This clearly reflects inequality in access to information amongst lower socio-economic groups and hence their greater vulnerability to abuse.

When consumers are provided with details of the implications on the type of information being shared when sharing one's smart meter data, they are significantly less willing to share their smart meter data. For example, a 2015 American survey found that when respondents were told that smart meter revealed a lot of personal information demand for smart meters decreased by up to 20% [135]. Similarly, a multi-stage longitudinal study in Germany, investigating different utility subscription models, found that a majority of respondents decided to change (~80%) or cancel (6%) their initial subscription choice once the corresponding privacy implications were described [136].

When looking specifically at the data sharing options within the DAPF, the Citizens Advice study found that only 43% of consumers were comfortable sharing half-hourly or real-time data (see Figure 3.1), dropping to 28% for those who did not have a smart meter [64]. Although we see that consumers do differentiate between data sharing frequencies many are unaware of their existing data sharing options.

Table 3.1 summarises selected and preferred data sharing options observed in the Citizens Advice study. A large proportion, 37%, of smart meter owners were unaware of their current data sharing options. In addition, 10% of respondents who did not have a smart meter could not decide on their preferred data sharing option. Additionally, of those who did not have a smart meter only 21% would want to share data at a half-hourly

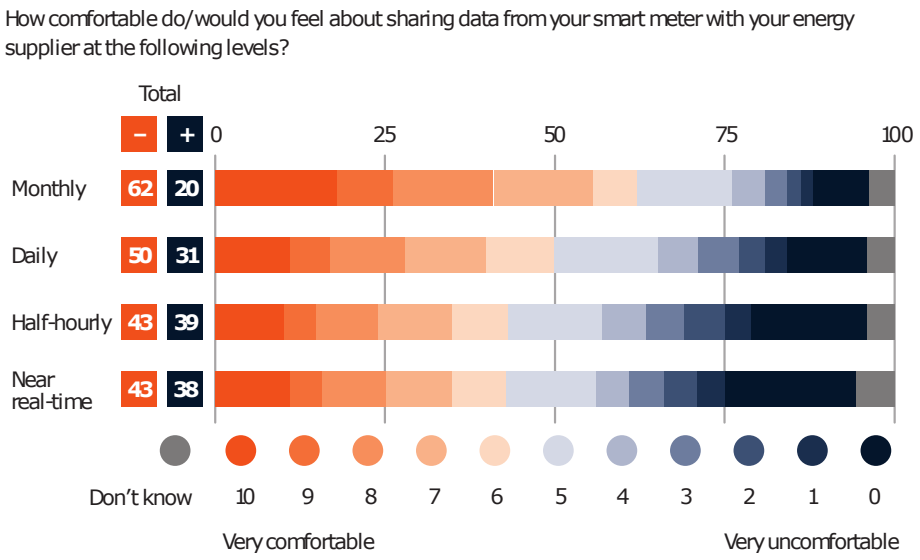


Figure 3.1: Comfort Levels at Different Data Resolutions. Reproduced from [64].

Table 3.1: Actual/Preferred Data Sharing Options. Summarised from [64].

Resolution	SM Owners %	Non-SM Owners %
Half-Hourly or Higher	14	21
Daily	20	17
Monthly	29	44
Don't Know	37	10

resolution. These findings suggest the there is some awareness as to the privacy risks but it also highlights the presence of information asymmetries and lack of engagement which were likely to hinder the ability to obtain informed consent.

Identifiable or Anonymous

When given the option of anonymisation[1], a significant proportion of consumers are more willing to share data. An OFGEM-commissioned survey from 2018 found that 41% of respondents would be more inclined to share high-resolution smart meter data[24]. The 2020 Ipsos Mori study also found that respondents found it important that data is grouped and anonymised so that individual homes cannot be identified [134].

[1]How and the extent to which anonymisation can be achieved for smart meter data will be discussed in detail in Chapter 4

3.1.2 Trust and Transparency – Who has Access and How Will it be Used?

Willingness-to-Share (WTS) is also highly dependent on who has access to the data. The Ipsos Mori study showed that 86% said it is important that data is used for improvements and is not sold or shared with other entities [134]. When asked about trust in a variety of organisations to handle their smart meter data, most respondents trusted OFGEM and other central bodies such as the DCC but trust levels for suppliers, their agents and other third parties were significantly less [24]. Another important factor, identified in the same study, is how the data will be used. Most consumers were happy to share their data if it benefits market operations and efficiency. Still, they were most sceptical of consumer facing uses, in particular if the data were to be used for targeted marketing. For example, a majority of consumers are willing to share their smart meter data in return for services such as information on suitable tariffs (61%), energy efficiency recommendations (57%), bill discounts (56%), or real-time feedback on energy consumption (54%).

However, we note that the above studies provided limited background on what personal information consumers are parting with when sharing their smart meter data. Trust and control play a significant role in determining consumers' attitudes towards data sharing [64], [91]. For example, an UK representative online survey on attitudes towards location data sharing highlighted that despite appreciation of the services, the levels of trust and sense of control over their data regarding delivery organisations was low[131]. Given the wide range of potential uses and misuses of smart meter data, it will be important to provide transparency and clarity on how and by whom smart meter data will be used.

3.1.3 Willingness-to-Pay/Accept for Privacy

As smart meter data can provide benefits to consumers, suppliers, and the wider electricity network, how these benefits are distributed is of significance. Overall consumers are either happy to share their consumption data, willing to share if details on how such data will be used, including on how it may benefit the system as well as benefit them personally, are provided, or are reluctant to share data under any circumstances [137]. A 2015 study found that when offered different electricity service contracts, those requiring smart meter data to be shared with third parties would require suppliers to provide a significant discount [27]. Interestingly, the study made a distinction between electricity usage (smart meter) data and personally identifying data, with consumers willing to pay £1.00/month to avoid sharing usage data but up to £3.11/month for personally identifiable data.

Wider privacy literature has shown that when eliciting consumers valuation of the privacy a large disparity between the willingness-to-pay and willingness-to-accept is observed[28]. Behavioural economic theory suggests that this may be because of an endowment effect, the idea that privacy should be guaranteed by default[138]. As a result willingness-to-accept studies tend to produce much higher privacy valuations. In addition, privacy valuations have been shown to be highly correlated with an individual's income[27]. It is therefore important to define the context within which an individual is making their valuation.

3.2 Survey

3.2.1 Aims and Objectives

As we see, existing studies either make distinctions between smart meter and personal information or respondents themselves make these distinctions. However, as shown in Chapter 2, this is not the case, indeed, smart meter data has embedded within it significant amounts of personal information. This information asymmetry, under a behavioural economics framework, results in bounded rationality, i.e. consumers are not fully informed and therefore make decisions which differ from those they would make were they fully informed[139]. We also find that existing studies have indicated that there is demand for anonymisation, with higher WTS when this option is offered, however its value has not been quantified in monetary terms.

Therefore, this section employs behavioural economics concepts, econometrics and social science methodologies to tackle these issues. Specifically, the central question addressed in our work is: what is the value of anonymisation and privacy? To this end, we develop the following hypotheses:

- H1: Consumers are willing to pay to protect their privacy. The WTP/A for anonymisation will be positive.
- H2: Consumers valuation of anonymisation will vary significantly. There will be significant heterogeneity in WTP/A across the population.
- H3: Information asymmetry lessens consumers' perceptions of privacy risks and therefore depresses their valuations of anonymisation. Consumers with more knowledge of the privacy implications of sharing smart meter data will have a higher WTP/A for anonymisation.

- H4: The endowment effect will impact consumers valuations of anonymisation. The WTA being higher that the WTP.

3.2.2 Survey Overview

To test these hypothesis, we employ a Discrete Choice Experiment (DCE), a technique widely used for valuing electricity related goods ranging from green electricity[140] to EV contracts [141]. Here, respondents are presented with choice tasks, each containing choices with a variety of different attributes, from which they have to select their preferred choice. By observing the choices made over multiple choice tasks and/or by multiple respondents (the respondent's stated preferences), we can develop an understanding of which attributes drive their decision making. When one attribute is monetary we can establish the trade-offs in monetary terms, resulting in a willingness-to-pay when the monetary attribute is a fee/payment to be made or a willingness-to-accept when it is a discount/payment received. Our study frames the choices as electricity contracts, differentiated by smart meter data sharing choices (whether data is anonymised and the frequency/temporal resolution of the data) and a change in their electricity monthly bill. Given the importance of presenting realistic scenarios to elicit accurate valuations, we use the consumers' actual bills as a reference. The relevant attributes for our study will be further detailed in Section 3.2.2.

The main gap we observe in the existing literature, in relation to consumer privacy valuations of smart meter data, is the information asymmetry and resulting bounded rationality of decision making. As such, we focus on how consumers would alter their valuations and choices if they knew the data dependence of privacy implications. Specifically, we aim to quantify how, if at all, willingness-to-pay for anonymisation changes depending on whether they are informed or not. To achieve this we employ a Randomised Control Trial (RCT), which is a technique that has been used extensively to test the impact of interventions, in combination with discrete choice experiments. For example, they have been used to test energy use behaviour change [142], and, similar to our setting, to test how privacy trade-offs are affected by knowledge of GDPR rights [143].

RCTs provide a means to control factors which are not specifically set out in the experiment itself. In our context this includes socio-demographic factors. To achieve this, participants are randomly assigned to a treatment group, those who will receive the intervention, or a control group, those who will not receive the intervention. By ensuring these two groups are similar in terms of a range of extra-experimental factors, we can gain

confidence that any differences in outcomes observed are due to the intervention. We employ the RCT to investigate the effect of information asymmetry. As such, we provide the control group with general information of the benefits of different smart meter data sharing options and the treatment group additional information on the privacy implications of their data sharing options. The following sections will describe the components of the survey including the details of the DCE and RCT.

Screening

Our survey consists of three parts, as outlined in Figure 3.2. The first collects socio-demographic information to determine eligibility, under our representativeness requirements. We are interested in quantifying the changes in WTP/A at the population level, to inform policy decisions under the SMIP, and at the individual level, to understand differentiating factors. We therefore want our sample to be representative of consumers engaging with the SMIP, specifically, GB energy bill-payers. The survey targeted a nationally representative sample of GB energy bill paying adults for both the control and treatment groups. In the absence of quotas for bill payers in GB, the quotas were based on the GB population as whole[27]. Quotas were included for gender, age, ethnicity, Socio-Economic Group (SEG) and region[2]. In addition, a soft quota for smart meter ownership was included to attempt to reflect proportions as of December 2020. To ensure response quality and realism of the online survey two exclusion criteria were included: a minimum completion time of 4 minutes and that respondents electricity bills could not be greater than 5 times national average (£300/month). The average bill (£57/month) was calculated based on OFGEMs average electricity consumption for a medium household in 2020 (2,900 kWh) [3] and the corresponding average electricity rates (23.5 p/kWh) [4]. This is followed by a participant information form explaining the academic nature of the survey before obtaining explicit consent to record and use respondents' data, including the publication of anonymised data as part of academic publications. Specifically, we provide a short overview of the overall study within which this survey is a part, an indication of the survey length and the types of information that will be elicited and how they can withdraw from the study, should they wish to.

[2]Details of quotas and their sources can be found in the note for Table 3.4

[3]https://www.ofgem.gov.uk/information-consumers/energy-advice-households/average-gas-and-electricity-use-explained

[4]https://www.ofgem.gov.uk/energy-data-and-research/data-portal/retail-market-indicators

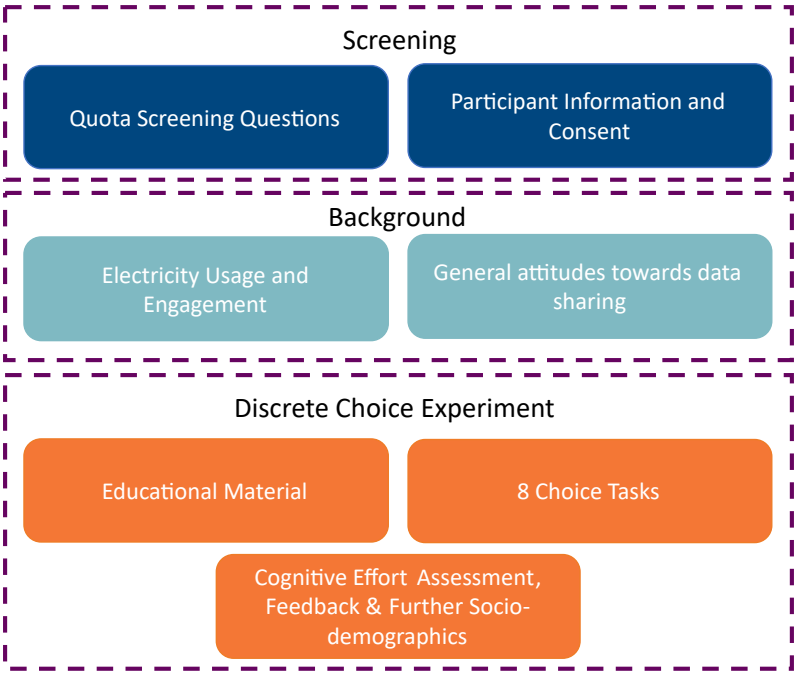


Figure 3.2: Overview of Survey Components

Background Information

The second section covers background information; asking respondents about their electricity supply characteristics and their existing attitudes to data sharing. Specifically, we ask respondents about their tariffs (standard variable tariff, fixed rate, pre-payment, time-of-use tariff or Economy 7/10), fuel supply (dual fuel or electricity only). In addition, for those with a smart meter, we ask them about their current data sharing options are (half-hourly, daily, monthly) and their level of engagement with their IHD (daily, 2-3 times a week, once a week, once a month, less than once a month, never). To establish respondents attitudes to data privacy and existing sharing preferences, prior to the DCE, we ask the following question[5]: *Typically, when you share your data in order to use a service or product, for example when you shop online, use an energy company or book a holiday online, which level of data sharing do you sign up to? Please tick all that apply.*

1. *Sharing the basic information the company requires to provide me with the service they offer*
2. *Allowing the company to use my information for marketing, research, forecasting etc.*
3. *Allowing my data and information to be passed to third parties.*

[5]Due to budget constraints a number of questions were removed from the original survey design. Specifically, two questions resembling those asked in [24] and [64]. The first asked respondents to rank different entities based on their levels of trust, and the second asked respondents to identify what data they believe to be inferable from smart meter data.

Discrete Choice Experiment

The survey is focused on eliciting consumers' preferences around privacy relating to smart meter data with the aim of quantifying the willingness-to-pay for privacy of smart meter data. The final section focuses on this through the DCE. As we showed in Chapter 2, the personal information that can be inferred from smart meter data is dependent on a number of factors: the frequency of data sharing (temporal resolution) and whether it has been aggregated across multiple households (spatial resolution). In addition, to evaluate the demand for PPTs we are also interested in the effect of anonymisation. To investigate the value placed on these different dimensions, we develop an unlabelled DCE, taking into account the choices currently available under the DAPF.

Table 3.2: Choice Attributes and Levels

Attribute	Levels	Description
Expected Change in Monthly Bill	-20% to + 20% in 5% intervals	Shown to respondents as £ amount linked to their actual monthly electricity bill if provided or based on an average bill of £57/month.
Anonymisation	Yes, No	Anonymised data cannot be linked to a particular person and therefore cannot be used to build profiles or identify individuals.
Frequency	Real-Time, Half-Hourly, Daily	The frequency/resolution of the smart meter data shared.

Each option consisted of three attributes: (1) the expected change in bill (negative for a discount and positive for a fee ranging from -20% to + 20% of the respondents electricity bill), (2) the frequency of data sharing (daily, half-hourly or real-time), and (3) whether data is anonymised (see Table 3.2 for further details). Following feedback from an initial pilot of 46 respondents, aggregation was excluded from the study given the difficulty in understanding the difference between aggregation and anonymisation, and limited scope for providing detailed information in an online survey format. Although, we note that this is an interesting finding in itself, as we will later show in Chapter 4, aggregation and anonymisation do not offer the same notion of privacy preservation.

To investigate the effect of background knowledge, and thus information asymmetry, on the willingness-to-pay for anonymisation the RCT was incorporated into the DCE. The mapping of data dependencies outlined in Chapter 2, based on a review of academic and grey literature, was used to select key personal information that can be inferred from smart meter data. The control group were given a general overview of the benefits of smart meter data sharing and basic definitional descriptions of the attributes. The treatment group were given further educational material about the implications of personal information being shared, and of the data options. The chosen characteristics and their dependence on the six potential data sharing options are summarised in Table 3.3. We selected a combination of energy related information, such as appliance usage, and non-energy related information, such as occupancy, income level and marital status.

Table 3.3: Attribute Restrictions and Privacy Implications

	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
Anonymisation		Anonymised			None	
Frequency	RT	HH	Daily	RT	HH	Daily
Expected Change in Monthly Bill	[-20%, 20%]	[-20%, 20%]	[0%,20%]	[-20%,-0%]	[-20%,-0%]	0%
No. of Profiles	9	9	5	5	5	1
Large Appliance Ownership				✓	✓	
Small Appliance Ownership				✓		
Appliance Usage and Routines				real-time	per half-hour	
Occupancy				real-time	per half-hour	per day
Household Details				✓	✓	✓
Income Level				✓	✓	✓
Marital & Employment Status				✓	✓	
Housing Details				✓	✓	✓

The experimental design consisted of 12 blocks of 8 choice tasks per respondent with two unlabelled alternatives in each choice. Several restrictions were placed on the design to ensure the exclusion of dominated alternatives: anonymised options are more expensive than non-anonymised options and sharing at a higher frequency is cheaper all else being equal. These are summarised in Table 3.3. In addition, we assumed that daily

sharing would come at no cost and that all other options would be priced relative to this. The resulting blocked fractional factorial design was generated using the SAS `%Choiceff` macro[144] and selected based on D-efficiency criterion[6]. The final design was generated using priors based on a pilot study of 43 respondents.

When introducing the DCE, all respondents (both the control and treatment groups) were provided with educational material relating to smart meter data and an explanation of the different attributes of the DCE. First, they were shown the following general statement about smart meters which mimics the type of information provided in other survey studies and promotional material disseminated by Smart Energy GB:

Smart meters are the new generation of electricity meters being rolled out across Great Britain. They show you how much energy you are consuming, in real-time, in pounds and pence. Your electricity consumption data can also be shared with your energy supplier which can help them operate more efficiently and pass on savings to you through reduced electricity bills.

In Great Britain, if you choose to install a smart meter, you have the option to choose how your electricity consumption data is shared and who can access it. By default electricity consumption data is only sent on a daily basis, similar to the way traditional electricity metering works.

However, to achieve some of the operational benefits, your electricity supplier may need access to more detailed data, for example, half-hourly or minute-by-minute readings.

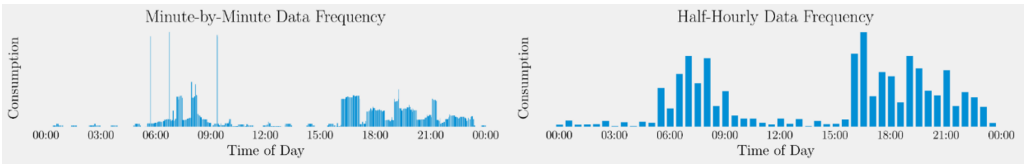
To establish a baseline understanding of respondents comfort levels with regards to sharing smart meter data they were then asked about their WTS: *How willing would you be to share your half-hourly electricity consumption data with your energy supplier?*

1. *Very Willing*
2. *Quite Willing*
3. *Neither Willing nor Unwilling*
4. *Not Very Willing*
5. *Not at all Willing*

Following this, the choice task attributes were explained. First, the different temporal resolutions or frequencies were described and an illustrative example for each was provided, as shown in Figure 3.3. Importantly, these descriptions did not include what, if any, personal information may be embedded within smart meter data. Instead they only mention that suppliers and other authorised parties would have access to their energy consumption data at the specified resolution.

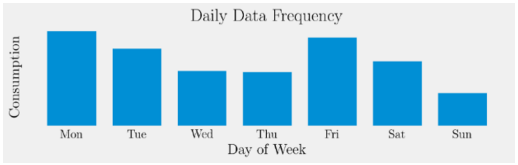
Next, a definition for anonymised data was provided: *Anyone with access is able to*

[6]Design details, including the code used and the resulting D-efficiency statistics can be found in Appendix B



(a) Minute-by-minute

(b) Half-Hourly



(c) Daily

Figure 3.3: Illustration of Smart Meter Data at Different Temporal Resolutions. Data was Generated using the CREST Demand Model to Simulate a 4 Person Household on a January Weekday[145].

extract insights and patterns from your electricity consumption data, but these cannot be linked to you personally. This ensures that even in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data cannot use it to identify or build a profile of you.

The treatment group is then provided additional information as to the specific personal information that is embedded within smart meter data and its dependence on the available data sharing options. Specifically, they are shown:

- A short description of the different type of personal information embedded within smart meter data, listed in Table 3.3 (full descriptions can be found in Appendix B).
- A labelled chart of the minute-by-minute data pointing out the energy usage patterns of different appliances (see Figure 3.4).
- A table showing the dependence of data resolution on the ability to extract the different types of personal information (see Table 3.3).

We note here that, strictly speaking, both groups receive some form of treatment, through the information provided above. Both groups are provided with the short description of potential benefits of smart meter data sharing. Although the control group is made aware of potential information embedded within smart meter data this is not described in terms of tangible personal attributes (e.g. employment status). As such, the control groups perceptions of smart meter data sensitivity rely on their existing knowledge (or lack there of). This mimics the setting in which most existing survey studies have been

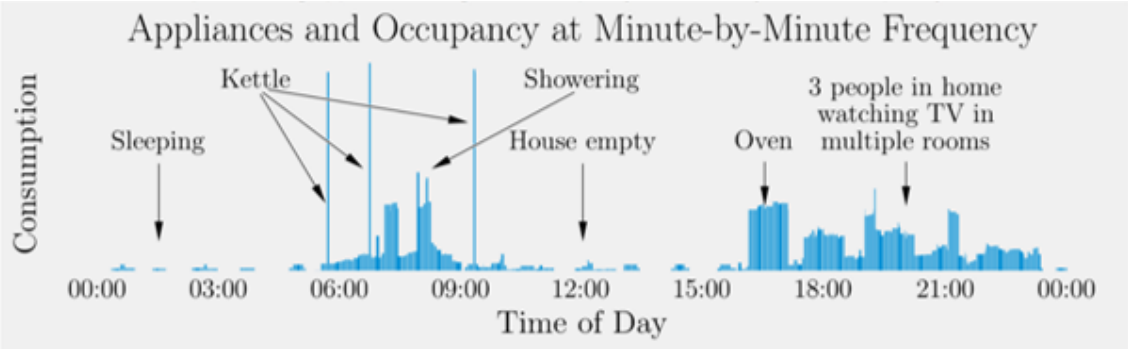


Figure 3.4: Labelled Minute-by-Minute Smart Meter Data. Labels taken from Underlying Appliance Data Produced by CREST Demand Model[145]

carried out and importantly the information landscape that consumers are actually faced with when installing a smart meter.

Finally, an example choice task, shown in Figure 3.5, was also included in the educational material. The control group made choices simply based on the first three rows (shown in white), which correspond to the attributes considered. The treatment group were also shown the privacy implications of each option (shown in the beige rows).

Which option (A or B) would you prefer? Base your choice on the options on this page only.

	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.85
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.
	✔	○

Figure 3.5: Example Choice Task

Following the introduction to the DCE, respondents were again asked about their WTS half-hourly data. This time, whether they were more likely or less likely to share after reading the educational material, or if it did not affect their earlier answer. We then repeated the question for sharing anonymous data:

- *Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **not anonymised** before being shared?*

1. *More likely*
2. *It makes no difference*
3. *Less likely*

- *Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **anonymised** before being shared?*
  1. *More likely*
  2. *It makes no difference*
  3. *Less likely*

The respondents were then presented with 8 choice tasks, each containing different scenarios of data sharing options. As we did not include a status quo option in the choice tasks, respondents were then asked the following question: *Thinking about the choices you have just made, would you opt to have a smart meter with one of the data sharing options you selected or choose not to have a smart meter?*

1. *Have a smart meter with one of the selected data sharing options*
2. *Not have a smart meter*

To measure whether respondents were able to process the educational material they were asked to answer three true/false statements, so-called manipulation checks, relating to the material. These were[7]:

- **Check 1**: A negative expected change in your monthly bill (shown in green) indicates a reduction in your electricity bill. **[FALSE]**
- **Check 2**: A half-hourly frequency means that your total electricity consumption for each half hour is sent to your supplier every 30 minutes. **[TRUE]**
- **Check 3**: Anonymisation of your consumption data ensures your information can be linked back to you in the event of data breach. **[FALSE]**

Respondents were asked to provide feedback as to whether they could comprehend the choice task, as well as an open-ended question. Finally, they were asked for additional demographic data, which included home ownership and income information. The full survey questionnaire and an example choice task set for both the control and treatment groups is provided in Appendix B.

[7]Correct answer shown in brackets.

3.2.3 Sample

The survey was fielded through an online survey platform programmed and conducted by Accent Market Research and their panel partner Sevanta ComRes. Sevanta ComRes develops a target population from general population studies, from which it draws a random set of respondents to create a target sample based on a matching algorithm. Participants were invited through an online platform and paid according to the expected time (approximately £0.50). The study protocol was reviewed and approved by the Imperial College London Research Governance and Integrity Team (RGIT) under SETREC number 21IC6603.

The survey was conducted between 25th March 2021 and 12th April 2021 during which a total of 2810 respondents started the survey. Of these, 598 were screened out based on the eligibility criteria, 94 were screened out for not reaching the minimum completion time, with remaining respondents completing the full survey in 9.95 minutes. This was similar across the control and treatment groups with average completion times of 9.54 minutes and 10.37 minutes, respectively. The distribution of completion times is shown in Figure 3.6a. In addition, 99 were removed to ensure the nationally representative quotas[8]. Screening was performed by Accent Market Research prior to data access. The sample received by us constituted 965 respondents; 477 in the control group and 488 in the treatment group.

[8]Remainder did not complete the survey and were therefore not included.

Table 3.4: Sample Quotas and Statistics Post-Filtering/Exclusions

		Control (n = 337)		Treatment (n = 349)		GB
		n	%	n	%	%
Age[1]	18-34	74	22.0	80	22.9	28
	35-54	111	32.9	126	36.1	34
	55-64	56	16.6	55	15.8	15
	65+	94	27.9	87	24.9	23
	Refused	2	0.6	1	0.3	
Gender[1]	Male	161	47.8	172	49.3	49
	Female	174	51.6	174	49.9	51
	Refused	2	0.6	3	0.9	
Ethnicity[2]	White	301	89.3	308	88.3	86
	Asian	21	6.2	20	5.7	8
	Black	6	1.8	9	2.6	3
	Mixed	5	1.5	9	2.6	2
	Other	4	1.2	3	0.9	1
	Refused	0	0.0	0	0.0	
Socio-Economic Group[3]	AB	95	28.2	80	22.9	27
	C1	91	27.0	103	29.5	28
	C2	54	16.0	77	22.1	20
	DE	94	27.9	85	24.4	25
	Refused	3	0.9	4	1.1	
Region[1]	England	289	85.8	282	80.8	87
	Wales	16	4.7	20	5.7	5
	Scotland	24	7.1	29	8.3	8
	Refused	0	0.0	0	0.0	
Smart Meter Ownership[4]	Yes	175	51.9	195	55.9	44
	No	161	47.8	152	43.6	56
	Don't Know	1	0.3	2	0.6	

Note: Percentages may not add up due to rounding. [1]2018 ONS Population Projections[146]. [2]2011 UK Government Ethnicity Facts and Figures[147]. [3]National Readership Survey Social Grades[148]. [4]December 2020 Quarterly Smart Meter Statistics[149, Table 5a]. Includes smart meters in smart and traditional mode.

Further exclusion and filtering criteria were included, namely:

- Non-traders - Respondents who selected the same option (A/B) for all 8 or 7 choice sets were excluded from the study. This removed 100 respondents.
- Manipulation checks - Only respondents who got at least two of the manipulation checks correct were included. This removed 225 respondents.

We found that the majority of respondents were able to answers the first two manipulation checks correctly (Check 1: 78%, Check 2: 84%), with similar proportions across the control (C) and treatment (TR) groups (Check 1: C-78% TR-79%, Check 2: C-85%, TR-83%). However for the final check only 56% of respondents answered correctly[9]. This is possibly due to the wording (differentiating between **can** and **can't**) rather than an understanding of the tasks. As these form the basis of our exclusion criteria, it was decided to filter the sample if respondents got at least two of the three checks correct. In addition, there is significant overlap between the respondents excluded based on the two criteria, suggesting the excluded individuals were indeed less engaged with the choice tasks. The criteria resulted in similar exclusions across the two groups. The final sample used in this study was 337 in the control group and 349 in the treatment group.

As shown in Table 3.4, both the control and treatment groups are broadly representative of GB for the socio-demographic quotas[10]. However both groups have a significant over-representation of smart meter owners, 52% and 56% respectively, compared to 44% for the population at the time[149]. Previous studies on smart meter acceptance and privacy have found that smart meter ownership is a significant indicator of privacy concerns[64]. Given that the installation of smart meters in an opt-in process in the UK smart meter owners are likely to be more trusting of their energy supplier to handle their data.

Although there were no specific quotas for respondents' electricity supply characteristics (beyond smart meter ownership) these may have an impact on respondents' WTP/A for smart meter data privacy. Specifically, consumers with a dual fuel supply will consumer less electricity than if they only had electricity, resulting in lower electricity bills. The resulting monthly savings shown to them during the DCE would be lower. At the same time, the potential privacy inferences may be smaller as some appliances (e.g. gas central heating and gas stoves) and resulting activities (e.g. cooking[11]) are not

[9] A full breakdown can be found in Table B.5 in Appendix B.

[10] A breakdown for the full sample prior exclusions can be found in Table B.1 in Appendix B.

[11] We note that it may still be possible to infer activities such as cooking from other usage patterns of other appliances, as discussed in Section 2.4.2.

included in the electricity usage data. Those on time-varying tariffs (e.g. Economy 7/10 or Octopus Agile) may potentially have a greater awareness of the connection between certain activities and their electricity consumption as they are actively incentivised to shift consumption to lower price periods. Table 3.5 summarises the breakdown of the electricity supply characteristics of the sample and the corresponding population level distribution/estimates. Definitive statistics on the supply type and tariffs are not available for all the categories considered. As such, we estimate these with the best available data as of April 2021. The supply type split is estimated based on the proportion of domestic customers connected to the gas grid in 2021[150]. The split between standard variable, fixed and other tariff types is available for April 2021 for non-prepayment customers[151, Tab 1]. The latest data available for the proportion of pre-payment meters/customers is based on a 2018 OFGEM Request for Information from suppliers[152, p. 49]. Based on these estimates we find a slight over-representation of respondents who only have an electricity connection. Additionally, we observe an under-representation of respondents with a non-standard tariff, such as time-varying tariffs[12].

Respondents' average monthly electricity bill plays a significant role in the DCE. Overall, 66% of respondents provided their (valid) electricity bills with an observed average of £65.82, with the average and proportion providing data being similar across the control (£67.50) and treatment (£64.30) groups, as shown in Figure 3.6b. Those who did not provide their monthly bill or who entered a bill greater than £300/month were assigned £57, the national average bill. Although, our sample average is within a similar range as the average bill used, the skewed nature of the distribution implies that we are likely introducing a systematic error. When including our imputation and exclusion criteria the average bill for the full sample is £62.89, with £65.30 and £60.60 for the control and treatment groups respectively [13]. We are therefore, likely, under-estimating respondents' bills and therefore the WTP/A, with the true effect likely larger due to the skewed nature of the bill distribution across the population. We attempt to address this by focusing on percentage changes in monthly bills rather than absolute values.

Finally, the results of the structured feedback showed that overall 68% of the total sample said they were able to understand all the choices, 61% found the options realistic and 61% found it easy to choose between the options presented[14]. Again, these proportions

[12]Those who do not know the tariff are likely on the default option - a standard variable tariff. Combining these, results in a fairly representative proportion for both the control and treatment groups.

[13]The average bill for the full sample is £62.80, with £63.80 and £61.90 for the control and treatment groups respectively.

[14]A full breakdown can be found in Table B.4 in Appendix B.

Table 3.5: Sample Supply Characteristics Post-Filtering/Exclusions

		Control (n = 337)		Treatment (n = 349)		GB[1]
		n	%	n	%	%
Supply Type	Dual Fuel	257	76%	240	69%	84.9%
	Electricity Only	70	21%	93	27%	15.1%
	Don't Know	10	3%	16	5%	
Tariff	Standard Variable	80	24%	77	22%	36.4%
	Fixed	165	49%	165	47%	28.7%
	Pre-Payment	35	10%	43	12%	16.7%[2]
	Time-of-Use	9	3%	10	3%	18.3%[3]
	Economy 7/10	12	4%	24	7%	
	Don't Know	36	11%	30	9%	

Note: Percentages may not add up due to rounding.

[1] Fuel type data based on proportion of domestic customers connected to the gas grid in 2021[150]. Tariff data from OFGEM for April 2021 [151, Tab 1].

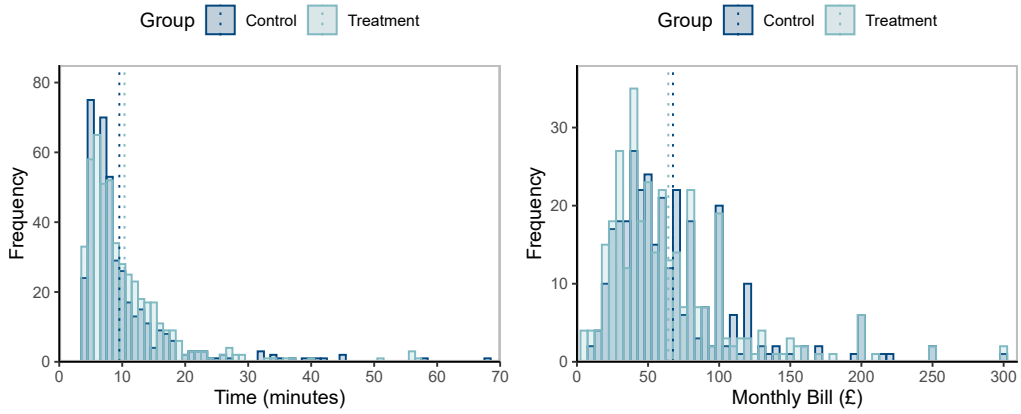
[2] Pre-payment meters estimated at 4.4 million based on latest available OFGEM data [152, p. 49].

[3] Category 'Other non-standard variable tariffs'. Split between Time-of-Use and Economy 7/10 not available.

were similar across both the control and treatment groups. As such, there is a large minority of respondents for whom the information may not have registered. Although this indicates that the results may not be reflective of actual perceptions for this sub-sample, it also suggests that obtaining informed consent given the complex privacy implications of sharing smart meter data is difficult.

3.2.4 Modelling Framework

The DCE provides us with data on respondents stated preferences over the given options. To estimate the willingness-to-pay/accept for the different data sharing options and attributes (anonymisation and frequency) two discrete choice theory models are explored, namely the Multinomial Logit Model (MNL) and the Mixed Logit Model (MXL). In this section we briefly explain the modelling framework assumed by these techniques (summarised from [153]) and detail our chosen model to estimate the WTP/A.



(a) Survey Completion Time (b) Provided Monthly Bills within Sample

Figure 3.6: Distribution of Survey Completion Time and Provided Monthly Bills by Control and Treatment Groups for Full Sample. Dotted Lines Indicate Sample Means.

As described in Section 3.2.2, a DCE presents respondents with options each with different attributes. Random utility theory states that each choice an individual makes results from the individual, n, choosing the alternative, j, with the highest utility, among a set of alternatives, $\mathcal{J}$. We assume here that the individuals derive utility (or dis-utility) from the attributes of each alternative. Utility is a constructed measure and therefore does not have a natural level or scale. As such, we can only derive meaningful information in relative terms. For example, half-hourly data sharing increases a respondents' utility by X *compared to* sharing real-time data. In order to quantify these utilities, we must specify an underlying discrete choice model i.e. a framework which translates options and their attributes into utility functions.

Multinomial Logit Model (MNL)

The MNL aims to estimate the average sample level attribute utilities. Under this model it is assumed that each individual, n, has some true (unobserved) utility for each option, j:

$$U_{n,j} = V_{n,j} + \epsilon_{n,j}, \quad \forall j \in \mathcal{J} \tag{3.1}$$

where, $V_{n,j}$ is the observed or modelled utility, and $\epsilon_{n,j}$ is the error term. A number of assumptions are required to obtain a tractable model that can then be estimated from the DCE data:

- Each $\epsilon_{n,j}$ is assumed to be I.I.D. over n and j[15], following a Gumbel distribution

[15]This effectively means we assume that we have sufficiently modelled the utility V_{nj} and any residual

(Type I Extreme Value) with a variance of $\frac{\pi^2}{6}$. Under this assumption the difference between two errors follows a logistic distribution.

- Utility functions are linear in attributes/parameters i.e. $V_{n,j} = \beta' x_{n,j}$ where, $x_{n,j}$ is a vector of observed attributes relating to alternative j and β is a vector of estimated utility parameters.

- Independence from Irrelevant Alternatives (IIA). A respondents' probability of choosing one choice over another is independent of any other alternatives being available. For example, in our setting if the first option is to share anonymised half-hourly data for a 5% discount and the second option is to share non-anonymised half-hourly data for a 15% discount the introduction of a third option, non-anonymised daily sharing without a discount, should not affect the utility difference between the original two options. Formally, the ratio of probabilities of two choices is independent of any other alternatives.

Under these assumptions the probability of individual n choosing alternative i is then given by:

$$P_{n,i} = \frac{e^{V_{n,i}}}{\sum_{j \in \mathcal{J}} e^{V_{n,j}}} \tag{3.2}$$

The parameters β are then estimated via maximum likelihood estimation (maximising the log-likelihood) to make the observed choice distribution match the modelled choice distribution. For panel data, where we observe multiple choices ($t \in \mathcal{T}$) from each individual the utility specification becomes $U_{n,j,t} = V_{n,j,t} + \epsilon_{n,j,t}, \forall j \in \mathcal{J}, \forall t \in \mathcal{T}$. If we assume $\epsilon_{n,j,t}$ is I.I.D. over t as well, then we have the same model.

Mixed Logit Model (MXL)

The MNL assumes that there is a single population level parameter we wish to estimate for each attribute. However, preferences can vary significantly, resulting in heterogeneity. To capture this we can use the MXL which estimates a distribution of individual respondents' attribute utilities. The true utility can then be formulated as:

$$U_{n,j} = \sum_{k \in \mathcal{K}} \beta_{n,k} x_{n,j,k} + \epsilon_{n,j}, \quad \forall j \tag{3.3}$$

where, $\beta_{n,k}$ now represents the parameter of attribute, k, of individual n. Each attribute parameter, β_k, is assumed to vary over the population following some pre-defined distribution with density $f_k(\beta_k|\theta_k)$. For example, if it follows a location-scale distribution such

errors are random and uncorrelated.

as the normal distribution (i.e. $\beta_k \sim N(\theta_{k,1}, \theta_{k,2})$) it can be decomposed into:

$$\beta_k = \theta_{k,1} + \theta_{k,2} Z_k \tag{3.4}$$

where, Z_k is the standard normal variate. Table 3.6 summarises the properties of the distributions which will be considered for our modelling[16]

Table 3.6: Parameter Distributions and their Properties (summarised from [155])

Distribution	Standard Variate (Z_k)	Decomposition (β_k)	Mean	Range
Normal	$N(0,1)$	$\theta_{k,1} + \theta_{k,2} Z_k$	$\theta_{k,1}$	Unbounded.
Log-Normal	$N(0,1)$	$e^{\theta_{k,1} + \theta_{k,2} Z_k}$	$e^{\theta_{k,1} + \frac{\theta_{k,2}^2}{2}}$	$\beta_{n,k} > 0$
Symmetric Triangular	$U(0,1)$	$\frac{\theta_{k,1}\left(Z_k^1 + Z_k^2\right)}{2} + \theta_{k,2}$	$\frac{\theta_{k,1}}{2} + \theta_{k,2}$	$\theta_{k,2} \leq \beta_{n,k} \leq \theta_{k,1} + \theta_{k,2}$
Log-Uniform	$U(0,1)$	$e^{\theta_{k,1} + \theta_{k,2} Z_k}$	$\frac{e^{\theta_{k,1} + \theta_{k,2}} - e^{\theta_{k,1}}}{\theta_{k,2}}$	$e^{\theta_{k,1}} < \beta_{n,k} < e^{\theta_{k,1} + \theta_{k,2}}$

Importantly, the MXL removes the need for a number of the assumptions required for a MNL. Specifically, it allows for preference variation across the population (this can also be alternatively viewed as allowing for correlated errors) and does not require the IIA. The flexibility of the MXL allows any random utility model to be approximated to any degree of accuracy. However, this comes at the cost of computational complexity and convergence guarantees.

The resulting probability of individual, n, choosing alternative i is then given by:

$$P_{n,i} = \int \frac{e^{V_{n,i}(\beta)}}{\sum_{j \in \mathcal{J}} e^{V_{n,j}(\beta)}} f(\beta) d\beta \tag{3.5}$$

Due to the integrals, there is no closed form for choice probabilities or the resulting log-likelihood function which we aim to maximise. The log-likelihood function, in general, is also not globally concave and therefore susceptible to local minima. As a result, the model needs to be estimated via Monte Carlo simulation, by drawing from $f(\beta)$ to estimate the integrand and repeating the estimation process with multiple start values. A detailed description of estimation methods can be found in [153, Chapter 8 & 10].

[16]This is not an exhaustive list of potential distributions. However it covers some of the most popular distributions considered in the discrete choice modelling literature[154].

Attribute and Utility Specification

In this section we formalise our model which builds on the MXL described above. We define the modelled/observed utility of each option, j, for each respondent, n, in each choice task, t, as:

$$\begin{aligned} V_{n,j,t} =& \alpha_{n,1} FeeP_{j,t} + \alpha_{n,2} DiscP_{j,t} + \beta_{n,1} HH_{j,t} + \beta_{n,2} D_{j,t} + \beta_{n,3} Anon_{j,t} \\ &+ \beta_{n,4} Anon_{j,t} \times HH_{j,t} + \beta_{n,5} Anon_{j,t} \times D_{j,t} + \epsilon_{n,j,t} \end{aligned} \tag{3.6}$$

where, $\alpha_{n,\star}$ are the coefficients of the monetary parameters, $\beta_{n,\star}$ are the coefficients for the non-monetary parameters and $\epsilon_{n,j,t}$ is the error due to unobserved factors. The parameters, $\alpha_\star$, and $\beta_\star$ are assumed to follow a pre-defined distribution for which the parameters θ will be estimated. For single parameter distributions (e.g. a symmetric zero-bounded triangular distribution) $\theta = \theta_1$. Similarly, for two parameter distributions (e.g. a normal distribution) $\theta = [\theta_1, \theta_2]$. For the MNL, the coefficients are fixed over n, simplifying the specification.

To account for the potential 'super-endowment' effect we split the expected bill changes between an effective fee (for an increase in the bill) and discount (for a decrease in the bill)[156]. The fee and discount were coded as continuous variables using percentage changes in the bill as opposed to £ values to ensure robustness against the skewed bill distribution observed in the sample. The frequency and anonymisation attributes were dummy coded. In the utility function specified above, $FeeP_{j,t}$ and $DiscP_{j,t}$ are the fee and discount in percentage points respectively. $HH_{j,t}$ indicates half-hourly non-anonymised data sharing, $D_{j,t}$ indicates daily non-anonymised data sharing, $Anon_{j,t}$ indicates anonymised data sharing, $Anon_{j,t} \times HH_{j,t}$ indicates the interaction between half-hourly and anonymised data sharing and $Anon_{j,t} \times D_{j,t}$ indicates the interaction between daily and anonymised data sharing. All utilities are measured against a reference, in this case real-time non-anonymised data sharing. We can then construct the change in utility compared to this reference by adding the corresponding attribute coefficients. For example, the utility of anonymised half-hourly data sharing for individual, n, (with no change in costs) would be $\beta_{n,1} + \beta_{n,3} + \beta_{n,4}$ greater than the utility derived from sharing real-time non-anonymised data.

The Willingness-to-Pay/Accept (WTP/A) can be estimated as the ratio, $\frac{\beta}{\alpha}$, of the data sharing attributes (frequency and anonymisation - β) and the cost variable (expected change in bill - α). As we aim to quantify both WTP and WTA, we perform the modelling in the preference space, as specified by the utility function in (3.6), so as to avoid the imposition of different distributional assumptions on the WTP/A values that would

occur in the WTP space[157]. For the MNL, calculating the ratios is achieved using the Delta method with a classical covariance matrix which accounts for the introduced error propagation as both α and β are estimated from data[158]. However, for the MXL, we would be estimating the ratio of two distributions. In most cases there are no closed-form expressions for the moments of this ratio. As a result, we employ the so-called Krinsky-Robb method[159]. This involves calculating the ratio and the associated confidence intervals via simulation. The distribution parameters ($\theta_{k,1}$, $\theta_{k,2}$) are first sampled and then used to generate the distribution of coefficients, β_k. The ratios between the coefficients are then computed and the confidence intervals are calculated via bootstrapping.

Base Model Specification

Given that privacy preferences vary across the population we measure, the willingness-to-pay/accept for the different data sharing options the MXL model is of particular interest. The choice of distributions can have a significant impact on estimates. As such, a number of popular distributions (normal, log-normal, log-uniform, and triangular) were explored for the fee and discount parameters. The coefficients of data sharing attributes were assumed to follow normal distributions, so as not to impose any restrictions on the sign and magnitude. To estimate the MXLs it is necessary to draw random samples to simulate the distributions. We use 1,000 draws generated using the Modified Latin Hypercube Sampling (MLHS) algorithm [160] to avoid potential collinearity issues associated with the use of Halton draws when more than 5 random parameters are estimated[161]. The models were operationalised using the `Apollo` R package[155] and estimated using [162].

Table 3.7 summarises the model estimates for the MNL and the MXL with different assumptions for the monetary parameters. The MNL (MNL_B) results have a number of insignificant parameters (Anon and Anon x HH) as well as one with the wrong (against expectation) sign (HH, though insignificant). This suggests that there may be significant heterogeneity in the utility for the different data sharing options among respondents. We see that allowing for heterogeneity using a MXL (MXL_F) significantly improves model fit as measured by the pseudo-R^2. Additionally, when allowing for heterogeneity in the monetary parameters the sign of all the data sharing options are correct and significance of the parameter estimates is improved. This is in line with previous studies on privacy where significant heterogeneity was observed[143], [161], [163].

Although the log-normal price model (MXL_{LN}) has the best fit (in terms of the pseudo-R^2 and information criteria) it leads to very high, implausible WTP/A estimates as shown

Table 3.7: Mixed Logit Models under Different Distributional Assumptions for Cost Parameters

		MNL_B	MXL_F	MXL_{TRI}	MXL_N	MXL_{LN}	MXL_{LU}
	Fee(%)	-0.069*** (0.006)	-0.104*** (0.008)	-0.294*** (0.024)	-0.278*** (0.027)	-1.644*** (0.148)	1.440*** (0.357)
	Discount(%)	0.045*** (0.004)	0.063*** (0.005)	0.161*** (0.013)	0.123*** (0.014)	-6.317*** (1.377)	4.410*** (1.129)
	HH	-0.118 (0.128)	-0.120 (0.171)	0.032 (0.182)	-0.179 (0.252)	0.160 (0.213)	0.133 (0.213)
$\theta_{k,1}$	D	0.598*** (0.110)	0.926*** (0.139)	0.923*** (0.155)	1.069*** (0.222)	1.006*** (0.171)	1.014*** (0.171)
	Anon	0.134 (0.135)	0.260 (0.187)	0.489* (0.200)	0.343 (0.273)	0.630** (0.231)	0.576* (0.232)
	Anon X HH	0.378 (0.253)	0.481 (0.333)	0.318 (0.355)	1.019* (0.494)	0.730+ (0.416)	0.765+ (0.417)
	Anon X D	-0.669*** (0.190)	-1.025*** (0.243)	-0.860** (0.268)	-0.771* (0.379)	-0.598* (0.305)	-0.606* (0.306)
	HH		0.620*** (0.084)	0.652*** (0.089)	0.825*** (0.133)	0.567*** (0.118)	0.599*** (0.117)
	D		0.025 (0.466)	0.069 (0.438)	0.777*** (0.223)	0.339 (0.229)	0.331 (0.231)
$\theta_{k,2}$	Anon		1.621*** (0.077)	1.686*** (0.085)	1.896*** (0.136)	1.485*** (0.112)	1.509*** (0.115)
	Fee(%)				0.295*** (0.027)	2.337*** (0.275)	6.288*** (0.676)
	Discount(%)				0.251*** (0.019)	8.293*** (2.071)	21.649*** (4.933)
Cost Disttribution		Fix	Fix	Triangle	Normal	Log-Normal	Log-Uniform
	n Ind	686	686	686	686	686	686
	n Obs	5488	5488	5488	5488	5488	5488
	AIC	6786	5999	5887	5396	5378	5381
	BIC	6832	6065	5953	5476	5457	5460
	LL	-3386	-2989	-2933	-2686	-2677	-2678
	pseudo-R2	0.110	0.214	0.229	0.294	0.296	0.296
	Adj pseudo-R2	0.108	0.212	0.226	0.291	0.293	0.293

Note: Standard errors shown in parentheses. + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

For MXLs HH (Half-Hourly), D (Daily), Anon (Anonymised) are distributed normally whereas Anon x HH and Anon x D are assumed to be fixed. Fee(%) and Disc (%) follow the distribution specified in the heading.

Triangle distribution is symmetric and zero bounded resulting in $\theta_{FEE,2} = \theta_{DISC,2} = 0$.

For all normally distributed parameters the $\theta_{k,1}$ and $\theta_{k,2}$ are the mean and standard deviation of the distributions, respectively.

For Log-Normal and Log-Uniform the $\theta_{k,1}$ and $\theta_{k,2}$ are the location and scale parameters the underlying normal and uniform distributions, respectively.

Table 3.8: Mean Willingness-to-Pay/Accept under Different Distributional Assumption on Cost Parameters

	MNL_B		MXL_F		MXL_N		MXL_{LN}		MXL_{TRI}		MXL_{LU}	
	WTP	WTA	WTP	WTA	WTP	WTA	WTP	WTA	WTP	WTA	WTP	WTA
HH	-1.7	-2.63	0.93	2.27	-0.1	11.35	26.34	5.35E+34	0.28	0.57	3.52	5.31E+9
	[1.99, -5.4]	[-8.23, 2.97]	[-2.42, 3.99]	[-5.51, 9.83]	[-55.51, 36.54]	[-82.6, 107.09]	[-94.85, 142.92]	[-1.53E+27, 2.17E+27]	[-2.59, 3.09]	[-4.61, 5.83]	[-5.17, 14.08]	[-1.96E+10, 4.00E+10]
D	8.63	13.32	5.98	13.91	3.18	-7.41	158.85	1.29E+35	8.80	15.90	25.70	4.15E+10
	[12.61, 4.66]	[9.00, 17.65]	[2.94, 9.28]	[7.77, 19.64]	[-31.6, 43.62]	[-112.15, 80.58]	[73.56, 321.25]	[-4.21E+24, 1.50E+28]	[5.82, 12.1]	[12.0, 20.0]	[12.65, 44.93]	[2.06E+9, 1.58E+11]
Anon X RT	1.94	2.99	3.72	8.85	-0.86	-4.58	133.38	-5.66E+34	4.63	8.53	14.75	2.29E+10
	[5.69, -1.82]	[-2.90, 8.87]	[0.51, 6.9]	[1.41, 17.12]	[-24.63, 21.56]	[-44.39, 40.92]	[33.06, 291.16]	[-5.33E+24, 7.77E+27]	[1.54, 7.69]	[2.77, 14.4]	[3.96, 30.38]	[6.88E+9, 1.02E+11]
Anon X HH	5.7	8.79	4.22	9.93	5.29	14.31	264.12	4.01E+34	7.93	14.5	37.49	5.81E+10
	[7.44, 3.96]	[6.29, 11.3]	[2.53, 5.92]	[5.98, 13.94]	[-39.34, 45.72]	[-86.85, 104.71]	[166.99, 491.09]	[-1.56E+23, 1.75E+28]	[6.19, 9.80]	[11.5, 17.8]	[19.36, 63.15]	[2.77E+9, 2.14E+11]
Anon X D	0.9	1.39	-2.25	-5.15	1.93	40.06	200.01	1.12E+35	5.18	9.64	25.16	3.91E+10
	[3.42, -1.61]	[-2.66, 5.45]	[-5.98, 1.1]	[-13.03, 2.87]	[-29.5, 44.76]	[-60.19, 63.11]	[118.75, 379.68]	[-2.10E+23, 1.61E+28]	[3.22, 7.00]	[5.31, 14.2]	[12.00, 44.68]	[8.95E+8, 1.43E+11]

Note: Mean WTP/A against reference of Non-Anonymised Real-Time data sharing, expressed in % change in monthly bill. RT (Real-Time), HH (Half-Hourly), D (Daily), Anon (Anonymised). 95% confidence interval shown in parentheses. MNL values generated using Delta method. MXL values generated using Krinsky-Robb Method with 100,000 draws of coefficients accounting for full covariance matrix, repeated 10,000 times.

in Table 3.8. A similar issue is observed for log-uniform prices (MXL_{LU}). The normal price model (MXL_N) provides more plausible estimates. However from a theoretical standpoint, the resulting WTP/A distributions have undefined moments (as the normal distribution allows for a zero cost coefficient)[154]. As a result, we select the triangular distribution (MXL_{TRI}), which provides an improved model fit compared to the MNL, plausible WTP/A estimates and a defined mean for the WTP/A distribution.

Given the heterogeneity, we also note that there may be respondents who did not consider certain attributes, effectively having a coefficient of zero for these attributes. This attribute non-attendance is not well accommodated when using uni-modal and long-tailed distributions like the log-normal[164]. Indeed, we see that a large proportion of respondents always choose cheaper options indicating no value in smart meter data privacy and a smaller group always choose the higher privacy option, as shown in Figure 3.7. The rest exhibit a trade-off which is skewed towards cheaper options. We do not

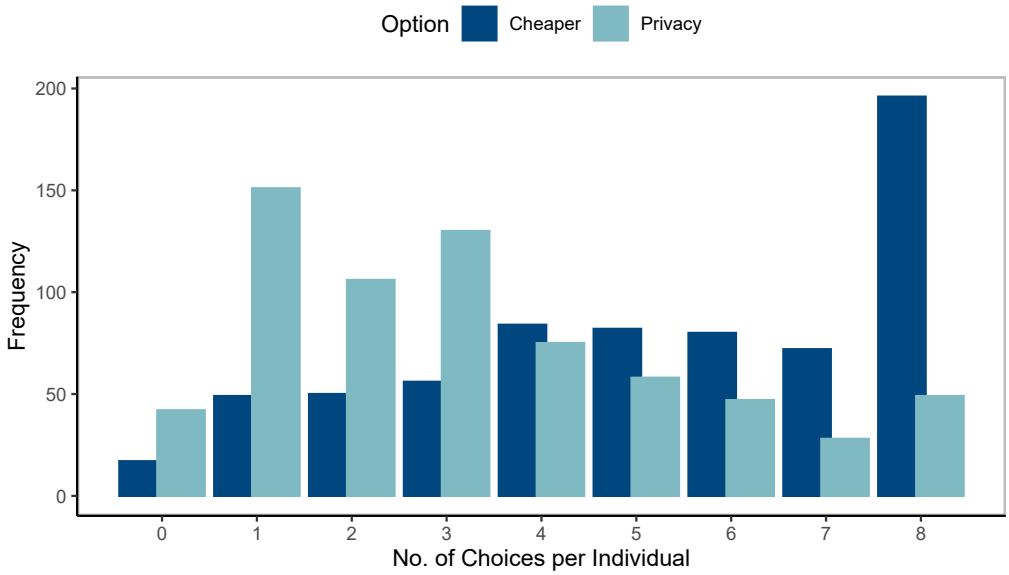


Figure 3.7: Histogram of Respondents by the Number of Times each Respondent Choose the Cheaper or Higher Privacy Option out of their 8 Choice Tasks.

investigate this effect further here, however, future work could investigate the use of flexible distributions to better accommodate the attribute non-attendance[164], [165]. We present results, based on the triangular MXL, first for the complete sample, including both the control and treatment group and then discuss the effect of the treatment in Section 3.3.4, by measuring interaction terms.

Interactions

The MXL allows us to model heterogeneity in preferences through a distribution, however, the base model does not tell us what the drivers of this heterogeneity are. In order to understand these, we include interaction terms such as respondents' socio-demographic characteristics. For example, we can quantify the effect of age on the WTP/A for anonymisation by modifying the specification of β_3 in (3.6) to:

$$\beta_3 = \theta_{3,1} + \theta_{3,2}Z_3 + \theta_{3,3}AGE^{55+} \tag{3.7}$$

where, AGE^{55+} is a dummy variable which is 0 if respondent n is below 55 and 1 if they are 55 or older.

However, MXLs with interaction terms increase complexity considerably and can cause differing results depending on which interaction terms are included[27]. As such, we choose to investigate heterogeneity, as well as the effect of the treatments using a MNL with interaction terms.

To investigate the drivers of heterogeneity we consider the following categories based on insights from existing survey work as well as other questions asked in our survey:

1. Attitude to Data Sharing (see Section 3.2.2): BA - only basic data sharing, MR - marketing and research and TP - sharing with third parties (base = TP).

2. Demographics: Age (base = under 55s), Gender (base = male/other) and SEG (base = DE).

3. Electricity Supply: Smart meter ownership (base = owns a smart meter), TV-Tariff - whether they are on a time-varying tariff such as Economy 7/10 or another time-of-use tariff (base = do not have a time-varying tariff), IHD - IHD owners who engage with their IHD more than once a week (base = IHD owners who engage once a week or less and those who do not own an IHD).

4. Level of Understanding: FEED - strongly or somewhat agreed that they were able to understand the choices (base = strongly or somewhat disagreed + neither agreed nor disagreed + didn't know), MANIP - got all three manipulation check statements correct (base = those who only got less than two correct).

5. Effect of the treatment: TR - treatment group (base = control group).

In each case, the base or reference group is the group expected to have a lower WTP/A based on existing literature. We would therefore expect to see positive coefficients for all

Table 3.9: Multinomial Logit Models Post Filtering/Exclusions

	MNL_B	MNL_{HET}	MNL_{TR}	MNL_{TRxSH}
Fee(%)	-0.069 (0.006)***	-0.072 (0.006)***	-0.072 (0.006)***	-0.072 (0.006)***
Discount(%)	0.045 (0.004)***	0.046 (0.004)***	0.046 (0.004)***	0.046 (0.004)***
Half-Hourly	-0.118 (0.128)	-0.147 (0.130)	-0.147 (0.130)	-0.146 (0.130)
Daily	0.598 (0.110)***	0.603 (0.111)***	0.601 (0.111)***	0.601 (0.111)***
Anon	0.134 (0.135)	-0.728 (0.170)***	-0.793 (0.173)***	-0.361 (0.177)*
Anon X HH	0.378 (0.253)	0.429 (0.257)+	0.429 (0.257)+	0.423 (0.257)
Anon X Daily	-0.669 (0.190)***	-0.661 (0.193)***	-0.659 (0.193)***	-0.664 (0.193)***
Anon X AGE 55+		0.184 (0.062)**	0.187 (0.062)**	0.184 (0.062)**
Anon X Female		0.368 (0.061)***	0.371 (0.061)***	0.361 (0.062)***
Anon X SEG C1C2		-0.066 (0.072)	-0.074 (0.073)	-0.071 (0.073)
Anon X SEG AB		0.105 (0.084)	0.107 (0.084)	0.097 (0.084)
Anon X No SM		0.236 (0.071)***	0.237 (0.071)***	0.231 (0.071)**
Anon X TV Tariff		0.476 (0.110)***	0.467 (0.111)***	0.462 (0.111)***
Anon X IHD		0.152 (0.082)+	0.147 (0.082)+	0.145 (0.082)+
Anon X FEED		0.134 (0.071)+	0.142 (0.071)*	0.147 (0.071)*
Anon X MANIP		0.114 (0.062)+	0.122 (0.062)*	0.126 (0.062)*
Anon X MR		0.539 (0.082)***	0.541 (0.082)***	
Anon X BA		0.463 (0.068)***	0.465 (0.068)***	
Anon X TR			0.113 (0.060)+	
Anon X TR X TP				-0.383 (0.085)***
Anon X C X TP				-0.345 (0.086)***
Anon X TR X BA+MR				0.250 (0.083)**
n Ind	686	686	686	686
n Obs	5488	5488	5488	5488
AIC	6786	6653	6651	6646
BIC	6832	6772	6777	6772
LL	-3386	-3309	-3307	-3304
pseudo-R^2	0.110	0.130	0.131	0.131
Adj pseudo-R^2	0.108	0.126	0.126	0.126

Note: Standard Errors shown in Parentheses.+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

these interaction terms. The MNLs (MNL_B - base model without interactions, MNL_{HET} - with interactions excluding treatment effect, MNL_{TR} - with interactions excluding treatment effect, and MNL_{TRxSH} - with all interactions) can be found in Table 3.9[17]. The next section analyses the results using the MXL with triangularly distributed cost parameters (MXL_{TRI}) and the MNLs with interactions (MNL_{HET},MNL_{TRxSH}).

3.3 Results

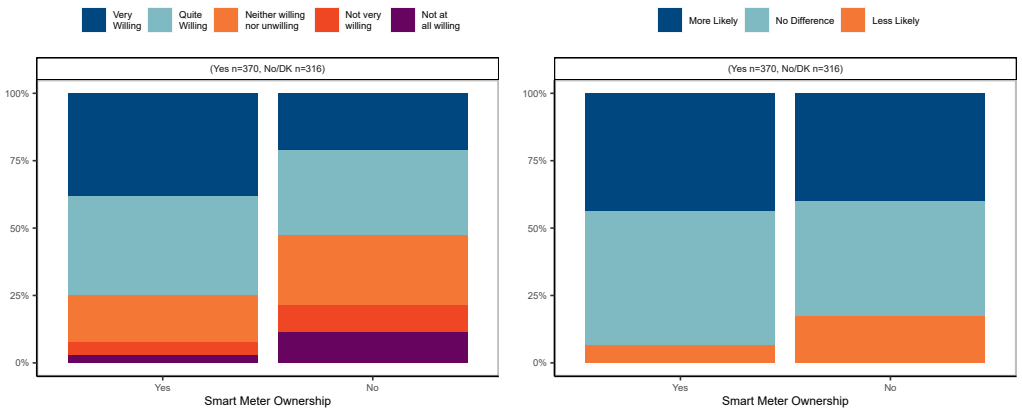
The following sections will discuss the results of the survey and their policy implications. We have applied a mixed-methods approach, complementing our quantitative analysis

[17]The corresponding MNLs for the full sample, without exclusions, can be found in Table B.3 in Appendix B.

from the discrete choice experiment with the respondents' answers to the open-ended feedback question.

3.3.1 The Value of Anonymisation

We begin by considering respondents' pre-treatment willingness-to-share, as this provides a direct comparison to existing studies. Prior to the treatment we find that 62% of respondents are willing to share their half-hourly smart meter data, with those with smart meters being more willing (see Figure 3.8). This matches closely with survey results obtained by OFGEM as part of their consultation on MHHS[24]. Post treatment, we find that 41% would be more willing to share their smart meter data if it were anonymised. Given that many of the uses of smart meter data, such as, improving load forecasting, tariff setting and network management do not require non-anonymised data, there is significant potential to widen access to high resolution smart meter data while ensuring consumer privacy using anonymisation.



(a) Pre-Treatment Willingness-to-Share

(b) Post-Treatment Change in Willingness-to-Share if Data were Anonymised

Figure 3.8: Respondents Willingness-to-Share their Half-Hourly Data and the Effect of Anonymisation Split by Smart Meter Ownership.

Although the general WTS smart meter data is high, this is mainly amongst smart meter owners[18]. Existing studies suggest that this is likely driven by the presumed benefits of sharing this data in terms of improving system operation and reducing environmental impacts[24], [137]. We saw this sentiment expressed by respondents[19]:

[18] Approximately 47% of households in GB had a smart meter at the time of the survey[149]

[19] Quotes are presented using the following coding - ID, Treatment (TR) or Control (C), Gender, Age. Where a respondents refused to answer the socio-demographic question we used R.

"Anything to help reduce cost and can provide ways to save energy."
(10310, C, M, 55-64)

Figure 3.9 shows the mean WTP (the average fee) for respondents to avoid sharing real-time non-anonymised data and the mean WTA (the average discount) required to incentivise respondents to share real-time non-anonymised data. For half-hourly data the WTP is small (0.3% of monthly bill or ~£0.16). This can be explained by relatively similar information that is shared under the two options albeit with significantly higher accuracy in the case of real-time data. Similar results have been observed in other studies where consumers don't tend to differentiate between half-hourly and real-time data (see Figure 3.1). Although much of the information is similar, access to real-time data can be very beneficial for novel business models currently being explored. Energy as a service, demand response pricing, remote control of assets and local energy system all require access to high resolution data to better predict usage patterns and identify flexibility.

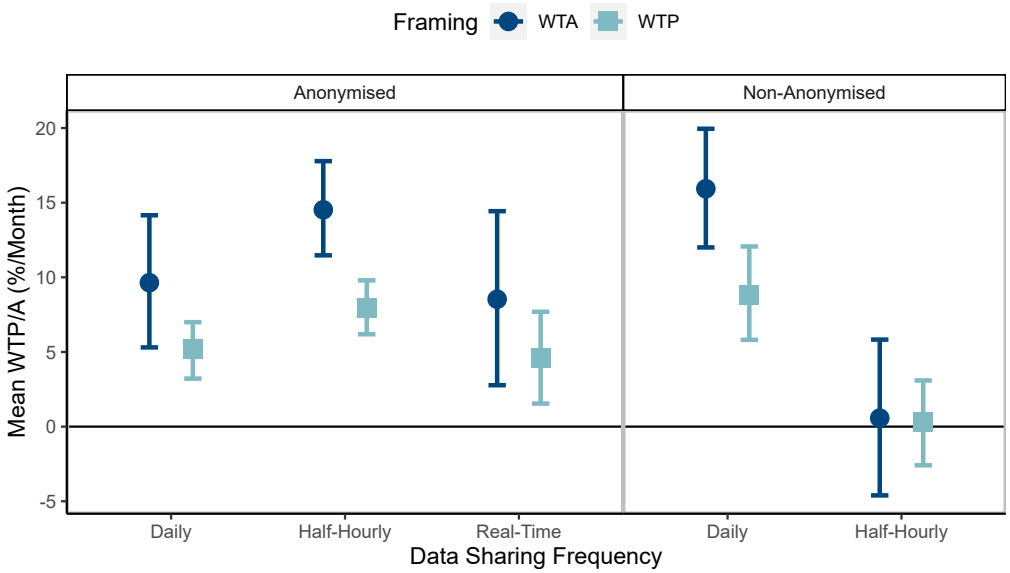


Figure 3.9: Mean Willingness-to-Pay/Accept, in % of Monthly Bill, for Different Data Sharing Options. Reference Option is Non-Anonymised Real-Time Data. Error Bars Represent 95% Confidence Intervals. Generated from MXL_{TRI} using Krinsky-Robb Method with 100,000 Draws of Coefficients with Full Covariance Matrix Simulated 10,000 Times.

There is a clear preference for sharing data at a daily frequency with consumers willing to pay on average up to 8.80% (~£5.01[20]) of their monthly bill and demanding a discount of up to 15.93% (~£9.08). The WTP figures are in line with extant consumer valuations for

[20]Based on an average bill of £57/month.

sharing personal information that vary between from £5.30[27] to £7.27[26], whereas the WTA figures are significantly higher. The differences in preferences regarding frequency of data sharing become less evident when data is anonymised, as illustrated by the overlapping confidence intervals. Daily sharing, even without anonymisation, has a high valuation, possibly suggesting it is viewed as sufficient protection.

However, we see a range of perspectives among the respondents expressed through the open-ended feedback question. We see opinions expressed at both extremes. Some saying they would not give up anonymity under any conditions, regardless of fees or discounts:

> "Given the choices, I found that I prefer NOT to have lots of information go to my supplier if this was NOT anonymised, even if that meant I would be charged more. I had thought I would be influenced entirely by cost, but when asked to choose, I found I didn't like the idea of that quantity of information being readily available when it could be identified directly to me/us as a household."
> (10492, TR, F, 75+)

> "I value my privacy at any cost even if it means losing out financially."
> (10967, C, M, 55-64)

Others who felt the incentives presented to them were insufficient:

> "Anonymisation has a big value, bigger than a discount."
> (10071, C, M, 18-34)

> "These are trifling amounts when one considers what is at stake : personal privacy.[...]"
> (10231, TR, M, 55-64)

As well as, some who asserted that anonymity should be provided by default:

> "[...] I wouldn't pay more for my electricity to stay anonymous. That should be free."
> (11045, C, M, 35-54)

> "It was difficult to decide when you had to pay a lot more for anonymisation. I don't feel you should pay more for it."
> (10124, TR, F, 35-54)

3.3.2 Privacy by Design

Obtaining informed consent in the case of high-resolution smart meter data, specifically, for uses outside of the regulated activities outlined in the DAPF, is difficult, as the extent of potential privacy infringements are not known[7], [91]. The DAPF is supplementary to existing data protection legislation (UK Data Protection Act 2018 and GDPR)[65]. As such, access to, as well as processing and usage of smart meter data, considered personal and identifiable information, is also subject GDPR. The ICO, which regulates data protection in the UK, has raised concerns regarding the current SMIP and specifically around the latest OFGEM consultations on MHHS[57]. The ICO also sets out guidelines on how to adhere to existing data protection regulations and are proponents of the Privacy by Design approach[166]. The Privacy by Design framework requires the default or zero price option to be the maximum privacy option. Indeed, we saw from the previous quotes that some respondents also expect this to be the case. In the case of smart metering this would be daily data sharing. Energy suppliers would then have to incentivise consumers through discounts or services to access more granular data. In such a policy framing the willingness-to-accept rather than the willingness-to-pay is of interest.

As discussed in other studies the endowment effect, exemplified by a higher WTA than WTP is particularly pronounced when in relation to privacy valuations[28]. On average, we find that the ratio is 1.83 [1.49, 2.26] for anonymisation of real-time data with similar ratio for the other data sharing options.

To investigate the effect of policy framing, we simulate choices based on the estimated utility distributions from the MXL_{TRI} model. Here, we again employ the Krinsky-Robb method simulation approach. Coefficients are sampled 1,000 times and used to build the utility functions for 10,000 draws of these coefficients. The expected market share is then represented by the average choices made over the coefficient sampling. Two scenarios are investigated: (1) daily data sharing is the default option with a discount offered to consumers for sharing higher resolution data, and (2) real-time sharing is the default with a fee demanded on consumers to share lower resolution data. We assume the half-hourly option attracts $1/2$ the discount of the real-time option in the first scenario or $1/2$ the fee of the daily option in the second scenario.

From Figure 3.10 it is clear that the choice of policy framing and whether or not to offer the option of anonymising data can have a large impact on data sharing. Without anonymisation and a default option of daily sharing (as is currently the case under the DAPF) we see that most consumers are unwilling to share high resolution data. A 20%

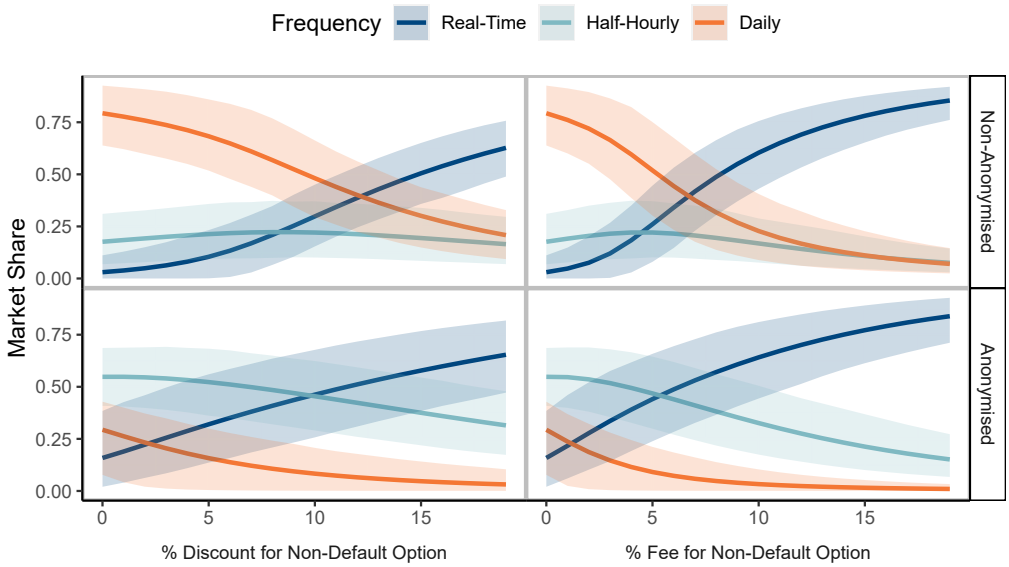


Figure 3.10: Expected Market under Different Framing Options by Fee/Discount and Anonymisation. Shaded Regions Represent 95% Confidence Intervals. Generated from MXL_{TRI} using Krinsky-Robb Method with 10,000 Draws of Coefficients with Full Covariance Matrix to Produce Utility Functions and Simulated 1,000 Times.

discount decreases this from 79.3% to 20%. If, however data is anonymised only 3% continue to prefer the daily option suggesting a significant increase in the availability of high resolution data is possible with anonymisation. When real-time data sharing is the default option the influence of monetary incentives are more pronounced. In both the anonymised and non-anonymised cases the share of consumers sticking to daily sharing if they have to pay an additional 20% on their electricity bills drops to less than 0.01%.

3.3.3 Understanding Preference Heterogeneity

The utility estimates for the MXL models in Table 3.7 suggest that there is significant heterogeneity among respondents for the different data sharing options. As depicted in Figure 3.11, the MXL_{TRI} model estimates a distribution of coefficients which we then use to estimate the WTP and WTA distributions. We note here that the WTA estimates are not only higher, but are also more widely spread, due to the increased variations in price sensitivities for a discount in comparison to a fee.

To attempt to explain the heterogeneity in WTP for the different data sharing options, we estimate a MNL model with a number of individual-specific interaction terms, as described in Section 3.2.4. Figure 3.12 shows the percentage point change in WTP/A for different sample subsets compared to a corresponding reference group.

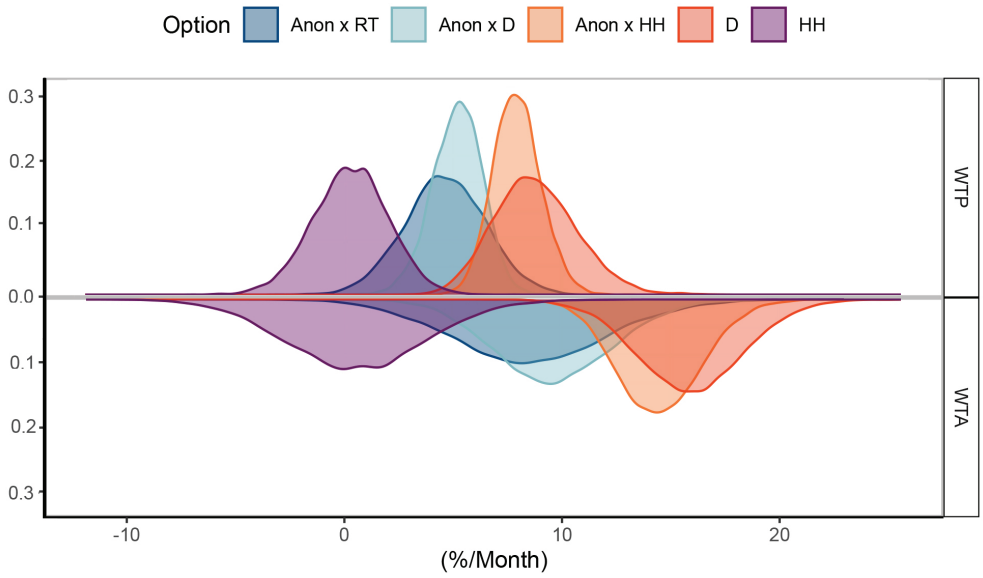


Figure 3.11: Simulated Willingness-to-Pay/Accept Distributions for Different Data Sharing Options. Reference Option Non-Anonymised Real-Time Data Sharing. Generated from MXL_{TRI} using Krinsky-Robb Method with 100,000 Draws of Coefficients with Full Covariance Matrix Simulated 10,000 Times.

We see that respondents who normally share only basic data when using services have a much higher value for anonymisation (an increase of up to 10.0% in WTA). Similarly, those who shared data for marketing and research purposes also had a higher WTP/A compared to those who shared data with third parties.

Those in higher SEGs (AB compared to DE) were found to have slightly higher WTP/A, although this is not observed for C1C2. This apparent lower level of concern for privacy among lower SEGs was also observed in previous studies. Although this is sometimes attributed to a lack of knowledge of the potential implications of data sharing, it is likely that financial constraints play also play a significant role in our study. This is exemplified by one of the respondents in our sample:

> "My options were based on costs, I can't afford to pay more."
> (10580, TR, F, 75+, SEG = DE)

Another interesting observation made in [91], was that those from lower socio-economic backgrounds were already heavily monitored and as such may have higher concerns, but felt they did not have any choices.

Conversely, on average, older respondents and women have higher WTP/A for privacy. Those aged 55 and above had a 4.0% point higher WTA than younger respondents,

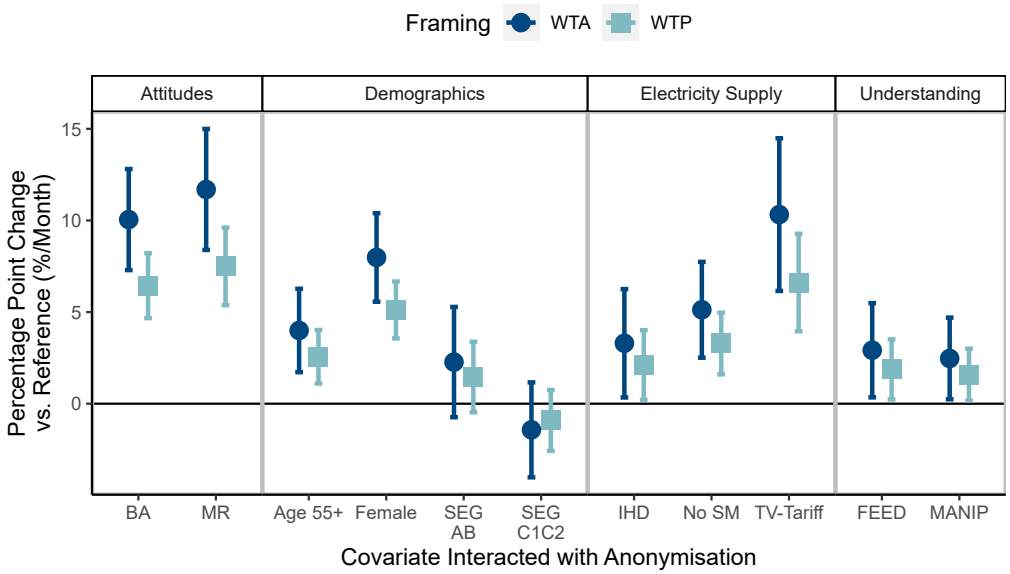


Figure 3.12: Effect of Respondent Characteristics on Willingness-to-Pay/Accept for Anonymisation. Interaction Effects shown in Percentage Point Change in % of Monthly Bill. Error Bars Represent 90% Confidence Intervals. Generated using Delta Method with a Classical Covariance Matrix from MNL_{HET}. Reference Group (left to right): TP, TP, AGE $<$ 55, Male, SEG DE, SEG DE, Engage with IHD once a week or less and those without an IHD, Smart Meter Owners, Non-TV-Tariff.

where as women had a 8.0% point increase in WTA compared to men. This is in line with existing literature on privacy in general and has been posited to be due to the lower levels of trust and unfamiliarity with technology[64]. As discussed, those who do not have smart meters (No SM) tend to have higher privacy concerns and therefore a higher WTP/A. We observe a 5.1% point increase in WTA for those do not have a smart meter as compared to those with a smart meter. This finding also indicates that privacy concerns, at least in part, play a role in consumers resistance to getting a smart meter.

Interestingly, those who are on time-varying tariffs (TV-Tariff) and those who engage with their IHD more than once a week (IHD) also have higher WTP/As. These groups are arguably most aware of the connection between their daily activities and their energy consumption. As such, they may be more aware of the personal information embedded within smart meter. Those with a time-varying tariff had a significantly higher WTA than those on single rate tariffs such as a standard variable tariff with a percentage point increase of 10.3%. Table 3.10 shows a breakdown of the level engagement with their IHD amongst those who knew they owned one. A majority of IHD owners in our sample, 57.7%, do engage with their IHD frequently. Those highly engaged with their IHD had a

Table 3.10: Engagement among IHD Owners Post-Filtering/Exclusions

	Control (n = 125)		Treatment (n = 156)	
	n	%	n	%
Daily	43	34.4	56	35.9
2-3 times a week	29	23.2	34	21.8
Once a week	19	15.2	25	16.0
Once a month	2	1.6	10	6.4
Less than once a month	13	10.4	11	7.1
Never	19	15.2	20	12.8

Note: Percentages may not add up due to rounding.

modest increase of 3.3 percentage points.

Finally, we considered the extent to which respondents understood the survey. First, based on their self-reported understanding (FEED), specifically those who agreed that they were able to understand the choices presented to them. Second, using the manipulation checks (MANIP), specifically those who got all three manipulation checks correct. We find that respondents with a greater level of understanding, both the self-reported and measured, have a small but significant (2.92% and 2.47% point increase for FEED and MANIP, respectively) increase in WTP/A compared to those with a lower level of understanding. This suggests that messaging and the ease with which options can be accessed and understood affect decision-making. The resulting information asymmetries are therefore likely impacting privacy perceptions. The next section will focus on the treatment to further explore and quantify the extent and effect of information asymmetries.

3.3.4 Information Asymmetry and Informed Consent

We now consider the effect of our treatment, specifically, providing respondents with information on the privacy implications of sharing smart meter data. A handful of existing studies showed that the WTS is affected by such interventions ([135], [136]), however we focus on the effects on WTP/A.

First, we note that regardless of sharing or payments, a significant proportion of smart meter owners are unaware of their current data sharing options, as shown in Table 3.11. Over 40% and 30% of respondents in the control and treatment groups, respectively, did not know what their current data sharing options were. This is not unique to our sample,

Table 3.11: Actual Data Sharing Choices amongst Smart Meter Owners

Resolution	Control (n = 175)		Treatment (n = 195)	
	n	%	n	%
Half-Hourly	26	14.9	42	21.5
Daily	47	26.9	57	29.2
Monthly	31	17.7	34	17.4
Don't Know	71	40.6	62	31.8

Note: Percentages may not add up due to rounding.

with similar results observed in the 2019 Citizens Advice survey[64]. In addition, we found that a significant proportion of smart meter owners did not have or did not know if they had an IHD (see Table 3.12). This finding also has implications for BEIS's cost-benefit analysis of the SMIP. The expected energy savings from informational feedback on energy usage constitutes a large proportion of the benefits. However, we see that up to 24% of smart meter owners are not receiving this informational feedback.

Table 3.12: IHD Ownership among Smart Meter Owners Post-Filtering/Exclusions

	Control (n = 175)		Treatment (n = 195)	
	n	%	n	%
Yes	125	71.4	156	80.0
No	43	24.6	35	17.9
Don't Know	7	4.0	4	2.1

Note: Percentages may not add up due to rounding.

We find that the sample generally has low privacy concerns with 48% and 51% of the control and treatment group regularly allowing their personal information to be accessed by third parties, respectively (see Table 3.13). Although there are no definitive statistics on the distribution of privacy concerns for the GB population, other studies on privacy, especially those conducted through face-to-face or telephone interviews, suggest this proportion is significantly smaller in the general population. For example, a 2018 study on consumer attitudes to data collection and use found that 81% were concerned about organisations selling their data to third parties and only 13% had no concerns[167].

The current DAPF and GDPR regulations are designed around consumers providing

Table 3.13: General Attitudes to Data Sharing across Sample

	Control (n = 337)		Treatment (n = 349)	
	n	%	n	%
Basic Information	321	95.2	317	90.8
Marketing and Research	219	65.0	234	67.0
Third Party Access	161	47.8	179	51.3

informed consent to data sharing. From the open-ended feedback we see that respondents in the control group, in particular, are not necessarily aware of the information embedded within smart meter data:

> "I don't really understand what a hacker will gain from learning about my electricity consumption"
> (10001, C, F, 55-64)

As discussed in Chapter 2, providing consumers with details of the implications of sharing smart meter data is difficult, if not impossible given the ability to link datasets. As such, the result of our treatment may be an underestimation of consumers privacy concerns.

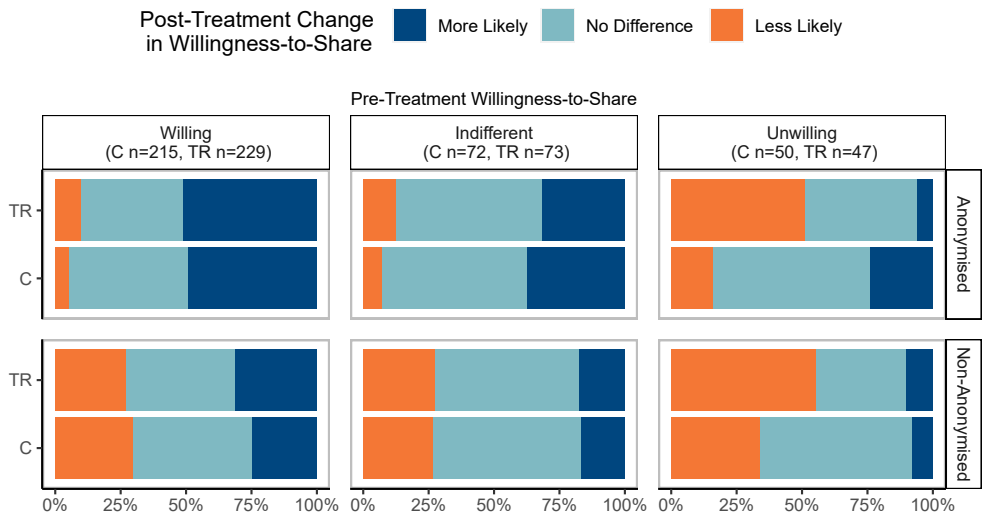


Figure 3.13: Post-Treatment Change in Willingness-to-Share Anonymised or Non-Anonymised Half-Hourly Data Split by Pre-Treatment Willingness-to-Share.

The WTS half-hourly data is shown in Figure 3.13. Of those who were willing or indifferent to sharing prior to being shown information on smart meter data, we observe

no significant difference, between the control (C) and treatment (TR) groups, in their post-treatment WTS. However, for those who were initially unwilling, we see a significant difference between the two groups in their post-treatment response. Those in the treatment group (56%) were less willing to share high resolution than those in the control group (32%). Interestingly, this difference is observed whether data are anonymised or not, suggesting that the option of anonymisation alone, may not necessarily alleviate the concerns of those who are particularly hesitant to share data.

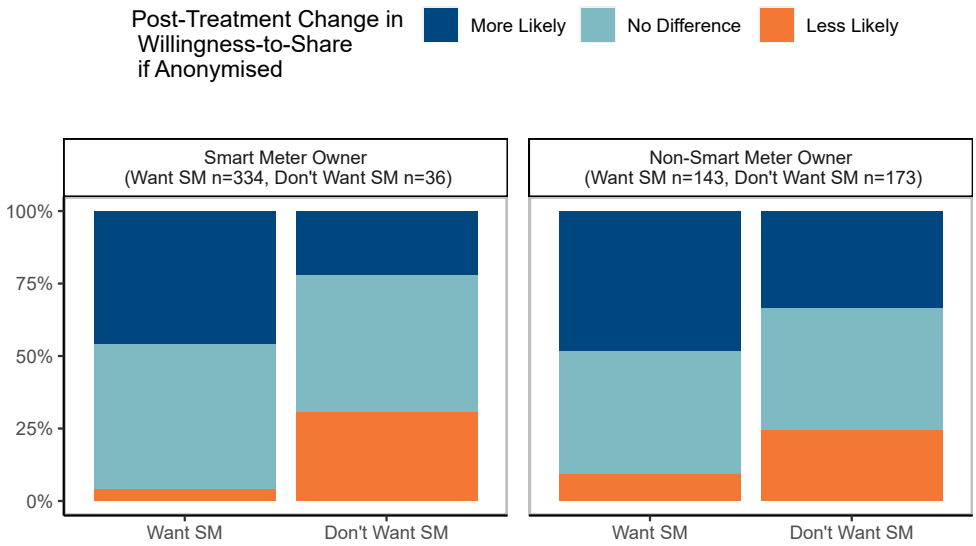


Figure 3.14: Post-Treatment Willingness-to-Share Anonymised Half-Hourly Data Split by Smart Meter Ownership and Actual Demand for Smart Meters.

Figure 3.14 shows respondents change in WTS, post treatment, if data were anonymised, by their demand for a smart meter. We see that anonymisation is attractive to people who want a smart meter, both those who already have one and as well as those who don't. However, those who do not want a smart meter are also not convinced by the anonymisation option. Some of the driving factors, which have been discussed in the literature [131], [168], are also reflected in the open-ended feedback. Specifically, wider issues of trust:

> "Being a retired IT Consultant, I am deeply concerned with the amount of personal data being collected about the population. The recent revelations about Credit Reference Companies breaching the Data Protection Laws comes as no surprise, I have long suspected that these companies and Financial Institutions have been buying and selling our information for a long time." (10479, TR, M, 75+)

General scepticism of smart meters was expressed as well:

> "This is extremely sinister and has huge implications for ripping off customers. Under no circumstances should anyone ever have a so-called smart meter." (11012, TR, M, 35-54)

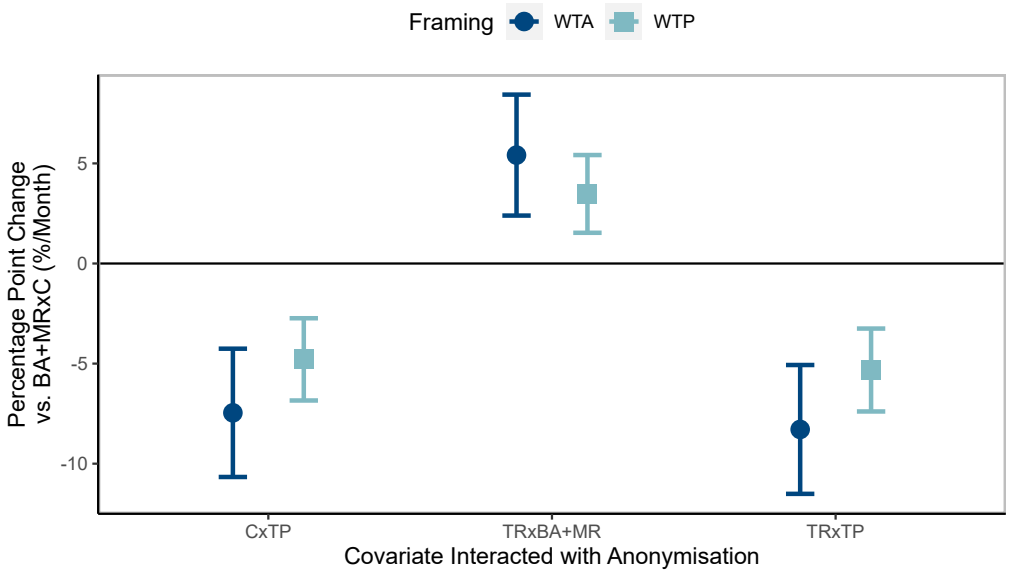


Figure 3.15: Effect of Treatment on Mean Willingness-to-Pay/Accept for Anonymisation by General Attitude Towards Data Sharing (BA+MR - Basic Sharing or for Market Research, TP - Sharing with Third Parties). Reference Group: Control Group with BA+MR. Error Bars Represent 90% Confidence Intervals. Generated using Delta Method with a Classical Covariance Matrix from MNL_{TRxSH}.

To measure the effect of information asymmetry on the WTP/A for anonymisation the MNL model with the individual-specific interaction terms discussed in the previous section was expanded with an additional interaction term for whether respondents were in the control (C) or the treatment (TR) group (MNL_{TR}). We find that those in the treatment group had a 2.45% increase in WTA compared to the control group. However, when we split this effect by respondents' general data sharing preferences, whether they shared data with third parties (TP) or not (BA+MR), we observe a significantly larger effect (MNL_{TRxSH}).. Figure 3.15 shows the effect of the treatment as a function of respondents general data sharing attitudes at a 90% confidence level. We see, that respondents who were willing to share data with third parties were not affected by the additional information and exhibited no value for anonymisation in either group. Specifically, we observe a 7.46% and 8.29% point reduction in WTA for the control and

treatment group, respectively.

Conversely, those not comfortable with sharing with third parties had a higher willingness-to-pay for anonymisation in the treatment group. We observe a 5.41% point increase in WTA in the treatment group compared to control group. Overall, the effect of the treatment was visible in this subgroup but was relatively small at the sample level. This is likely due to a number of factors, especially the difficultly in relaying such information. However, the open feedback questions also provide some interesting observations. We find that contrary to our initial hypothesis, some respondents were reassured when they were told about the privacy risks involved:

> "I would prefer to save money and maybe being anonymised is not so necessary as I first thought."
> (10230, TR, F, 75+)

At the same time, other respondents did become more concerned by the implications of data sharing:

> "I found the minute by minute example frightening to know that so much can be known about what goes on in your home [...]."
> (10244, TR, F, 55-64)

These dynamics may have also played a part in diminishing the effect of the treatment at the sample level.

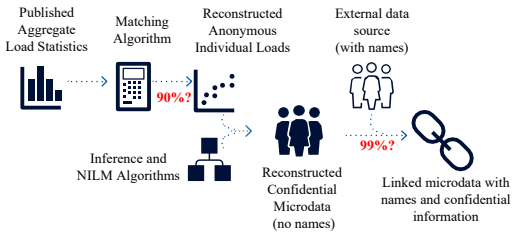
3.4 Discussion

This survey provides the first estimate of consumers' willingness-to-pay/accept for privacy and anonymisation in the context of smart meter data. It specifically couples the data sharing options currently available to consumers in the UK and what the implications of these options are, in terms of personal information shared. We see that there is significant heterogeneity among consumers, ranging from being indifferent to others with high concerns unwilling to share data under any circumstances. A data market must therefore be able provide consumers with a range of choices as to how much and with what level of detail they wish to share their data. We see that anonymisation can increase the level of high resolution data sharing but still requires monetary dis/incentives to reach, for example, 80% (the level of adoption required for many of the potential benefits of smart metering[2]). Additionally, we see that the presence of information asymmetry

may be resulting in underestimations of consumer privacy concerns. The application of mixed methods, with qualitative quotes complementing the quantitative analysis, was important in providing a more holistic picture.

At the same time, we note a number of limitations of the survey. As our survey was fielded online to a panel who regularly complete such surveys and in the process provide significant amounts of personal data this may be inducing significant bias. Different methods (e.g. telephone or face-to-face interviews) may provide a more representative sample in privacy terms. Respondents were paid the equivalent of £3/hour, due to budgetary constraints, which may not have been sufficient to obtain a quality sample. In addition, we note that the average comprehension for our DCE was fairly low (68%). There is a need to further develop the education material to be more accessible. It is important to keep in mind that informed consent should maintain a delicate balance between provision of essential information and information overload[169]. Finally, the role of trust needs to be explored further as we saw from the open ended feedback.

We see these methods as important tools to informing the design of a data market. Although we focused on the endowment effect and information asymmetries, there are number of other behavioural economics concepts that have a bearing on data valuations. For example, we considered the framing effects but did not account for status quo biases. This is especially relevant given the general apathy in the energy sector[4]. Further, the DAPF also offers a number of other options such as the ability to share with third parties, the effect of which needs to be explored further. This is of particular importance given the role of trust and the intended use of the data which were also highlighted in respondents open-feedback responses. Although we focused on monetary incentives in the form of changes in monthly bills, most existing methods employed by suppliers are based on offering additional services in return for data sharing. The effect of these services could also be included to further unpack how consumers value their privacy as opposed to their smart meter data. In addition, we looked at the effect of the data sharing frequency but did not consider variations in when that data is shared. For example, it has been posited that consumers may value the privacy of past events less than current events[14]. Quantifying this effect may be useful for developing Privacy-Preserving Techniques[170]. The next chapter reviews such Privacy-Preserving Techniques, looking at how they can incorporate heterogeneity, as observed through our survey, and increase transparency, in addition to their technical suitability for smart meter data.



CHAPTER 4

Privacy-Preserving Techniques

The survey study showed that consumers do have privacy concerns around smart meter data and that consumers do value anonymity. These were heightened when consumers were better informed. However, the survey results also highlight issues around communicating privacy risks to consumers. As such, obtaining informed consent for sharing smart meter data is difficult. In addition, there is significant heterogeneity suggesting a one-size fits all approach could severely limit the benefits of sharing or result in avoidably high costs. Privacy-Preserving Techniques could overcome these challenges by providing a secure means to share high-resolution smart meter data without compromising on consumer privacy [171]. This chapter describes different potential Privacy-Enhancing Techniques (PET) and Privacy-Preserving Techniques (PPT), and assesses their suitability for smart meter data applications and forms part of [Paper B]. These techniques can be broadly categorised into the following:

- Data obfuscation, altering data to provide anonymity or remove information which may be considered sensitive.
- Encryption and cryptographic methods, limiting access to authorised entities through intra-organisational permissions controls or other enhanced encryption techniques.
- User demand shaping, changing actual consumption to hide certain information which may be considered sensitive.
- Distributed/federated data processing, performing analysis on data in a distributed manner thereby limiting the sharing of raw data.

The terms privacy-enhancing and privacy-preserving are distinct, but often used interchangeably in the privacy literature. Privacy-Enhancing Techniques encompass methods

which limit the potential risks of privacy infringements, whereas Privacy-Preserving Techniques are those which focus on ensuring privacy throughout the data processing pipeline and may employ Privacy-Enhancing Techniques to achieve this. In addition, the term privacy-enhancing is criticised within the privacy literature as it implies an otherwise privacy-invasive system can be made private simply with the introduction of a PET[172]. However, this is in contrast to a Privacy by Design approach, which advocates that privacy measures should be fully integrated components of a system rather than add-ons[55]. In a similar vein, it has been argued that PETs imply that the data processor or users would be able to decide whether to grant a data owner privacy[172], as opposed the data owner being entitled to privacy protection of their data. As such, we choose to adopt the term Privacy-Preserving Technique throughout.

4.1 Privacy Properties

Many of these techniques have been investigated in the context of smart metering and have been considered at various stages of the SMIP by both BEIS and OFGEM but are yet to be implemented [31], [69]. As techniques vary in how and what aspect of privacy they protect, it is important to have consistent metrics which to assess their suitability for smart meter data. We develop a number of criteria including privacy guarantees, which formalise different aspects of privacy protection, and other desirable properties based on the use cases and privacy risks mapped in Chapter 2 and the consumer concerns identified Chapter 3. These are summarised in Table 4.1. Although each criteria is not necessarily equally important, we consider consider them all while assessing the suitability of PPTs for smart meter data.

Table 4.1: Assessment Metrics for Privacy-Preserving Techniques

Privacy Guarantees (Adapted from [173])	
Anonymity	Does the technique ensure that no individual can be singled out or identified from the data?
Linkability	Does the technique ensure that no individuals' data can be linked across different datasets?
Limiting Inference	Does the technique ensure that personal information cannot be inferred from the data?
Minimising Data Breaches	Does the technique minimise the potential privacy infringements which may arise in the event of a data breach?
Desirable Properties	
Access Individual Level Data	Can the technique still provide access to individual level data, allowing for personalised services to be offered?
Trusted Third Party	Can the technique operate without the need for a trusted third-party such as the DCC?
Easily Integrated	Can the technique be integrated within the SMIP without a significant overhaul of the framework?
Preserve Data Utility	Does the technique preserve the accuracy and utility of the data/results it outputs?
Accommodate Preference Heterogeneity	Can the technique provide different levels of privacy to different users, given the heterogeneity of privacy concerns among consumers?

4.2 Privacy-Preserving Techniques

4.2.1 Pseudonymisation

Pseudonymisation involves replacing identifiable features and data with unique identifiers which are consistent across the dataset. Under GDPR regulations, pseudonymised data is still considered personally identifiable information. In order to be considered truly anonymous, data must be stripped of all identifiable features but this can lead to data becoming unusable for meaningful analysis. Pseudonymisation provides a middle ground by reducing the possibility of re-identification while maintaining data quality. This is

formalised through the notion of K-anonimity, which provides a quantifiable assessment of the risk of linking attacks (see Box 4.1 for details). The ICO code of practice provides guidance on techniques and how to evaluate whether the risk of re-identification is sufficiently remote [174]. The UK Anonymisation Network also published the Anonymisation Decision Making Framework which provides practical steps to thinking about the context for each use case to understand the level of risk, manage the risk and effectively inform the anonymisation steps which need to take place [175]. For smart meter data this may include, among other things, masking names, addresses and the MPAN (a meter identifier linked to the supply point).

Pseudonymisation has been widely used across various industries and was considered as a privacy-preserving option in OFGEM's latest consultation on data access for energy suppliers [69]. It could, with modifications, be implemented within the UK's existing data collection framework [176]. It would allow access to individual household consumption data without, necessarily, revealing any personal information. However smart meter data has a lot of additional personal information embedded within it. These 'quasi-identifiers' can be used to re-identify individuals when linked to other databases such as billing and account information held by suppliers (see Box 4.1).

Academic literature has shown the risk of re-identification remains high even in country-scale location datasets [177]. For smart meter data, it has been shown that given a combination of pseudonymised high-resolution data (e.g. HH consumption data) and low-resolution data (e.g. daily or monthly consumption data extracted from billing), it is possible to de-pseudonymise the data and identify individuals with high accuracy (up to 99%) [178]. Two real-life examples of such linking attacks include:

- US Census Data: A combination of gender, birth dates, and postcodes were used to identify 87% of people in the U.S. [179].
- Netflix prize dataset: Even when the dataset on movie ratings of their 500,000 subscribers had names removed and ratings faked, subscriber records and other sensitive information could be identified [180].

Although K-anonymity provides a mathematical framework and guarantee of privacy, it does so by making an assumption on the background knowledge a potential attacker may have. However, with the presence embedded information within smart meter data and the availability of other commercial or public datasets make it difficult to account for potential linkages. Without accounting for this, the risk of re-identification is not fully captured within the K-anonymity framework.

Box 4.1: Linkage Attacks.(Examples adapted from [181])

A linkage attack is a type of privacy leakage where an individual's information can be discovered by obtaining information from another source. In other words, a Sensitive Attribute (SA) can be matched with a Key Attributed (KA) through a unique Quasi-Identifier (QID) information (whether it is anonymised or not). A linking attack can therefore be used to de-anonymise pseudonymised data. For example, Figure 4.1 shows two datasets, one pseudonymised the other is not. Due to the unique initial letters of the Postcode the individual "Carol" can be linked across the datasets and their information leaked.

Original Data			Anoymised Data	
KA	QID	SA	QID	SA
Owner Name	*Postcode*	*Data*	*Postcode*	*Data*
Alex	OL2 2XL	5.8	OL2 ***	5.8
Ben	LE3 9EE	4.5	LE3 ***	4.5
Carol	AB41 8UQ	6.7	AB41 ***	6.7
David	OL2 5EE	6.2	OL2 ***	6.2
Emma	LE3 1SR	5.3	LE3 ***	5.3

Figure 4.1: Illustrative Linking Attack using Postcodes.

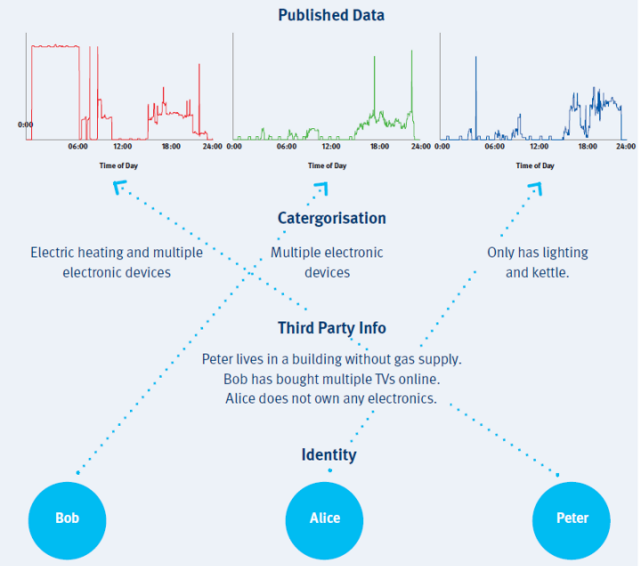


Figure 4.2: Illustrative Linking Attack using Smart Meter Data.

For an example using smart meter data we look at Figure 4.2, which shows three pseudonymised electricity load profiles. Using third party or other public information, appliance information could be uncovered. This would allow the data to be linked back to a particular individual using NILM. A dataset is said to have K-anonimity if for each identifier in a dataset, the information for each person contained in the release cannot be distinguished from at least K-1 other individuals whose information also appear in the dataset. In smart meter example, if Alice has electric heating, only Bob's data can be inferred, and the privacy of Peter and Alice is ensured. As such, this dataset is 2-pseudonymised.

4.2.2 Aggregation

Aggregation can be achieved by aggregating consumption data either across multiple periods of time (temporal) or the addition of data pertaining to multiple households (spatial). Aggregation reduces the privacy impact by helping to conceal the pattern of electricity usage inherent in one property's detailed consumption data, either by losing the detail within a sum totalled over a period of a time, or by masking the detail amongst other properties.

Spatial Aggregation

Although no set standards are available on the number of households which must be aggregated to provide privacy protection, a number of schemes have been proposed:

- 15/15 rule - The California Public Utility Commission Decision 14-05-016 defines a limited procedure for anonymising energy data. It is restricted to monthly average consumption by zip code. If at least 15 meters are aggregated together and no single meter comprises more than 15% of the total energy consumption, then the data are considered "anonymous" [182].

- DNOs in the UK, as part of their privacy plans, have chosen to aggregate households at the feeder level [67]. They require, for example, a minimum of 5 households to be aggregated at each low-voltage feeder. If a feeder has less than 5 households it must either be combined with other feeders, or the data cannot be accessed. This was based on an assessment of the coverage a DNO would get across their network [134].

Spatial aggregation provides a simple mechanism to provide access to high-resolution data as it can be easily integrated into the existing data architecture of the SMIP. For example, one supplier's aggregation framework follows three simple steps; data is requested from the relevant smart meters via the DCC, the individual level data sent back via the DCC is decrypted, validated, and temporarily stored, an aggregation script is run and then the individual level data is deleted [134]. When a sufficient number of households (greater than 20) are aggregated, the diversity of load can significantly reduce the ability of existing load disaggregation algorithms and other inference techniques to accurately extract the personal information embedded within smart meter data[117].

However, there are still a number of vulnerabilities presented by the technique itself as well as the nature of the smart meter roll-out. Given that suppliers and third parties offer

their services across different geographical locations, aggregation schemes which rely on location, as is being employed by DNOs, may result in restrictions on accessing data where there are few customers [69]. Additionally, aggregated data is vulnerable to reconstruction attacks, where individual level data can be deduced from aggregate information. An assessment of the simple aggregation schemes proposed in the DNO privacy plans showed that it is still possible to determine individuals' information, without complex algorithms, when the level of aggregation is below 10 households [183]. The study did not consider the availability of additional information or complex reconstruction methods thus providing a lower bound on the level of aggregation required. A recent publication has shown that with some additional aggregate information on consumption patterns, such as the average change in consumption between each half-hour, it is possible to reconstruct the entire individual level database with high accuracy [184].

Under the current framework suppliers and other third parties can incentivise customers to provide non-anonymised high-resolution data through the different data options available to customers. If a significant proportion of customers choose to provide such data, then the release of aggregated data would no longer protect those who do not consent to providing such information. As the proportion of those consenting increases, it is possible to deduce the individual consumption data of those who have not consented with greater accuracy. As a result, the privacy protections provided by spatial aggregation depends on the level of aggregation, others' privacy preferences and the type of aggregate data released.

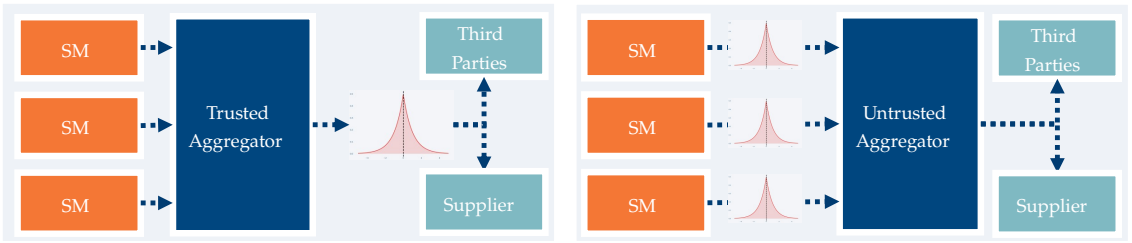
Temporal Aggregation

The resolution or frequency of smart meter data significantly impacts what information may be inferable from smart meter data. Additionally, many of the quantified benefits of smart metering laid out in BEIS's cost-benefit analysis do not require sharing of high-resolution data. As such, restricting data access to daily or monthly resolutions can protect from load disaggregation algorithms and limit the information that can be extracted from smart meter data. However, restricting access to higher resolution data precludes many of the envisioned benefits of smart grids, local energy systems and time-of-use tariffs.

4.2.3 Differential Privacy

Differential Privacy is a technique, rooted in cryptography, which ensures the anonymity of individuals within a dataset (e.g. the HH smart meter data of a suppliers' consumers)

while allowing aggregated information and statistics about the dataset to be shared (e.g. the average consumption for each half-hour). It works by ensuring that the aggregated information released about the dataset is not altered if an individual is included or excluded from the dataset. It overcomes a number of vulnerabilities posed by using pseudonymisation or aggregation alone as it protects against the worst case; it places no limits on the background knowledge someone analysing the data might have [185]. As such, it protects against even the most extreme collusion attacks i.e. even if the data of all individuals in a dataset except one were known, the differentially private aggregation will not reveal the data of that remaining individual.



(a) Global DP (b) Local DP

Figure 4.3: Dataflow for Differentially-Private Systems

The noise injection needed to achieve this, however, introduces a trade-off between the privacy being offered to individuals and the accuracy of the data being shared [186]. The amount of noise to be introduced is inversely proportional to the number of consumers to be aggregated[1]. Simulations suggest that high accuracy can be maintained when the number of customers are of the order of thousands [29]. This will be explored in more detail in Chapter 5. However, with application of specific knowledge and practical assumptions on the availability of background knowledge, an analyst can significantly reduce the amount of noise that needs to be introduced [187]. Nevertheless, determining the appropriate level of noise injection to provide privacy remains a topic of research [188]. DP has been gaining traction across multiple industries with real-world implementations including:

- Tracking emoji and text usage by Apple [189].
- The Community Mobility Reports published by Google providing insights into the effect of measures to combat Covid-19 [190].
- The 2020 US census is being published through a differentially-private mecha-

[1]In the case of global differential privacy.

nism[191].

- Electricity analytics company Recurve has implemented an open-source DP scheme adapted to smart metering [192].

Differential Privacy also has other desirable properties:

- Differentially-private algorithms have been developed for complex tasks e.g. clustering or general machine learning.
- As opposed to other data obfuscation and encryption techniques, the level of privacy offered can be controlled through a privacy budget.
- Variations of DP called heterogeneous or personalised DP allows for each individual within the database to be provided with a different level of privacy based on their preferences [193].
- It can be implemented with a trusted aggregator (e.g. DCC) performing the aggregation and noise addition, known as global DP, or with an untrusted aggregator where noise is added at the smart meter, known as local DP[194]. The different architectures are shown in Figure 4.3. The level of noise addition for local DP tends to be higher then global DP as aggregation effects cannot be leveraged.

4.2.4 Homomorphic Encryption

The SMETS 2 specifications provide details of the encryption capabilities required by smart meters in the UK [62]. This ensures that only authorised parties have access to consumption data. This is handled centrally through the DCC and the Smart Energy Code [195]. Although encryption provides a level of security and limits access to data it does not in itself ensure privacy. A promising PPT is homomorphic encryption [196]. It allows data analysts to perform arithmetic operations (e.g. addition, subtraction, multiplication, and division) on encrypted data without having to first decrypt it, ensuring that underlying data cannot be accessed while maintaining the accuracy of results [32]. Furthermore, it requires neither secure communication channels nor a trusted third party [197]. Such techniques can be categorised into the following [33]:

- Partial (e.g. Paillier or ElGamal) which allow only certain operations, such as addition and multiplication to be performed [198].
- Full which allow all operations to be performed. However, existing schemes result in high computational complexity, rendering applications impractical [199].

Cryptographic techniques such as homomorphic encryption typically suffer from high computational complexity, key distribution issues, overhead, and poor scalability, preventing practical applicability in a smart meter setting where computational and bandwidth resources are limited. Additionally, cryptographic techniques are vulnerable to statistical attacks and power analysis[33].

4.2.5 Multi-Party Computation

Multi-Party Computation (MPC) allows multiple parties to calculate an agreed upon functions without sharing their private data. The only information that can be inferred about the data is whatever could be inferred from seeing the output of the function alone. MPC protocols have been developed for a range of functions including basic statistical aggregations such as sum, count, min/max, histograms and percentiles.

MPC is achieved using cryptographic techniques or protocols. These vary depending on the specific function or set of functions considered. For example, aggregation may use secret sharing schemes[200], or set intersections can be computed securely and privately using Bloom filters[201]. Depending on the method chosen, MPC can work without the need for a trusted third party.

MPC can provide similar guarantees as Homomorphic Encryption however it does not require complex algorithms. It can use the traditional cryptographic system known as Advanced Encryption Standard easing integration within existing systems with encryption capabilities. In addition, it is safe against quantum attacks making it more future proof[202].

Multi-Party Computation-based aggregation and settlement calculation protocols using smart meter data have been proposed in the academic literature[200], [203]. It has also been used real world applications, namely, calculating aggregate demographics and salary information from employers in Boston for a gender pay gap study [204].

4.2.6 User Demand Shaping

In contrast to the above techniques, user demand shaping functions behind the meter by altering actual consumption patterns. Appliances and activities have characteristic consumption profiles. Changing the consumption data seen by the smart meter can hide specific appliances, or more generally limit the amount of information inferred from the consumption data recorded at the smart meter. This is achieved through smart control of flexible assets such as batteries[33]. There are a number of advantages to such an

approach:

- Does not require a trusted third party to verify or perform the privacy-preserving actions.
- Protects against physical hacking and unauthorised monitoring.
- Preserving accuracy of data collected and transmitted by smart meter.

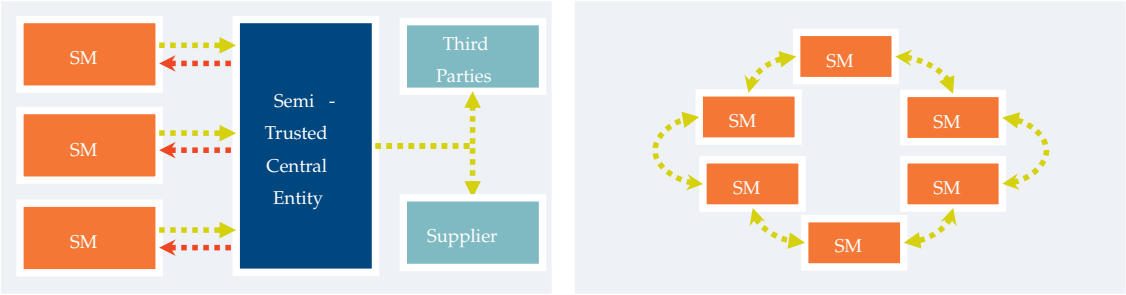
User demand shaping, as an approach, provides customers with direct control over their privacy. However by altering the actual consumption of a household the value of the data to make operational insights is greatly diminished. Overall, the main disadvantages of this approach are:

- Requires a battery or generation onsite,
- Reduces potential financial benefits that can be leveraged from battery flexibility[30],
- Does not provide anonymity,
- By altering the load seen at the grid the accuracy with which inferences about flexibility and usage can no longer be made [205].

4.2.7 Federated Data Processing

Existing data architectures require data to be collected from smart meters and then processed and analysed centrally. For example, when determining how much energy to buy from the wholesale market on the consumers' behalf, a supplier may want to use historical consumption data from its customers to forecast future consumption to aid their decision-making. They would collect the data, train a centralised model, and then use the trained model to forecast future consumption. While centralised techniques may have worked well in the past, this is becoming computationally challenging with the trend toward Big Data in recent years.

Federated Learning and privacy-preserving forms of Peer-to-Peer (P2P) distributed learning (sometimes called 'no-peek' [206]) are distributed computing techniques that allow such analysis to be performed at many computing nodes without having to first centralise the data. Not only does this reduce the computational burden at one location, it also reduces the communication requirements as large datasets no longer need to be sent to the cloud, and reduces the potential attack surface as data is no longer kept in one place. Instead, all raw data is kept locally; models are trained locally; and only model updates are shared between participants/with a central entity (Figure 4.4).



(a) Federated Learning (FL) (b) Peer-to-Peer (P2P)

Figure 4.4: Dataflow for Edge-based Data Processing Systems. Red Lines Indicate Model Transfer. Green Lines Represent Model Parameter Transfer.

However, even though raw data is not shared directly, FL and P2P distributed learning models can indirectly leak information during their model update stage. This may allow the local raw datasets to be reconstructed by malicious parties [207], [208]. Thus, unlike DP (Section 4.2.3), FL and P2P distributed learning do not provide a mathematical guarantee of privacy. In order to reduce the information revealed in model updates, FL and P2P distributed learning are often combined with some of the other privacy-preserving mechanisms discussed in this chapter, for example: homomorphic encryption [209] and/or DP. A review of such combinations in the context of collaborative forecasting can be found in [210]. Additionally, while there may be some accuracy loss when privacy-preserving distributed learning techniques are applied instead of a centralised approach, although most studies have found this to be negligible [207].

Federated Learning is being introduced into a number of fields where privacy and intellectual property rights are of particular concern. Notable examples include:

- Keyboard query suggestions for Google devices[211];
- A platform for multi-institutional collaboration to develop medical imaging diagnostic tools for brain tumours[212];
- MELLODDY, an EU-funded project to develop a platform for collaboration between pharmaceutical companies for drug discovery [213].

Furthermore, FL is being actively researched and applied to smart meter data for forecasting and clustering applications [47], [214], [215]. When applied to smart grids, FL is often applied in conjunction with edge computing: a distributed computing paradigm where computation is performed at the 'edges' of the network closer to where the data

is collected. In the context of smart metering, such devices include: smart meters, IHDs, and CADs which could connect to the HAN within the home.

4.3 Suitability For Smart Metering

The properties of each technique are summarised in Table 4.2. Traditional techniques such as pseudonymisation and aggregation do not provide provable guarantees of privacy without making assumptions on the available background information an attacker may have. It may seem reasonable to assume that an attacker does not have access to any individuals data (identifiers in the case of pseudonymisation, individual high-resolution consumption data in the case of aggregated data) under current regulatory arrangements. However, additional data streams are already being generated e.g. publicly available feeder level data from DNOs, increasing potential disclosure risks. Despite this, both aggregation and pseudonymisation are attractive options as they can be integrated and implemented with minimal change to the existing SMIP architecture. Homomorphic encryption overcomes some of the shortcomings of traditional methods but is computationally intensive and limited in terms of how data can be processed. Federated Learning and user demand shaping offer a decentralised approach to privacy preservation. However, neither provide anonymity but instead alter the data being sent. The current smart metering infrastructure does not have the computational resources required for these techniques. The need for interoperability may also hinder suppliers' ability to develop novel machine learning algorithms if such techniques were to be integrated into the SMIP.

Differential Privacy is considered the gold standard for privacy, and it provides provable guarantees of privacy and anonymity while also allowing the privacy-utility trade-off to be explored, introducing flexibility. Although it only allows for access to aggregate data, many of the applications for smart meter data can be performed using aggregated data with marginal improvements provided by access to individual level data. It has already been widely implemented in practice by both private entities such as Apple and Google and public entities such as the US Census Bureau. Specifically, it meets most of the privacy guarantees and desirable properties which are motivated by the use cases and privacy risks mapped in in Chapter 2 and importantly, the consumer concerns identified Chapter 3.

Our consumer-centric approach leads us to DP, however we note that although we present these techniques as separate, many can be combined to attain the properties of

the different techniques. For example, in [216] the authors combined DP, FL and MPC to develop a privacy-preserving FL protocol for financial applications. By combining the three techniques they leverage the distributive properties of FL, remove the need for a trusted aggregator through MPC and provide privacy guarantees in even the face of extreme collusion attacks through differentially private noise addition.

The next section will explore how the US Census Bureau has implemented DP for the 2020 US census, the associated challenges and lesson for implementing a similar approach for the UK's smart metering programme.

Table 4.2: Properties of Privacy-Preserving Techniques

		Pseudo-nymisation	Aggregation	Homomorphic Encryption	User Demand Shaping	Differential Privacy	Federated Learning	Multi-Party Computation
Privacy Guarantees	Anonymity	⋆	⋆	⋆		✓		
	Invulnerable to Linking					✓		
	Invulnerable to Inference				✓	✓		
	Minimise Data Breaches		⋆	✓	✓	✓	✓	✓
Desirable Properties	Individual Level Data	✓			⋆	⋆		
	No Trusted-Third Party			✓	✓	✓	✓	✓
	Easily Integrated	✓	✓			✓		✓
	Preserve Data Utility	✓	⋆	⋆	⋆		⋆	⋆
	Preference Heterogeneity	✓			✓	✓	✓	

Note: ⋆ indicate properties that the privacy-preserving technique is purported to have and which, in some cases, does have for practical purposes. However, these properties are not evidenced by theoretical guarantees.

4.4 US Census – A Case Study of Differential Privacy

4.4.1 Background

The US Census collects demographic information on all persons in the US on a decennial basis. This data is used to support a number of essential functions such as budget allocations and apportioning seats in the House of Representatives. Hence, the accuracy of such data is paramount. However, the Census Bureau is also legally obligated to ensure respondents' privacy and data confidentiality, prohibiting the release of any personally identifiable information[217].

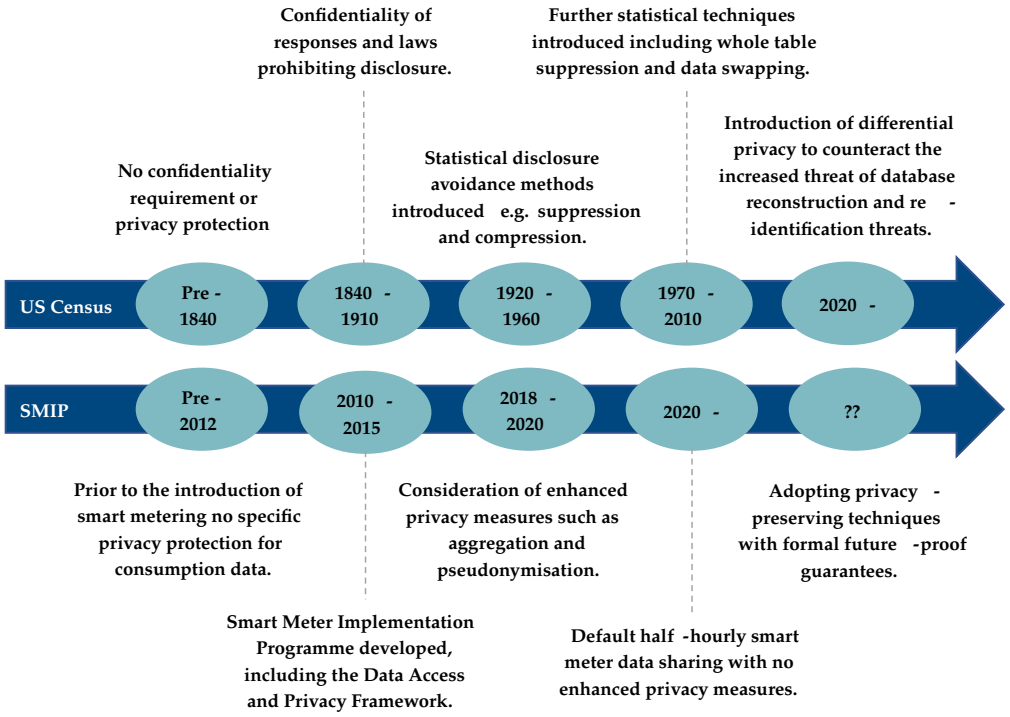


Figure 4.5: Evolution of Privacy Protection for US Census and SMIP. US Census timeline adapted from [191].

The Census Bureau has constantly been updating its privacy and data protection protocols. As shown in Figure 4.5, starting with confidentiality for businesses and laws banning census takers from disclosing information in the late 1800s, through the 1900s, a number of aggregation and suppression techniques were introduced to account for the risks of indirect disclosure. More recently, new techniques such as data swapping and top-coding were introduced to reduce the number of tables that would need to be suppressed [191]. However, advances in computing power and the availability of third-party data sources have meant that existing techniques are no longer able to protect against indirect

disclosure[218].

In 2016, the Census Bureau conducted an internal experiment to quantify the potential risk of disclosure of personal information under the existing publishing protocols. They used published aggregate data tables from the 2010 census to[219]:

- Reconstruct the underlying individual records (block ID, sex, age, race, ethnicity) of all 308 million individuals,
- Re-identify the individuals by matching the reconstructed database with commercially available data which include names and addresses.

The results of this exercise were striking. As shown in Figure 4.6, the individual level database could be reconstructed completely with 100% accuracy for 46% of the population and 71% of the population when allowing for a +/- 1 year age range. In addition, the individual variables (block ID, sex, age, race, ethnicity) were unique for more than 50% of the population. Finally, 38% of people's records could be exactly linked to the third-party data sources and their names and addresses identified. Given that aggregate data tables are published publicly, and no authorisation is required to access them, this was deemed a real and immediate threat.

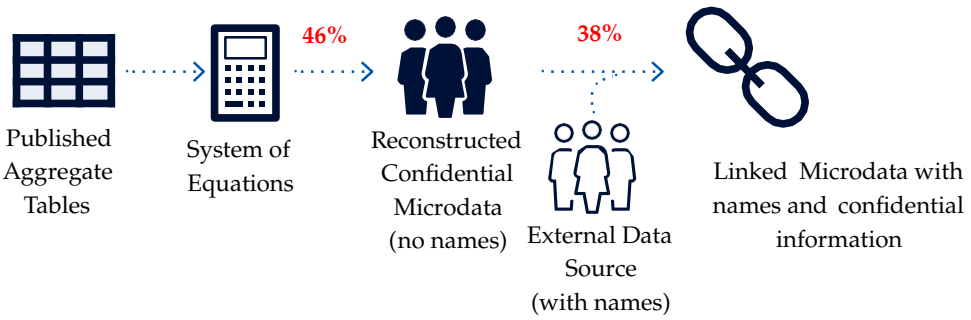


Figure 4.6: Database Reconstruction and Re-identification Attacks on US Census Dataset. Adapted from [220, Slide 12].

4.4.2 Implementation

This experiment showed that traditional statistical methods could not ensure that the Census Bureau was meeting its legal and ethical obligations. As a result, in 2018, they adopted DP for the 2020 Census to protect confidentiality. The Office for National Statistics (ONS) in the UK is also considering the use of DP for future UK Censuses [221]. To implement DP the Census Bureau developed a bespoke algorithm, the TopDown algo-

rithm, which accounts for the specific nature of census data and optimises accuracy while ensuring privacy [222]. There were four main challenges:

1. Determining the privacy/accuracy trade-off: As DP introduces noise it necessarily reduces accuracy. However, determining the appropriate trade-off and hence privacy budget (ϵ) remains a question for policymakers. The Census Bureau developed a framework to assess the cost and benefits based on the WTA privacy loss and the cost of increasing accuracy in terms of foregone privacy[223]. For example, the misallocation of funds due to errors in the population counts can be used to quantify the monetary implications for data accuracy. Similarly, a reference for the WTA for privacy loss can be drawn from existing law on losses incurred due to identity theft or stated preference surveys. The Census Bureau has been actively working with other government departments and demographers who use census data to determine of the optimal privacy budget in an ongoing and iterative process[224].

2. Allocating the privacy budget: Census data is collected and summarised at four levels of hierarchy – state, county, tract, and block. Each level has fewer and fewer people within it, from millions at the state level to thousands or less at the block level. The amount of noise that DP must add to ensure privacy is inversely proportional to the number of people. As a result, the algorithm was designed to provide block-level aggregates with more of the privacy budget.

3. Consistency and invariants: When adding noise to data it is possible to end up with some cases producing nonsensical results. For example, one may end up with -10.5 White females in a block when tabulating counts by ethnicity and sex. In addition, if computed separately, the sum of the population of each county in a state would not match the total population of the state. To overcome this, significant amounts of post-processing was incorporated into the TopDown algorithm. This resulted in additional accuracy loss, which was more than the error introduced by the noise introduced to achieve DP.

4. Product catalogue and computation: The census produces a large but finite set of tabulations (e.g. population counts by age and sex or household size counts at each level). Given the requirements described above it is necessary to compute the tables and relevant noise addition all at once for all states. As a result, the algorithm is computationally intensive and requires dedicated computing infrastructure to perform.

Aside from the technical challenges and specific features of the census data, the Census Bureau has also faced issues in convincing and communicating the new approach with data users. There has been significant debate amongst social scientists [225] and a court case [226] in Alabama. The accuracy of data still varies across geography (e.g. urban vs. rural) as well as ethnicity[227].

4.4.3 Lessons for Smart Metering

Integrating DP for smart meter data in the UK would require addressing many of the same challenges faced by the US Census Bureau as well as a number of specific issues given the nature of smart meter data and the SMIP structure.

1. Disclosure Risks and Threats: Aggregate census data is made publicly available and does not require authorisation for data access. Conversely, access to smart meter data is governed by strict permissions controls allowing only authorised parties registered with the DCC which reduces risks as these are currently heavily regulated entities (e.g. energy suppliers and network operators). The transition to a more dynamic domestic electricity sector will result in more, potentially unregulated, entities such as aggregators and switching websites, and their agents having access to smart meter data. Similarly, the current emphasis on widening access to data for public interest purposes and creating Open Data platforms, could result in comparable risks as publicly available datasets such as the census.

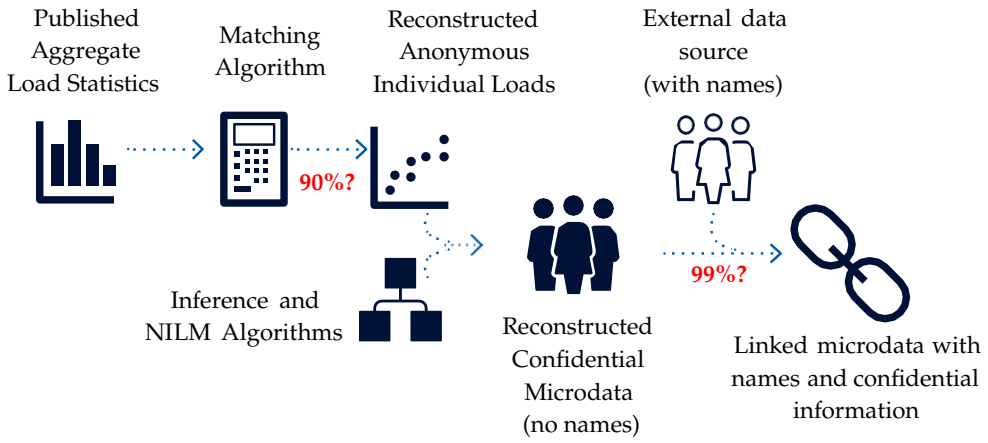


Figure 4.7: Potential Database Reconstruction and Re-identification Attacks on SM Data.

Additionally, census data directly includes socio-demographic information which can be used for re-identification whereas smart meter data contains only consumption information. However, significantly more socio-demographic information can

be extracted from smart meter data with relatively high accuracy (over 90% depending on temporal resolution as detailed in Table 2.4) using standard statistical inference techniques as shown in Section 2.4. As a result, access to individual high resolution smart meter data could be linked to other publicly or commercially available datasets (e.g. using low resolution billing data, with up to 99% accuracy[178]) in a similar way to the census data, as shown in Figure 4.7.

2. Who should implement it? Several entities have been proposed to act as trusted processors of smart meter data. These include energy suppliers, network operators, Elexon (settlement body), the DCC and the ONS [228]. A trusted aggregator which collects data from individual smart meters, applies DP and then sends this data to authorised parties could follow a similar top-down framework as the Census Bureau. However, the SMIP does not currently provide access to unencrypted smart meter data to these potential trusted processors, so would require a fundamental redesign. Alternatively, it is expected that Elexon will have access to individual smart meter data as part of the MHHS, which would allow them to apply appropriate privacy protections. However, future energy markets may potentially be highly decentralised with peer-to-peer transactions. In such a scenario, a local DP model would be more appropriate with suppliers and network operators implementing their own versions of DP. This is similar to the current framework for DNOs who submit their privacy plans for approval from OFGEM. However, it is important to note that this would require a technical oversight mechanism to ensure that the proposed DP implementations actually provide meaningful privacy protection[229]. For example, Apple has been criticised for marketing their products as differentially-private but the large privacy budget potentially provides little actual protection[230]. One solution to this would be to require companies to publish their code as the Census Bureau has done[231], and specifically, to detail the privacy budgets they use[232].

3. Developing a product catalogue: Unlike the US census data, there is currently no predefined catalogue of data products that a user can access. However, there are a set of operational use cases for smart meter data which could provide a basis for developing such a catalogue. For example:

    - Network operators need data at different levels of spatial granularity; the total national demand, demand at the Grid Supply Point (GSP) level and the feeder

level,

- Suppliers need data on their total customer base and demand for customers on different types of tariffs,
- BEIS needs usage data split by different sectors at different temporal resolutions (monthly and yearly) for publications and analyses which rely on DUKES data.

An outline of the size and definitional scope of potential smart meter data products is provided in Figure 4.8.

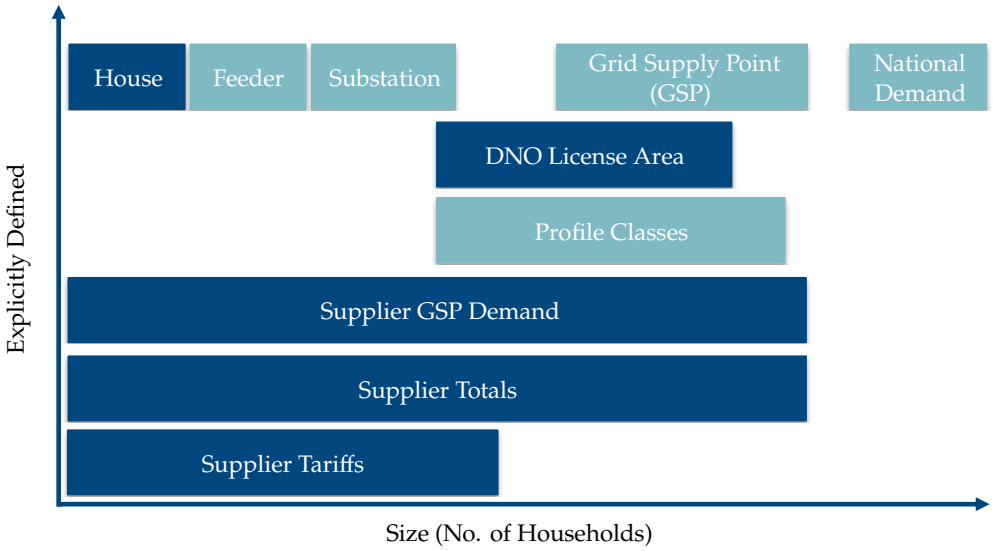


Figure 4.8: Illustrative Data Catalogue for the UK Electricity Network. Dark Blue Elements Require Smart Meter Data, whereas Light Blue Elements are Already Metered Independently.

4. Dynamic nature of smart meter data: The US census is a static database, data is collected once and the obligation to protect privacy is time-limited to 72 years (after which the individual level records are made public)[233]. It is possible to determine a privacy budget and allocate that budget over time. In contrast smart meter data is constantly generated and the data are correlated in time. As a result, it would be necessary to determine a time limit for which the privacy protection must apply. This could be based on existing data retention storage limitation policies companies have to follow under GDPR (Article 5 1.e). The notion of Discounted Differential Privacy (DDP) has been proposed to distribute the privacy budget by considering the differences in the sensitivity of historical (e.g. one year old) and real-time data[170].

5. Determining the appropriate trade-offs: Determining the trade-off between privacy and data utility remains an open question and is highly application and domain specific. In most existing implementations of DP, the trade-off has been chosen based on an assessment by the implementing party. Essentially it has, to date, remained a policy choice. The Census Bureau has focused on the acceptable level of accuracy loss for its' data users (government departments and academia) but not investigated individuals' attitudes towards protecting their privacy. Given the two-way communication offered by smart metering and the dynamic nature of the electricity market, a market-based mechanism to determine the revealed preferences of consumers, who generate the data, and data users (supplier, network operators etc.) could be introduced[234].

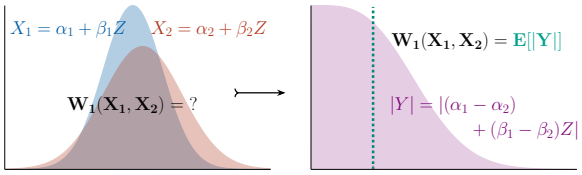
4.5 Discussion

This chapter reviewed a number of PPTs focusing on their suitability for smart meter data and the SMIP. Many PPTs have been proposed for smart meter data ranging from general techniques, such as DP, to the domain specific user demand shaping. PPTs vary in their definitions of privacy (e.g. anonymity, limiting inference), which may be continuous (e.g. the privacy budget in DP) or discrete (e.g. homomorphic encryption), and hence so do the privacy infringements (e.g. membership inference, non-intrusive load monitoring, linking attacks or data breaches) they protect against. Therefore, we developed a set of privacy guarantees, and desirable properties a suitable PPT should posses. We found that DP, in particular, is a prime candidate to achieve this given that it can be implemented either centrally or in a distributed manner, provides provable, tuneable and future proof guarantees of privacy protection and has been implemented in other sectors. It has already been widely implemented in practice by both private entities such as Apple and Google and public entities such as the US Census Bureau.

Although it only allows for access to aggregate data, many of the applications for smart meter data can be performed using aggregated data with marginal improvements provided by access to individual level data. By way of example, we use the US Census's recent incorporation of DP for the 2020 census to explore the challenges of implementing DP and lessons for the SMIP. Overall, we see that smart meter data under the existing SMIP has similarly conceivable disclosure risks to those faced by the US Census data. The US Census Bureau faced significant challenges in deciding the appropriate privacy-utility trade-off, specifically, how to determine the privacy budget, ϵ, and how to split it across

the multiple statistics they release as part of the census. This would also be an issue with regards to smart meter data, and would require the development of a data catalogue which maps out different levels of aggregation and types of groupings that one would want to publish. In addition, we identify two unique challenges for smart meter data, which require further investigation. Namely, who should implement a differentially-private smart metering system and how to account for the dynamic nature of smart meter data.

In order to begin tackling these challenges, in particular, the issue of determining the appropriate privacy-utility trade-off, the remaining chapters in this thesis propose a data market mechanism. The next chapter first investigates the drivers of smart meter data value/utility and then proposes a valuation metric to embody them.



CHAPTER 5

Valuation of Differentially-Private Data

One of the key recommendations we propose to improve smart meter data privacy of the SMIP is the introduction of Privacy-Preserving Techniques. Specifically, DP offers tuneable and future proof guarantees of privacy protection. However, the noise addition needed to achieve this introduces a trade-off between privacy and utility/value. This motivates the use of a market mechanism to determine the appropriate trade-off. Indeed, the concept of a data market has been gaining traction with a multitude of disciplines[235], including a growing interest in markets for smart meter and other energy domain data[16]. Broadly, these frameworks consist of:

- A valuation mechanism, which may be as simple as the data quantity or a more complex mechanism such as the value of a loss function for a specific task (e.g. mean squared error of a forecast).
- A procurement/pricing mechanism, which translates the valuation into a procurement decision, whether to buy the data or not, and how much to pay for it (e.g. a cooperative game formulation with payments determined using the Shapley value, the average marginal contribution to a valuation metric).

This chapter develops a valuation mechanism for differentially-private smart meter data, focusing on the specific issues and dynamics identified in the previous chapters. As we have shown in Chapter 3, many consumers are willing to pay for their privacy with compensation depending on the level of privacy afforded, while others would require compensation before sharing their data irrespective of the privacy implication, if given the option. In addition, a significant proportion of consumers do not trust retailers or

third parties to handle their data[24]. In order to determine the financial value of smart meter data, we need to determine its effect on revenue generating activities. As discussed in Chapter 2, smart meter data has many potential uses/benefits for different entities and across a variety of tasks. These observations raise five main challenges for developing a data valuation mechanism for smart meter data:

1. How can we determine the drivers of data value?
2. How can the value of data be quantified without first accessing it?
3. How can we ensure the data valuation mechanism is also privacy-preserving and does not require a trusted party?
4. How can we model the effect of differentially-private noise addition on data value?
5. How can we ensure that the data valuation mechanism is able to represent data value across multiple potential uses/benefits?

To tackle the challenges above we propose a model agnostic valuation metric, namely, the Wasserstein Distance[1]. First, in Section 5.1 we explore the main drivers of value for differentially-private smart meter data by focusing on the retailers' energy procurement problem. The majority of this section forms [Paper A]. This is followed by Section 5.2, which outlines the theoretical underpinnings for why the WD is an appropriate data valuation metric and its advantages over other metrics. This forms part of [Paper E]. Section 5.3 details methods to privately compute the WD. This includes existing methods for empirical data distributions and novel closed-form expressions for location-scale distributions. Section 5.3.2 provides improved bounds and exact expressions for the effect of DP on data in the WD. These two sections form part of [Paper C].

5.1 Drivers of Value

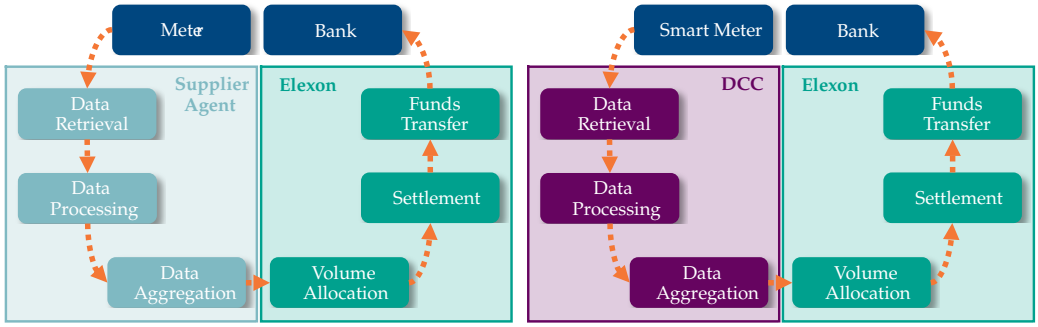
5.1.1 Background

Energy procurement is a core function for a Load-Serving Entity (LSE) (e.g. electricity retailer) and this can have a significant impact on system operation, through the scheduling and activation of reserves. It also provides a clear mechanism for how data, and specifically data quality, affects monetary quantities such as the LSE's profit. In contrast, some of the other benefits outlined in Section 2.3, such as, automated readings are independent

[1]When referring to the WD we specifically mean the 1-Wasserstein distance, unless otherwise specified.

of data quality and have already been valued. We note that many of the remaining operational uses of smart meter data such as tariff setting[236] can be viewed as an extension of the energy procurement problem. In addition, demand elasticity estimation[237] and clustering[238] can be incorporated as additional pre/post processing steps with the ultimate value derived from their effects on energy procurement. A comprehensive review of operational uses cases can be found in [20].

A LSE is required to forecast its consumer group's aggregated load and then purchase sufficient energy in advance as described in Section 2.3.1. To produce load forecasts the LSE needs historical consumption data of its consumers, which may vary in levels of aggregation or temporal resolution. Currently, meter data are collected and processed on an ad-hoc basis by LSEs or a contracted thrid party (e.g. supplier agents) and then sent to the settlement body, Elexon (see Figure 5.1a). In the UK, the introduction of smart meters will mean that, Elexon would be able to access data directly from the DCC, with LSEs no longer required in the process[69](see Figure 5.1b).



(a) Current (b) Potential Future

Figure 5.1: UK Metering and Settlement Process. Reproduced from Elexon [239].

Concurrently, the move towards MHHS will mean that LSEs will have to forecast HH consumption as opposed to daily volumes. This raises the question as to whether LSEs should have access to individual consumers HH data and how that access should be provided for forecasting purposes. Given the significant amount of personal information embedded within HH smart meter data, DP provides a means to balance access and privacy.

Extant literature on DP for smart meter data has focused on its effect on forecasting accuracy or risk of disclosure[29], [240]. However the resulting impact on energy procurement costs and LSE profits has, to the best of our knowledge, not been investigated. To this end, we propose a framework to assess drivers of value of differentially-private smart

meter. The framework consists of three alternative settlement and forecasting schemes: one in which HH data sharing is mandatory, one in which HH data are not shared and one where HH is shared but privacy is preserved using DP. To compare the different schemes a forecasting and procurement strategy for an LSE is developed. It consists of a short-term load forecasting mechanism and an optimal procurement strategy for the LSE, in the day-ahead and balancing markets. A case study using actual smart meter data and historical market prices is used to quantify the value of differentially-private smart meter data.

5.1.2 Differentially-Private Smart Meter Data

In Section 4.2.3 we provided an overview of DP, its attributes and usage. Here, we focus on the technical exposition of DP and its application to smart meter data. Current applications of DP can be split into two main functions; to protect an individual from identification [29], [186], [241], [242] or to protect against NILM algorithms from identifying appliances[243] (in this case each appliance can be considered an individual with smart meter data being an aggregation, see Figure 2.5). We choose to focus on the former, as this ensures that aggregate smart meter data can still be used to gain valuable information regarding consumption patterns and allow for segmentation without being able to specifically attribute any of these to a particular individual. We note that there is a growing body of work on adapting machine learning algorithms [244] and optimisation problems [245] to incorporate DP. These task-specific approaches can lead to significantly better performance in terms of privacy-utility trade-off but are not applicable to our scenario. In addition, by restricting our analysis to aggregate data we maintain the original definition of and use case for DP. As a result, we avoid the issues around sequential composition when individual level data is used to train machine learning models[229], [246].

We aim to understand the effect of using differentially-private smart meter data in a range of applications rather than understanding the performance of a specific differentially-private algorithm. To this end, the classical DP mechanism proposed in the literature is adapted, by drawing on existing literature, to address two main differences between privacy of smart meter data and other forms of data namely; the heterogeneity of privacy concerns identified in Chapter 3 and continuous reporting of data discussed in Section 4.2.3.

First, we define smart meter data (e.g. an LSE's consumer group) as an evolving

dataset E[170]:

$$E \coloneqq \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,T} \\ x_{2,1} & x_{2,2} & \dots & x_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,T} \end{bmatrix} \in \mathcal{E}^{N \times T} \subseteq \mathbb{R}^{N \times T} \tag{5.1}$$

where, N is the number of smart meters/individuals in the dataset and T is the length of the dataset.

Next, we assume that, as the dataset evolves, we are interested in the output of some aggregation mechanism Y (e.g. the mean) which takes $E(t)$ as input. Formally:

$$Y = [y_1, y_2, \dots, y_T] \in \mathcal{Y} \subseteq \mathbb{R}^T \tag{5.2}$$

where, y_t is the mechanism output at time instance t.

Differential Privacy aims to allow the computation and publishing of Y without compromising the privacy of any individual $i \in \mathcal{N}$ in the dataset E. It ensures this by providing probabilistic guarantees that, for any two neighbouring datasets (datasets that differ by one individual), the mechanism output is indistinguishable. We formalise the notion of neighbouring datasets and indistinguishablity in the context of aggregate smart meter data as follows.

Definition 5.1.1. *(Neighbouring Datasets). Two datasets E and E' are neighbouring datasets, shown by $E \sim E'$, if they differ by at most one row.*

Definition 5.1.2. *(Differential Privacy) A randomised mechanism M is (ϵ, δ)-differentially private for $\epsilon > 0, \delta \in (0, 1)$, if for any pair of neighbouring datasets E,E' and any output $Y \in \mathcal{Y}$ the following holds:*

$$\mathbb{P}[M(E) \in Y] \leq exp(\epsilon)\mathbb{P}[M(E') \in Y] + \delta \tag{5.3}$$

Here, ϵ is the privacy budget or leak that can be interpreted as the probability of identification and δ is the probability of failure. If $\delta = 0$ we obtain the stricter definition of DP, ϵ-DP. Although there are a number of approaches to achieve DP (e.g. randomised response[242], exponential mechanism[247]), we focus on additive noise mechanisms. To convert the deterministic mechanism, Y, into a randomised mechanism, M, we can introduce carefully tuned zero-mean noise to its output:

$$m_t = y_t + \xi_t \tag{5.4}$$

The noise, ξ_t, needs to be calibrated to the sensitivity of the specific mechanism. The sensitivity is the largest change of the output with respect to any changes in mechanism inputs.

Definition 5.1.3. *(Global sensitivity). The global sensitivity Δ_n of a mechanism Y is defined as:*

$$\Delta_n = \sup_{E,E':E\sim E'} \|Y(E') - Y(E)\|_n \tag{5.5}$$

For example, if Y is the mean per time step, we can define l_1 global sensitivity as the range of individual load values in a given time period t:

$$\Delta_{1,t} = \frac{\max_i (E_{i,t}) - \min_i (E_{i,t})}{N} \tag{5.6}$$

.

The global sensitivity allows us to produce two popular additive noise mechanisms, the Laplace and Gaussian mechanism. The Laplace mechanism uses the l_1 sensitivity and produces an ϵ-DP mechanism. The Gaussian mechanism uses the l_2 sensitivity and produces an (ϵ, δ)-DP mechanism. Again formally:

Definition 5.1.4. *(Laplace Mechanism). $M_{Lap}(x, f, \epsilon)$ provides ϵ-DP for a function $f(x)$[185, Definition 3.3]:*

$$M_{Lap}(x, f, \epsilon) = f(x) + \xi \tag{5.7}$$

where, $\xi \sim Lap(\alpha_l = 0, b = \frac{\Delta_1}{\epsilon})$, Δ_1 is the l_1 global sensitivity and ϵ is the privacy budget.

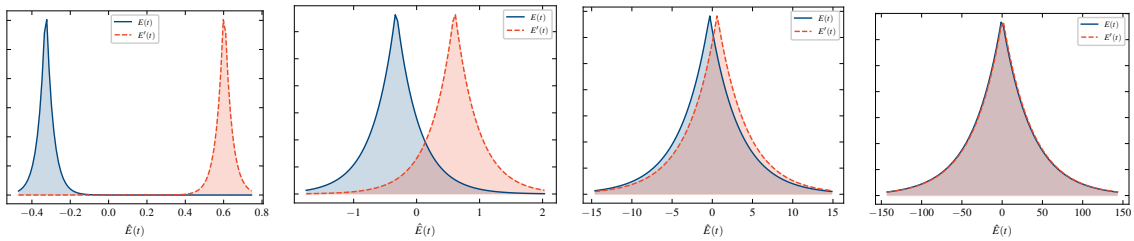
Definition 5.1.5. *(Gaussian Mechanism) $M_N(x, f, \epsilon, \delta)$ provides (ϵ, δ)-DP for a function $f(x)$[185, Theorem A.1]:*

$$M_N(x, f, \epsilon, \delta) = f(x) + \xi \tag{5.8}$$

where $\xi \sim N(0, \frac{2\ln(1.25/\delta)\Delta^2}{\epsilon^2})$, Δ_2 is the l_2 global sensitivity, ϵ is the privacy budget and δ is the probability of failure.

Although these additive noise mechanisms provide a simple method to achieve privacy, the noise which is introduced can be significant depending on the sensitivity of the function of interest Δ and the chosen privacy parameters, ϵ and δ. Figure 5.2 shows the output distribution of the Laplace mechanism for varying levels of the privacy budget, ϵ. We see that as the privacy budget reduces, the difference between the two distributions is no longer discernible, however, the dispersion of the distributions increases significantly.

Most extant literature on the application of DP and its variants to smart meter data have assumed that the privacy budget is fixed and that queries are independent. A



(a) $\epsilon = 100$ (b) $\epsilon = 10$ (c) $\epsilon = 1$ (d) $\epsilon = 0.1$

Figure 5.2: Illustrative Example of Laplace Mechanism

detailed review can be found in [240]. However, smart meter data and the resulting queries are continuously generated and updated. Consequently, the privacy loss defined by DP is accumulated across each query, requiring the addition of increasing amount of noise to ensure (5.3) holds. Over time this degrades data quality and renders new data useless[170]. Techniques have been proposed to overcome this limitation such as selective sampling based on time series dynamics[248] which improves performance but are still sensitive to the number of queries and would not guarantee the specified privacy budget over an infinite time horizon.

[170] overcomes this, providing a bounded mechanism by introducing the notion of DDP. It draws upon the concept of discounting from economic theory to propose that data further from the past is less sensitive than current data.

Definition 5.1.6. *(Discounted Differential Privacy). A Laplace mechanism M is* (ϵ, γ)*-differentially private for* $\epsilon > 0$ *and* $\gamma \in (0, 1]$ *under exponential discounting if [170, Corollary 12]:*

$$b = \frac{\Delta f_t}{\epsilon\,(1-\gamma)} \tag{5.9}$$

where, b is the scale of the Laplace noise, ϵ *is the privacy budget as before and* γ *is the discount rate (a measure of how much one values past data).*

If an individual does not place any value on the privacy of past data then $\gamma = 0$ and the mechanism is equivalent to standard DP (i.e. the accumulated privacy loss is not considered) whereas if one places high value on the privacy of past data then $\gamma \longrightarrow 1$ and the required noise tends to infinity.

Finally, to account for heterogeneous privacy preferences we can generalise the DP definition in (5.3).

Definition 5.1.7. *(Heterogeneous Differential Privacy) [52, Definition 1] A randomised mechanism M, is* ϵ*-differentially private for* $\epsilon > 0$*, if for any pair of neighbouring datasets E,E' that*

differ in the $i-th$ row and any output $Y \in \mathcal{Y}$ the following holds:

$$\mathbb{P}[M(E) \in Y] \leq exp(\epsilon_i)\mathbb{P}[M(E') \in Y] \tag{5.10}$$

Again, like conventional DP there are a number of ways to achieve heterogeneous DP. For example, [193] proposes a two-stage approach consisting of a weighted sampling stage followed by the Laplace mechanism. A minimum privacy level is selected which is afforded to all individuals and then sampling is used to provide individuals with higher privacy requirements (lower ϵ_i) additional privacy protection. Other formulations include modifications of the Laplace mechanism which introduce weightings to the noise scaling[52], [249].

5.1.3 Domestic Load Forecasting and Settlement

Non-Half-Hourly Settlement (NHHS)

In the absence of HH smart meter data, electricity settlement is based on system-wide Daily Load Coefficient (DLC)s which are published ahead of time. DLCs are standardised load profiles that specify the amount of annual consumption a specific consumer group (domestic, SME etc.) consumes in a particular half hour. These are generated based on HH measurement taken from a sample of consumers within each consumer group (for details see [80]). A LSE is only required to forecast daily demand (E^d) and is therefore insulated from HH changes while still exposed to HH prices (see Table 5.1).

Table 5.1: Settlement Schemes

	NHHS	**HHS -** DLC^{sys}	**HHS -** E^{hh}	**HHS –** $DDP(\epsilon, \alpha)$
Forecast Input	E^d, DLC^{sys}	E^d, DLC^{sys}	E^{hh}	$E^{hh} + Lap(\epsilon, \alpha)$
Forecast Parameter	E^d	E^{hh}	E^{hh}	E^{hh}
Settlement	$E^d DLC^{sys}$	E^{hh}	E^{hh}	E^{hh}
Risk Exposure	E^d, λ^{bal}	E^{hh}, λ^{bal}	E^{hh}, λ^{bal}	E^{hh}, λ^{bal}

Half-Hourly Settlement (HHS)

To assess what the underlying value of sharing HH data would be we present three alternatives (see Table 5.1): a scheme in which data sharing is mandatory i.e. the LSE has access to all its consumers aggregate unaltered HH data (HHS - E^{hh}), a scheme where

only aggregate daily data are shared (HHS - DLC^{sys}) and a scheme where aggregate HH data are shared but is privacy-protected using DDP (HHS - $DDP(\epsilon, \gamma)$). Under all these schemes settlement is based on actual HH consumption but the data available for forecasting purposes differs. An overview of the dataflows for each scheme is shown in Figure 5.3. The next section details the forecasting and procurement models used.

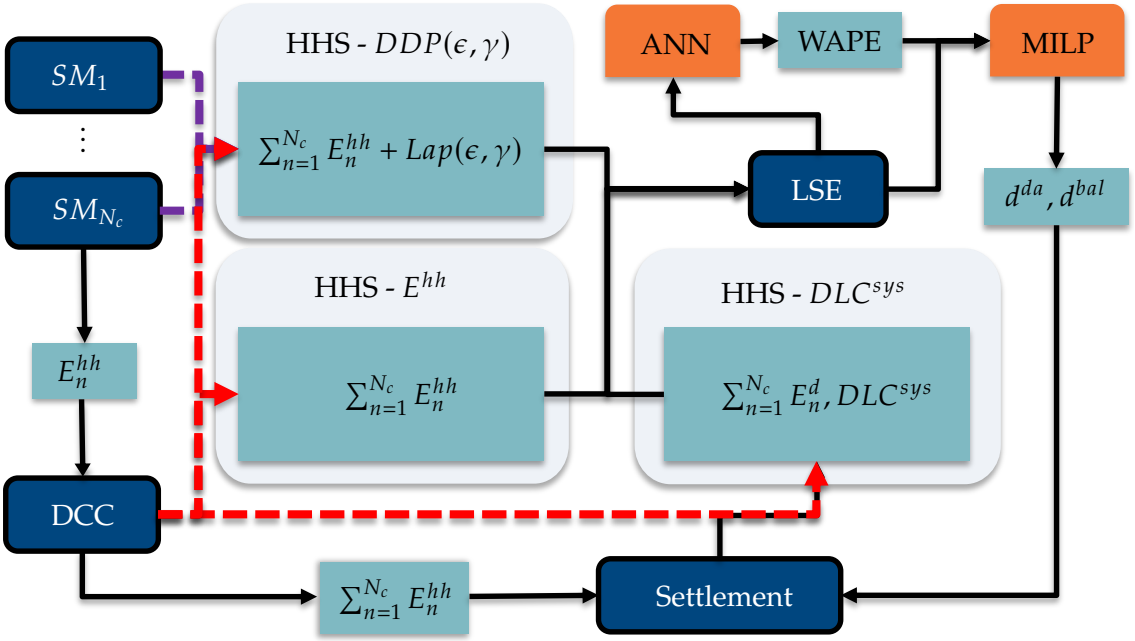


Figure 5.3: Overview of Settlement Schemes[2]

5.1.4 Model Definition

Short-Term Load Forecasting

Artificial Neural Networks (ANN) have been widely used and perform well for short-term forecasting applications and are able to capture both linear and non-linear dependencies[250], [251]. We use a simple ANN consisting of three layers: input layer, hidden layer, and output layer where the hidden layer has four hidden neurons. The following features are considered:

$$X = [W, WKD, SP, E_{t-h}, E_{t-h-1}, E_{t-2h+1}, E_{t-2h}, E_{t-3h}] \tag{5.11}$$

where E_{t-*h} are the lagged/historical load values and h is the number of periods in the day (48), W is the week in the year, WKD is the day of the week, and SP is the

[2]Laplacian noise can be constructed by summing n I.I.D. Gamma distributions allowing for decentralised noise addition at the smart meter[194].

settlement period. The model is implemented in Python using the MLPRegressor model in Scikit-Learn[252].

Load-Serving Entity Procurement Problem

A LSE needs to procure energy to meet its customer group's load by participating in long-term trading, day-ahead and intra-day markets, and settling any imbalances between purchase volumes and actual consumption in the balancing market. Here we focus on day-ahead and balancing markets.

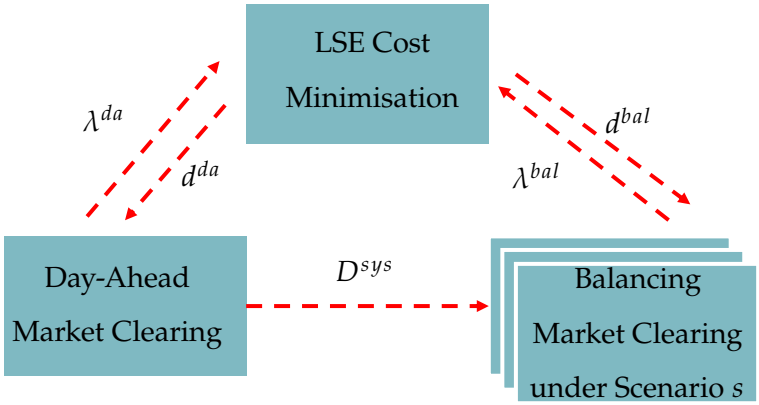
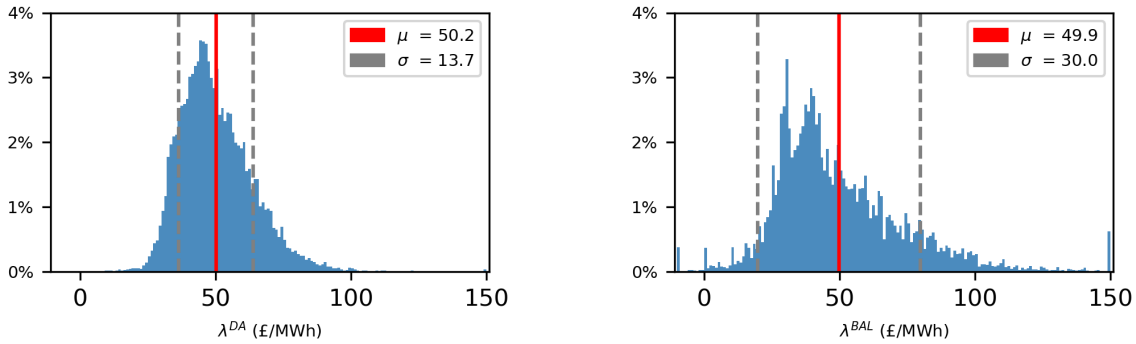


Figure 5.4: Bi-Level Structure of the LSE Procurement Problem

The LSE's procurement strategy can be formulated as a two-stage risk-constrained stochastic program, similar to [253](see Figure 5.4). We assume the LSE is a price-making market entity in both the day-ahead and balancing market and account for risk-aversion by including the optimisation of the Conditional Value-at-Risk (CVaR).



(a) Day-Ahead Prices

(b) Balancing Prices

Figure 5.5: GB Electricity Market Prices from 2017 to 2019[254]

We note that based on historical data mean day-ahead and balancing prices in the UK, are similar, however as shown in Figure 5.5 balancing prices are significantly more

volatile and skewed. The CVaR represents the expected value of the worst-case scenario costs that could occur given that the costs exceed a specified percentile $1-\alpha$ allowing a LSE to account for this asymmetric risk. The formulation is as follows:

$$\min_{d^{da},d^{bal}} \overbrace{\underbrace{\sum_t \lambda_t^{da} d_t^{da}}_{Day-Ahead} + \underbrace{\sum_s \pi_s \sum_t \lambda_{s,t}^{bal} d_{s,t}^{bal}}_{Balancing}}^{\text{Expected Cost}(\hat{\Omega})} + \underbrace{\beta \left[\zeta + \frac{1}{1-\alpha} \sum_s \pi_s \eta_s \right]}_{CVaR} \tag{5.12}$$

s.t.

$$d_t^{da} + d_{s,t}^{bal} = d_t^{fore} + d_{s,t}^{err}, \forall s, \forall t \tag{5.12a}$$

$$\sum_t \left[\lambda_t^{da} d_t^{da} + \sum_t \lambda_{s,t}^{bal} d_{s,t}^{bal} \right] - \zeta \leq \eta_s, \forall s \tag{5.12b}$$

$$\eta_s \geq 0, \forall s \tag{5.12c}$$

where d_t^{da} and $d_{s,t}^{bal}$ are the volumes procured by the LSE in the day-ahead and balancing market respectively, λ_t^{da} and $\lambda_{s,t}^{bal}$ are the market prices, π_s is the scenario probability, β is the risk-aversion factor, α is the CVaR confidence level, ζ and η_s are auxiliary variables to calculate CVaR, d_t^{fore} and $d_{s,t}^{err}$ are the day-ahead forecast load and the realised error. Given that the LSE is a price-making entity, λ_t^{da} and $\lambda_{s,t}^{bal}$ are dependent on the total demand. These can be modelled as piece-wise linear curves:

$$\lambda_t^{da} = \sum_b \lambda_b^G u_{t,b}^{da}, \forall t \tag{5.12d}$$

$$\lambda_{s,t}^{bal} = \sum_f \lambda_b^F u_{s,t,f}^{bal}, \forall s, \forall t \tag{5.12e}$$

$$D_t^{sys} - \frac{\Delta}{2} \leq \sum_b u_{t,b}^{da} \tilde{D}_b^{sys} \leq D_t^{sys} + \frac{\Delta}{2}, \forall t \tag{5.12f}$$

$$D_{s,t}^{imb} - \frac{\Delta}{2} \leq \sum_f u_{s,t,f}^{bal} \tilde{D}_f^{imb} \leq D_{s,t,}^{imb} + \frac{\Delta}{2}, \forall s, \forall t \tag{5.12g}$$

where $D_t^{sys} = D_t^{da} + d_t^{da}$, the total system demand in period t and $D_{s,t}^{imb} = D_{s,t}^{bal} + d_{s,t}^{bal}$, the total system imbalance in scenario s, $\tilde{D}_b^{da}$ is a discretisation of the system demand into b increments of Δ (similarly for $\tilde{D}_f^{bal}$) and $u_{t,b}^{da}$ and $u_{s,t,f}^{bal}$ are binary variables that select the appropriate demand level. The resulting model is a Mixed-Integer Quadratic Program (MIQP) due to the products of binary and continuous variables ($u_{t,b}^{da} d_t^{da}$ and $u_{s,t,f}^{bal} d_{s,t}^{bal}$). These can be linearised by replacing the bi-linear terms with a new variable on which a number of constraints are imposed giving an exact Mixed-Integer Linear Program (MILP) reformulation[255]. For example the term $u_{t,b}^{da} d_t^{da}$ can replaced by an auxiliary variable,

$C^{da}_{t,b}$, and four additional constraints:

$$C^{da}_{t,b} \leq u^{da}_{t,b} d^{max}_{t,b}, \forall t, \forall b \tag{5.12h}$$

$$C^{da}_{t,b} \leq d^{da}_{t}, \forall t, \forall b \tag{5.12i}$$

$$C^{da}_{t,b} \geq d^{da}_{t} - (1 - u^{da}_{t,b}) d^{max}_{t,b}, \forall t, \forall b \tag{5.12j}$$

$$C^{da}_{t,b} \geq 0, \forall t, \forall b \tag{5.12k}$$

The MILP reformulation is implemented in FICO^{TM} Xpress through the Python API.

Assessment Metrics

To gauge the difference between the load profile of the LSE's consumer group and the rest of the system we employ the Kullback-Leibler Divergence (KLD). Its value can be interpreted as the information gain achieved if HH data from the consumer group (DLC^{c}) is used instead of system level data (DLC^{sys}). We assume the average weekly DLC variations are normally distributed. In this context it can be defined as follows[256]:

$$KLD = \sum_{t \in T^{w}} \left[\log \frac{\sigma^{sys}_{t}}{\sigma^{c}_{t}} + \frac{\sigma^{c^2}_{t} + \left(\mu^{sys}_{t} - \mu^{c}_{t} \right)^2}{2\sigma^{sys^2}_{t}} - \frac{1}{2} \right] \tag{5.13}$$

where μ and σ are the mean and standard deviation of the DLC for each HH of the week respectively. The resulting values are provided in nats (natural unit of information).

To measure the accuracy of the forecasts we employ the Weighted Absolute Percentage Error (WAPE) metric as it exhibits more stable behaviour for values close to zero. This is especially relevant in the presence of Photovoltaics (PV) and battery storage as net load profiles can be negative.

$$WAPE = \frac{\sum_{t} \left| E^{act}_{t} - E^{fore}_{t} \right|}{\sum_{t} E^{act}_{t}} \tag{5.14}$$

5.1.5 Case Study

Data

We use smart meter data from the CER Behavioural Trials which includes 6010 residential and SME consumers for a period of 75 weeks[257][3]. For 50% of consumers synthetic PV [258] and EV [259] load profiles are added to better reflect the increased load diversity expected in the future. Day-ahead and balancing market bidding curves are generated

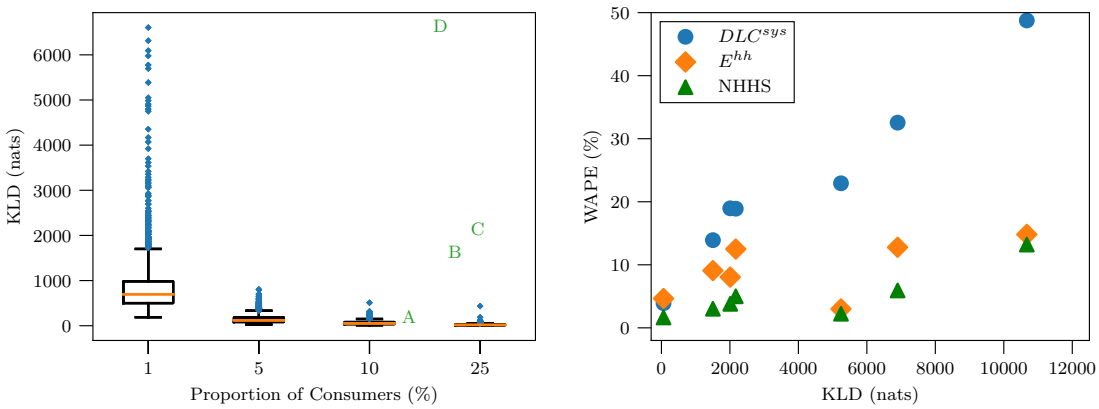
[3]After data processing to remove periods and meters for which less than 95% of data was available.

based on historical UK market data for 2018 from Elexon[254]. The WAPEs of the ANN described in Section 5.1.4 are used to generate 50 demand forecast scenarios, assuming they are normally distributed and then scaled to a representative UK system level based on the share of total load of consumers in the CER dataset. Finally, to develop a set of distinct load profiles we cluster the 6010 consumers into 8 groups using K-Means clustering of the average weekly DLCs for each consumer across a week. As we use a mixture of data sources (smart meter data from Ireland in 2010-11, synthetic PV and EV loads and market prices from GB in 2018), there are likely to be inconsistencies between price and load dynamics. However, the aim of this case study is to highlight the drivers of data value within the proposed framework, specifically those inherent to the data, rather than quantifying exact figures. As such, we argue that the dataset is appropriate to for this purpose. The framework could be used with a more consistent dataset (e.g. the SERL database[95] and market prices for GB) to develop exact figures for data value in the GB context.

Figure 5.6a shows the KLDs for randomly sampled meters under various proportions of consumers. When the meters are a small proportion of the total consumers the KLDs of the aggregated load can be large but as the proportion increases the KLDs decrease significantly. We argue that as LSEs begin to offer more innovative and targeted tariff mechanisms KLDs could be large even for large groups of consumers as they change consumption patterns based on particular time-varying incentives. We select four of the clustered groups (shown in grey in Figure 5.7) to test the framework. To add context we plot the resulting clusters on Figure 5.6a (labels A-D). From Figure 5.6b, which shows the forecast error, it is clear that as the KLD of a consumer group increases, having HH data for that specific group results in a greater reduction in forecasting error.

Results

Forecast Accuracy for Different Consumer Group The top plots in Figure 5.7 show the weekly average DLCs for the selected consumer groups. The bottom plots show the WAPE under each scheme. It is clear that when the average DLC of the consumer group is similar to the system (Group A), reflected in a low KLD, the WAPE is low even in the case where HH data are not provided (DLC^{sys}). As a result providing HH data using the DDP mechanism does not increase utility. However, as the group KLD increases, there is an increase in the difference between the WAPE with HH data and the WAPE without access to HH data. This provides a range within which a LSE is able to explore

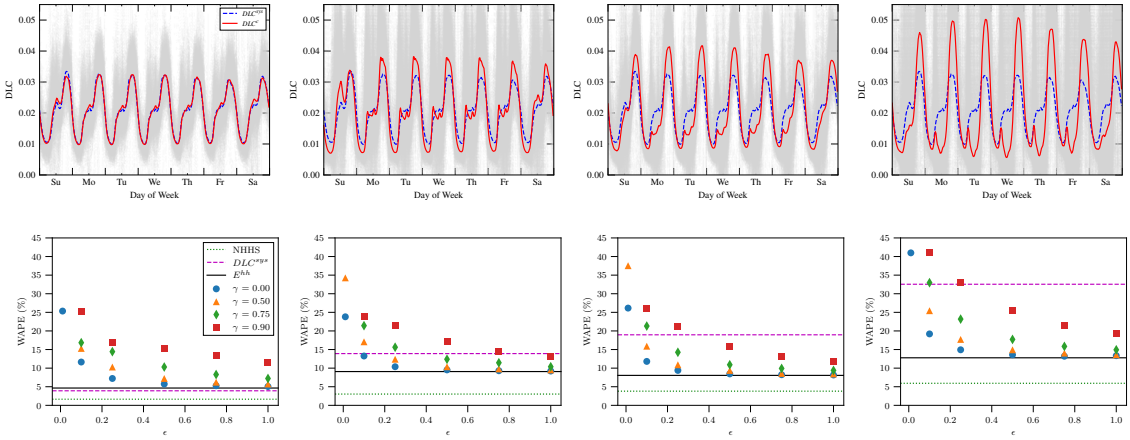


(a) KLD for Different Market Shares Sampled 2000 Times. Labels A-D Represent the 4 Cluster Groups Under Consideration.

(b) WAPE under Different Settlement Schemes for the 8 Cluster Groups

Figure 5.6: Distribution of KLD and WAPE across the Dataset used for the Case Study.

the privacy-utility trade-off. For example for Group C the LSE would be able to gain a 5% reduction in WAPE while providing a ($\epsilon = .25, \gamma = .75$) level of privacy. This shows that privacy-preservation can be achieved without significantly degrading data utility.



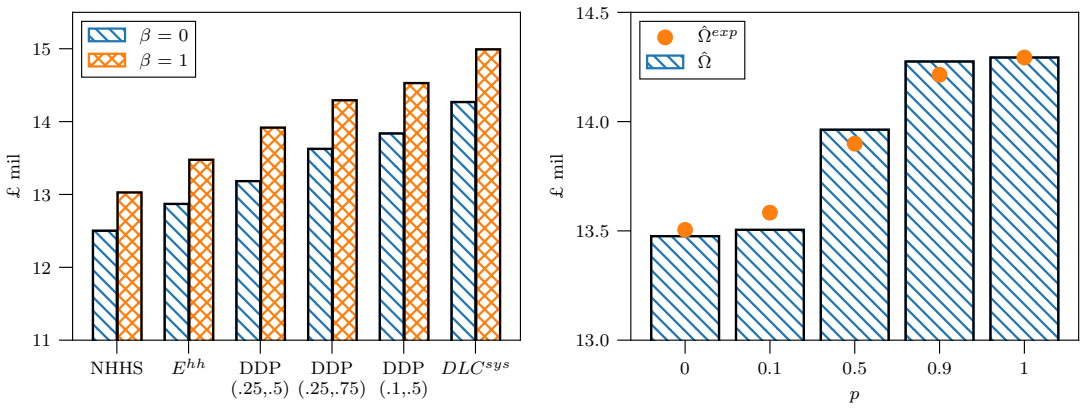
(a) Group A ($KLD = 62$) (b) Group B ($KLD = 1504$) (c) Group C ($KLD = 2010$) (d) Group D ($KLD = 6899$)

Figure 5.7: Cluster Group Weekly Average DLC (top), WAPE for Different Settlement Mechanisms (bottom).

Market Value Figure 5.8a shows the procurement costs based on scenarios generated for the different settlement schemes for Group C. A LSE can make significant cost savings while still providing consumers with privacy. On average we see that a 1% increase in WAPE results in a 0.8 - 1% cost reduction. A greater reduction is observed in the CVaR as a 1% increase in WAPE results in a 2-3% reduction in CVaR. The value of better forecasting

accuracy and hence HH data are also highly dependent on the market dynamics. At peak times uncertainty is more expensive, as there is less flexibility in the balancing market when overall demand is high, resulting in larger cost differences between the schemes.

Heterogeneous Privacy Preferences As privacy concerns vary, we investigate how the costs change when only a fraction of the consumer group has privacy concerns. Assuming a proportion of the consumer group p has privacy concerns we generate forecasts separately for them using DDP, with the load for the remaining consumers modelled using E^{hh}. Figure 5.8b shows the resulting procurement costs using this method for $DDP(.25, .75)$. The dots represent the weighted average cost ($\hat{\Omega}^{exp}$) that would be expected based on the proportion p. Splitting consumers based on privacy concerns can improve overall data quality and reduce overall procurement costs when p is low. However a trade-off is observed between reduced data degradation, as less noise is added, and benefits of aggregation, which smoothens the load profile.



(a) Under Different Settlement Schemes

(b) Under Different Privacy Proportions

Figure 5.8: Expected Procurement Cost for Cluster Group C. $\beta = 0$ and $\beta = 1$ indicate a risk-neutral and risk-averse strategy respectively.

5.1.6 Summary

We investigated the value of smart meter data by applying a framework consisting of a DDP model to ensure individuals cannot be identified from aggregated data, a short-term load forecasting method using ANN to quantify the impact of data availability and privacy protection on the forecasting error, and an optimal procurement problem, to assess the market value of the privacy-utility trade-off introduced by DDP.

Results show that, aside from the day-ahead and balancing market prices, the value

of smart meter data is driven by:

- Reference data: when the load profile of a LSE's consumer group differs from the reference data (e.g. system average), an increasingly relevant scenario with the introduction of dynamic tariffs, and distributed storage and generation, there is significant value in sharing data while retaining individual consumer privacy.
- LSE risk-aversion: risk-aversion leads to higher average costs, however the improvement in load forecasting accuracy generated by data sharing has a greater impact on the CVaR. This indicates that a more risk-averse LSE accrues greater benefits from data sharing as they value uncertainty reduction more.
- Level and distribution of privacy concerns: The level of privacy concerns and resulting ϵ and γ have a significant impact on value. As global DP protects all individuals equally this can lead to overly conservative results as the privacy parameters are determined by the individual with highest privacy requirements/concerns. As shown in Chapter 3, there is significant heterogeneity which could be leveraged to reduce the global sensitivity parameter (Δ) by optimising the noise introduced by the DDP mechanism and explicitly incorporating heterogeneous privacy preferences would further improve performance.

The analysis conducted above provides a basis from which to develop a valuation metric which embodies the determinants of data value for smart metering. The remainder of this chapter develops such a metric.

5.2 Wasserstein Distance as a Valuation Metric

5.2.1 Data Valuation Metrics

Existing Approaches

Data valuation is a broad subject with many proposed valuation techniques and metrics. These range from simple quantity based approaches (i.e. size of a dataset) to more complex techniques that attempt to determine quality (e.g. accuracy of the data) or task specific performance (e.g. improvement in loss function of a machine learning task). Given the breadth of the subject we defer to [16] for a comprehensive review of different data valuation and pricing schemes with a specific focus on energy data. Instead, we focus on assessing existing techniques which address one or more of the challenges we

set out at the beginning of the chapter and the drivers of value identified in the previous section. We also note that although we focus on 'data' valuation, our analysis also applies to information valuation (e.g. a forecast) and model valuation (e.g. the contribution of individuals' model updates in a federated learning setting).

As summarised in Table 5.2, we consider existing valuations techniques based on:

- Value Consolidation: How the value of each individual's data, $V(X_i)$, is consolidated to provide the value of a dataset, $V(X)$. For example, it may be assumed to be additive ($V(X) = \sum_i V(X_i)$) or based on marginal value ($V(X_i) = f(X) - f(X_{-i})$, where X_{-i} is the full dataset excluding individual i).
- I-O: Whether value, V, corresponds to the input (I) space (e.g. $V(X_i)$ or $V(X_i, X_j)$), output (O) space (e.g. $V(f(X_i))$ or $V(f(X_i), f(X_j))$) or in both (I-O) domains (e.g. $V(X_i, X_j) = f(g(X_i, X_j))$.
- Differential Privacy: Does it explicitly consider the effect of DP?
- Private: Is value computed in a privacy-preserving manner?

Refs. [38], [39], [260], [261] propose cooperative game frameworks where the value of a dataset, V, is determined by its impact on the performance/loss $L(X)$ of a particular model (e.g. mean squared error of a linear regression). They are able to capture the combinatorial nature of data value by using the Shapely Value (average marginal contribution) and provide an accurate assessment of performance for a particular task. Indeed, [39] studies a simplified version of the retailer energy procurement discussed in the previous section and provides a framework for valuing forecasts. However, these approaches assume the buyer or a trusted third-party (e.g. market platform) has access to the data prior to purchase. While this can be overcome by adapting the framework to provide performance information in a privacy-preserving and/or distributed manner (e.g. [262], which applies multiplicative randomisation to protect data privacy in a distributed forecasting mechanism), the results still correspond only to the output space and are therefore task specific. In addition, valuation based on performance for a particular task is prone to manipulation through model mis-specification.

In contrast, [43]–[49] use a range of statistical distances between the input data and a reference distribution as the valuation metric[4]. Ref. [44] provide an encrypted and

[4]One exception is a recent work which proposes data valuation based on features' impact on the KLD between the output predictive distribution with and without the features for a Bayesian linear regression task [263].

differentially private aggregation protocol and data market where value is determined by the JSD between the data and a uniform distribution as the reference since it can be considered the least informative. However, as detailed in [264] a uniform distribution as a reference is not theoretically grounded. Instead, as in [43], which also considers the JSD, the true data distribution or an a priori distribution should be used as a reference. The use of statistical distances also raises issues around how value should be consolidated across individuals. For example, [43] assumes the JSD is additive, although, again, this is not theoretically grounded. Similarly, [48], [265], which use the Earth Mover's Distance (EMD) or WD, calculate a weighted average of individual distances.

Table 5.2: Data Valuation Metrics

Valuation		Value Consolidation	DP	Private	I/O	
Type	Metric					
Loss Function	$L(X)$	Combinatorial	⋆[1]		O	[37], [260], [261]
	$L(X)$	Combinatorial			O	[38], [39]
Statistical Distance	JSD	Additive		✓	I	[43]
	JSD	Individual		✓	I	[44]
	EMD/WD	Individual		⋆[2]	I-O	[45], [46]
	EMD/WD	Weighted Average		⋆[2]	I-O	[48], [265]
	EMD/WD	Marginal Value	✓	⋆[2]	I-O	[266]
	KLD	Combinatorial			O	[263]
Other	Shannon Entropy + Non-Noise Ratio	N/A			I-O	[267]
	Variance Reduction	Additive			I-O	[215], [268]
Proposed	EMD/WD	Probabilistic Bounds	✓	✓	I-O	

[1] Data/model buyers can specify their noise sensitivity functions and data owners can specify their privacy preference functions. However, the functions are not explicitly linked to model accuracy[261].

[2] Although private computation is not explicitly considered, the EMD/WD can be computed privately, as will be shown in Section 5.3.

By shifting to the input space we obtain a task-agnostic notion of value, however

we lose guarantees in the output space. To overcome this, [45] bounds the loss of a federating learning algorithm by bounding the weight divergence using the EMD/WD. Refs. [46], [48], [265], simulate the input-output relationship to estimate parameters for a pre-specified transfer function. Ref. [267] employs a similar strategy except they use the Shannon entropy, a measure of information, and the non-noise ratio, a measure of data accuracy as the valuation metric. Data value in this case is independent of any reference data. Refs. [215], [268] focus specifically on the retailer energy procurement problem. They develop a framework to link the forecast uncertainty to the expected procurement costs and propose the reduction in standard deviation of the forecast errors, provided by data, as a value metric. In this framework, value is dependent on the assumed distribution of forecast errors. Ref. [266] frames data valuation as a Distributionally-Robust Optimisation (DRO) problem, foregoing the need to make any distributional assumptions. Instead, differentially-private data and the EMD/WD distance between the differentially-private data and original data are shared and value is determined by shadow prices. It provides a link between input data quality through the EMD/WD to the objective function of an optimisation problem.

Ref. [266] is therefore the only method which explicitly models the effect of DP of data value. Ref. [261] proposes a two-sided marketplace where data/model buyers can specify their noise sensitivity functions and data owners can specify their privacy preference functions. However, the functions are not explicitly linked to value or performance. Refs. [43], [44] use DP to ensure the valuation process is privacy-preserving but do not consider the implications of differentially private noise addition on data value. Although private computation is not explicitly considered in these studies, the EMD/WD can be computed privately as will be shown in Section 5.3.

The DRO framework set out in [266] fulfils most of the criteria, however, it requires access to the (differentially-private) data in addition to the EMD/WD distance in order to determine value. This requires data to be shared prior to purchase and assumes a set privacy level, eliminating the possibility to incorporate compensation-dependent privacy levels. Methods relying purely on statistical distances provide a means to obtain a task-agnostic notion of value and can be computed efficiently and privately. However, existing data valuation frameworks using statistical distances do not provide a method to relate input value to output value for a range of tasks. They also lack the ability to explicitly model the effect of DP. As we see from the studies summarised above, a number of different distances have been used, however determining the most suitable remains an

open question. The remainder of this chapter will outline a framework based on the WD to address these shortcomings.

Statistical Distances

A statistical distance provides a measure of how close two statistical objects, such as two probability distributions, are. In our context this is framed as a method to determine the closeness of an individuals' smart meter to some reference distribution (e.g. the system average or the LSE's consumer bases' average from Section 5.1). There are a wide range of distances with different properties, providing insights along different dimensions of the probability distributions. A comprehensive review including the relationship between these distances can be found in [269]. We focus on five popular distances/divergences which have been proposed in the context of data valuation:

- Wasserstein Distance (WD): $d_W(X_1, X_2) = \int_{-\infty}^{\infty} |F_{X_1}(x) - F_{X_2}(x)| dx$
- Kullback-Leibler Divergence (KLD): $d_{KLD}(X_1, X_2) = \sum_{x \in \Omega} f_{X_1}(x) \log \frac{f_{X_1}(x)}{f_{X_2}(x)}$
- Jensen-Shannon Divergence (JSD): $d_{JSD}(X_1, X_2) = \frac{1}{2} d_{KLD}(X_1, X_m) + \frac{1}{2} d_{KLD}(X_2, X_m)$ where, $f_{X_m} = \frac{f_{X_1} + f_{X_2}}{2}$
- Kolmogorov-Smirnov Metric (KS): $d_{KS}(X_1, X_2) = \sup_{x \in \mathbb{R}} |F_{X_1}(x) - F_{X_2}(x)|$
- Total Variation Distance (TVD): $d_{TVD}(X_1, X_2) = \frac{1}{2} \sum_{x \in \Omega} |f_{X_1}(x) - f_{X_2}(x)|$

where, X_1, X_2 are univariate probability distributions defined on a countable space Ω. f_{X_i} and F_{X_i} denote the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the univariate distribution X_i respectively.

To compare the statistical distances above, we set out a number of desirable properties for their use as data valuation metrics. These are summarised for the distances under consideration in Table 5.3. First, whether the distance is a metric and therefore obeys four axioms[5]: (1) identity of indiscernibles $d(X_1, X_1) = 0$, (2) symmetry $d(X_1, X_2) = d(X_2, X_1)$, (3) triangle inequality $d(X_1, X_3) \leq d(X_1, X_2) + d(X_2, X_3)$, and (4) non-negativity $d(X_1, X_2) \geq 0$. Next, whether the distance is non-saturating. This may be seen as a useful property as it can be used to model the law of diminishing returns, a common assumption in data

[5]We aim to use the statistical distance as a proxy for relative difference in performance, $L(X_1) - L(X_2) = -(L(X_2) - L(X_1))$, as such it is desirable for the distance to be symmetric and equal to zero when $X_1 = X_2$. In addition, as we are motivated by the combined value of multiple data, additivity or in this case subadditivity (triangle inequality) is a useful feature. Indeed, we use this and the non-negativity property to develop approximation bounds in Section 5.2.2.

valuation[267]. However, we argue that it is restrictive for the valuation metric itself to exhibit these dynamics and should instead be modelled explicitly as a function of data quantity not data quality.

Another important consideration is whether the distance is defined when the two distribution under considerations have disjoint or non-overlapping supports. This is especially relevant for empirical data, as the resulting distributions will be bounded. Finally, as we see from the above definitions, these statistical distances can be defined either in terms of the CDF's of distributions or their PDF's. When working with empirical data, distances defined based on CDFs are more attractive as they avoid the need to estimate PDFs, either by placing distributional assumptions on the data or using distribution-free methods such as kernel density estimation, which can be computationally intensive and prone to significant error for smaller datasets.

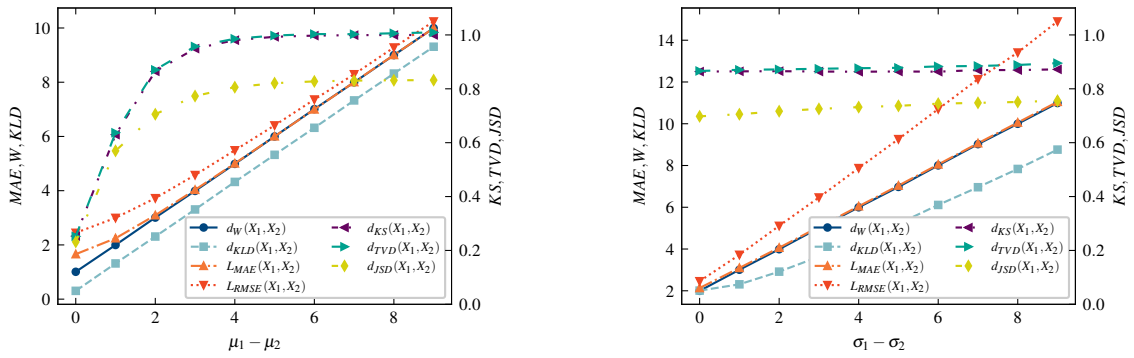
Table 5.3: Comparison of Different Statistical Distance/Divergences

Measure	Metric	Non-Saturating	Disjoint Supports	Input
KLD		✓		f_{X_1}, f_{X_2}
JSD	✓[1]		✓	$f_{X_1}, f_{X_2}, f_{\frac{X_1+X_2}{2}}$
KS	✓		✓	F_{X_1}, F_{X_2}
EMD/WD	✓	✓	✓	F_{X_1}, F_{X_2}
TVD	✓		✓	f_{X_1}, f_{X_2}

[1] The square root of the JSD is a metric.

The KLD, which we offered as a potential measure of value in Section 5.1 is non-saturating but does not meet any of the other criteria namely, it is not a metric as it is not symmetric and does not obey the triangle inequality, it requires PDFs to calculate and is infinite when the supports are not the same. The JSD, which can be seen as a symmetrisation of the KLD, overcomes some these issues as it is a metric and is defined for disjoint supports. However, the JSD still requires PDFs to calculate and also requires the calculation of a mid-point distribution $f_{\frac{X_1+X_2}{2}}$. In addition, the JSD is bounded and thus exhibits saturating behaviour. The TVD has similar limitations. The WD and KS metric rely on CDFs, are defined for disjoint supports and are metrics. However, the KS metric is bounded and saturating. Therefore, the WD exhibits all the desired characteristics while also providing meaningful non-saturating values[270] and is finite and defined

even when neither measure is absolutely continuous with respect to the other[271]. In addition, the WD takes into account the metric space i.e. the actual distance between points in the the two distributions rather than their distance in probability. Figure 5.9 shows the value of the statistical distances under consideration between two Gaussian distributions X_1 and X_2. We also include two common loss functions, (1) L_{MAE}, the mean absolute error, and (2) L_{RMSE}, the root mean squared error, to illustrate the drawbacks of saturating behaviour.



(a) $X_1 \sim N(\mu_1, 4), X_2 \sim N(0, 1)$

(b) $X_1 \sim N(2, \sigma_1^2), X_2 \sim N(0, 1)$

Figure 5.9: Dynamics of Various Statistical Distances for Gaussian Data

The Wasserstein Distance has been used in a range of applications such as generative adversarial networks [272], distributionally robust optimisation [266], bounding generalisation errors of machine learning models[273], and recently in probabilistic forecasting of wind power[274]. Interestingly, the WD also provides a natural way to link the input and output space. The dual formulation of the WD provides an alternative interpretation as the error in the expected value of 1-Lipschitz functions, h, due to the approximation of one distribution by another[275]:

$$d_W(X_1, X_2) = \sup_{\|h\|_{Lip} \leq 1} \left| \int h dX_1 - \int h dX_2 \right| \tag{5.15}$$

Although this definitional equivalence relates specifically to the WD, the ability to develop such bounds can be extended to other distances using relationships between distances (see [269, Figure 1]). Indeed, a connected line of work on developing theoretical performance guarantees for data-driven decision making in non-I.I.D. settings, has developed such bounds based on KS and TVD[276]. They show that unlike WD based bounds, the KS and the TVD based bounds are dependent on the diameter of the probability space and may therefore be less tight in general.

5.2.2 Wasserstein Distance based Data Valuation

We choose the WD as our data valuation metric due to the theoretical advantages over other distances laid out above. We now outline our proposed data valuation framework, which is developed with the valuation of aggregate smart meter data in mind, but can be applied more generally to other settings for valuing aggregate data. We assume there is a target or true distribution, X_T, which is an aggregation of N data sources, $X_1, \ldots, X_N$, owned by individuals $i \in \mathcal{N}$. Here, we consider the aggregation mechanism is the Euclidean barycenter $\left(X_T = \frac{1}{N}\sum_{i\in\mathcal{N}} X_i\right)$, however, this could be extended to other aggregation mechanisms such as the Wasserstein barycenter[277].

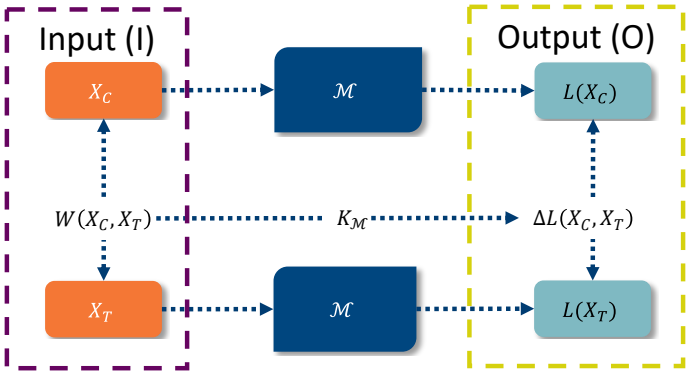


Figure 5.10: Overview of Proposed Valuation Framework

Under existing valuation frameworks such as cooperative games, value for a particular task $\mathcal{M}$, is based on calculating the difference in performance achieved with some subset of data sources $C \subseteq \mathcal{N}$ and the performance achieved without the data[38], [260]. We instead focus on the difference in performance achieved with a subset distribution $X_C = \frac{1}{|C|}\sum_{i\in C} X_i$ and the target distribution X_T. This allows us to incorporate the fact that, in our setting, the best performance for any task is achieved with access to all data sources i.e. when the target distribution is used. We define the performance difference as[6]:

$$\Delta L_{\mathcal{M}}(X_C, X_T) = L_{\mathcal{M}}(X_C) - L_{\mathcal{M}}(X_T) \tag{5.16}$$

where, $L_{\mathcal{M}}(X_C) = \mathbb{E}[l_{\mathcal{M}}(x_C)]$ and $l(\cdot)$ is a loss function/performance metric (e.g. mean absolute error).

Input-Output Relationship

We aim to value the individual data sources, X_i, which may be used for a variety of different tasks, using only information in the input domain, namely the WD between the

[6]Also referred to as the regret in machine learning literature, in this case w.r.t. a target distribution X_T.

data sources and the target distribution, $W(X_i, X_T)$, as shown in Figure 5.10. By placing mild assumptions of Lipschitz continuity on the loss function, $l(\cdot)$, we are able to develop a theoretically grounded relationship between the WD between two distributions and the expected difference in the loss function.

Definition 5.2.1. *(Lipschitz Continuity). Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ where, $d_{\mathcal{X}}$ denotes a metric on the input set $\mathcal{X}$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ where, $d_{\mathcal{Y}}$ denotes a metric on the output set $\mathcal{Y}$, a function $l : \mathcal{X} \rightarrow \mathcal{Y}$ is Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x_1 and x_2 in $\mathcal{X}$:*

$$d_{\mathcal{Y}}\left(l(x_i), l(x_2)\right) \leq K d_{\mathcal{X}}(x_1, x_2) \tag{5.17}$$

where, the smallest such K is known as the Lipschitz constant.

Many common loss functions, such as the logistic loss, Mean Pinball Loss (MPL) and hinge loss are Lipschitz continuous for any input. Whereas, the mean squared error is only Lipschitz continuous for a bounded input space. For regression models and neural networks, we need to ensure the loss function is Lipschitz continuous in both its data input and its parameters. This can be seen as a form of regularisation, which requires a constrained optimisation procedure. Indeed, Lipschitz regularisation has been investigated to control generalisation error[273], [278]. As such, requiring Lipschitz continuity is not necessarily restrictive and also provides desirable generalisation properties.

We can now formalise the relationship between the difference in performance for a class of K-Lipschitz loss function and the WD.

Theorem 5.1. *(Lipschitz Bound). Given a K-Lipschitz loss function, $l(x_i)$, for a task $\mathcal{M}$, the difference between the loss obtained using X_1 or X_2 is bounded by the Wasserstein distance between them[42]:*

$$|L_{\mathcal{M}}(X_1) - L_{\mathcal{M}}(X_2)| \leq K \cdot W_1(X_1, X_2) \tag{5.18}$$

where, $L_{\mathcal{M}}(X_i) = \mathbb{E}[l_{\mathcal{M}}(x_i)]$.

Proof. The proof follows directly from the definition of Lipschitz continuity and the dual representation of the WD in (5.15) with the metrics $d_{\mathcal{X}}, d_{\mathcal{Y}}$ being the l_1 norm. □

From a data valuation perspective the Lipschitz bound shows us that, if we minimise the WD between a data source X_i and the target data distribution X_T, we maximise performance.

Value Consolidation

As mentioned above, existing valuation mechanisms which use statistical distances assume data value is additive or simply value each data source in isolation. However, as highlighted in the case study in Section 5.1, the value of a particular data source is affected by the other data available and is therefore combinatorial. To capture this effect we would need to calculate the WD for each combination of data sources, which would be computationally intensive. However, we note that the Wasserstein distance converges as a function of the number of data sources [279]. As the WD value establishes an upper bound on the performance difference, we can take advantage of aggregation effects. To this end, we are able to develop a Hoeffding-type bound on the WD for any combination/subset C of data sources using only the WDs between the individual data sources X_i and the target distribution X_T.

Theorem 5.2. *(Hoeffding Bound). Given a target distribution X_T, which is an aggregation of N data sources $X_1, \ldots, X_N$, the WD between any aggregated subset distribution X_C (with $C \subseteq \mathcal{N}$) and the target distribution is upper bounded, for a given confidence level δ, by:*

$$t_{\delta,N}(C) \leq \sqrt{\left(\frac{N-n_c}{N}\right)\frac{\sum_{i\in C} W\left(X_i, X_T\right)^2 \ln\left(\frac{2}{1-\delta}\right)}{2n_c^2}} \tag{5.19}$$

where $n_c = |C|$

Proof. As the Wasserstein distance is a metric, we can apply the triangle inequality to upper bound the WD of any subset C of data sources:

$$W\left(\frac{1}{n_c}\sum_{i\in C} X_i, X_T\right) \leq \frac{1}{n_c}\sum_{i\in C} W(X_i, X_T)$$

Using the above result and the non-negativity of the WD we can view the WD of any subset C as a bounded random variable on the interval $\left[0, \frac{1}{n_c}\sum_{i\in C} W(X_i, X_T)\right]$.

We can then obtain a zero-concentration bound by applying the Hoeffding inequality and noting that $\mathbb{E}\left[W(\sum_{i=1}^{N} X_i, X_T)\right] = 0$:

$$P\left\{W\left(\sum_{i\in C}\frac{1}{n_c}X_i, X_T\right) \geq t\right\} \leq 2\exp\left(\frac{-2n_c^2t^2}{\sum_{i=1}^{n_c} W\left(X_i, X_T\right)^2}\right)$$

Then for a specified confidence level $\delta \in [0, 1)$ the resulting upper bound on the WD of any subset C is:

$$t_\delta(C) \leq \sqrt{\frac{\sum_{i\in C} W\left(X_i, X_T\right)^2 \ln\left(\frac{2}{1-\delta}\right)}{2n_c^2}}$$

Lastly, to account for the finite sample N we introduce a finite sample correction factor of $\left(\frac{N-n_c}{N}\right)$ [280]. □

5.2.3 Case Study: Synthetic Aggregates

To evaluate the performance of the WD as a data valuation metric, we setup a case study using synthetic data distributions. The proposed valuation framework introduces two layers of approximation:

- Lipschitz bound as an approximation/proxy for the actual value of a K-Lipschitz loss function,
- Hoeffding bound as an approximation of the WD for any subset of the data sources.

We investigate the WD's ability to capture value for a number of estimation tasks both in absolute (i.e. the tightness of the bounds) and relative terms (i.e. correlations between the actual value and the WD). As noted in Section 5.2.1, similar bounds could be developed for other statistical distances. We therefore also include a comparison to these other distances. Finally, to understand the implications of the approximation on payments in a data market we compare the Shapley payoff allocation proportions across the different tasks and distances considered.

Experimental setup

As the aim of using the Lipschitz bounds is to provide a task agnostic valuation metric, we investigate three use cases and associated loss functions, commonly applied to smart meter data; (1) mean estimation/Root Mean Squared Error (RMSE), (2) quantile estimation/MPL (including median/Mean Absolute Error (MAE)) and (3) newsvendor cost/NV. We test three different data distributions with different properties; Gaussian (symmetric and unbounded), Uniform (symmetric and bounded) and Exponential (asymmetric and unbounded). We generate 8 data sources over 50 trials with location and scale parameters distributed uniformly. The target distribution is the Euclidean barycenter of the 8 data sources and we calculate the distances and loss function values for all 255 combinations of data sources. The experiment parameters as summarised in Table 5.4.

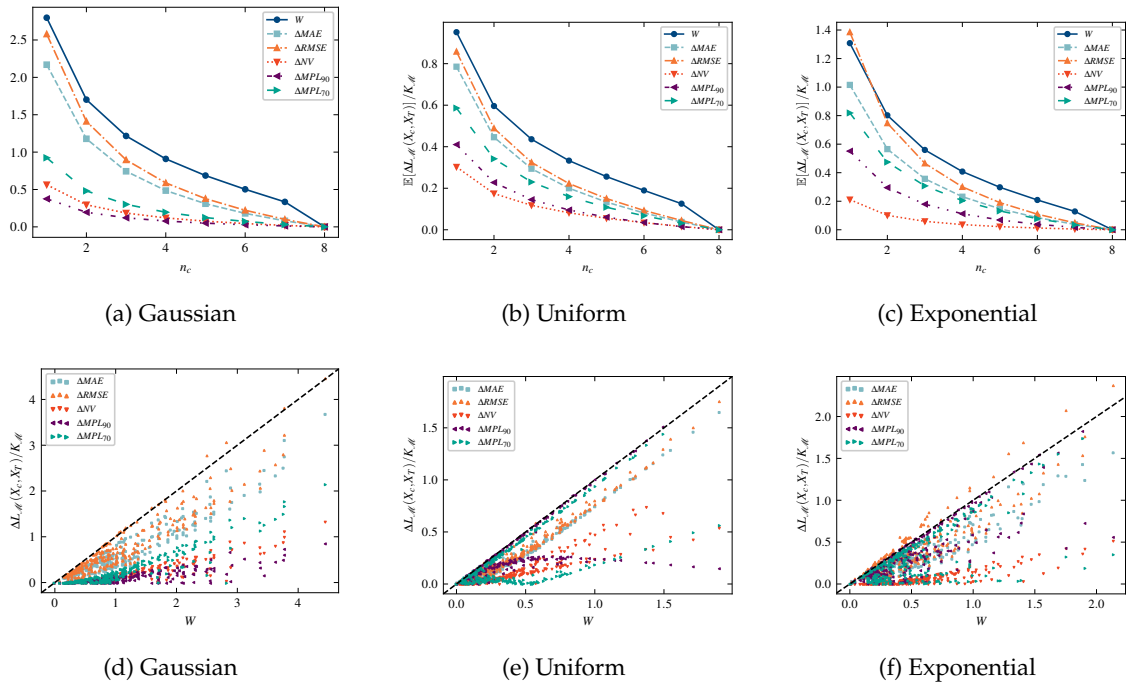
Table 5.4: Experimental Parameters for Data Valuation

Parameter	Value/Range
N	8
Trials	50
Distributions	Gaussian, Uniform, Exponential
Loc, Scale	$\alpha_i \sim U(10, 20)$, $\beta_i \sim U(1, 5)$
Loss Functions	RMSE, MAE, MPL_q q = {0.1, 0.2, . . . , 0.8, 0.9}, NV
Distances	WD, KLD, KS, TVD, JSD

Results

Lipschitz Bounds To assess the performance of the Lipschitz bounds, we are interested in the difference $\Delta L_{\mathcal{M}}(X_c, X_T) - K \cdot W(X_c, X_T)$, where, $\Delta L_{\mathcal{M}}(X_c, X_T)$ is the difference between the loss function for task, $\mathcal{M}$, when using a subset of the data, X_c, and using the target distribution, X_T. The top plots in Figure 5.11 show the expected value of the WD, $\mathbb{E}[W]$ (over the 50 trials), and $\Delta L_{\mathcal{M}}(X_c, X_T)$, divided by the associated Lipschitz constant, $K_{\mathcal{M}}$. The bottom plots show the loss function values, again divided by $K_{\mathcal{M}}$, against the WD for a particular trial. The dashed line represents the Lipschitz boundary with values below the line satisfying the bound and values above it violating the bound. From the top plots we see that the Lipschitz bound is tight (i.e. the difference is small), on average, for the MAE and RMSE across all distributions. However, for unbounded distributions (Gaussian and Exponential), the RMSE is in fact not Lipschitz, resulting in some coalition values violating the Lipschitz bound, as shown in bottom plots. For the MPL and NV the Lipschitz bound is much looser. This is due to the dynamics of the Lipschitz constant for the MPL, which will be discussed in more detail in Chapter 7.

Correlations Figure 5.12 shows the correlation performance between the distances and loss functions considered. The top plots show the average correlation coefficients ρ over the 50 trials and the bottom plots show whether using the WD results in a higher correlation with a target metric (y-axis) or the another source metric (x-axis) has a higher correlation. We see that the correlation between the WD and the loss functions is between 0.7 and 1.0. Overall, we see that the WD has higher correlations for almost all the considered loss functions. The KLD has higher correlations with MAE and RMSE for Gaussian data and the MPL for lower quantiles (10 & 20) for Uniform data. For exponential



(a) Gaussian (b) Uniform (c) Exponential

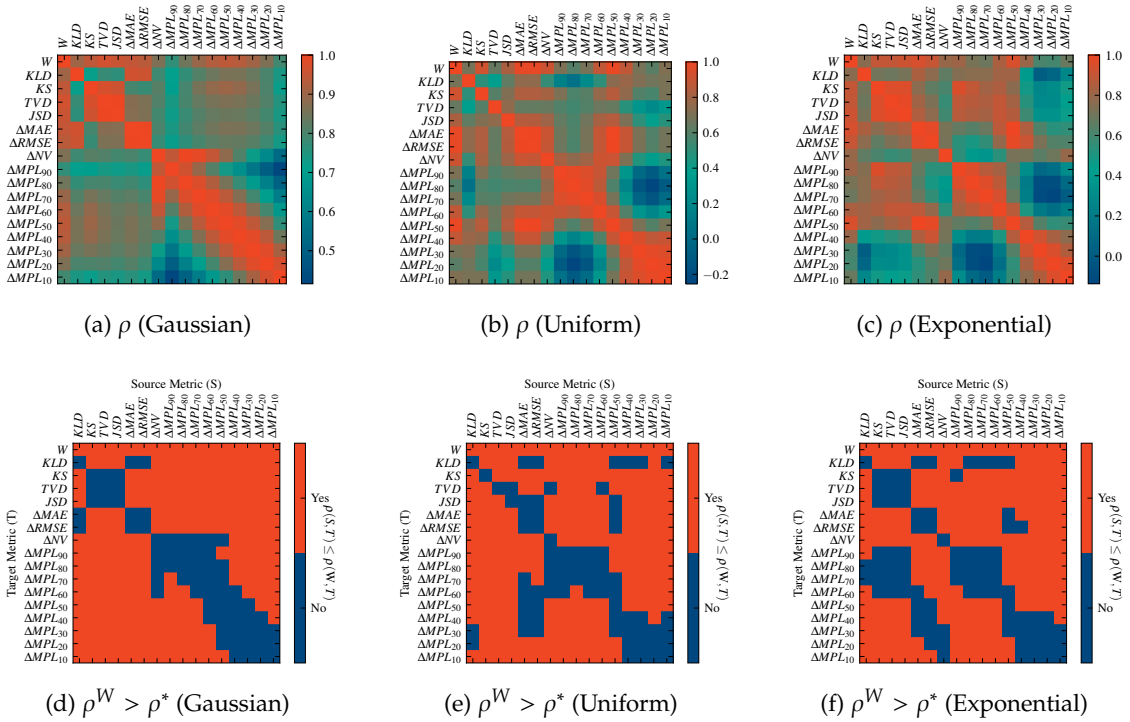
(d) Gaussian (e) Uniform (f) Exponential

Figure 5.11: Performance of Lipschitz Bounds for Different Data Distributions. Top: Mean of Metrics as Function of Coalition Size. Bottom: Scatterplot of all Coalitions against Wasserstein distance.

data we see that, for higher quantiles (60 to 90), the other distances out-perform the WD. Although we see that the best distance varies, the WD is consistent with high correlations across loss functions and distribution.

Shapley Values Figure 5.13 shows the Shapley allocation proportions for each data source using different (a) distances and (b) loss functions for Gaussian data[7] for a particular trial. For the loss functions the value function used was $V = \max_{s \subseteq \mathcal{N}}(L(X_s)) - L(X_c)$. Similarly, for distances the value function is $V = \max_{s \subseteq \mathcal{N}}(d(X_s, X_T)) - d(X_c, X_T)$. We see that across the distances the allocation proportions are quite similar with differences less than 5%. In contrast, the allocations exhibit significantly more variation across loss functions. Figure 5.13c shows the average differences or mis-allocations using different distances across the loss functions. The KLD performs better for MAE and RMSE and KS is better for higher quantiles. Overall, the WD results in either the least or second least in mis-allocations, suggesting a more stable notion of value. The results are broadly in line with the correlation analysis, i.e. higher correlations lead to lower average mis-allocations.

[7]Similar dynamics are observed for Uniform and Exponential data and are therefore omitted.



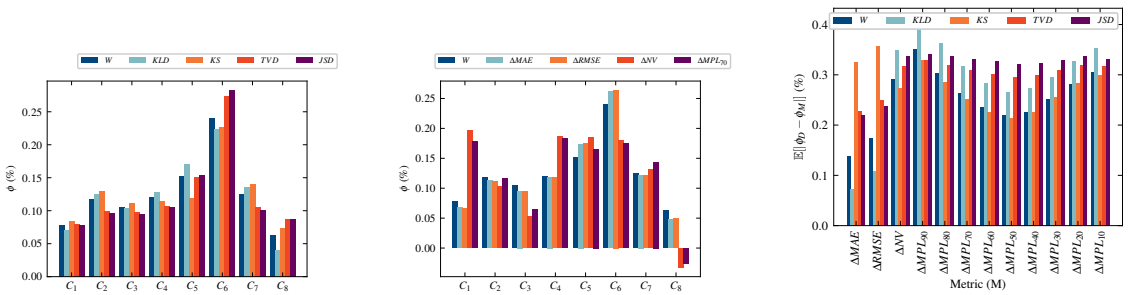
(a) ρ (Gaussian) (b) ρ (Uniform) (c) ρ (Exponential)

(d) $\rho^W > \rho^*$ (Gaussian) (e) $\rho^W > \rho^*$ (Uniform) (f) $\rho^W > \rho^*$ (Exponential)

Figure 5.12: Correlations between Distances and Loss Functions.

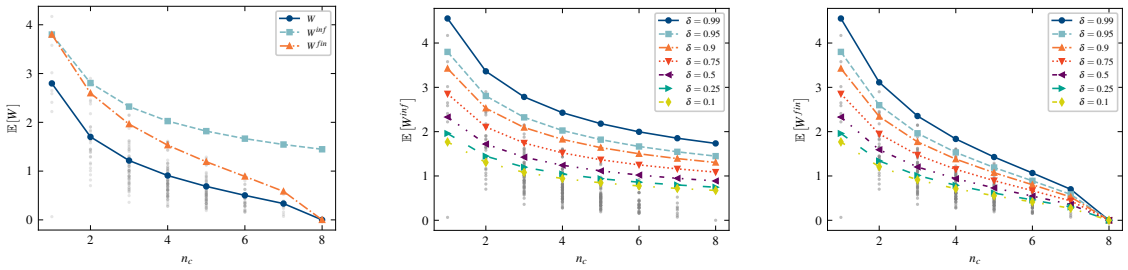
Hoeffding Bounds Figure 5.14a shows the expected value of the Hoeffding bounds and the actual WD. The grey dots are the WDs for each coalition of the given size. We include the Hoeffding bounds with (W^{fin}) and without (W^{inf}) the finite population correction as these have implications on the complexity of the market formulations that will be discussed in Chapter 6. The finite population formulation ensures convergence to zero with a full dataset whereas the infinite formulation maintains a non-zero bias which is more significant for smaller datasets. Figure 5.14b and 5.14c show the effect of tuning the confidence level δ on the Hoeffding bounds.

Ultimately, we aim to develop a data market which maximises data value. This



(a) Distances (b) Loss Functions (c) Average Mis-Allocation

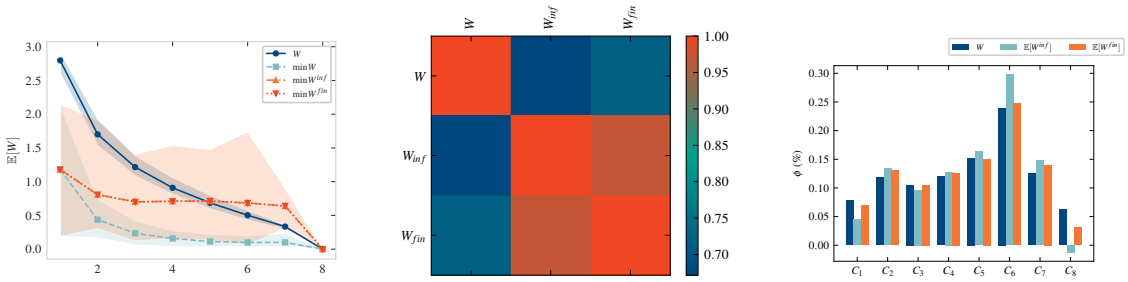
Figure 5.13: Shapley Allocations for Gaussian Data.



(a) Expected Value ($\delta = 0.95$) (b) Effect of δ on W^{inf} (c) Effect of δ on W^{fin}

Figure 5.14: Hoeffding Bounds for Gaussian Data. Grey Dots are Actual WDs for each Potential Coalition.

translates to minimising the WD. The Hoeffding bound offers an attractive option over which to optimise, as it captures aggregation effects without needing to calculate the WD for each combination of data sources. To this end, we compare the average minimisers of the actual WD, achieved when calculating each combination, and the minimisers of the Hoeffding bounds. As shown in Figure 5.15a, using the Hoeffding bounds improves average performance when compared to random selection (the average WD, W, for a given coalition size shown in dark blue), when the coalition size is small compared to the total number of data source ($n_c \leq 5$ in this case). As such, access to combinatorial information results in better performance overall, represented by $\min W$.



(a) Hoeffding Minimiser (b) Correlation (c) Shapley Allocation

Figure 5.15: Hoeffding Bound Performance with $\delta = 0.95$

The correlations between the actual WD and the Hoeffding approximations are $\rho(W, W^{fin}) = 0.72$ and $\rho(W, W^{inf}) = 0.67$. Again, we see this also effects the Shapley allocations, although the finite Hoeffding bound provides similar allocations to the actual WD. We also note that the Hoeffding bound approximation results in a bias. Although the Hoeffding bound accounts for the aggregation effects i.e. the convergence of the data distribution to the target distribution as $n_c \to N$, it does not capture the combinatorial effects of the

aggregation itself e.g. given two distributions which are individually far away from the target but when aggregated may be much closer to the target distribution.

5.3 Privacy in the Wasserstein Distance

We have shown that the WD as a valuation mechanism can be used to capture the relevant drivers of value for smart meter data and approximates value across tasks with Lipschitz functions. In addition, it does not require access to the underlying data as in [266], only the WD itself. The two remaining challenges of privacy-preserving calculation and modelling of the effect of differential privacy are explored in this section. First, we derive novel closed-form expressions for the WD as a function of distributional parameters. We then show how these can be used to develop exact expression and bounds on the effect of differentially-private noise addition on the WD. Finally we show how existing privacy-preserving techniques can be used to compute the WD in a privacy preserving manner when data sources are defined by distributional parameters as well as for empirical data.

5.3.1 Analytical Expressions for the Wasserstein Distance

The analytical definition of the WD (also called the Kantorovich distance, Mallow's distance, L_p-metric and the Earth Mover's Distance for special cases) is given by[281, Definition 6.1]:

Definition 5.3.1. *(p-WD). Let (X, d) be a Polish metric space, and let $p \in [1, \infty)$. For any two marginal measures μ and ν on X , the WD of order p between μ and ν is given by:*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{1/p} \tag{5.20}$$

$$= \inf\{\mathbb{E}[d(X, Y)^p]^{1/p}, \mu = F_\mu(X), \nu = F_\nu(Y)\} \tag{5.21}$$

where, $\Pi(\mu, \nu)$ denotes the collection of all measures on $\mathcal{X}$ with marginals μ and ν. The set $\Pi(\mu, \nu)$ is also called the set of all couplings of μ and ν.

In the univariate case, the WD simplifies to a function of the difference between the quantile functions ($F_X^{-1}(q) = \inf\{x : F_X(t) \geq q\}, q \in (0, 1)$):

$$W_p(X, Y) = \left(\int_0^1 |F_X^{-1}(q) - F_Y^{-1}(q)|^p dq \right)^{1/p} \tag{5.22}$$

When $p = 1$, our distance of choice, an alternative expression in terms of the CDF can be obtained:

$$W_1(X, Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| dt \tag{5.23}$$

In general, the WD does not admit closed-form expressions. There are two exceptions: (1) when X and Y are Gaussian then $W_2^2(X, Y)$ admits a closed-form expression and (2) when the distributions are univariate ($d = 1$)[275]. Although (5.23) provides a practical method for calculating the univariate 1-Wasserstein distance in many applications, a general closed-form or analytical expression directly in terms of distributional parameters remains desirable. Indeed, a closed-form/analytical representation would be more computationally efficient and convenient as it does not require the evaluation of an integral[282] and bypasses the need to conduct Monte-Carlo simulations.

It would also provide exact expressions for the effect of noise addition in differentially-private data analysis where the Wasserstein distance may be used as a metric of data utility. In this case it is also desirable to be able to compute the WD privately to avoid potential privacy infringements. However, to the best of our knowledge, a closed-form representation, in terms of distributional parameters, for the 1-Wasserstein distance for many widely used distributions (e.g. Gaussians) is not available either in the multivariate case or the univariate case. This section will focus on the 1-Wasserstein distance between independent univariate distributions belonging to a location-scale family.

Definition 5.3.2. *(Location-scale Distribution) For $\alpha \in \mathbb{R}$ and $\beta \in (0, \infty)$, let $X = \alpha + \beta Z$. The two-parameter family of distributions associated with X is called the location-scale family associated with the given (standard) distribution of $Z \sim (0, 1)$ if its CDF is a function only of $\frac{x-\alpha}{\beta}$:*

$$F_X(x \mid \alpha, \beta) = F\left(\frac{x - \alpha}{\beta}\right) \tag{5.24}$$

with the standard CDF defined as:

$$F_Z(x) = \Phi_Z(x) \tag{5.25}$$

Consequently, its quantile function can be expressed as:

$$F_X^{-1}(q \mid \alpha, \beta) = \alpha + \beta\Phi_Z^{-1}(q) \tag{5.26}$$

where $\Phi_Z^{-1}(q)$ is the quantile function for the standard distribution Z.

For the avoidance of confusion we specify that, α is the location parameter and β the scale parameter whereas μ denotes the usual mean, σ the standard deviation, and Σ the covariance matrix in the multivariate case. When referring to a specific random variable within the location-scale family, with a slight abuse of notation $\Phi_D(x)$ and $\Phi_D^{-1}(q)$ denote the standard CDF and quantile functions for a distribution $Z \sim D(0, 1)$, respectively (e.g. N for the Gaussian or Lap for the Laplace). Lastly, we use the term closed-form

throughout to mean either truly closed-form expressions (i.e. those based solely on elementary functions) or analytical expressions (i.e. those which include functions such as the Gamma function, $\Gamma(k)$, or the standard Gaussian CDF, $\Phi_N(x)$, which can be efficiently computed from lookup tables).

Using the above representations for univariate distributions, we provide improved distribution specific closed-form upper bounds and exact expressions for the 1-Wasserstein distance between two independent univariate location-scale distributions. The expressions are based solely on the location and scale parameters (α, β) of the two distributions in question and the standard quantile function $\Phi_Z^{-1}(x)$.

There are a number of well established bounds on the 1-Wasserstein distances for distributions with finite first and second moments (see Appendix C for details). The closed-form bounds provide good linear and quadratic approximations, however, it remains desirable to obtain an exact expression for the 1-Wasserstein distance. To this end, we present the Theorem below which is summarised graphically in Figure 5.16 and formally below:

Theorem 5.3. *Given two univariate independent random variables* $X_1 = \alpha_1 + \beta_1 Z$ *and* $X_2 = \alpha_2 + \beta_2 Z$ *the 1-Wasserstein distance between them is:*

$$W_1(X_1, X_2) = E[|Y|] \tag{5.27}$$

where, $Y = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2) Z$.

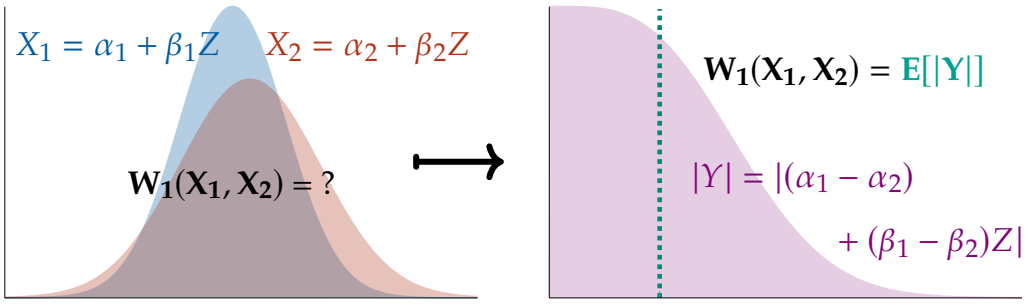


Figure 5.16: The 1-Wasserstein distance $W_1(X_1, X_2)$ between independent univariate location-scale distributions is equivalent to the mean of a folded distribution ($|Y|$) within the same family (Z) whose underlying location and scale are equal to the difference of the locations and scales of the original distributions.

Proof.

$$W_1(X_1, X_2) = \int_0^1 |F_1^{-1}(q) - F_2^{-1}(q)|dq \tag{5.28}$$

$$= \int_0^1 |\Big(\alpha_1 + \beta_1 \Phi_Z^{-1}(q)\Big) - \Big(\alpha_2 + \beta_2 \Phi_Z^{-1}(q)\Big)|dq \tag{5.29}$$

$$= \int_0^1 |\alpha_1 - \alpha_2 + (\beta_1 - \beta_2)\, \Phi_Z^{-1}(q)|dq \tag{5.30}$$

$$= \int_0^1 F_{|Y|}^{-1}(q)dq \tag{5.31}$$

Note that the integrand defines the quantile function, $F_{|Y|}^{-1}(q)$, of a folded/absolute value random variable distributed as Z. Specifically, the underlying random variable $Y = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)\, Z$. Using the substitution $q = F_{|Y|}(x)$ and $dq = F'_{|Y|}(x)dx$:

$$= \int_{-\infty}^{\infty} x f_{|Y|}(x)dx \tag{5.32}$$

$$= E[|Y|] \tag{5.33}$$

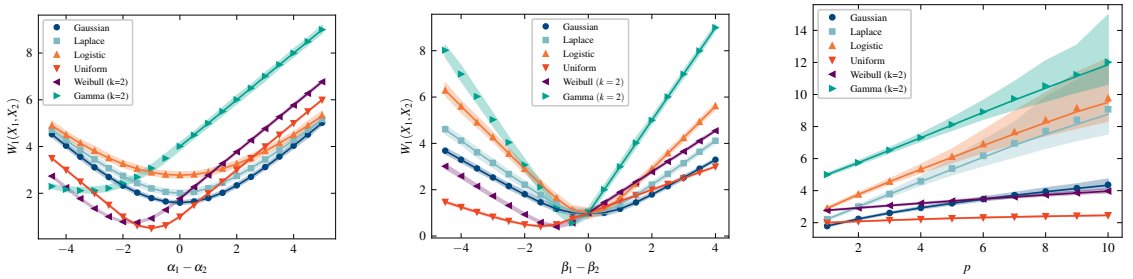
□

For conventionally non-negative distributions such as the Weibull distribution, the 1-Wasserstein distance is simply $E[|Y|] = E[Y]$. However, in the general case the mean of the folded variable is not readily available in the literature. As such, Figure 5.17 does not include the closed-form values (solid lines) when $\alpha_1 - \alpha_2 < 0$ or $\beta_1 - \beta_2 < 0$ for the Gamma or Weibull distributions. We show a basic example of how we can extend closed-form expressions for the uniform case in Appendix C.

Figure 5.17 shows the 1-Wasserstein distance for selected distributions. The empirical distance was calculated using the `Python Optimal Transport (POT)` package[283]. The empirical 1-Wasserstein distance (marker) is averaged over $N_r = 10^2$ simulations with $N_s = 10^4$ samples in each simulation. The shaded area indicates the 95% confidence interval. The closed-form expressions based on Theorem 5.3 are represented by the solid lines. We also provide a list of closed-form expressions for the 1-Wasserstein distance between selected location-scale distributions in Table C.1 in Appendix C.

We note that Theorem 5.3 can be extended to p-Wasserstein distances, using (5.22), with the resulting value being $E[|Y|^p]^{1/p}$. Figure 5.17c shows a Monte-Carlo simulation of the higher order WDs.

We highlight the use of Theorem 5.3 for the 1-Wasserstein distance between two univariate Gaussians. It is the mean of a folded Gaussian, providing an expression as a function of distributional parameters and the standard normal CDF, $\Phi_N(x)$.[284, Equation 7].



(a) $W_1(X_1 = 5+5Z, X_2 = \alpha_2+3Z)$ (b) $W_1(X_1 = 1+5Z, X_2 = \beta_2 Z)$ (c) $W_p^p(X_1 = 1+5Z, X_2 = 3+Z)$

Figure 5.17: Wasserstein distances for selected location-scale distributions

Corollary 5.4. *Given two univariate independent Gaussians* $X_1 \sim N(\mu_1, \sigma_1^2)$ *and* $X_2 \sim N(\mu_2, \sigma_2^2)$ *the 1-Wasserstein distance is equal to the mean of a folded Gaussian* $E[|Y|]$ *where* $Y \sim N(\mu_y = \mu_1 - \mu_2, \sigma_y^2 = (\sigma_1 - \sigma_2)^2)$*:*

$$W_1(X_1, X_2) = |\mu_y| \left[1 - 2\Phi_N\left(-\frac{|\mu_y|}{|\sigma_y|}\right)\right] + |\sigma_y|\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_y^2}{2\sigma_y^2}\right) \tag{5.34}$$

This section explored the properties of the 1-Wasserstein distance in the univariate case. We provided an exact analytical expression for the 1-Wasserstein distance between independent univariate location-scale distributions based solely on distributional parameters and special functions such as the standard Gaussian. Further work is needed to determine whether this approach can be extended to the multivariate case given that the theorems presented in this section rely on the monotony of transport in the univariate case, which allows the WD to be expressed in terms of quantile functions.

5.3.2 Differential Privacy in the 1-Wasserstein Distance

In this section we provide improved closed-form bounds for the 1-Wasserstein distance between a differentially-private data distribution $X_{DP} = X_1 +$ DP-noise and a reference (non-private) data distribution X_2. In particular, two popular noise addition mechanisms; (1) the Laplace mechanism and (2) the Gaussian mechanism, are studied. Additionally, in the case the data distributions X_1 and X_2 are Gaussian, a common assumption, we provide an exact expression and an approximation for the Gaussian and Laplace Mechanisms respectively.

Gaussian Mechanism

Based on Definition 5.1.5 of the Gaussian mechanism the differentially-private data distribution $X_{DP} = X_1 + X_N$. An upper bound based on the triangle inequality and Jensen's inequality is provided in [285, Theorem 2]:

$$W(X_1 + X_N, X_2) \leq W(X_1, X_2) + \frac{\sqrt{2\ln(1.25/\delta)}\Delta}{\epsilon} \tag{5.35}$$

Using Corollary 5.4, we generate improved upper bounds and in the case X_1 and X_2 are Gaussian, we can provide an exact expression for the 1-Wasserstein distance between differentially-private data and a reference distribution under the Gaussian mechanism. We first present the improved upper bound in Proposition 5.5.

Proposition 5.5. *Given a (ϵ, δ)-DP data distribution $X_{DP} = X_1 + X_N$, where, $X_N \sim N(0, \frac{2\ln(1.25/\delta)\Delta^2}{\epsilon^2})$, and a reference (non-private) data distribution X_2, the 1-Wasserstein between them is upper bounded by:*

$$W(X_1 + X_N, X_2) \leq W(X_1, X_2) + \frac{2\Delta}{\epsilon}\sqrt{\frac{\ln(1.25/\delta)}{\pi}} \tag{5.36}$$

Proof. As the WD is a metric, it obeys the triangle inequality:

$$W(X_1 + X_N, X_2) \leq W(X_1, X_2) + W(X_N, \delta_0) \tag{5.37}$$

where, δ_0, is the dirac delta distribution concentrated at 0. The second term can be reduced to:

$$W(X_N, \delta_0) = \sigma_N\sqrt{\frac{2}{\pi}} \quad \text{using (5.27)} \tag{5.38}$$

$$= \frac{2\Delta}{\epsilon}\sqrt{\frac{\ln(1.25/\delta)}{\pi}} \tag{5.39}$$

□

Next, if X_1 and X_2 are also Gaussian, we can provide an exact expression as X_1 and X_N are independent and the resulting differentially-private data is distributed as $X_{DP} \sim N(\mu_1, \sigma_1^2 + \frac{2\ln(1.25/\delta)\Delta^2}{\epsilon^2})$.

Corollary 5.6. *Given a (ϵ, δ)-DP data distribution $X_{DP} = X_1 + X_N$, where $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_N \sim N(0, \frac{2\ln(1.25/\delta)\Delta^2}{\epsilon^2})$, and a reference (non-private) data distribution $X_2 \sim N(\mu_2, \sigma_2^2)$, the 1-Wasserstein between them is:*

$$W(X_1 + X_N, X_2) = |\mu_y|\left[1 - 2\Phi\left(-\frac{|\mu_y|}{|\sigma_y|}\right)\right] + |\sigma_y|\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_y^2}{2\sigma_y^2}\right) \tag{5.40}$$

where, $\mu_y = \mu_1 - \mu_2$, $\sigma_y = \sigma_{DP} - \sigma_2$ and $\sigma_{DP} = \sqrt{\sigma_1^2 + \frac{2\ln(1.25/\delta)\Delta^2}{\epsilon^2}}$.

Laplace Mechanism

Based on Definition 5.1.4 of the Laplace mechanism the differentially-private data distribution is $X_{DP} = X_1 + X_L$. A similar upper bound to (5.35) for the Laplace mechanism is also provided in [285, Theorem 2]:

$$W(X_1 + X_L, X_2) \leq W(X_1, X_2) + \frac{\sqrt{2}\Delta}{\epsilon} \tag{5.41}$$

However, similar to Proposition 5.5, in the univariate case it is possible to obtain a tighter upper bound.

Proposition 5.7. *Given a ϵ-DP data distribution $X_{DP} = X_1 + X_L$, where, $X_L \sim Lap(0, \frac{\Delta}{\epsilon})$, and a reference (non-private) data distribution X_2. The 1-Wasserstein between them is upper bounded by:*

$$W(X_1 + X_L, X_2) \leq W(X_1, X_2) + \frac{\Delta}{\epsilon} \tag{5.42}$$

Proof. Again by the triangle inequality:

$$W(X_1 + X_L, X_2) \leq W(X_1, X_2) + W(X_L, \delta_0) \tag{5.43}$$

where, δ_0, is the dirac delta distribution concentrated at 0. The second term can be reduced to the following by applying Theorem 5.3 and the mean of the folded Laplace [286, Proposition 2.3]:

$$W(X_L, \delta_0) = \mathbb{E}[|Y|], \quad Y \sim Lap(0, \beta_y) \tag{5.44}$$

$$= |0| + |\beta_y| \exp\left(-\frac{|0|}{|\beta_y|}\right) \tag{5.45}$$

$$= |\beta_y| \tag{5.46}$$

$$= \frac{\Delta}{\epsilon} \tag{5.47}$$

□

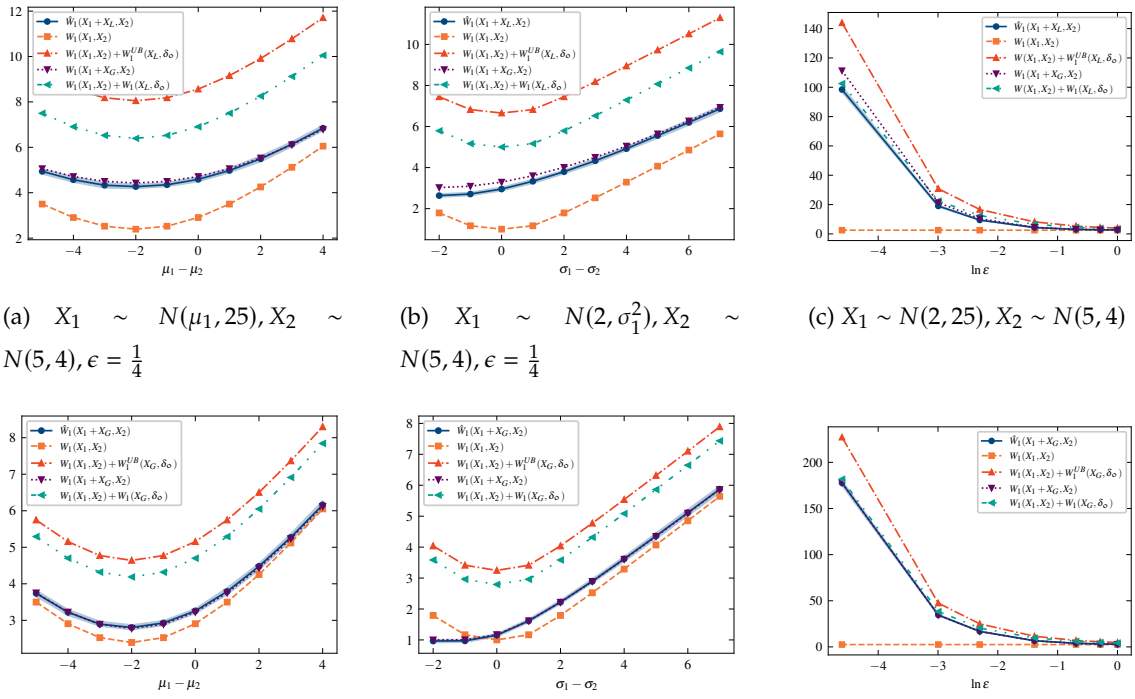
If X_1 and X_2 are Gaussian, we can provide an additional bound that is better than Proposition 5.7 for larger ϵ. The actual differentially-private data distribution will follow a Gaussian-Laplace distribution $X_{DP} \sim NL(\mu_1, \sigma_1, 1/b_l, 1/b_l)$ [287]. The mean and variance are $E[X_{DP}] = \mu_1$ and $Var[X_{DP}] = \sigma_1^2 + 2b_l^2$ respectively. Interestingly, for a given b_l the Gaussian-Laplace is also a location-scale distribution. However, applying Theorem 5.3 would require the computation of the following quantity, $\int_0^1 |\beta_{DP}\Phi_{NL}^{-1}(q) - \sigma_2\Phi_N^{-1}(q)|dq$

(where, $\beta_{DP}(\sigma_1, b_l)$ is the scale parameter given that $X_{DP} = \mu_1 + \beta_{DP} Z_{NL}$), for which there is no closed form or convenient lookup table. Instead we observe that the Gaussian-Laplace can be approximated with high accuracy by a Gaussian with the same mean and variance. This is especially true when b_l is smaller than or comparable to σ_1.

Observation 5.8. *Given an ϵ-DP data distribution $X_{DP} = X_1 + X_L$, where, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_L \sim Lap(0, \frac{\Delta}{\epsilon})$, and a reference (non-private) data distribution $X_2 \sim N(\mu_2, \sigma_2^2)$ the 1-Wasserstein between them can be approximated by the 1-Wasserstein between Gaussians. If $\sigma_1 \gtrapprox \sqrt{2}\frac{\Delta}{\epsilon}$ the following holds:*

$$W(X_{DP}, X_2) \approx |\mu_y| \left[1 - 2\Phi_N\left(-\frac{|\mu_y|}{|\sigma_y|}\right)\right] + |\sigma_y| \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_y^2}{2\sigma_y^2}\right) \tag{5.48}$$

where, $\mu_y = \mu_1 - \mu_2$, $\sigma_y = \tilde{\sigma}_{DP} - \sigma_2$, and $\tilde{\sigma}_{DP} = \sqrt{\sigma_1^2 + 2\frac{\Delta^2}{\epsilon^2}}$.



(a) $X_1 \sim N(\mu_1, 25), X_2 \sim N(5, 4), \epsilon = \frac{1}{4}$

(b) $X_1 \sim N(2, \sigma_1^2), X_2 \sim N(5, 4), \epsilon = \frac{1}{4}$

(c) $X_1 \sim N(2, 25), X_2 \sim N(5, 4)$

(d) $X_1 \sim N(\mu_1, 25), X_2 \sim N(5, 4), \epsilon = 1$

(e) $X_1 \sim N(5, \sigma_1^2), X_2 \sim N(5, 4), \epsilon = 1$

(f) $X_1 \sim N(2, 25), X_2 \sim N(5, 4)$

Figure 5.18: Differential Privacy in the 1-Wasserstein Distance using Laplace (top) and Gaussian (bottom) Additive Noise Mechanisms

Figure 5.18 illustrates the improved bounds provided above. We assume $\Delta = 1$ and $\delta = 10^{-2}$. The empirical 1-Wasserstein distance ($\hat{W}_1$) averaged over $N_r = 10^2$ simulations with $N_s = 10^4$ samples in each simulation. The shaded area indicates the 95% confidence interval. The bounds provided in (5.35) and (5.41) are denoted $W_1(X_1, X_2) + W_1^{UB}(X_L, \delta_0)$

and $W_1(X_1, X_2)+W_1^{UB}(X_G, \delta_0)$ respectively. The improved bounds $W_1(X_1, X_2)+W_1(X_L, \delta_0)$ and $W_1(X_1, X_2) + W_1(X_G, \delta_0)$ perform significantly better than the previous bounds, especially for smaller privacy budgets(ϵ). The Gaussian approximation (5.48) of the Laplace Mechanism denoted $W_1(X_1 + X_G, X_2)$ in Figures 5.18(a)-(c) performs well compared to the upper bounds for larger privacy budgets(ϵ), where, $\sigma_1 \gtrapprox \sqrt{2}\frac{\Delta}{\epsilon}$.

In this section, the effect of differentially-private noise addition on the 1-Wasserstein distance was modelled explicitly providing a tighter upper bound for both the Laplace and Gaussian mechanism when the data are univariate and an exact expression in the case where the distributions are also Gaussian.

5.3.3 Private Computation

We have shown that when the data under consideration are within the same location-scale family we can obtain closed-form representations in terms of distributional parameters. As a result, the computation of the WD is equivalent to calculating aggregate sums of parameters. This can be efficiently calculated using one or a combination of the Privacy-Preserving Techniques (e.g. DP, homomorphic encryption or MPC) discussed in Chapter 4.

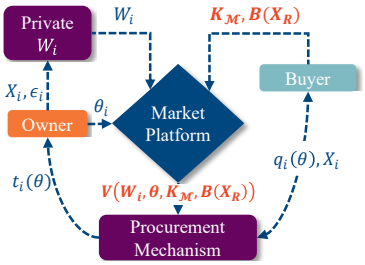
For empirical data where placing distributional assumptions may be undesirable, the WD between two discrete one dimensional distributions can still be calculated privately and efficiently. Ref. [201] shows that by encoding the discrete data histograms as two-dimensional points, it can be shown that the WD is equivalent to the size of the set intersection between the two histograms. Further, the size of a set intersection or cardinality can be efficiently calculated in a privacy-preserving manner, using existing MPC techniques. The problem, namely Private Set Intersection-Cardinality (PSI-CA), has been extensively studied in the MPC literature[288], [289]. Specifically, techniques such as bloom filters allow for private computation with or without a trusted central entity. MPC has also been proposed for calculating correlations among data from IoT devices in a privacy-preserving manner, also in the context of data valuation[290]. We note that the use of DP would ensure consistency across the valuation framework. However, this would require the privacy budget to be split between the WD calculation and the aggregate smart meter data. We note that there are also differentially private algorithms to compute the PSI-CA [291].

5.4 Discussion

This chapter proposed a data valuation framework based on the WD with a focus on applications for differentially private smart meter data. First, we identified the drivers of value for differentially private data through a forecasting and energy procurement framework. We showed that value is highly contextual, depending on the availability of reference data, risk appetites and the level and distribution of privacy concerns, as well as energy market conditions. We then presented the WD, a statistical distance, as a theoretically grounded data valuation metric. As a measure of difference between data distributions in the input space, it is a task agnostic valuation metric. As such, it induces the inevitable trade-off between accuracy for a particular task and applicability, as a measure of value, across multiple tasks. We provided justification for choosing the WD over other potential statistical distances based on its statistical properties. We also showed, through simulations with synthetic data, that the WD performs better than the other distances considered, across the tasks and distributions.

Conventional data valuation mechanisms use performance improvement for a particular task as a indication of value. We show that the WD can also be used in this context as it provides a natural upper bound for tasks with Lipschitz loss functions. We then also tackle the combinatorial nature of data value by presenting a linear approximation scheme for the WD using the Hoeffding bound. We showed that these approximations do degrade performance. However, we see that Shapley allocations for a range of tasks are broadly consistent using these approximations versus using the actual task performance. The effect of these approximations will be investigated in more detail in the context of a data market mechanism in the next chapter.

Finally, we provide novel closed-forms for the WD for location-scale distributions. Although these are limited to univariate data, this is sufficient for application to smart meter data. The closed-form expressions are used to endogenously model DP in the WD, which allows us to provide improved bounds and an exact expression for the impact of DP. We also show how the WD can be computed in a privacy-preserving manner, using existing MPC techniques, both for location-scale distributions defined by their distributional parameters and for empirical data. Combining MPC and DP, for this purpose, would ensure consistent privacy guarantees across data valuation as well as data usage. Optimising this split of privacy budget, ϵ, across the two functions would be an interesting area for further research.



CHAPTER 6
Market for Differentially-Private Data

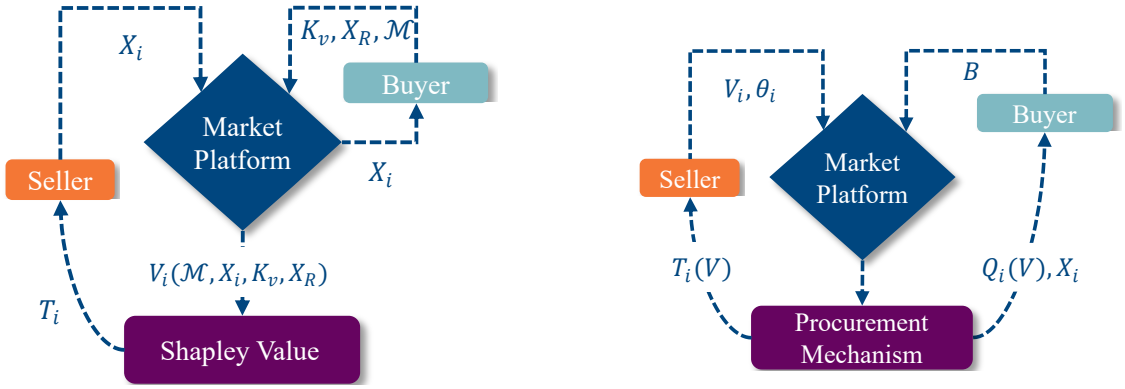
Having established the WD as an appropriate data valuation metric, we now shift our attention to using it to develop a data market or procurement mechanism. We propose three procurement mechanisms under two different scenarios. For task-agnostic data procurement we propose a budget feasible mechanism. For task-specific data procurement we propose an endogenous budget feasible mechanism as well as a mechanism which optimises both data value and payments. In each case, we provide reformulations of the mechanism design problem as Mixed Integer Second Order Conic Program (MISOCP), which can be solved using commercial solvers.

Section 6.1 starts by providing an overview of existing market mechanisms for data procurement and positioning our proposed mechanisms. Section 6.2 outlines our proposed procurement mechanisms. Section 6.3 investigates the performance of the proposed mechanisms on synthetic Gaussian data, extending the case study presented in Section 5.1.5. The majority of this chapter forms part of [Paper E].

6.1 Limitations of Existing Data Market Mechanisms

We limit ourselves to a market structure with a single data buyer (e.g. the energy retailer), multiple data sellers (e.g. consumers with smart meters) and a market platform (which could be the buyer). Consequently, a procurement mechanism (shown in Figure 6.1) takes as inputs a data valuation metric (V), sellers' reserve prices or willingness-to-sell (θ) and a buyers' willingness-to-buy or budget (B), and outputs allocations (Q i.e. the sellers'

data has been purchased), and payments for data sellers (T)[1].



(a) Cooperative Games

(b) Incentive Mechanism Design

Figure 6.1: Data Procurement Mechanisms

Here, like with the data valuation mechanism, we focus on developing a data procurement mechanism which accounts for the challenges identified in previous chapters. Of particular relevance are the facts that; some consumers do not trust retailers or third parties to handle their data or to share the benefits of data sharing with them, consumers' WTP/A for privacy is dependent on the level of privacy offered, and some consumers would demand compensation for data sharing regardless of privacy concerns. This motivates a procurement mechanism which is able to model consumers WTP/A (non-zero seller reserve prices), including its dependence on privacy levels (the privacy budget, ϵ, under DP). Again, as with the valuation mechanism, the procurement mechanism should be able to operate without the need for a TTP to determine payments.

Although we present the WD as an indicator of value across a broad range of tasks, in many settings the buyers' willingness-to-buy for data is dependent on the actual or expected improvement offered by the data for a particular task (e.g. through a reduction in uncertainty). Namely, data procurement decisions impact the costs/profits for a particular task, while at the same time costs/profits impact the data procurement decisions. This decision-dependent structure requires the modelling of the buyers' budget endogenously. This is closely linked to the field of optimisation under decision-dependent uncertainty[292]. For example, in [293] a two-stage robust scheduling problem is proposed, where the operator is able to purchase forecasts from distributed prosumers in the first stage, to reduce uncertainty in the second stage. Although we do not adopt a robust or distributionally robust decision-dependent modelling framework, our valuation

[1]Here, we present a general structure in order to aid the exposition of existing data procurement mechanisms. Our proposed model is formalised in Section 6.2

framework (WD and Lipschitz bounds) could be incorporated into such a framework. In addition, in order to establish a baseline for this endogenous structure, when no data is purchased, we need a method to incorporate reference data (X_R).

As summarised in Table 6.1, the criteria for assessing existing data market mechanisms are:

- Trusted-Third Party (TTP): Does the mechanism require a TTP to compute payments?
- Reserve Prices (RP): Does the mechanism model reserve prices for sellers?
- Endogenous Differential Privacy (EDP): Does the mechanism allow for the endogenous modelling of privacy parameters (ϵ)?
- Endogenous Budget (EBD): Can the mechanism incorporate the dependence of a buyer's budget on data procurement decisions?
- Reference Data (X_R): Can the mechanism incorporate the presence of (free) reference data or public information?

Cooperative Games Refs. [37], [38], [260] model the data procurement mechanism or data market as a cooperative game frameworks. Here, it is assumed that by sharing data and entering into a coalition, value (e.g. forecasting accuracy) will be improved. The coalition with all participants is called the grand coalition and is usually assumed to be the coalition with maximum value. The aim of the market platform in a cooperative game is to determine a payment policy which ensures that each participant is no worse off in the grand coalition than in another subset (incentive compatibility in this setting) or by not participating in the game at all (individual rationality).

The above studies, based on cooperative games, do not model reserve prices (θ), instead they assume a non-negative utility is sufficient to ensure participation. For example, in [39], which models the energy procurement problem with a market for information of scheduleable load, the authors note a post allocation adjustment would be required to account for privacy concerns. These techniques also assume the buyer[38], [39] or a TTP[37], [260] have full access to data and model. We note that there are methods to determine payoffs in a distributed (e.g. the blockchain based consensus and verification framework proposed in [294]) or privacy preserving manner (e.g. using DP[295]). They use the Shapley value to determine payoff allocations and a coefficient, K_v, to convert

Table 6.1: Existing Data Market Mechanisms

Type	Mechanism	Valuation Metric	TTP	RP	EBD	EDP	X_R	
Cooperative Game	Comb. Auction	$L(X)$	Yes		✓			[37], [260]
			Yes		✓		✓	[38], [39]
			Yes	✓	✓	★[1]		[261]
Incentive Mechanism Design	Prior-Free	$\epsilon(\theta)$	No	✓		✓		[50]
		JSD[6]	No	✓				[43]
	Bayesian	ϵ	No	✓		★[1]		[51]
		MSE($\epsilon(\theta)$)	No	✓	✓	✓	★[1]	[52]
Other	Revenue Sharing	JSD	No	✓		★[1]	✓	[44]
	Marginal Cost/Revenue	$\Delta\sigma$	Yes	✓	✓			[268]
	Regression	LASSO	Yes	✓	★[1]		✓	[40]
Proposed	Bayesian	$W(\epsilon)$	No	✓	✓	★[1]	✓	

[1] ★ indicates that the criterion is modelled, but not in the form required.

this into monetary values. This is calculated using the value obtained by each coalition. Hence, they are able to capture the endogenous budget dynamics. However, this is computationally intensive, and importantly is dependent on the model complexity of the underlying task. We note that this is an active area of research with approximation schemes, similar to the Hoeffding bound presented in the previous chapter[2], that improve computational tractability [295], [296]. We also note that these frameworks assume the buyer provides a model to the platform with which the Shapley values are calculated. As a result, the method is open to manipulation by buyers if they choose to mis-specify

[2]Specifically, the Hoeffding bound is used to determine the sampling rate for calculating coalition values to obtain an estimate of the Shapley value with a desired accuracy level.

their model to depress data value (e.g. using an incorrect lag structure, not appropriately transforming the data for an ARX model or the use of a less complex model). As a result, these approaches limit the buyer to purchasing the model output (e.g. forecast) rather than the actual data.

Ref. [260] presents a two-sided market where buyers have private values, which are elicited through an incentive mechanism design approach (which we explore further in the next paragraph), and sellers are assumed to have no privacy concerns or reserve prices. This has been applied in the energy domain to procurement of wind power forecast trading in[37]. Ref. [261] extends [260] by allowing sellers privacy preferences to be considered and returns differentially-private trained models to the buyer. The framework assumes sellers have monotonic compensation functions which are dependent on their privacy preferences and the Shapley value calculated on the original data. Similarly, buyers price functions are based on their budget, Shapley coverage and noise sensitivity. The effects of noise, introduced by DP, are therefore formulated in terms of subjective costs of the buyers and sellers rather than being connected explicitly to model accuracy. Both [38], [39] explicitly model the availability of reference information X_R, a baseline forecast performance achieved by a retailer/central agent without any additional data from data sellers.

Incentive Mechanism Design An alternative approach is incentive mechanism design. Here it is assumed sellers have private reserve prices (θ) and some value (V). The aim of the market platform is to optimise an objective, which could be the value subject to a budget (B), or a combination of payments made to sellers (T) and the procured value. As the reserve prices are assumed to be private, the market platform aims to develop an allocation and payment policy, which ensures truthful bidding by sellers (incentive compatibility) and that sellers are no worse off by participating in the market (individual rationality). When the aim of the platform is a function of payments, a Bayesian approach is required, i.e. a prior distribution over the sellers' reserve prices is needed to obtain incentive compatibility.

Incentive mechanism design approaches inherently model the reserve prices. The frameworks in [43], [50], [51] present budget feasible mechanisms i.e. mechanisms which maximise some notion of value subject to an exogenous budget provided by the buyer. In [50], [51] the platform aims to purchase data so as to maximise the privacy budget, ϵ, assuming data is I.I.D., and do not require a TTP as the platform only needs access to ϵ and θ. In [43], the platform aims to minimise the harmonic mean of the JSD of the procured

data (non-I.I.D. setting), where the JSD for each seller is made differentially private using the exponential mechanism. As these mechanisms rely on an exogenous budget, they do not account for the decision-dependent or endogenous structure.

Ref. [52] presents a platform which aims to optimise a mean estimator for I.I.D. data sellers with heterogeneous, compensation-dependent privacy parameters, $\epsilon(\theta)$. The objective of the platform is to minimise the sum of the MSE and payments, and therefore models the endogenous budget dependence and the effect of DP. The formulation assumes that data sellers gain, both in monetary terms through payments, as well as in terms of an improvement in the mean estimate provided by the platform compared to using their own data. As such, in the reference case each seller incurs a cost rather than the buyer or platform.

Other Approaches There are a number of other mechanisms which do not fall into the above categories. Ref. [44] proposes a revenue sharing model for aggregated differentially private data, where payments (T) are determined by the JSD, the seller's reserve price, and the size of the dataset. In addition, it is assumed the platform takes a commission from each seller. Similar to [51], it is assumed that sellers have fixed privacy preferences. Data is first made ϵ-differentially private, after which the JSD is calculated. However, no analytical connection between the JSD and ϵ is provided, impeding the modelling on endogenous privacy preferences. In [268], the authors propose a market clearing mechanism for the retailer energy procurement problem. The retailers' demand curve is based on the valuation mechanism developed in [297]. The supply curve is generated by modelling data sellers' reserve prices, as a function of a privacy sensitivity parameter. However, the mechanism relies on the assumption of perfect competition and thus does not consider incentive compatibility. Ref. [40] proposes a regression market where payments are based on the Least Absolute Shrinkage and Selection Operator (LASSO). The data sellers' reserve prices are translated into the penalty parameters of the LASSO regression. By incorporating the data market within the regression problem it provides the data buyer with a means for selecting useful features directly, without the need to calculate a separate metric like the Shapley value.

Proposed Mechanism To achieve the criterion described above we present three mechanisms, using Bayesian incentive mechanism design, for different scenarios:

1. **Exogenous Budget**: The buyer aims to maximise data value under an exogenous budget (B). Here we focus on task agnostic procurement. We extend the mecha-

nisms proposed in [43] and [51] which deal with the non-I.I.D. setting without DP and the I.I.D. setting with fixed heterogeneous privacy preferences, respectively. Our first contribution is the use of the WD to capture both heterogeneity/intrinsic value and DP effect. Second, we propose to optimise over the Hoeffding bound as opposed to assuming data is additive.

2. **Endogenous Budget Constraints**: The buyer aims to maximise data value with endogenous budget constraints. Here we assume the buyer still provides an exogenous budget, $B(X_R)$, where this represents the maximum they are willing to pay. However, the buyer also wants to ensure that payments made for data procurement will lead to at least as much of an increase in value, again in monetary terms, for a specific task. Here, we employ the Lipschitz bound to connect data value to task specific value. Our contributions here are two-fold; we provide a novel method to account for the decision-dependent structure of data procurement for a specific task, and unlike existing cooperative game frameworks we avoid the need to share the model for the specific task, requiring only the Lipschitz constant $K_{\mathcal{M}}$. This provides the buyer with model privacy, provides resistance of model mis-specification, and avoids the need to actually train the model during the valuation and procurement process.

3. **Joint Optimisation**: The buyer aims to optimise data value and payments simultaneously. Here, instead of optimising data value alone subject to an endogenous budget constraint, as above, we assume the buyer aims to optimise the data value and payments for a specific task. This scenario is closely linked to the work of [52]. In terms of contributions, in addition to those mentioned for the endogenous budget constrained mechanism, we broaden the scope of the platform problem to minimising the expected loss of any Lipschitz loss function and payments, as opposed to the mean squared error in [52]. We also include the effect of reference data the buyer may possess or which they could obtain at zero cost (e.g. public datasets). Although the mechanism we present here is limited to data sellers with fixed privacy preferences, our formulation leaves open the possibility of extending the framework to account for endogenous DP[3].

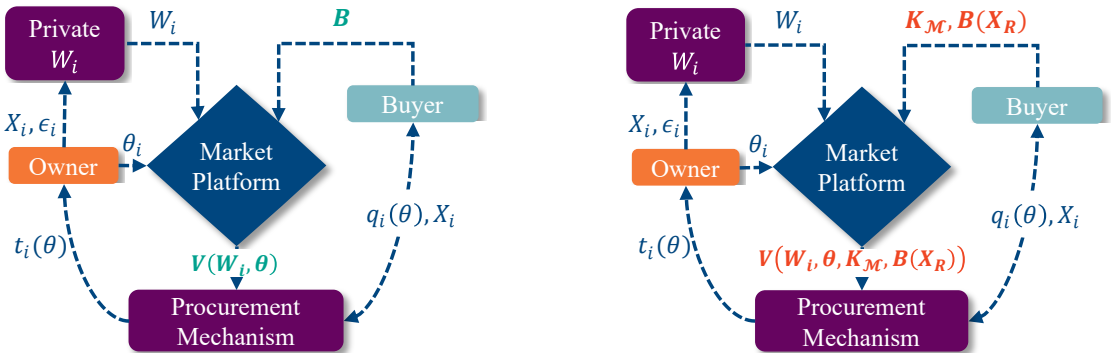
[3] As shown in Section 5.3.2, the effect of DP can be incorporated directly into the WD to allow the modelling of compensation dependent ϵ.

6.2 Incentive Mechanism for Differentially-Private Data

6.2.1 Modelling Framework

In this section we formalise the three proposed mechanisms. We start by describing the modelling framework, namely Bayesian incentive mechanism design, which is common to the three propose mechanisms. Following this we detail the differing objectives and budget constraints of each proposed mechanism.

Most of the analysis for a Bayesian optimal mechanism detailed in [298] and in [52], are directly applicable to our proposed mechanism. However, as will be described below, our problem is not separable leading to a more complex solution method. For completeness and notational consistency, we present our adaptation of the modelling framework and relevant results in [52], [298] in full. We use the notation established in [298].



(a) Exogenous Budget

(b) Endogenous Budget Constraints/Joint Optimisation

Figure 6.2: Proposed Data Procurement Mechanisms

Data Owners/Sellers

We assume we have N data sellers, where $i \in \mathcal{N}$. Each seller has a private dataset, $X_i \in \mathcal{X}$, and a private reserve price for their data, $\theta_i \in \left[\underline{\theta}_i, \bar{\theta}_i\right]$, with $0 \leq \underline{\theta} < \bar{\theta} < \infty, \quad \forall i$. The reserve price vector is defined as $\theta := (\theta_1, \ldots, \theta_N)$ on a joint probability space $\Theta := \left[\underline{\theta}_i, \bar{\theta}_i\right] \times \ldots \left[\underline{\theta}_N, \bar{\theta}_N\right]$. We assume that the seller's valuation is drawn from a distribution with a PDF, $f_i(\cdot)$, and corresponding CDF, $F_i(\cdot)$. In addition, we assume the distributions of θ_i are independent but not necessarily identically distributed[4]. The seller, i, with reserve price, θ_i, receives a payment, t_i, with probability, q_i. Their resulting utility

[4]This corresponds to the distribution of WTA estimations we conduct in Chapter 3.

is therefore:

$$u_i = t_i - \theta_i q_i \tag{6.1}$$

The true/target data distribution, X_T, is an aggregation[5] of the N sellers' data $\mathcal{X}$[6]. The value of each seller's data is differentiated by the WD, $W_i(X_i, X_T)$, a non-negative metric, capturing the similarity between X_i and X_T. Assuming fixed privacy preferences, each seller has a given privacy requirement, ϵ_i, which must be fulfilled if their data is procured. We can incorporate the effect of this privacy preference, using the upper bounds developed in Section 5.3.2, directly into their individual WD, W_i:

$$W_i = W(X_i, X_T) + W(X_{DP}, \delta_0) \tag{6.2}$$

where, X_{DP} is the additive noise mechanism used to achieve DP. $W(X_{DP}, \delta_0)$ is defined in Proposition 5.7 for the Laplace Mechanism and in Proposition 5.5 for the Gaussian Mechanism.

Traditionally, W_i indicates the i-th order WD. However, as we focus solely on the 1-Wasserstein distance, for notational convenience, we abuse the standard notation slightly and omit the dependence on the target distribution X_T, by defining it as $W_i \coloneqq W_1(X_i, X_T)$. We also denote the vector of all individual WDs as $W \coloneqq [W_1, \ldots, W_N]$. We assume the platform has access to these distances. As detailed in Section 5.3, this can be achieved while preserving data privacy through private calculation.

Data Buyer

We assume there is a single data buyer procuring data from the sellers, in order to obtain the target distribution, X_T. In the exogenous budget mechanism, the buyer has a budget, B. In the endogenous and joint case, the buyer has some reference data X_R (e.g. public dataset) available to them and a corresponding benchmark performance value of their model/task, $B_{\mathcal{M}}(X_R)$, using this reference data. We assume the performance metric, $L_{\mathcal{M}}(\cdot)$, is $K_{\mathcal{M}}$-Lipschitz. To preserve privacy, the buyer cannot access the seller's data X_i.

In addition, to maintain computational tractability, the WD of any subset of procured data is be approximated using the Hoeffding bound in Theorem 5.2, which only requires the computation of individual WDs, W_i. This probabilistic bound allows the buyer to set their risk preferences by choosing a confidence level, δ, for the approximation. The

[5]For example, the Euclidean aggregate $X_T = \frac{1}{N}\sum_{i\in\mathcal{N}} X_i$ or it could be the Wasserstein barycenter $X_T = \arg\min_{X_T} \frac{1}{N}\sum_{i\in\mathcal{N}} W_1(X_i, X_T)$ as considered in the forecast trading mechanism in [299].

[6]We also consider the case where N is infinite.

confidence level determines the probability that the Hoeffding bound is greater than the actual WD. As the WDs are calculated privately, using MPC, and the raw data $\mathcal{X}$ is not shared with the platform prior to procurement, the platform could be the buyer and need not be a separate trusted entity.

Market Platform

The platform receives the individual WD, W_i, reserve prices/privacy sensitivities, θ_i, and privacy budgets, ϵ_i, from each data seller. We assume here that there is no known statistical relationship, a priori, between W and θ, which the platform could exploit[7]. In the exogenous budget case the buyer provides their budget B. In the endogenous or joint case the buyer provides their benchmark performance, $B_{\mathcal{M}}(X_R)$, a task specific Lipschitz constant, $K_{\mathcal{M}}$, and their risk preference, δ. The information flow for theses scenarios is depicted in Figure 6.2.

Table 6.2: Summary of Proposed Mechanisms

	Objective	Budget	Inputs	Use Case
Exogenous Budget	$V(W)$	B	θ, W, B	Task agnostic procurement
Endogenous Budget	$V(W, K_{\mathcal{M}}, \delta)$	$B_{\mathcal{M}}(X_R)$ $-V(\cdot)$	$\theta, W, K_{\mathcal{M}}$, $B_{\mathcal{M}}(X_R), \delta$	Task specific welfare maximisation
Joint Optimisation	$V(W, K_{\mathcal{M}}, \delta) + \sum t_i$	$B_{\mathcal{M}}(X_R)$ $-V(\cdot)$	$\theta, W, K_{\mathcal{M}}$, $B_{\mathcal{M}}(X_R), \delta$	Task specific profit maximisation

The platform runs a data acquisition mechanism to select and remunerate data sellers while optimising the buyers' aim. We summarise the objectives, budget constraints, inputs and use cases for each of the proposed mechanisms in Table 6.2. For the exogenous mechanism, the buyer aims to minimise V subject to the external budget, B. This models the task agnostic procurement of data, where V simply represents how close the procured data is to the target data distribution. In terms of energy domain applications, this could represent an entity aiming to obtain a representative sample of smart meter data which

[7]We do however, investigate the effect of correlations between W and θ on performance, in the case studies at the end of this chapter

will then be used to a variety of tasks, for example a research institution with a fixed grant. The objective remains the same in the endogenous budget case, however, the budget is now dependent on V. This represents a scenario where the buyer aims to maximise welfare (minimise V) while ensuring welfare gains, through data procurement, are commensurate with the associated procurement costs as well as ensuring costs do not exceed a reference budget, $B_{\mathcal{M}}(X_R)$. A potential application for this mechanism would be a data procurement mechanism within an energy collective model, where the buyer would be the community manager, a central coordinating, non-profit entity[300]. In the joint optimisation case, the buyer aims to minimise V and the associated data costs subject to the same budget as the endogenous case. This models a buyer aiming to maximise profits and therefore has applications in our motivating example, the retailer energy procurement problem.

As we take a Bayesian approach to the mechanism design problem, we also assume that the platform has access to the distribution of seller valuations, $f_i(\theta_i)$. In addition, as the revelation principle applies, we focus on direct revelation mechanism where sellers reporting their reserve prices truthfully is a Bayesian Nash Equilibrium[301]. We restrict ourselves to deterministic mechanisms, i.e. once reserve prices are reported the mechanism determines, with certainty, which data has been procured[8]. We introduce ex-post Incentive Compatibility (IC) constraints to ensure this equilibrium, by requiring that each seller has no incentive to misrepresent their reserve prices when others report truthfully. In addition, we include ex-post Individual Rationality (IR) constraints to ensure participation does not leave any seller worse off. Lastly, we include ex-interim Budget Feasibility (BF) constraints to ensure that payments made to data sellers do not exceed a budget[9]. In summary, our mechanism aims to ensure that data cannot be purchased from a seller unless they are compensated with at least their reserve price, the total payments made to sellers do not exceed the buyer's budget (in expectation over sellers' valuations), and truth telling is a dominant strategy.

[8]This is motivated by the fact that data sellers are interested in their ex-post rather than their expected payments. As such, even for a stochastic mechanism we would need to ensure payments are sufficient for participation for all potential outcomes, otherwise the platform would need to re-adjust payments ex-post[301].

[9]We argue that an ex-interim budget constraint is reasonable in our setting, as the buyer will be procuring data repeatedly. In addition, where budgets are derived based on task specific performance (the endogenous budget and joint optimisation mechanisms), we envision the task itself to be stochastic in nature, with data procurement reducing uncertainty.

6.2.2 Data Procurement Mechanism

The platform's task is to determine which sellers' data to buy and how much to pay them to maximise the benefit. Formally, we can define the following mechanism:

Definition 6.2.1. *(Data Procurement Mechanism) In the fixed privacy scenario we define a direct mechanism as a tuple (q, t, V) where:*

- *For all $i \in \mathcal{N}$, $q : \Theta \to (0,1)^N$ is a function which maps reserve prices, θ, to a selection probability, q_i* [10]
- *For all $i \in \mathcal{N}$, $t : \Theta \to \mathbb{R}_+^N$ is a function which maps reserve prices, θ, to payments, t_i*
- *$V(\cdot) : W \to \mathbb{R}^+$ is a function which maps the sellers individual WDs, W, to the expected value of the WD for a subset/combination of data, $V(W, q(\theta), K_{\mathcal{M}}, \delta)$*

Having defined the parameters of our mechanism, we now develop the platform's task as an optimisation problem. Let $q_i(\theta)$ and $t_i(\theta)$ denote the ith components of q and t, respectively and let the subscript $-i$ denote the vector excluding the ith component. Although the platform's problem is similar for the three mechanisms we propose, we start by highlighting the differences in formulation.

Exogenous Budget In the exogenous budget mechanism, the platform's problem can be formulated as:

$$\min_{q,t} \quad \int_{\Theta} V(W, q(\theta)) f(\theta) d\theta \tag{6.3}$$

$$\text{s.t.} \quad U_i(\theta_i|\theta_i) \geq 0, \quad \forall i \in \mathcal{N}, \forall \theta_i \tag{6.3a}$$

$$U_i(\theta_i|\theta_i) \geq U_i(\theta_i'|\theta_i), \quad \forall i \in \mathcal{N}, \forall \theta_i, \theta_i' \tag{6.3b}$$

$$\int_{\Theta} \sum_{i \in \mathcal{N}} t_i(\theta) f(\theta) d\theta \leq B \tag{6.3c}$$

where,

$$U_i(\tilde{\theta}_i|\theta_i) \coloneqq \int_{\Theta_{-i}} \left(t_i(\tilde{\theta}_i, \theta_{-i}) - \theta_i q_i(\tilde{\theta}_i, \theta_{-i}) \right) f_{-i}(\theta_{-i}) d\theta_{-i} \tag{6.4}$$

which denotes the expected utility of a seller with reserve price, θ_i, if they report a reserve price, $\tilde{\theta}_i$, and all other sellers report truthfully. The aim of the platform is to minimise V, the expected WD of the procured subset of data, by determining the optimal selection probabilities, q, and payments, t. V depends on the selection, q, as such the platform aims

[10] As we restrict our focus to deterministic mechanisms, q is in fact a vector of binary selection decisions.

to minimise the expected V, over the joint seller valuation space, Θ. The first constraint (6.3a) represents individual rationality, essentially we ensure that when seller i reports their true reserve price, θ_i, their utility must be non-negative. The next constraint,(6.3b), encodes incentive compatibility. Here we ensure that for a seller with a true reserve price, θ_i, their utility when reporting some other reserve price θ_i' is lower or equal to the utility achieved when reporting truthfully. Finally, (6.3c) describes the budget feasibility constraint. The sum of expected payments t_i (over seller valuations, Θ), must be less than the exogenous budget parameter B.

Endogenous Budget In a similar vein, the platform's problem for the endogenous budget mechanism is:

$$\min_{q,t} \quad \int_{\Theta} V(W, q(\theta), K_{\mathcal{M}}, \delta) f(\theta) d\theta \tag{6.5}$$

$$\text{s.t.} \quad U_i(\theta_i|\theta_i) \geq 0, \quad \forall i \in \mathcal{N}, \forall \theta_i \tag{6.5a}$$

$$U_i(\theta_i|\theta_i) \geq U_i(\theta_i'|\theta_i), \quad \forall i \in \mathcal{N}, \forall \theta_i, \theta_i' \tag{6.5b}$$

$$\int_{\Theta} \sum_{i \in \mathcal{N}} t_i(\theta) f(\theta) d\theta \leq B_{\mathcal{M}}(X_R) - \int_{\Theta} V(W, q(\theta), K_{\mathcal{M}}, \delta) f(\theta) d\theta \tag{6.5c}$$

The only difference, compared to the exogenous budget mechanism is the modification of the budget constraint. The dependence of V on $K_{\mathcal{M}}$ does not affect the optimisation as it is a scaling factor. The dependence on δ will be discussed in the following section.

Joint Optimisation Finally, for the joint optimisation mechanism the platform's problem can be formulated as:

$$\min_{q,t} \quad \int_{\Theta} \left[V(W, q(\theta), K_{\mathcal{M}}, \delta) + \sum_{i \in \mathcal{N}} t_i(\theta) \right] f(\theta) d\theta \tag{6.6}$$

$$\text{s.t.} \quad U_i(\theta_i|\theta_i) \geq 0, \quad \forall i \in \mathcal{N}, \forall \theta_i \tag{6.6a}$$

$$U_i(\theta_i|\theta_i) \geq U_i(\theta_i'|\theta_i), \quad \forall i \in \mathcal{N}, \forall \theta_i, \theta_i' \tag{6.6b}$$

$$\int_{\Theta} \sum_{i \in \mathcal{N}} t_i(\theta) f(\theta) d\theta \leq B_{\mathcal{M}}(X_R) - \int_{\Theta} V(W, q(\theta), K_{\mathcal{M}}, \delta) f(\theta) d\theta \tag{6.6c}$$

In this case the constraints are identical to the endogenous case however, we introduce the expected payments into the platform's objective. We see that in all cases the incentive compatibility constraint result in a infinite dimensional problem, as we need to ensure it holds for any reserve price realisations within the joint support, Θ. In order to obtain a tractable, solvable optimisation problem we need to reformulate the above representations. In the rest of this section we focus on the joint optimisation mechanism, however,

the results and reformulations are also applicable for the exogenous and endogenous budget mechanism.

6.2.3 Problem Reformulation

Objective Reformulation

We start by defining the following quantities:

$$T_i(\theta_i) \coloneqq \int_{\Theta_{-i}} t_i(\theta_i, \theta_{-i}) f_{-i}(\theta_{-i}) d\theta_{-i} \tag{6.7}$$

$$Q_i(\theta_i) \coloneqq \int_{\Theta_{-i}} q_i(\theta_i, \theta_{-i}) f_{-i}(\theta_{-i}) d\theta_{-i} \tag{6.8}$$

to be the expected payment and the expected probability of selling for seller i, respectively. Using the above quantities and the definition of the sellers' utility the platform's problem can be defined as:

$$\min_{q,t} \quad \int_{\Theta} \left[V(W, q(\theta), K_{\mathcal{M}}, \delta) + \sum_{i \in \mathcal{N}} t_i(\theta) \right] f(\theta) d\theta \tag{6.9}$$

$$\text{s.t.} \quad T_i(\theta_i) - \theta_i Q_i(\theta_i) \geq 0, \quad \forall i \in \mathcal{N}, \forall \theta_i \tag{6.9a}$$

$$T_i(\theta_i) - \theta_i Q_i(\theta_i) \geq T_i(\widetilde{\theta}_i) - \theta_i Q_i(\widetilde{\theta}_i), \quad \forall i \in \mathcal{N}, \forall \theta_i, \widetilde{\theta}_i \tag{6.9b}$$

$$\int_{\Theta} \sum_{i \in \mathcal{N}} t_i(\theta) f(\theta) d\theta \leq B_{\mathcal{M}}(X_R) - \int_{\Theta} V(W, q(\theta), K_{\mathcal{M}}, \delta) f(\theta) d\theta \tag{6.9c}$$

The IR, (6.9a), and IC, (6.9b), constraints are identical to those of a buyer in the standard single-item auction problem ([302, ¶ 4]). Assuming the value of each data source is fixed (i.e. $W_i = W_i(\theta_i)\ \forall \theta_i$), we can apply Myerson's Lemma directly. This allows us to characterise the payments, t, in terms of the selection probabilities, q. As a result, the platform's problem can be formulated as follows[11]:

$$\min_{q} \quad \mathbb{E}_{\Theta} \left[V(W, q(\theta), K_{\mathcal{M}}, \delta) + \sum_{i \in \mathcal{N}} q_i(\theta_i) \psi_i(\theta_i) \right] \tag{6.10}$$

$$\text{s.t.} \quad Q_i(\theta_i) \geq Q_i(\widetilde{\theta}_i), \quad \forall i \in \mathcal{N}, \theta_i, \widetilde{\theta}_i, \theta_i < \widetilde{\theta}_i \tag{6.10a}$$

$$T_i(\theta_i) = \psi_i(\theta_i), \quad \forall i \in \mathcal{N} \tag{6.10b}$$

$$\mathbb{E}_{\Theta} \left[V(W, q(\theta), K_{\mathcal{M}}, \delta) + \sum_{i \in \mathcal{N}} q_i(\theta_i) \psi_i(\theta_i) \right] \leq B(X_R) \tag{6.10c}$$

where, $\psi_i(\theta_i) = \theta_i + \frac{F_i(\theta_i)}{f_i(\theta_i)}$ is the virtual cost of seller i. The first constraint, (6.10a), is the monotonicity requirement for the selection rule, which ensures that the selection probability is greater when reporting a reserve price, θ_i, than the selection probability

[11] A complete proof can be found in Appendix D.1.

when reporting a reserve price $\bar{\theta}_i$, if $\theta_i \leq \bar{\theta}_i$. Constraint (6.10b) is the payment rule, and (6.10c) is the budget feasibility constraint re-written in terms of virtual costs.

Next, we characterise the function V. We aim to select a subset of data X_C which minimises performance loss compared to the target data X_T, while ensuring that we do not exceed the budget $B_{\mathcal{M}}(X_R)$. First, we obtain an upper bound on the performance by applying the Lipschitz and Hoeffding bounds developed in the previous chapter.

$$V(W, q(\theta), K_{\mathcal{M}}, \delta) = L\left(\frac{1}{|C|}\sum_{i\in C} q_i(\theta_i)X_i\right) - L\left(X_T\right) \tag{6.11}$$

$$\overset{(a)}{\leq} K_{\mathcal{M}} W\left(X_C, T\right) \tag{6.12}$$

$$\overset{(b)}{\leq} C(K_{\mathcal{M}}, \delta, N)\frac{\sqrt{f(N,q)\sum_{i\in C} q_i(\theta_i)W_i^2}}{\sum_{i\in C} q_i(\theta_i)} \tag{6.13}$$

where, (a) results from Theorem 5.1, and (b) results from Theorem 5.2. $C(K_{\mathcal{M}}, \delta, N)$ is a constant, and $f(N, q)$ is a function dependent on whether we assume an finite or infinite population for the Hoeffding bound.

To avoid infeasibility due to (6.10c), we drop the budget feasibility constraint, (6.10c), and reformulate the objective as:

$$\min_q \min\left(B_{\mathcal{M}}(X_R), \mathbb{E}_\Theta\left[C(K_{\mathcal{M}}, \delta, N)\frac{\sqrt{f(N,q)\sum_{i\in\mathcal{N}} q_i(\theta_i)W_i^2}}{\sum_{i\in\mathcal{N}} q_i(\theta_i)} + \sum_{i\in\mathcal{N}} q_i(\theta_i)\psi_i(\theta_i)\right]\right) \tag{6.14}$$

Finally, to model the inner minimum within the objective function in (6.10) we introduce an additional selection probability q_0, which represents the probability of not buying any data from the sellers and instead relying solely on the reference data, X_R. If $q_0 = 0$ then $\sum_{i\in\mathcal{N}} q_i \geq 1$, which indicates the platform has chosen to procure at least one dataset. Conversely, if $q_0 = 1$ then $\sum_{i\in\mathcal{N}} q_i = 0$, which indicates that the platform chooses not to buy any additional data. We assume here that the reference data X_R is available at zero cost, although reference data costs could easily be included with an addition term, $q_0 t_0$. The resulting platform problem is:

$$\min_q \quad \mathbb{E}_\Theta\left[q_0 B_{\mathcal{M}}(X_R) + C(K_{\mathcal{M}}, \delta, N)\frac{\sqrt{f(N,q)\sum_{i\in\mathcal{N}} q_i(\theta_i)W_i^2}}{\sum_{i\in\mathcal{N}} q_i(\theta_i)} + \sum_{i\in\mathcal{N}} q_i(\theta)\psi_i(\theta_i)\right] \tag{6.15}$$

$$\text{s.t.} \quad Q_i(\theta_i) \geq Q_i(\widetilde{\theta}_i), \quad \forall i \in \mathcal{N}, \theta_i, \widetilde{\theta}_i, \theta_i < \widetilde{\theta}_i \tag{6.15a}$$

$$T_i(\theta_i) = \psi_i(\theta_i), \quad \forall i \in \mathcal{N} \tag{6.15b}$$

where, q_0 is the probability of selecting the reference data and not procuring any additional data.

We have now reformulated the platform's problem solely in terms of reserve prices and selection probabilities. However, (6.15) is still infinite dimensional due to the monotonicity constraints.

Characterising the Optimal Mechanism

We now aim to obtain a point-wise optimisation problem which can be solved using conventional techniques (e.g. convex optimisation). Specifically, we follow the approach of [52], where the monotonicity requirement is initially dropped and then shown to hold. Using this, we provide a solution method for both the infinite and finite case. We reformulate the problem in (6.15) into a MISOCP, which can be solved using commercial solvers.

Infinite Population Under the infinite population assumption:

$$C^{inf} = K_{\mathcal{M}} \sqrt{\frac{\ln\left(\frac{2}{1-\delta}\right)}{2}}, \quad f(N,q) = 1 \tag{6.16}$$

Ignoring the monotonicity constraint in (6.15) allows us to cast a point-wise optimisation problem for a given cost vector θ. As $B(X_R)$ is a constant, the problem can equivalently stated as the minimisation of the expected WD of the procured data and the virtual costs:

$$\min_{q} \quad q_0 B_{\mathcal{M}}(X_R) + C^{inf} \frac{\|q_i W_i)\|}{\sum_{i\in\mathcal{N}} q_i} + \sum_{i\in\mathcal{N}} q_i \psi_i \tag{6.17}$$

$$\text{s.t.} \quad \sum_{i=0}^{N} q_i \geq 1 \tag{6.17a}$$

$$q \in \{0,1\}^{N+1} \tag{6.17b}$$

where, q_0 is the selection probability for the reference data. The norm representation arises from the fact that q_i is binary. As a result, $q_i = q_i^2$ implying $\sqrt{\sum_{i\in\mathcal{N}} q_i W_i^2} = \sqrt{\sum_{i\in\mathcal{N}} q_i^2 W_i^2}$ which is $\|q_i W_i\|$.

The above problem can be reformulated as a MISOCP:

$$\min_{q,s,z} \quad q_0 B_{\mathcal{M}}(X_R) + C^{inf} s + \sum_{i\in\mathcal{N}}^{N} q_i \psi_i \tag{6.18}$$

$$\text{s.t.} \quad \|q_i W_i\| \leq \sum_{i\in\mathcal{N}}^{N} z_i \quad \Longleftrightarrow \quad \left(\sum_{i\in\mathcal{N}}^{N} z_i, q^T W_1(X)\right) \in \mathcal{Q}^{N+1} \tag{6.18a}$$

$$0 \leq z_i \leq M q_i, \quad \forall i \in \mathcal{N} \tag{6.18b}$$

$$0 \leq s - z_i \leq M(1 - q_i), \quad \forall i \in \mathcal{N} \tag{6.18c}$$

$$1 \leq \sum_{i=0}^{N} q_i \leq N \tag{6.18d}$$

$$q \in \{0,1\}^{N+1}, s \in \mathbb{R}_+, z \in \mathbb{R}_+^N \tag{6.18e}$$

where, $M > \|W\|$.

An auxiliary variable, s, is introduced to represent the fractional objective $\|q_i W_i\| \leq s \sum_{i\in\mathcal{N}} q_i$. We assume here that either $q_0 = 1, \sum_{i\in\mathcal{N}} q_i = 0$ or $q_0 = 0, \sum_{i\in\mathcal{N}} q_i \geq 1$. To ensure this and avoid the trivial solution, we introduce an additional constraint $\left(\sum_{i=0}^{N} q_i \geq 1\right)$. The bi-linear term $s \sum_{i\in\mathcal{N}} q_i$ can be linearised exactly using Glover Linearisation as $\sum_{i\in\mathcal{N}} q_i$ is the sum of binary variables[303]. To achieve this N auxiliary continuous variables z are introduced. A similar approach can be used to obtain the MISOCP formulations for the exogenous and endogenous budget mechanisms. These can be found in Appendix D.3.

We dropped the monotonicity requirement to obtain a point-wise optimisation problem, however, to ensure that we maintain IC and IR we must show that the platform's optimisation problem is indeed monotonic. By making mild assumptions (Assumption 6.1) on the virtual costs ψ this can be ensured. A full proof can be found in Appendix D.2.

Assumption 6.1. *For all $i \in N$, the virtual cost $\psi_i(\theta) = \theta + \frac{F_i(\theta)}{f_i(\theta)}$ is increasing in θ.*

As discussed in [52], Assumption 6.1 is standard in mechanism design, in particular for procurement auctions such as ours. Distributions which have this property are also called regular distributions, and include Gaussians, uniform and exponential distributions.

Lastly, we characterise the payments to each data seller i as:

$$t_i(\theta_i) = q_i^* \psi_i(\theta_i) = q_i^* \left(\theta + \frac{F_i(\theta)}{f_i(\theta)}\right) \tag{6.19}$$

where q_i^* is the solution to (6.17).

Finite Population If instead, we include the finite population correction:

$$C^{fin} = K_{\mathcal{M}} \sqrt{\frac{\ln\left(\frac{2}{1-\delta}\right)}{2(N-1)}}, \quad f(N,q) = N - \sum_{i\in\mathcal{N}} q_i \tag{6.20}$$

The point-wise optimisation problem, ignoring monotonicity, under the finite population assumption becomes:

$$\min_{q} \quad q_0 B_{\mathcal{M}}(X_R) + C^{fin}\sqrt{\frac{(N - \sum_{i\in\mathcal{N}} q_i)\sum_{i=1}^{N} q_i W_i^2}{(\sum_{i\in\mathcal{N}} q_i)^2}} + \sum_{i\in\mathcal{N}} q_i \psi_i(\theta_i) \tag{6.21}$$

$$\text{s.t.} \quad \sum_{i=0}^{N} q_i \geq 1 \tag{6.21a}$$

$$q \in \{0,1\}^{N+1} \tag{6.21b}$$

In order to convexify the objective of the original problem, we introduce a number of auxiliary variables and make substitutions:

1. First, note that $N - \sum_{i=1}^{N} q_i = \sum_{i=1}^{N}(1 - q_i)$, resulting in the numerator within the square root being $\sum_{i=1}^{N} W_i^2 q_i \left(\sum_{j=1}^{N}(1 - q_j)\right)$.
2. The binary products $q_i \sum_{j=1}^{N}(1 - q_j)$ are linearised by introducing auxiliary binary variables $r_{i,j}$ and constraints (6.22b)-(6.22d). The resulting objective term is $\sqrt{\sum_{i=1}^{N}\sum_{j\neq i} W_i^2 r_{i,j}}$. Note, that as $q_i(1 - q_i) = 0$ we only require $N^2 - N$ auxiliary binary variables.
3. Lastly, note that $\sqrt{\sum_{i=1}^{N}\sum_{j\neq i} W_i^2 r_{i,j}}$ is equivalent to a matrix norm (as r is binary) which, similar to (6.18), can be reformulated as a second order conic constraint. This is achieved by introducing a continuous variable s to linearise the objective and z to linearise the binary-continuous products.

The resulting reformulation as a MISOCP is given by:

$$\min_{q,r,s,z} \quad q_0 B_{\mathcal{M}}(X_R) + C^{fin} s + \sum_{i\in\mathcal{N}} q_i \psi_i(\theta_i) \tag{6.22}$$

$$\text{s.t.} \quad \|Wr\| \leq \sum_{i\in\mathcal{N}} z_i \quad \Longleftrightarrow \quad \left(\sum_{i\in\mathcal{N}} z_i, r^T W\right) \in \mathcal{Q}^{N+1} \tag{6.22a}$$

$$r_{i,j} \leq q_i, \quad \forall i \in \mathcal{N} \tag{6.22b}$$

$$r_{i,j} \leq 1 - q_j, \quad \forall j \in \mathcal{N} \tag{6.22c}$$

$$r_{i,j} \geq q_i - q_j, \quad \forall i \in \mathcal{N}, j \in \mathcal{N}/i \tag{6.22d}$$

$$0 \leq z_i \leq M q_i, \quad \forall i \in \mathcal{N} \tag{6.22e}$$

$$0 \leq s - z_i \leq M(1 - q_i), \quad \forall i \in \mathcal{N} \tag{6.22f}$$

$$1 \leq \sum_{i=0}^{N} q_i \leq N \tag{6.22g}$$

$$q \in \{0,1\}^{N+1}, r \in \{0,1\}^{N^2-N}, s \in \mathbb{R}_+, z \in \mathbb{R}_+^N \tag{6.22h}$$

Finally, given the equivalence of the reformulation with the original problem. We ensure optimality and the desired properties of incentive compatibility, individual rationality and budget feasibility in the same manner as the infinite case. Specifically, showing that the selection is monotonic an that payments are defined by the payment identity, using the optimal selection, q^*, obtained by solving (6.22). The MISOCP formulations for the finite exogenous and endogenous budget mechanisms can be found in Appendix D.3.

Determining $B_{\mathcal{M}}(X_R)$ and Budget Feasibility

All the proposed mechanisms aim to minimise the WD between the procured subset of data, $X_C = \frac{1}{|C|}\sum_{i\in C} q_i X_i$, and the target distribution, X_T, subject to a budget constraints. The buyer is required to provide an external budget, B or $B_{\mathcal{M}}(X_R)$ and the mechanism ensures budget feasibility with respect to this external budget. In the joint optimisation and endogenous budget mechanisms, we aim to ensure that the expected performance loss, in monetary terms, by using X_C instead of X_T and the associated payments to procure X_C is less than the performance loss achieved with the (free) reference data, X_R. As such, we can define the external budget as $B = L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_C)$. However, as we do not have access to $L_{\mathcal{M}}(X_C)$, we develop a lower bound on the budget:

$$B = L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_C) \tag{6.23}$$

$$= [L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)] - [L_{\mathcal{M}}(X_C) - L_{\mathcal{M}}(X_T)] \tag{6.24}$$

$$\overset{(a)}{\geq} [L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)] - K_{\mathcal{M}} W(X_C, T) \tag{6.25}$$

$$\overset{(b)}{\geq} [L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)] - C(K_{\mathcal{M}}, \delta, N)\frac{\sqrt{f(N,q)\sum_{i\in C} q_i(\theta_i)W_i^2}}{\sum_{i\in C} q_i(\theta_i)} \tag{6.26}$$

where, (a) results from Theorem 5.1, and (b) results from Theorem 5.2.

Ideally, $B_{\mathcal{M}}(X_R) = L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$, however, we do not have access to the $L_{\mathcal{M}}(X_T)$, as this would also violate the data privacy of the sellers. The buyer is therefore forced to estimate $B_{\mathcal{M}}(X_R)$, using for example, historical performance, or theoretical problem-specific bounds. We explore the implications of under or over-estimation below:

- **Lower Bound**, if $B_{\mathcal{M}}(X_R) < L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$, the mechanism retains budget feasibility. As the lower bound results in an under-estimation of the budget, the buyer ends up with less data than they could have procured.
- **Upper Bound**, if $B_{\mathcal{M}}(X_R) > L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$, the mechanism can no longer provide budget feasibility guarantees. The over-estimation will lead to the buyer purchasing

more data than they should. If the resulting performance and payments are higher than $L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$, the buyer will be worse off than if they simply used the reference data. However, as the WD provides an upper bound on the performance loss over-estimation does not necessarily lead to budget infeasibility. A natural choice for an upper bound would be the Lipschitz bound, $K_{\mathcal{M}}W(X_R, X_T)$.

In both cases, the cost of estimation error is borne by the buyer, thus incentivising the buyer to produce accurate estimates of $B_{\mathcal{M}}(X_R)$. Data sellers, on the other hand, are ensured a payment above their reserve prices, thereby maintaining individual rationality. This ensures a seller/user-centric approach. If we wish to maintain budget feasibility, we could develop a privacy-preserving protocol to calculate $L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$. The accuracy would be dependent on the technique and the particular performance metric. We note, of course, that such a technique could then be used to create a privacy-preserving cooperative game framework. However, we argue that our approach still provides benefits, in terms of computational costs, in this scenario. A cooperative game still requires the calculation of each coalition value whereas we would only require the calculation of one term $L_{\mathcal{M}}(X_T)$. The computational advantages are particularly pronounced when the underlying model is computationally intensive.

6.3 Case Study: Procuring Gaussian Aggregates

The three proposed mechanisms serve different purposes and are therefore not directly comparable. As such, we set up two different case studies, the first to assess our exogenous budget mechanism against existing exogenous budget mechanism and the second to assess the endogenous budget mechanisms (including the joint optimisation mechanism) against a centralised benchmark. Given the considerable difference in binary variables required in the MISOCP reformulations of the finite and infinite formulations we include both in the comparisons. We use the same synthetic Gaussian data as in Section 5.2.3. The simulations were implemented in Python using CVXPY. They were solved using Gurobi 9.5.0 on a DELL XPS 15 with an 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz 2.30 GHz processor and 64GB RAM.

6.3.1 Exogenous Budget

To evaluate the performance of the proposed exogenous budget mechanism, using the finite (FIN) and infinite (INF) formulations, we compare them against the following existing approaches and benchmarks:

- Central (CEN) - Here we assume full access to the value of each coalition. The mechanism selects the coalition with maximum value (minimum WD) that is budget feasible. It provides a benchmark of best possible performance.

- Random ($RAND$) - Here we assume a coalition is selected at random from all budget feasible coalitions. The value is therefore the average value of budget feasible coalitions. This provides a worst case benchmark in the absence of an optimal selection criteria.

- Single Minded Query (SMQ)[51] - This is a Bayesian mechanism, similar to our exogenous budget mechanism, which aims to maximise (reserve price independent) value, V, subject to ex-interim budget feasibility. It assumes sellers' values are additive ($V = \sum_i V_i$), resulting in a separable problem where each data seller receives a take-it-or-leave-it offer. The offers are determined by solving an auxiliary, convex, optimisation problem, which, for uniformly distributed reserve prices, is a SOCP. If the sellers' reserve price is lower then the offer, θ^*, the sellers' data is purchased. We run this mechanism assuming the value of each data seller is the reciprocal of its distance from the target distribution, $V_i = \frac{1}{d_i}$.

- Greedy Knapsack ($PTAS$) [43] – This provides a polynomial time approximation scheme to the same problem as above, in a prior-free environment. Again, sellers' values are assumed to be additive and we use $V_i = \frac{1}{d_i}$. First sellers are sorted, in ascending order, by their cost per unit value ($g_i = \theta_i d_i$). The mechanism then finds the largest index k which satisfies $g_k \leq \frac{B}{\sum_{i \leq k} \frac{1}{d_i}}$. All sellers $i \leq k$, are selected and paid $p_i = \min \left\{ \frac{B}{\sum_{i \leq k} \frac{1}{d_i}}, g_{k+1} \right\} \frac{1}{d_i}$.

We investigate the performance of the above mechanisms for the different statistical distances discussed in Section 5.2.1, different budget levels, and correlations, $\rho(\theta, d)$, between reserve prices and the value metric (distance)[12]. The reserve prices are assumed to follow a uniform distribution ($U(0,1)$) and the budget levels are multiples of the maximum reserve price. The full list of sensitivities can be found in Table 6.3.

Next, we focus on our main contribution for exogenous budget mechanisms. Namely, we investigate the effect of incorporating both the heterogeneity or non-I.I.D. nature

[12]Extant literature suggests that consumer valuations of personal data are not necessarily linked to its' actual value but other considerations, such as privacy. We therefore consider the full range of potential correlations. We also note that the budget feasible mechanisms considered here, can be seen as variants of a knapsack problem. The computational complexity of such knapsack problems varies significantly, depending on correlations between weights (reserve prices in our case) and value[304].

Table 6.3: Experimental Parameters for Exogenous Budget Mechanisms

Parameter	Value/Range
N	8
Trials	50
Data Distribution	Gaussian
Loc, Scale	$\alpha_i \sim U(10, 20)$, $\beta_i \sim U(1, 5)$
Distances	WD, KLD, KS, TVD, JSD
Mechanisms	CEN, RAND, SMQ, PTAS, INF, FIN
Reserve Prices	$\theta \sim U(0, \bar{\theta})$, $\bar{\theta} = 1$
Correlation	$\rho \in \{-1, 0, 1\}$
Budget	$B \in \{0.1\bar{\theta}N, 0.2\bar{\theta}N, \ldots, \bar{\theta}N\}$
Privacy Budgets	$\epsilon \sim U(0, \bar{\epsilon})$, $\bar{\epsilon} \in \{0.1, \ldots, 100\}$, $\delta^{dp} = 10^{-15}$

of the data and DP. We compare the performance of the finite formulation for four different scenarios for the WD; (1) DP only $V_i = 1/\epsilon_i$, (2) Non-I.I.D. only $V_i = W(X_i, X_T)$, (3) Exact DP (only for Gaussians) $V_i = W(X_i + X_{DP}, X_T)$ and (4) Upper bound on DP $V_i = W(X_i, X_T) + W(X_{DP}, \delta_0)$. Here we also consider the effect correlations, however, in this case we simulate correlations, $\rho(\theta, \epsilon)$, between reserve prices, θ, and the privacy budget, ϵ, rather than the WDs. We assume reserve prices and privacy budgets are both distributed uniformly, and DP is achieved using the Gaussian mechanism. To show the significance of our unified metric we sweep the upper bound of the uniform distribution of ϵ. We note, however, that the Gaussian mechanism only provides meaningful privacy guarantees when $\epsilon \in (0, 1)$, and the probability of failure, $\delta^{dp} \gg 1/N$[229].

6.3.2 Endogenous Budget Mechanisms

For the endogenous budget and joint optimisation mechanisms there are no directly comparable mechanisms to evaluate performance against. Instead, we compare against benchmark values. We assume that the buyer is looking to buy data for a specific task and aims to minimise the relevant performance metric/loss function, $L_{\mathcal{M}}$. For example, for median estimation the buyer is aiming to minimise the MAE. We then compare our proposed mechanisms against:

- Central Actual ($CEN_{\mathcal{M}}$): The buyer has access to the value of the performance metric

for each coalition of data sellers and selects the optimal[13] budget feasible coalition.

- Central Distance (CEN_W): The buyer has access to the WD for each coalition of data sellers and selects the optimal[14] budget feasible coalition.

We run the mechanisms for a range of loss functions and reserve price-distance correlations $\rho(\theta, W)$. We assume the budget provided by the buyer is fixed at $B(X_R) = L_{\mathcal{M}}(X_R) - L_{\mathcal{M}}(X_T)$. Instead, we vary the upper bound on the distribution of reserve prices, $\bar{\theta}$. Finally, we investigate the effect of risk by adjusting the confidence level, δ, of the Hoeffding bound. The full set of parameters are summarised in Table 6.4.

Table 6.4: Experimental Parameters for Endogenous Budget Mechanisms

Parameter	Value/Range
N	8
Trials	50
Data Distribution	Gaussian
Loc, Scale	$\alpha_i \sim U(10, 20)$, $\beta_i \sim U(1, 5)$
Loss Functions	RMSE, MAE, $\text{MPL}_{\text{q=0.9,0.8}}$
Objectives	Endogenous Budget, Joint Optimisation
Mechanisms	$CEN_{\mathcal{M}}, CEN_W, INF_W, FIN_W$
Reserve Prices	$\theta \sim U(0, \bar{\theta})$, where, $\bar{\theta} \in \{0, 0.2, \ldots, 2.4\}$
Correlation	$\rho \in \{-1, 0, 1\}$
Confidence Level	$\delta \in \{0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$

6.3.3 Levels of Approximation

The proposed data valuation and procurement mechanisms introduce a number of approximations and bounds to achieve the desired modelling and computational properties. For example, the inclusion of reserve prices to model consumers' WTP/A, the use of the WD instead of the performance metric for a particular task to provide a model agnostic and privacy-preserving data valuation metric or the Hoeffding bound to avoid the calculation of the WD for each coalition. As such, the mechanism will not perform as well as,

[13]Minimum $L_{\mathcal{M}}(X_C) - L_{\mathcal{M}}(X_T)$ for the endogenous budget mechanism and minimum $L_{\mathcal{M}}(X_C) - L_{\mathcal{M}}(X_T) + \sum_{i \in C} t_i$ for the joint optimisation mechanism.

[14]Minimum $W(X_C, X_T)$ for the endogenous budget mechanism and minimum $K_{\mathcal{M}} W(X_C, X_T) + \sum_{i \in C} t_i$ for the joint optimisation mechanism.

for example, a cooperative game mechanism. To understand the levels of approximation, we investigate performance (value of performance metric) under different assumptions:

- Centralised Optimal (Shapley)- This is the performance achieved, $L(X_T)$ if the buyer had access to all data, and the procurement costs are determined using the Shapley Value as described in [39]. The buyer is included as an additional player in the cooperative game, with value being zero in coalitions which do not include the buyer. Implicitly, this assumes data sellers' have no reserve prices and there are no privacy concerns.

- Centralised IR (CEN_{IR}) - Here we assume there is a fixed external budget $B(X_R)$ and sellers have reserve prices. The mechanism selects the coalition with minimum, $L(X_C)$, subject to $t_i \geq q_i\theta_i, \forall i$.

- Centralised IC (CEN_{IC}) - The scenario is the same as above, however the mechanism needs to satisfy incentive compatibility. This is achieved by ensuring $t_i \geq q_i\psi_i, \forall i$. The difference between CEN_{IC} and centralised optimal scenario can be viewed as the observed 'Price of Anarchy', the cost of selfish behaviour[305].

- Centralised W (CEN_W) - Here we replace the actual task specific performance with the WD and associated Lipschitz constant. We still assume access to the value for each combination. This combines the cost of not having a TTP and desire for a task agnostic valuation metric.

- Centralised W + DP (CEN_{DP}) - Here we include the effect of differential privacy on the WD and can be thought of as the cost of privacy.

- FIN/INF - Finally, we have the proposed joint optimisation mechanism where the WDs for each coalition are replaced by the Hoeffding bound. The difference compared to the above scenario can be thought of as the cost of computational efficiency.

The cost differences are illustrative, as they vary depending on input values, however, it provides an overview of the effect of the approximations induced and a basis for studying the trade-offs introduced.

6.3.4 Results

Exogenous Budget Markets

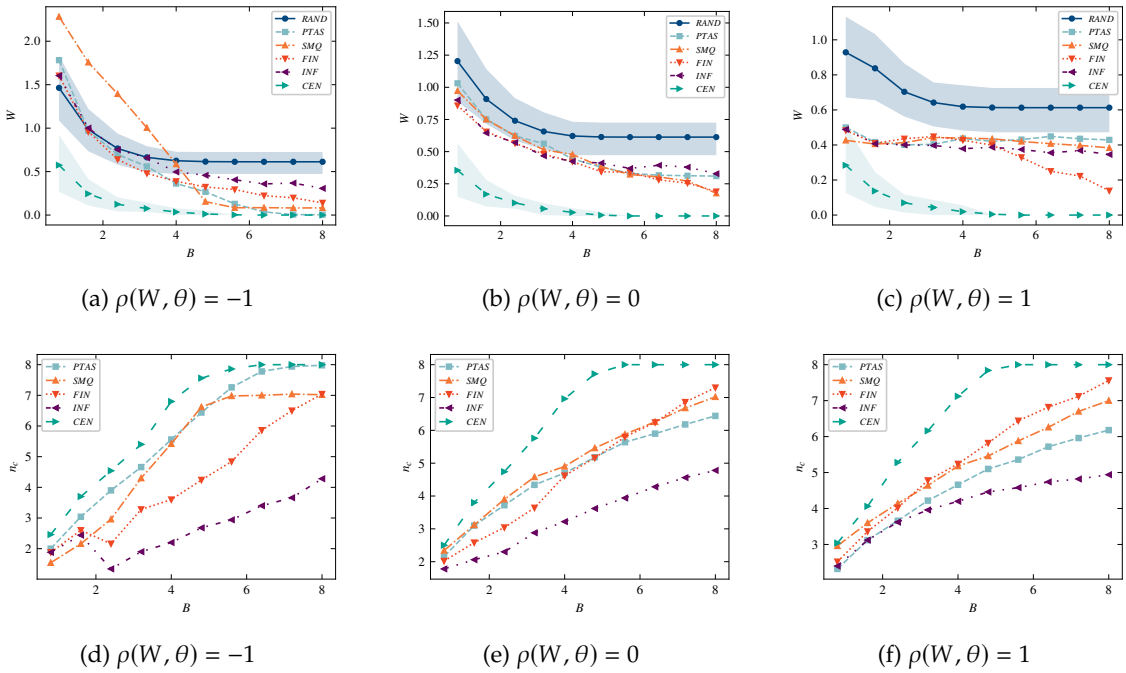
Comparison of Market Mechanisms Figure 6.3 shows the performance of the different budget feasible mechanisms considered for minimising the WD. The average WD of the selected coalition across the 50 trials is represented by the lines. In addition for the two benchmarks (CEN, $RAND$), we include the 95% confidence intervals. We see that overall, performance improves when the $\rho(W, \theta) = 0$ or 1. This is expected, as the later scenario assumes sellers with a higher WD (i.e. lower value) have higher costs, resulting in higher value per unit cost. In addition, performance of all mechanisms is generally between the two benchmarks, with the exception of SMQ in the case where WD and cost are negatively correlated. In this case SMQ picks coalitions of smaller size, because of the assumptions used to develop the mechanism.

SMQ assumes value is additive, resulting in a separable problem, where the aim is to determine individual payment thresholds. The payment thresholds are determined by maximising the expected value, based on the reserve price distributions $f_i(\cdot)$ without considering the actual reserve prices, θ_i. As a result, the mechanism allocates some of the budget to sellers which are not selected. The drawback of this effect is most pronounced in a budget constrained scenarios, i.e. for low B (with $\rho(W, \theta) = -1$ being the extreme case). However, the separability of the problem also allows the mechanism to drop incentive compatibility. As such, it is able to purchase more data, when the budget constraints are higher (i.e. less budget is 'wasted' on un-selected sellers), as the payments $t_i \ngeq q_i\psi_i$. As a result, for $B \geq 5$, for the negatively correlated scenario, SMQ performs much better.

$PTAS$ instead aims to minimise the average value per unit cost. As it accounts for the actual reserve prices it performs better than SMQ in the negatively correlated case. Additionally, like SMQ, it assumes value is additive but also assumes a prior-free environment meaning payments do not need to ensure $t_i \ngeq q_i\psi_i$.

Our proposed mechanisms, FIN and INF, perform consistently across budgets and correlations. As we account for the combinatorial nature of the problem and the actual reserve prices the mechanisms provide stable performance. FIN achieves a lower WD, than the other mechanisms, in the uncorrelated and positively correlated scenarios. This is due to the explicit modelling of the aggregation effect, through the 1/n term in the Hoeffding bound. This is particularly pronounced when the budget is higher. INF performs worse than the others expect in the positively correlated case. This is because it

underestimates the coalition size effect, however, it is computationally more efficient than FIN.



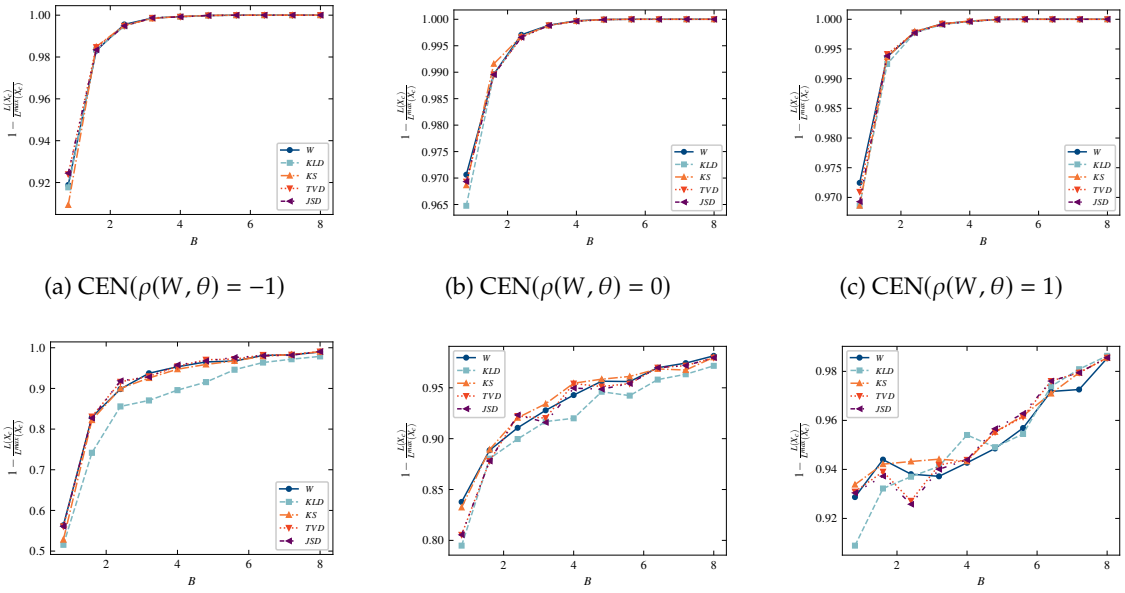
(a) $\rho(W,\theta) = -1$ (b) $\rho(W,\theta) = 0$ (c) $\rho(W,\theta) = 1$

(d) $\rho(W,\theta) = -1$ (e) $\rho(W,\theta) = 0$ (f) $\rho(W,\theta) = 1$

Figure 6.3: Exogenous Budget Mechanisms under Different Value-Price Correlation. Performance (top) and Average Number of Data Sellers Selected (bottom).

Comparison of Statistical Distances As discussed in Chapter 5, we choose the WD as our data valuation metric due to its theoretical properties, however, it is possible to use other distances. Figure 6.4 shows the improvement in loss (RMSE) for mean estimation, as a percentage of the worst case loss in the dataset $L^{max}(X_c) = \max_c L(X_c)$, using different distances. In the benchmark case (CEN) we see that performance is very similar across distances. However, using our proposed mechanism, FIN, the KLD performs worse than the other distances considered. Overall, performance is consistent across distances suggesting the selection of distance should be based on task specific or other theoretical properties.

Effect of Differential Privacy One of the main motivations for choosing the WD over other distances, is that we can model both the non-I.I.D. nature of the data and the effect of DP. This is the main contribution we make to the class of exogenous budget feasible mechanisms. Figure 6.5 shows the RMSE for mean estimation following data procurement using FIN. We assume the budget constrained scenario and compare the performance of the different metrics for increasing ϵ. When ϵ is small (i.e. high privacy preferences)

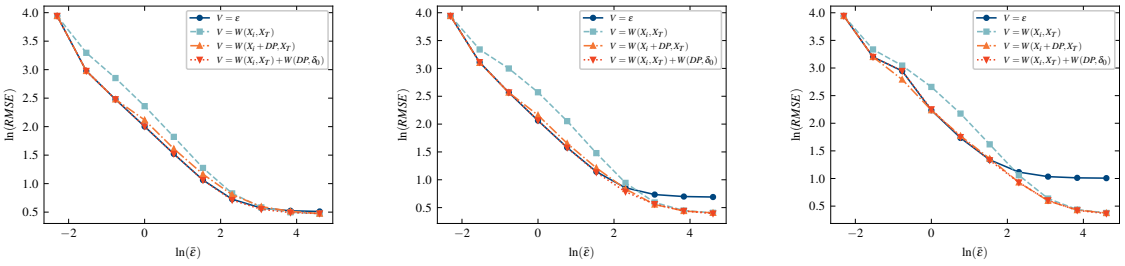


(a) CEN($\rho(W,\theta) = -1$) (b) CEN($\rho(W,\theta) = 0$) (c) CEN($\rho(W,\theta) = 1$)

(d) FIN ($\rho(W,\theta) = -1$) (e) FIN ($\rho(W,\theta) = 0$) (f) FIN ($\rho(W,\theta) = 1$)

Figure 6.4: Accuracy of Exogenous Budget Mechanisms for Mean Estimation using Different Distances

this is the main driver of value differentiation, as a result the methods which consider the effect of DP perform better than using only the WD. Conversely, when ϵ is higher the non-I.I.D. nature is the main driver and the methods which include the WD perform better. This is more pronounced when prices, θ, and privacy budgets, ϵ, are uncorrelated or negatively correlated. Both combined approaches provide good performance across privacy budgets and correlation scenarios.



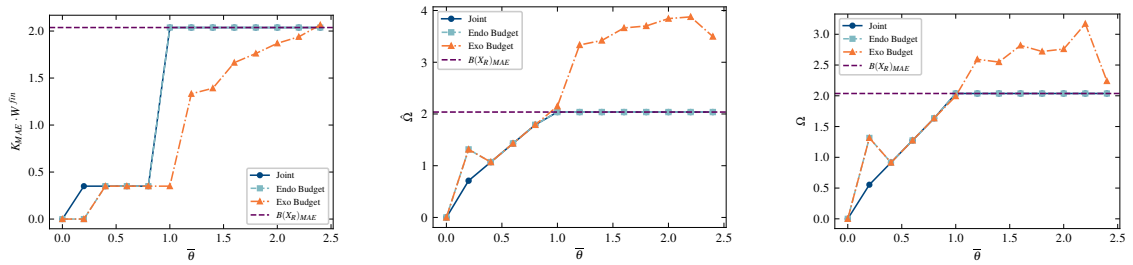
(a) $\rho(\epsilon,\theta) = -1$ (b) $\rho(\epsilon,\theta) = 0$ (c) $\rho(\epsilon,\theta) = 1$

Figure 6.5: Effect of Differential Privacy and Data Heterogeneity (B= 1.6)

Endogenous Budget Mechanisms

Objectives We now shift to the results for the endogenous budget mechanisms. Figure 6.6 illustrates the dynamics of the proposed mechanisms; exogenous budget, endogenous

budget and joint optimisation, for median estimation (i.e. minimising the mean absolute error) for a single trial of the finite formulation, FIN, assuming the value and reserve prices are negatively correlated. Figure 6.6a shows the modelled loss, $K_{\mathcal{M}} \cdot W_c^{fin}$, determined by the Hoeffding bound, for the procured subset of data, X_c. We see that as reserve prices, $\bar{\theta}$, increases the WD of the procured data increases. For the endogenous budget and joint optimisation mechanisms this does not exceed the reference budget $B(X_R)_{MAE}$, however, for the exogenous budget mechanism the reference budget is exceeded. Next, Figure 6.6b and 6.6c, show the expected cost $\hat{\Omega} = K_{\mathcal{M}} \cdot W_c^{fin} + \sum_{i \in C} t_i$ and actual cost $\Omega = (L_{\mathcal{M}}(X_c) - L_{\mathcal{M}}(X_T)) + \sum_{i \in C} t_i$, respectively. We see that the exogenous budget mechanism exceeds the reference budget and is therefore not budget feasible in this context. However, the other two mechanisms maintain budget feasibility, even in terms of actual costs as the Hoeffding bound and Lipschitz bound provide an upper bound on the actual loss. The endogenous budget can result in a lower loss, as we see in Figure 6.6a, but the overall costs (inc. payments) may be higher. The endogenous budget mechanism is useful in scenarios where the aim is to maximise task performance while maintaining decision-dependent budget feasibility. However, if the buyer also aims to minimise payments then the joint optimisation approach is most relevant, and will be the focus of the remaining results.



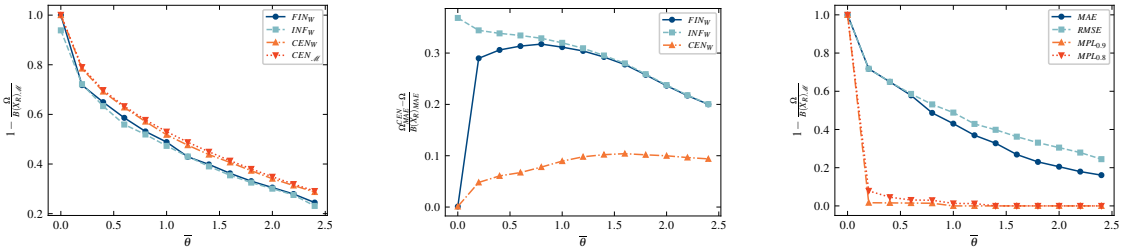
(a) Expected Loss ($K_{MAE} \cdot W^{fin}$) (b) Expected Cost ($\hat{\Omega}$) (c) Actual Cost (Ω)

Figure 6.6: Median Estimation (MAE) under Different Objectives (FIN, $\rho(W, \theta) = -1$)

Benchmarks and Tasks Having detailed the dynamics of the mechanisms, we now look at the average performance of the joint optimisation mechanism, comparing it against benchmarks and across tasks. Figure 6.7a shows the percentage improvement in cost compared to the reference $B_{MAE}(X_R)$ for median estimation. We see that the central mechanism using the WD, CEN_W, is very similar to the central mechanism using the actual loss values, CEN_{MAE}. FIN_W and INF_W perform slightly worse than the central case for median estimation. Figure 6.7b shows the cost difference, in percentage terms, compared to CEN_{MAE}, again for median estimation. First, we note that the infinite

formulation, INF_W, may not select all data sources even when they are free, resulting in a non zero cost difference for $\bar{\theta} = 0$. As the reserve prices increase we see similar cost differences for FIN_W and INF_W. We see that the cost difference peaks at around $\bar{\theta} = 1$ for INF and FIN before decreasing. This is due to the bias introduced by minimising the Hoeffding bound, shown in Figure 5.15a. Indeed, we do not observe this in the central mechanism using the WD, CEN_W.

Figure 6.7c shows the average percentage error across three different task types; median estimation (MAE), mean estimation (RMSE), quantile estimation (MPL). We see that the finite formulation has similar performance for mean and median estimation but performs badly for quantile estimation (both 90th and 80th). This is due to the tightness of the Lipschitz bound, also mentioned in Section 5.2.3. The MPL, the loss function for quantile estimation, is asymmetric resulting in an overly conservative Lipschitz constant (especially for Gaussian data). This effect and methods to improve performance will be investigated in more detail, in the next chapter.



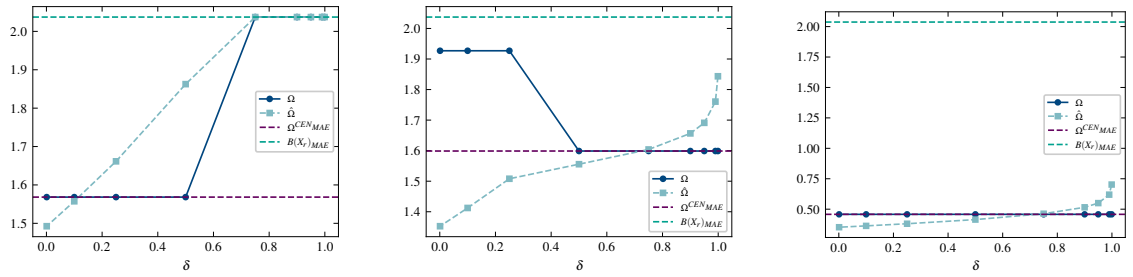
(a) Percentage Improvement across Benchmarks (MAE)

(b) Percentage Error vs. CEN (MAE)

(c) Percentage Error across Tasks

Figure 6.7: Average Joint Optimisation Performance across Different Tasks and Benchmarks ($\rho(W, \theta) = 0$)

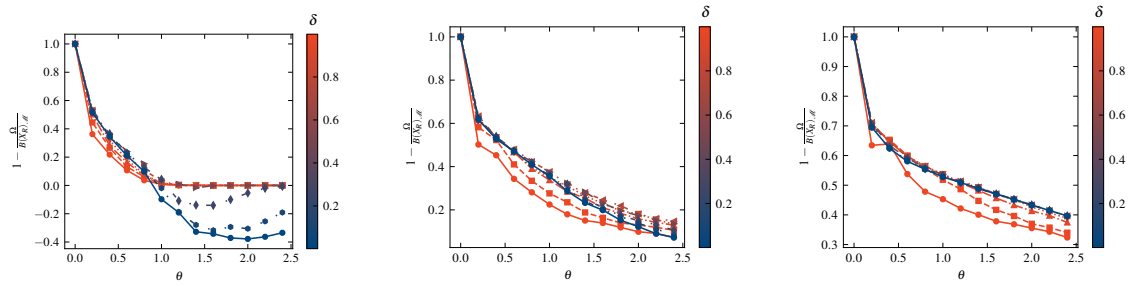
Risk One method to tackle the conservativism of the approach is to adjust the confidence level of the Hoeffding bound, δ. By reducing the δ, we are reducing the upper bound on the loss, effectively assuming each data source is more valuable. This can lead to an improvement in the procurement decisions of the mechanism but comes at increasing risk of underestimating the bound. We note that the confidence level indicates that the probability the Hoeffding bound is below the true WD is $1 - \delta$. As such, it does not tell us the probability of being below the actual loss, although this is less than $1 - \delta$. Figure 6.8 shows how changing δ affects the modelled and actual procurement costs for median estimation. We plot the actual cost Ω, the modelled cost $\hat{\Omega}$, the cost achieved in the



(a) $\rho(W, \theta) = -1$ (b) $\rho(W, \theta) = 0$ (c) $\rho(W, \theta) = 1$

Figure 6.8: Risk-Adjustment using δ adjustment for Median Estimation ($\overline{\theta} = 1.4$)

central case CEN_{MAE} and the reference budget $B(X_R)_{MAE}$. For $\rho(W, \theta) = -1$, we see that reducing the confidence level still ensures the Lipschitz bound and we are able to achieve the central optimal result when $\delta \leq 0.5$. However, for $\rho(W, \theta) = 0$ reducing δ results in an underestimation of the actual loss. We therefore, end up with increased overall costs. Lastly, when $\rho(W, \theta) = 1$, we still underestimate the actual loss when $\delta \leq 0.4$ but this does not lead to a change in the overall cost.



(a) $\rho(W, \theta) = -1$ (b) $\rho(W, \theta) = 0$ (c) $\rho(W, \theta) = 1$

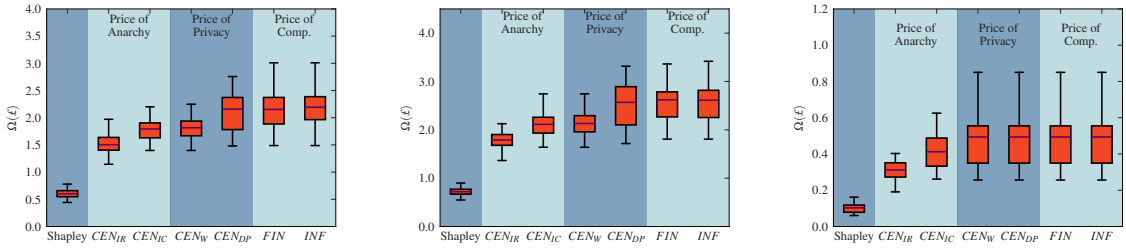
Figure 6.9: Average Performance with Risk-Adjustment for Median Estimation ($\overline{\theta} = 1.4$)

As the effectiveness of the risk adjustment depends on the tightness of the Lipschitz bound as well as correlations between value and reserve prices, we investigate the average effects across 50 trials. Figure 6.9 plots the improvement percentage against the reference budget for different values of δ. For the negatively correlated scenario, decreasing δ results in negative percentages for higher reserve prices. This indicates that in this budget constrained environment, on average, reducing conservatism leads to underestimation of the actual loss, an increase in the overall cost and loss of budget feasibility. Conversely, for the uncorrelated and positively correlated scenarios, reducing conservatism leads to an improvement in overall cost. The correlations are an indication of how much the budget constraints are limiting the selection of valuable data. As such, the negatively correlated

scenario, represents the worst-case in this respect and, hence, also results in the highest risk of over procurement.

Levels of Approximation

Finally, we present illustrative examples of the levels of approximation introduced by our mechanism for three tasks; median, mean and quantile estimation. Figure 6.10 shows boxplots of the actual cost (Ω) for each level of approximations under the following conditions; $\bar{\theta} = 0.8$, $\bar{\epsilon} = 5$, $\rho(\epsilon, \theta) = 0$ and $\delta = 0.95$. For all tasks we see that at each level of approximation the mean costs (purple line) increase or saturate. In this



(a) Median Estimation (b) Mean Estimation (c) 80th Quantile Estimation

Figure 6.10: Waterfall Charts Mapping Levels of Approximation

example, for mean and median estimation the largest effect is the inclusion of sellers' reserve prices (CEN_{IR}) and ensuring incentive compatible payments(CEN_{IC}), or the 'Price of Anarchy'. We see that the switch from using actual losses, to using WD (CEN_W) has a minimal impact on costs, suggesting it is a good valuation metric for mean and median estimation[15]. This, together with the effect of ensuring privacy-preservation of the procured data by adding differentially-private noise, models the price of privacy. Finally, to improve computational tractability, the Hoeffding bounds are used, resulting in the FIN and INF mechanisms. This further increase in costs is the cost of computational efficiency. Interestingly, we see that FIN and INF result in more concentrated costs for mean and quantile estimation, likely due to the reduced set of feasible coalitions these mechanisms can select. For quantile estimation, after introducing incentive compatibility, the costs saturate. Saturation occurs as the approximation results in an overly conservative estimate and the expected costs are higher than the budget $B(X_R)_{\mathcal{M}}$. As mentioned above this relates to the tightness of the Lipschitz bounds and will be explored further in the next chapter.

[15]Although the MSE is not strictly Lipschitz, we assume a Lipschitz constant, $K_{RMSE} = 1$. As shown in Figure 5.11, on average this holds for Gaussian data used in this case study.

6.4 Discussion

This chapter proposed three data procurement mechanisms, based on Bayesian incentive mechanism design, which build upon the data valuation mechanism proposed in Chapter 5. First, we identify the gaps in existing data procurement mechanisms. Cooperative game frameworks provide accurate data valuation for a particular task, but require a trusted third party to access sellers' data and the buyer's model. As a result, there is no privacy protection offered to either data sellers or the buyer and the method is open to model manipulation by the buyer. In addition, most of these mechanisms do not allow for the modelling of reserve prices or the effects of privacy, with [261] being the exception. On the other hand, incentive mechanism design approaches do allow for modelling reserve prices. However, existing budget feasible approaches either assume data is differentiated by privacy preferences but is I.I.D. (e.g. [51]), or is non-I.I.D. without consideration of privacy preferences(e.g. [43]). The exogenous budget mechanism we propose solves this by using the WD as the valuation metric, endogenously capturing both the non-I.I.D. nature of data and the effect of privacy preferences.

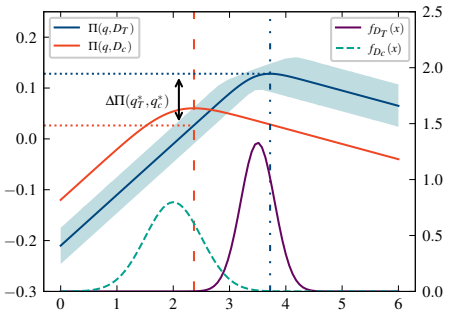
We show that our proposed exogenous budget mechanism results in more stable performance than existing budget feasible mechanisms, with the finite formulation outperforming existing mechanism when the valuation metric and reserve prices are uncorrelated or positively correlated. In addition, we showed that by using the WD to capture both the effect of DP and the non-I.I.D. nature of data, our proposed valuation metric performs better than considering only DP or the non-I.I.D. setting.

Cooperative games are able to capture the decision-dependent nature of data procurement using the Shapley value. However, they calculate the value of each coalition of data sellers, which is computationally intensive, growing exponentially in the number of data sellers. By introducing the Hoeffding bound we are able to account for aggregation effect while only requiring the calculation of individual WDs. The endogenous budget constrained and joint optimisation mechanisms use this to model the decision-dependent structure which is not captured in existing exogenous budget mechanism. We reformulate these mechanisms as MISOCPs which can be solved using commercial solvers. Although generic algorithms (e.g. branch and bound/cut) for MISOCP are efficient, there may be possibilities to exploit the problem structure (e.g. using decomposition techniques) to improve computational efficiency further. These mechanisms are closely linked to [52], however, by using the WD, the proposed mechanisms can be used for a wide range of

tasks, unlike [52], which is specific to mean estimation. Here we also include the ability to explicitly incorporate the effect of reference information on the procurement process, through the budget constraints.

We note, however, that ensuring feasibility of the budget constraint requires knowledge of the model performance with the target distribution. Alternatives to maintain privacy are proposed and the implications of each on budget feasibility are discussed. Lastly, our approach lends itself to extension to compensation-dependent privacy settings as in [52] as opposed to methods where privacy protected data is shared prior to procurement as in [266]. Our mechanism considered individual rationality, incentive compatibility and budget feasibility as the required market properties. However, there are a variety of other criteria and properties that may be incorporated/considered. For example, whether payments are symmetric i.e. do data owners with the same value receive the same compensation[38]. In addition, we focused on the single buyer-multiple seller model, however the mechanisms could be extended to a multiple buyer multiple seller environment. This, more realistic scenario, may have implications for the derived data valuations and payments due to the introduction of correlations across buyers. This has been shown to lead to value depression[306], [307].

The case studies showed that the performance of the endogenous budget and joint optimisation mechanism is heavily influenced by the tightness of the Lipschitz bound. We showed the dynamics of the mechanisms and explored the effect of risk adjustment on procurement performance and budget feasibility. The level of conservatism and the bias introduced by the Hoeffding bound affect performance differently, depending on the tightness of the Lipschitz bound, and the correlations between valuation metric and reserve prices. The next chapter, which proposes a joint energy and data market, explores this issue in more detail.



CHAPTER 7

A Joint Energy and Data Market

The data valuation and procurement mechanisms presented in Chapter 5 and 6 are generic, and can be applied to any aggregate dataset. One of the main aims of this thesis, however, is to develop such a mechanism specifically for smart meter data. We, therefore, return to the retailers' energy procurement problem, the initial motivating example. The joint optimisation mechanism allows us to develop the notion of a joint energy and data market. Specifically, we aim to optimise the retailers' total profit from day-ahead energy procurement and the procurement of smart meter data, which is used to inform the day-ahead procurement through load forecasting.

First, in Section 7.1, we start by reviewing the drivers of data value in this context and detail existing mechanisms which specifically focus on the retailer energy procurement problem. Next, Section 7.2 outlines a simplified version of the retailer energy procurement problem considered in this chapter. In Section 7.3, we detail the joint energy and data market framework. We then apply the framework to two different case studies, the first on forecast procurement and the second on actual smart meter data procurement, in Section 7.4. This chapter forms the majority of [Paper F].

7.1 Background

In order to determine the value of smart meter data, in monetary terms, we need to determine its effect on revenue generating activities. As discussed in previous chapters smart meter data has many potential benefits for different entities and across a variety of tasks. To illustrate the application of the WD based data market framework proposed in Chapter 6, we revisit the electricity retailer's energy procurement problem.

Section 5.1 provided an overview of the day-ahead energy procurement problem and

discussed the main drivers of value. Specifically, the retailer's expected profit is dependent on:

- The bidding quantity in the day-ahead market.
- The level of risk aversion of retailer.
- The accuracy the retailer's day-ahead electricity load forecast.
- The balancing market prices and the day-ahead prices/bids.

The case study showed that differentially-private smart meter data can significantly improve the retailer's day-ahead load forecast. The extent to which this is the case depends on the privacy budget and resulting noise addition introduced by DP as well as the availability of other 'reference' load data.

In this chapter we aim to explore how these factors can be incorporated into the data and energy procurement decisions of the retailer. We focus on a simplified version of the energy retailers' procurement problem, namely where the retailer is a risk-neutral price-taker, and propose a joint energy and data market. That implies a data market which accounts explicitly for the effect of data procurement decisions on the energy market.

7.1.1 Existing Approaches

As discussed in Section 5.1, the retailer's energy procurement problem has been studied extensively with many different formulations incorporating a variety of market assumptions (e.g. price-maker/taker, single or dual balancing prices) and retailer assumptions (e.g. risk aversion, tariff setting). However, the valuation of data within this context is a relatively new endeavour. Generic data market approaches were discussed in the previous chapter. Here we expand on two mechanisms, proposed specifically for the retailer energy procurement problem, in order to compare our proposed valuation and procurement mechanism.

The cooperative game framework proposed in [39], where a retailer has Gaussian day-ahead load forecasts for its customers, made up of scheduled and unscheduled load. The retailer can purchase information on scheduled load from its customer in order to reduce forecast uncertainty. In this scenario, the addition of some customers information may lead to an increase in the forecast error (due to changes in the mean forecast). As a result of this non-monotonicity, the mechanism does not ensure positive Shapley values and as a consequence cannot guarantee individual rationality, even in the absence

of privacy concerns. In [215], the authors propose the data value rate: a distribution dependent coefficient of value per unit reduction in forecast uncertainty, measured by the standard deviation of the forecast errors. The retailer selects the distribution by fitting distributional parameters to historical forecast errors and assesses the fit based on the KS metric. The data value rate is then used to develop a marginal demand curve for the retailer[268]. The supply curve is generated by modelling data sellers' reserve prices as a function of a privacy sensitivity parameter. The market clearing mechanism assumes perfect competition and therefore does not account for potential incentive compatibility issues. Importantly, both of these methods value forecasts, specifically reductions in forecast uncertainty, rather the smart meter data itself.

In addition to specific mechanisms proposed for the retailer procurement problem there are a number of other closely linked studies. In [308], that propose a joint energy and data market for a two-stage robust scheduling problem. The second stage uncertainty is modelled as an elliptical uncertainty set. Prosumers (entities that both consume and produce electricity) submit parameters, both energy bidding information and the reduction in uncertainty offered by their data, to the market operator. The market operator then simultaneously solves the robust energy scheduling problem and determines payments for the data market, which is modelled as a cooperative game. Privacy and reserve prices for prosumers data are not considered and the reduction to the uncertainty set offered by the data of each prosumer coalition is assumed to be known. Another two-stage robust scheduling problem is proposed in [293], where the market operator is able to purchase forecasts from distributed prosumers, in the first stage, to reduce uncertainty, which then is modelled via a polyhedral uncertainty set, in the second stage. Here reserve prices are included, in the form of prediction costs, however, the privacy of the forecasts in not considered. Lastly, although it is a generic data valuation framework, [266] is readily applicable to our problem, given the availability of closed-form solutions for the distributionally-robust variant of retailer energy procurement problem with a WD based ambiguity set [309].

7.1.2 Proposed Framework

Our proposed joint energy and data market framework differs from the existing approaches discussed above in that it focuses on ensuring data privacy, modelling reserve prices and most importantly values smart meter data directly, rather than forecasts, forecast accuracy, or changes in uncertainty sets. We make the following contributions:

- Application of a privacy-preserving data valuation mechanism which overcomes the need for a trusted centralised entity required in existing techniques, to the energy procurement problem.
- Application of an integrated forecasting and optimisation framework for the energy procurement problem allowing for direct data valuation rather than valuations based on forecasts or disjoint methods which value data based on forecast performance.

7.2 Retailer Energy Procurement Problem

As in previous chapters, we focus on the day-ahead energy procurement, however unlike in Section 5.1, we assume the retailer is a price-taker. This considerably simplifies the problem. We detail this problem, from [39], for notational consistency and ease of exposition.

7.2.1 Optimal Bidding Quantity

The energy retailer procures energy from the (day-ahead) wholesale market, q, at a price λ^w and receives a fixed rate of λ^r for the amount of energy consumed by its customers. It is assumed that the actual customer demand D is a random variable with a PDF, $f_D(y)$, and a CDF, $F_D(y)$, as shown in Figure 7.1a. Unlike most other retailer industries, electricity supply and demand must always match. As such, any differences between the energy procured in the wholesale market and the actual amount consumed by customers must be settled in the real-time/balancing market. We assume a balancing market with a dual pricing scheme with a system selling price, λ^+, and system buying price, λ^-. The retailers' profit function is:

$$\Pi(q, D) = \begin{cases} \lambda^r D - \lambda^w q - \lambda^-(D - q), & D \geq q \\ \lambda^r D - \lambda^w q + \lambda^+(q - D), & D < q \end{cases} \tag{7.1}$$

If the realised demand, D, is more than the energy procured in the wholesale market, the retailer must purchase the difference in the balancing market at the system buying price, λ^-. If instead, the realised demand is less than the wholesale energy procurement the retailer will sell the excess at the system selling price, λ^+. We maintain the same relational assumption between prices as in [39].

Assumption 7.1. *(Arbitrage free environment). We assume the day-ahead, λ^w and balancing market, λ^-, λ^+ prices adhere to the following inequalities. (1)$\lambda^+ \neq \lambda^-$,(2) $\lambda^+ \leq \lambda^w$, (3)$\lambda^- \geq \lambda^w$*

These inequalities setup an arbitrage-free setting. Assumption 7.1 implies that energy generation in the balancing market is more costly than planning production in the day-ahead markets. The final inequality indicates that the cost of settling a shortage is greater than the cost of settling a surplus of energy in the balancing market.

The aim of the retailer is to maximise their expected profit by choosing the day-ahead bidding quantity, q. The problem can be expressed as:

$$\max_q E_D[\Pi(q, D)] = \lambda^r \mu_D - \lambda^w q - \lambda^- E_D \left[D - q\right]^+ + \lambda^+ E_D \left[q - D\right]^+ \tag{7.2}$$

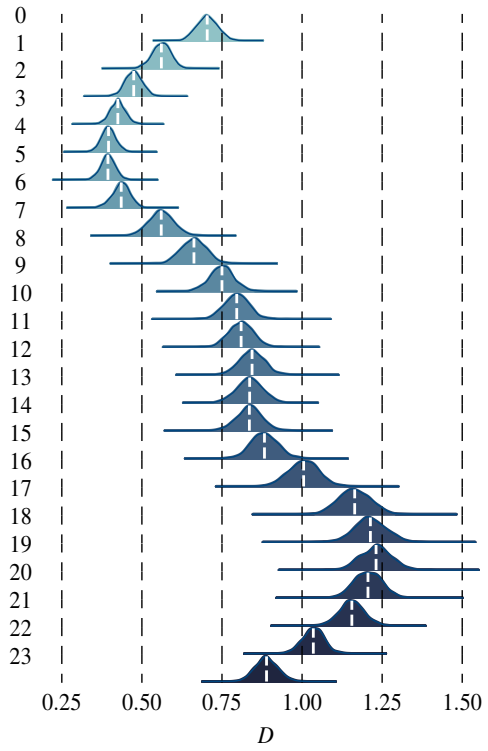
where, $\mu_D = E_D[D]$, and $[\cdot]^+$ denotes $\max(\cdot, 0)$.

As described in [39], the problem can be cast as a classical newsvendor problem and admits a closed-form solution. The optimal bidding quantity is a specified quantile of the demand forecast:

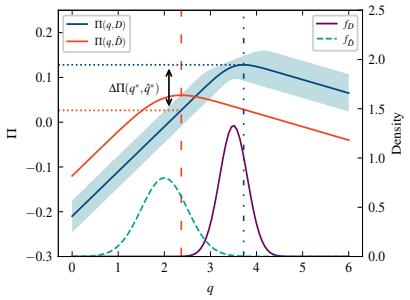
$$q^* = F^{-1}(\tau), \quad \text{where } \tau = \left(\frac{\lambda^- - \lambda^w}{\lambda^- + \lambda^+}\right) \tag{7.3}$$

We see that, τ, the critical fractile is a function, solely, of the market prices. It accounts for the asymmetry in penalties in the balancing market. However, this solution assumes the retailer knows the true demand distribution, which is not usually the case. Instead, the retailer must estimate the demand distribution, using, for example, historical data or through probabilistic forecasting. As a result, the retailer will make a procurement decision, $\hat{q}^*$, based on their estimated demand distribution $\hat{D}$, delivering a sub-optimal profit of $\Pi(\hat{q}^*, D)$. In our context, the retailer would produce day-ahead demand forecasts using reference data (e.g. national demand or load profiles provided by Elexon) or smart meter data from a subset of its customers.

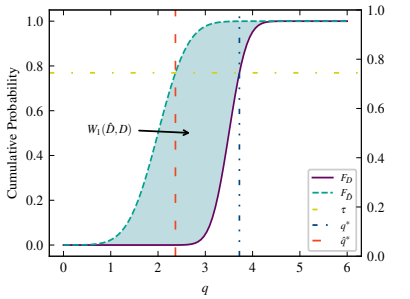
Figure 7.1 illustrates the dynamics between the demand distribution and the retailer's profit for Gaussian data. Figure 7.1b shows the true expected profit (blue) and the estimated expected profit as a function of the order quantity, q. In addition, we plot the PDFs for the true and estimated distribution. Figure 7.1c shows the CDFs of the respective demand distributions. We see that profit is maximised by the quantity corresponding to the critical fractile, for each distributions. Naturally, however, the optimal quantity assuming the estimated demand, $\hat{D}$, will result in a lower actual profit given the true distribution is D. We are, therefore, interested in the difference between the actual profit obtained when using the true demand distribution and the estimated demand distribution, $\Delta\Pi(q^*, \hat{q}^*)$. Intuitively, the difference in profit is related to some notion of difference between the forecast demand distributions. As we have shown in Chapter 5, we can use the WD to represent this difference.



(a) Illustrative Hourly Forecast Distribution over a Day



(b) Profit Distribution and Forecast PDF for a Single Timestep



(c) Forecast CDF for a Single Timestep

Figure 7.1: Effect of Distributional Shift on Profit. Plots (b) and (c) adapted from [39].

7.2.2 Integrated Forecasting and Procurement

The analysis above assumes we have access to demand forecasts and are then able to make optimal day-ahead procurement decisions based on these forecasts. This provides a method to value forecasts based on their effect on the retailer profit. However, in our framework we aim to value the smart meter data rather than the forecasts produced with it. In the next section, we therefore proceed to develop an explicit modelling framework which allows for the valuation of smart meter data itself, rather than the forecasts produced by them.

Forecasting Model

In order to establish a link between data and the retailer's profit, we need to establish and explicitly model how the data is used to produce a demand forecast. Therefore, instead of assuming access to the demand forecast distribution, D, we adopt the scenario described in [310], where we have access to some historical data $S_{N_s} = \{(d_1, \mathbf{x}_1), \ldots, (d_n, \mathbf{x}_n)\}$, where,

d_i is the demand[1] and $\mathbf{x}_i$ is a vector of covariates or features (e.g. settlement period, day of the week, and lagged demand values, X_{t-*}). This data can be leveraged to generate probabilistic day-ahead demand forecasts.

There are a variety of load forecasting techniques, ranging from model based approaches such as, ARIMA models, to more generic methods such as ANN (used in Section 5.1). As our main aim is developing a connection between the historical data and the retailer profit we choose to use ANNs which were also used in earlier case studies[2]. We stress that the choice of forecasting technique does not affect our framework, as long as the output is Lipschitz with respect to the input, $\mathbf{x}$. A comprehensive tutorial of probabilistic forecasting techniques can be found in [312]. We note one advantage of ANNs is that they are data-driven models and can model any continuous function, including non-linearities. It also does not place any distributional assumptions on the forecast errors. Specifically, we focus on the fully-connected feed-forward ANN with a single hidden layer, which can be represented as[310]:

$$\hat{y}(\mathbf{x}) = o(\mathbf{W}^{(2)} a(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) \tag{7.4}$$

where, the input nodes $\mathbf{x}$ are connected to the hidden layer nodes via a weight matrix, $\mathbf{W}^{(1)}$. The output of the hidden layer is connected to the output layer by $\mathbf{W}^{(2)}$. The bias vectors at each node are $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)}$. Finally, $o(\cdot)$ and $a(\cdot)$, represent the activation functions (e.g. ReLu or Sigmoid). The Lipschitz constant of an ANN depends on the weight matrices, and therefore on the training data used[3]. To ensure a particular Lipschitz constant we need to limit the norm of the weight matrices, either using a constrained, or regularised optimisation procedure [278].

[1]We note that this need not be the true demand. Indeed, the main motivation for the data valuation and procurement framework presented in the previous chapters, is that the retailer does not have access to the true historical demand (the smart meter data of its customers) but some other historical demand data.

[2]We also tested the simpler linear Autoregressive with Exogneous inputs(ARX) forecasting model but found that, for the smart meter dataset considered, it was unable to appropriately capture the load dynamics leading to biases. This was especially prominent for smaller coalitions of smart meters and when forecasting higher quantiles. These shortcomings are likely related to the increased irregularity of load data for smaller coalitions and the size of the data set with only spans 75 weeks. As such, there may have been insufficient data points to capture seasonal effects, especially for higher quantiles, which need significantly more data for accurate estimation[310], [311].

[3]Most common activation functions are 1-Lipschitz and by the composition property of the Lipschitz constant, they do not affect the Lipschitz constant of the network.

Optimisation

Now that we have a forecasting model to estimate the demand distribution, we investigate ways to characterise the resulting forecast distribution and optimal bidding quantity. As described in [310], there are a number of ways to approach the forecasting and procurement problem (see Figure 7.2). Disjoint approaches first estimate demand using a forecasting model to produce (mean) point forecasts, $\hat{y}_t(x)$ and forecast errors, ϵ_t. These are then used to construct the demand distribution, either using a model-based approach, where a distribution is fitted, using parameters θ, to the forecast errors, $\hat{F}_{\epsilon_t}(\hat{\theta})$, or the data-driven sample average approximation (SAA) method, where empirical forecast errors, $F^{-1}_{\epsilon_t}$, are used directly[4]. However, the disjoint nature of these approaches means we still do not have a direct relationship between the historical input data and the retailer's profit.

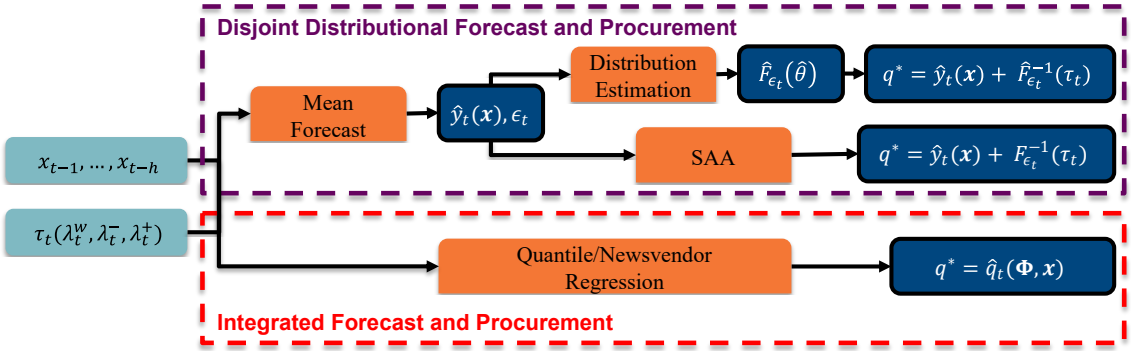


Figure 7.2: Forecasting and Procurement Frameworks

We can instead, adopt an integrated forecast and optimisation framework where we forecast the optimal bidding quantity directly. This is shown in Figure 7.2 below the two disjoint methods discussed in the previous paragraph. The integrated approach was introduced in [313] for linear forecasting models. They show that the resulting problem can be reformulated as a linear program in the case without regularisation, and as a SOCP when a l_2-norm based regularisation is introduced to avoid over-fitting. The framework was extended by [310], to general forecasting models, including non-linear machine learning techniques such as ANNs and decision trees. Most recently, [314] further extends the framework to include a more generalised class of newsvendor problems[5]. The authors also showed that for our setting, the integrated forecasting and optimisation approach is equivalent to a quantile regression problem. The integrated

[4]We note there are numerous other methods, such as, bootstrapping, however, the focus of this work is the overall framework, rather the specific techniques used. A comprehensive overview of techniques can be found in [310], [312].

[5]Non-linear in the bidding quantity, q.

retailer energy forecasting and procurement problem is equivalent to a quantile regression problem for the τ-th quantile, between historical demand data, d_i, and a set of features, $\mathbf{x}_i$[6]:

$$\min_{\mathbf{\Phi}} \mathbb{E}_D[\Pi(q(\mathbf{\Phi}, \mathbf{x_i}), D)] = \frac{1}{N_s} \sum_{i=1}^{N_s} (\tau - 1) \left[d_i - q_i(\mathbf{\Phi}, \mathbf{x_i})\right]^+ + \tau \left[q_i(\mathbf{\Phi}, \mathbf{x_i}) - d_i\right]^+ \tag{7.5}$$

where, $q_i(\mathbf{\Phi}, \mathbf{x_i})$ is the output of the ANN-based forecasting model with weight matrix, Φ, and N_s is the size of the dataset of historical values.

7.3 Joint Energy and Data Market

The integrated approach provides a direct relationship between historical data and the retailer's procurement costs/profits. We can now apply the proposed data valuation and procurement mechanism developed in Chapter 5 and 6. Overview of the dataflow of the proposed and existing, cooperative game, frameworks is provided in Figure 7.3. We assume the energy retailer has N_m customers, each with a smart meter. The customers, $i \in \mathcal{N}_m$, each have their historical smart meter data, X_i, a fixed privacy preference, ϵ_i, and a reserve price, θ_i. In addition, the retailer is sent the individual WDs, W_i, between the historical smart meter data of each individual, and the target distribution. As before, we assume the true distribution is the Euclidean barycenter of all the retailers customers. The WDs are calculated privately using MPC. The retailer provides the Lipschitz constant, K_t, which is specific to the forecasting and optimisation framework used and time dependent, the reference budget, $B_{NV}(X_R)$, and a confidence level, δ.

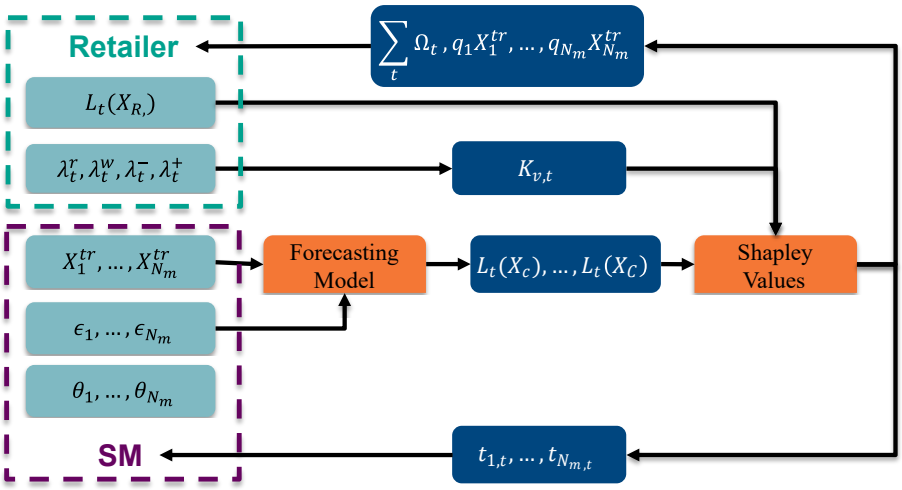
The following sections detail how the three main components of our mechanism are determined for the joint energy and data market. Namely, the WDs, the Lipschitz constant and reference budget. As discussed, in Chapter 6, the tightness of the Lipschitz constant and the choice of reference budget impact the conservatism of the mechanism and budget feasibility[7]. We propose a number of approximation schemes to investigate the trade-offs between maintaining budget feasibility and potentially improving profits.

7.3.1 Wasserstein Distance

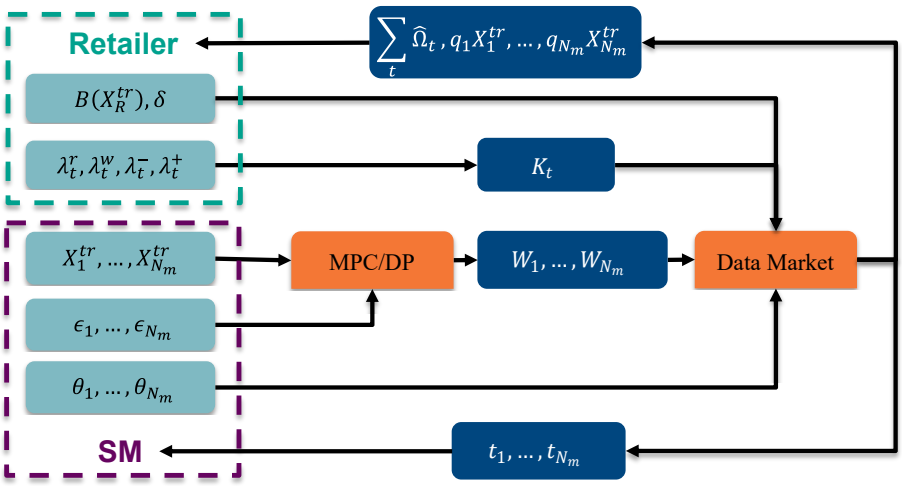
We assume the full historical smart meter dataset for each customer, is a univariate distribution, X_i^{tr}. This can also be thought of as the training set. The individual WDs

[6] A full proof can be found in Appendix D.4.

[7] Budget feasibility here refers to the combined energy and data market profit being greater than or equal to the profits received without data procurement.



(a) Existing (Cooperative Game)



(b) Proposed

Figure 7.3: Data Market Frameworks

are then defined as $W_i = W(X_i^{tr}, X_T^{tr})$, where, $X_T^{tr} = \frac{1}{N_m} \sum_{i \in \mathcal{N}_m} X_i^{tr}$. We assume that we operate in a big data regime, $N_s >> 1$[285]. That is to say, the training datasets sufficiently represent the true data distributions of each customer. We expect the WDs calculated using the training data to accurately reflect the difference across the true or test distributions:

Assumption 7.2. *Assuming a 'big data' regime the WD between the training set distribution, X_i^{tr}, and the test set distribution, X_i^{te}, is $W(X_i^{tr}, X_i^{te}) \approx 0$, for all customers $i \in \mathcal{N}_m$.*

We note, though, that our framework can easily be extended to account for discrepancies between the training and test sets. Indeed, the Wasserstein distance has been used to bound the generalisation error of models, in a similar manner to our use[273]. The convergence of a training set to the true distribution can be bounded in the Wasserstein

distance using concentration inequalities[315]. These concentration bounds[8] could be incorporated by applying the triangle inequality:

$$W(X_i, X_T) \leq W(X_i^{tr}, X_T^{tr}) + W(X_i^{tr}, X_i^{te}) + W(X_T^{tr}, X_T^{te}) \tag{7.6}$$

We note here the connection to Wasserstein based DRO. Indeed, one of the main uses of DRO is to optimise while account for potential train-test discrepancies ($W(X_T^{tr}, X_T^{te})$) [316]. However, our setting differs from the DRO framework in that we do not have access to X_T^{tr}. Instead, we have access to some reference data, X_R^{tr}. We could formulate a DRO problem with the ambiguity set defined by $W(X_R^{tr}, X_T^{tr})$. DRO will then optimise against the worst-case distribution within the Wasserstein ball described by $W(X_R^{tr}, X_T^{tr})$. However, we know that the target distribution results in the increased, and indeed, optimal profit and is therefore not the worst case but rather the best case[9].

7.3.2 Lipschitz Constant

The integrated forecasting and optimisation framework provides a function connecting the input data, $\mathbf{x}$, to the retailer's profit, Π. This allows us to determine a Lipschitz constant to the entire framework[10]:

$$K = 2K_f \max(\lambda^u, \lambda^o) \tag{7.7}$$

where, K_f is the Lipschitz constant of the forecasting method.

7.3.3 Calibrating Conservatism

The Hoeffding bound confidence level, δ, provides a principled method to calibrate the conservatism of the valuation, by adjusting the risk tolerance level. However, in some cases this is insufficient. The Lipschitz bound provides a worst-case bound on the performance loss compared to the target distribution. As such, this may result in overly conservative estimates. This was particularly pronounced for the newsvendor problem using Gaussian data (see Section 5.1.5). We, therefore, propose two methods to calibrate the Lipschitz constant, in such scenarios. The Lipschitz constant is applicable for any distribution. However, problem specific information can significantly tighten this bound (e.g. bounds on demand, restrictions on the potential demand distributions).

[8]The concentration inequalities require the calculation of distribution dependent constants as shown in [315, Theorem 2].

[9]We note a relatively new field of distributionally-optimistic optimisation[317], which could provide a fruitful avenue for future development.

[10]A full proof can be found in Appendix D.4

Transfer Function One could adopt a transfer function approach, similar to [267], where the relationship could be determined using, for example, the reference benchmark data available to the retailer or based be on simulations with synthetic data. This empirical Lipschitz constant is:

$$\hat{K} = \max_{i,j} \frac{\Delta\Pi(X_i, X_j)}{W(X_i, X_j)} \tag{7.8}$$

where, X_i, X_j, are synthetic/reference distributions. The above method may still lead to conservative estimates of the performance loss. Instead, we could adopt a probabilistic approach and use the average ratio between profit difference and the WD.

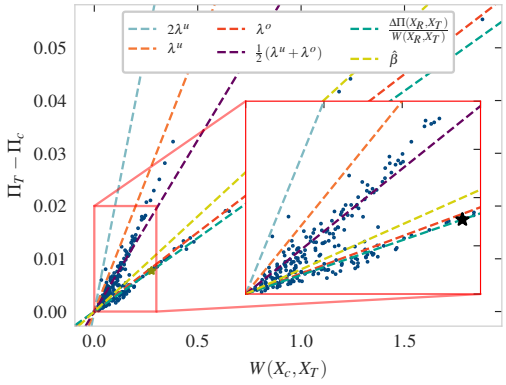


Figure 7.4: Transfer Function Assuming Linear Relationship

Figure 7.4 shows the relationship between the WD and profit differences for real smart meter data[11]. We include lines whose slope represent different cost parameters in newsvendor cost function. We see that values are broadly clustered between two lines, λ^o and λ^u. We also include a number of other options; the slope estimated from the data, $\hat{\beta}$, or based on some alternative reference dataset, X_R (indicated by the black star in the figure). There are very few coalitions whose ratio of profit difference to WD exceeds λ^u, suggesting the true Lipschitz constant, $2\lambda^u$, is overly conservative, in this case. However, using one of the other values prioritises estimation performance over theoretical guarantees, specifically, we forego the (probabilistic) guarantees of budget feasibility.

Lipschitz Relaxation Alternatively, we could relax the definition of Lipschitz continuity by introducing the notion of locally Lipschitz continuity. Essentially, we limit the input range considered, in order to provide a tighter Lipschitz bound. Formally:

Definition 7.3.1. *(Locally Lipschitz). A function* $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$ *is locally Lipschitz at*

[11]This data is used for the case study on smart meter data procurement in the next section.

$x_0 \in A$ if there exist constants $\xi > 0$ and $M \in \mathbb{R}_+$ such that:

$$|x - x_0| \leq \xi \Rightarrow |f(x) - f(x_0)| \leq M|x - x_0| \tag{7.9}$$

The improvement attained through this Lipschitz relaxation depends on the demand distribution considered, and the input attribute we choose to limit (e.g. the difference in bidding quantities, means of the distribution). To illustrate the effect of the relaxation, we analyse the Gaussian newsvendor problem. If we assume the demand distribution is Gaussian we can obtain a closed-form expression for the retailer profit (adapted from [318]):

$$\begin{aligned}\mathbb{E}[\Pi(q, D)] = &(\lambda^r - \lambda^w)\,\mu_D + \lambda^u\,(\mu_D - q) \\ &- (\lambda^u + \lambda^o)\left\{(\mu_D - q)\,\Phi\left(-\frac{(\mu_D - q)}{\sigma_D}\right) + \sigma_D \phi\left(\frac{\mu_D - q}{\sigma_D}\right)\right\}\end{aligned} \tag{7.10}$$

where, ϕ and Φ are the standard Gaussian PDF and CDF, respectively.

The Lipschitz constant is the maximum absolute gradient of the profit function, where:

$$\frac{\partial \Pi}{\partial q} = (\lambda^u + \lambda^o)\,\Phi\left(\frac{q - \mu_D}{\sigma_D}\right) - \lambda^u = \lambda^o - (\lambda^u + \lambda^o)\,\Phi\left(\frac{\mu_D - q}{\sigma_D}\right) \tag{7.11}$$

We see that the Lipschitz constant is dependent on how far the order quantity is from the mean of the demand distribution and the spread of the distribution itself. Placing bounds on these values would allow us to develop a locally Lipschitz condition which could provide better approximations. Alternatively, we could bound the difference between bid, q, and the optimal bid, q^*.

Proposition 7.3. *The Gaussian newsvendor is locally Lipschitz with respect to the bidding quantity q, if $|q - q^*| \leq \xi$, with the Lipschitz constant:*

$$K_{nv}^{\xi} = \begin{cases} \lambda^u - (\lambda^u + \lambda^o)\,\Phi\left(\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right), & \lambda^u > \lambda^o \\ \lambda^o - (\lambda^u + \lambda^o)\,\Phi\left(-\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right), & \lambda^u < \lambda^o \end{cases} \tag{7.12}$$

Proof. To determine the local Lipschitz constant we are interested in the maximum absolute gradient of the profit function. First we note that, the optimal bid is $q^* = \mu_D + \sigma_D \Phi^{-1}(\tau)$ from (7.3). Next, due to the monotonicity of $\Phi(x)$, we can focus on the extreme cases in the local neighbourhood defined by ξ.

- For $q = q^* + \xi$, the profit gradient is positive. Thus the profit gradient is simply:

$$\left|\frac{\partial \Pi}{\partial q}\right| = \lambda^o - (\lambda^u + \lambda^o)\,\Phi\left(-\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right) \tag{7.13}$$

- For $q = q^* - \xi$, the profit gradient is negative. The absolute profit gradient is then:

$$\left|\frac{\partial \Pi}{\partial q}\right| = -\frac{\partial \Pi}{\partial q} \tag{7.14}$$

$$= (\lambda^u + \lambda^o)\,\Phi\left(-\frac{\xi}{\sigma_D} - \Phi^{-1}(\tau)\right) - \lambda^o \tag{7.15}$$

$$= \lambda^u - (\lambda^u + \lambda^o)\,\Phi\left(\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right) \tag{7.16}$$

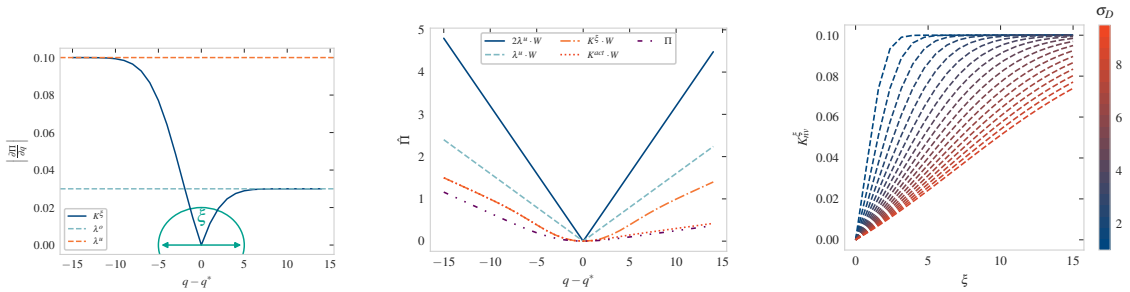
The maximum gradient for a given ξ is then $\max\left((7.13), (7.16)\right)$. The difference between (7.13),(7.16) is:

$$\Delta = \lambda^u - \lambda^o - \overbrace{(\lambda^o + \lambda^u)\left[\Phi\left(\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right) - \Phi\left(-\Phi^{-1}(\tau) - \frac{\xi}{\sigma_D}\right)\right]}^{H(\xi)} \tag{7.17}$$

We note that as $\xi \to \infty$, $H(\xi) \to 0$. At $\xi = 0$, $H(\xi) = \lambda^o - \lambda^u$. Due to monotonicity for $\xi \geq 0$, $H(\xi) \in [\lambda^u - \lambda^o, 0)$ or $(0, \lambda^u - \lambda^o]$. We then consider two cases for the underage and overage costs.

- $\lambda^u > \lambda^o$: $H(\xi) \leq \lambda^u - \lambda^o \implies \Delta \geq 0 \implies \max\left((7.13), (7.16)\right) = (7.16)$
- $\lambda^o > \lambda^u$: $H(\xi) \geq \lambda^u - \lambda^o \implies \Delta \leq 0 \implies \max\left((7.13), (7.16)\right) = (7.13)$

□



(a) Profit Gradient w.r.t. q (b) Wasserstein Approximation (c) Dependence on σ_D

Figure 7.5: Lipschitz Relaxation for Gaussian Newsvendor ($\mu_D = 10, \sigma_D = 3$)

Figure 7.5, shows the effect of using the Lipschitz relaxation. The plot on the left shows the change in the gradient as a function of the difference between the bid and the optimal bid, $|q - q^*|$. The middle plot shows the performance of the WD based bounds of the true profit, Π, using different constants. If we use the global Lipschitz constant, $2\lambda^u$, or even, λ^u, we obtain fairly loose, linear bounds. If, however, we used the locally Lipschitz constant, K^{ξ}, we obtain a tighter bound, especially for $q - q^* < 0$, as the

underage costs are higher than overage costs in this example. Lastly, if we knew the actual difference, $|q - q^*|$, we obtain a much tighter bound. We note that the above approach still requires an assessment of the demand uncertainty, σ_D. In the last plot we show the effect of the underlying demand uncertainty on the locally Lipschitz constant. We see that increasing the uncertainty reduces the gradient and as a result lowers the constant. As such, we can obtain an upper bound by selecting a theoretically motivated minimum standard deviation (e.g. load forecasts at higher levels of aggregation, such as, substation or national demand would be expected to have lower volatility).

The analysis of the Gaussian newsvendor above provides an insight into the dynamics affecting the performance of our WD based valuation mechanism for the newsvendor problem. Overall the tightness of the Lipschitz bound is affected, principally by the asymmetry of the cost/profit function of the newsvendor problem. Indeed, if $\lambda^u = \lambda^o$, the problem would be equivalent to median estimation, for which our mechanism performs well, as seen from the case studies in the previous chapters. In addition to the cost asymmetries, the demand distribution itself, specifically the spread of the PDF, which corresponds to the rate of change of the locally Lipschitz constant, plays a significant role in determining the tightness of the Lipschitz bound[12].

7.3.4 Reference Budget

Thus far we have assumed $B(X_R) = \Pi(X_T) - \Pi(X_R)$, which ensure budget feasibility. However, this requires knowledge of $\Pi(X_T)$ which, as discussed in Section 6.2.3, may not be an appropriate assumption. If we do not know this quantity, we can use bounds or approximations. To ensure budget feasibility we would need a lower bound on $B(X_R)$. However, in general, the reference budget can only be bounded as follows:

$$0 \leq B(X_R) \leq K \cdot W(X_R, X_T) \tag{7.18}$$

Introducing distributional or problem-specific information could be used to develop a useful lower bound. For example, limiting the difference between optimal bid and reference bid, as discussed in the context of the Lipschitz constant. An alternative is to forego budget feasibility and use the WD, which provides an upper bound, so as to maintain consistency within our framework.

[12]We note that this matches observations in [313, Theorem 4], regarding the concentration of demand across $q - q^*$.

7.4 Application to Smart Meter Data

To evaluate the performance of the proposed joint energy and data market we conduct two case studies. The first focuses on procuring forecasts, which allows us to compare our WD based valuation mechanism against existing techniques. The second implements the full integrated forecasting, optimisation and procurement mechanism using real smart meter data.

7.4.1 Case Study: Forecast Procurement

For this case study we use the framework presented in [39]. It involves an energy retailer aiming to maximise their profit, as described in Section 7.2, by procuring scheduled load information, thereby reducing forecast uncertainty. The information procurement is setup as a cooperative game and the payoff allocations are made using the Shapley value. We briefly outline the framework below.

Framework and Experimental Setup

The retailer has m customers, each with some scheduled load $L_i^s \sim N\left(\sum_{i\in M}\mu_i^s, {\sigma_i^s}^2\right)$ and unscheduled load $L_i^u \sim N\left(\sum_{i\in M}\mu_i^u, {\sigma_i^u}^2\right)$. The retailer produces a forecast (using e.g. historical data), which contains the schedulable and unschedulable load components:

$$D_\emptyset \sim N\left(\sum_{i\in M}\mu_i^u + \mu_i^s, \sum_{i\in M}{\sigma_i^u}^2 + {\sigma_i^s}^2\right) \tag{7.19}$$

The retailer can purchase information on scheduled load, from a subset or coalition, C, of customers, resulting in new load forecast distribution:

$$D_c \sim N\left(\sum_{i\in M}\mu_i^u + \sum_{i\in M/C}\mu_i^s + \sum_{i\in C} l_i^s, \sum_{i\in M}{\sigma_i^u}^2 + \sum_{i\in M/C}{\sigma_i^s}^2\right) \tag{7.20}$$

where, l_i^s is the realised scheduled load.

The best load forecast is achieved when the retailer has scheduled load information from all its customers, the grand coalition of the cooperative game. The load forecast is then[13]:

$$D_M \sim N\left(\sum_{i\in M}\mu_i^u + l_i^s, \sum_{i\in M}{\sigma_i^u}^2\right) \tag{7.21}$$

[13]We note that in this case study, the target distribution is not strictly an aggregation i.e $D_M \neq \frac{1}{N}D_i$. As a result, the Hoeffding bound convergence dynamics do not strictly apply in this case. We therefore, cannot guarantee to confidence level, δ, of the probabilistic Hoeffding bound. However $W(D_c,\ D_M) \to 0$ as $c \to M$.

The retailers' profit under each coalition is then:

$$\Pi(q_c, D_M) = \lambda^r D_M - \lambda^w q_c - \lambda^-[D_M - q_c]^+ + \lambda^+[q_c - D_M]^+ \tag{7.22}$$

where, $q_c = \mu_c + \sigma_c \Phi^{-1}(\tau)$, with μ_c and σ_c being the mean and standard deviation of the coalition load forecast, respectively, as defined in (7.20). The retailer is considered a player, 0, in the cooperative game. The value of a particular coalition is then:

$$V(C) = \begin{cases} 0 & \text{if } 0 \notin C \\ \mathbb{E}[\Pi(q_c, D_M)] - \mathbb{E}[\Pi(q_\emptyset, D_M)] & \text{if } 0 \in C \end{cases} \tag{7.23}$$

The payoff allocation for each consumer, ϕ_i, is then determined using the Shapley Value.

We consider a number of alternative valuation mechanisms to compare the performance of our WD based mechanism:

- Actual Profit Difference,($\Delta\Pi$)[39], as above.
- Data Value Rate ($\Delta\sigma$)[297, Table 1]: Value is determined by the reduction in the standard deviation of the forecast, $\Delta\sigma$, and a distribution specific data value rate, C_σ.

$$V_{\Delta\sigma} = C_\sigma \left(\overbrace{\sqrt{\sum_{i \in M} \sigma_i^{u^2} + \sigma_i^{s^2}}}^{\sigma_{D_\emptyset}} - \overbrace{\sqrt{\sum_{i \in M} \sigma_i^{u^2} + \sum_{i \in M/C} \sigma_i^{s^2}}}^{\sigma_{D_c}} \right) \tag{7.24}$$

  where, $C_\sigma \approx \frac{\sqrt{2\pi}}{4} \left(\lambda_t^- - \lambda_t^+\right) \left[\tau_t \ln \frac{1}{\tau_t} + (1 - \tau_t) \ln \frac{1}{1-\tau_t}\right]$.

- Distributionally-Robust Data Valuation [14] ($DRO_{c,M}$): The authors formulate the data valuation problem as a DRO problem. To adapt it to our setting, we assume the retailer has access to the distribution for each coalition, D_c, and the WD between the coalition the the grand coalition, $W(D_c, D_M)$. The resulting value function is:

$$V_{DRO_{c,M}} = \mathbb{E}[\Pi(q_c, D_c)] - \lambda^u W(D_c, D_M) - \mathbb{E}[\Pi(q_\emptyset, D_M)] \tag{7.25}$$

- Wasserstein w.r.t Reference ($W_{c,\emptyset}$): In this case the retailer can access the WD between the reference forecast and the coalition $W(D_\emptyset, D_c)$. The retailer, then aims

[14] We use the closed-form solution for 1-Wasserstein distance DRO newsvendor problem provided in [309]. We note this requires two additional assumptions; with the demand support is $\Xi \in (0, \infty)$ (not strictly applicable in this scenario), and $\lambda^u > \lambda^o$.

to maximise the difference between their reference data and the coalition resulting in[15]:

$$V_{W_{c,\emptyset}} = K \cdot W\left(D_{\emptyset}, D_c\right) \tag{7.26}$$

- Wasserstein w.r.t. Target ($W_{c,M}$): This method corresponds to our proposed valuation mechanism. We choose to use the zero-Shapley policy[319]. This is because negative coalition values are not necessarily an indication of actual reduction in profit but may be a reflection of the conservatism of our approach. As such, we assume negative coalitions provide no information, i.e. have a value of 0. Within this method we also consider the effect of the finite ($W_{c,M}^{fin}$) and infinite ($W_{c,M}^{inf}$) Hoeffding formulations.

$$V_{W_{c,\emptyset}} = \max\left(\Pi(q_M, D_M) - \Pi(q_{\emptyset}, D_M) - KW\left(D_M, D_c\right), 0\right) \tag{7.27}$$

To investigate the effect of DP we extend the framework as follows. We assume each consumer has some privacy preference ϵ_i and once the retailer purchases the scheduled load information, it will be sent with differentially private noise, using the Gaussian mechanism[16]. The retailer's forecast with subset C of purchased customers data then becomes:

$$D_c^{dp} \sim N\left(\sum_{i\in M}\mu_i^u + \sum_{i\in M/C}\mu_i^s + \sum_{i\in C} l_i^s, \sum_{i\in M}\sigma_i^{u\,2} + \sum_{i\in M/C}\sigma_i^{s\,2} + \sum_{i\in C}\sigma_i^{dp\,2}\right) \tag{7.28}$$

where σ_i^{DP} is the standard deviation of the DP noise.

The grand coalition forecast with DP is:

$$D_M^{DP} \sim N\left(\sum_{i\in M}\mu_i^u + l_i^s, \sum_{i\in M}\sigma_i^{u\,2} + \sigma_i^{DP\,2}\right) \tag{7.29}$$

where, $\sum_{i\in M}\sigma_i^{DP\,2} = \gamma\sum_{i\in M}\sigma_i^{s\,2}$, and γ, is a non-negative noise multiplier.

Here we investigate two methods to model the effect of DP under the Wasserstein valuation framework:

- $W(D_c^{dp}, D_M^{dp})$: Exact expression for the forecast variance, by employing Corollary 5.4. This assumes the data is known to be Gaussian.
- $W(D_c, D_M^{dp})$: Upper bound on forecast variance, with no distributional assumption, using Proposition 5.5.

[15]We note here that this provides no guarantees of obtaining a better forecast or increased profits, only that it will be different from the reference forecast $D_{\emptyset}$.

[16]Chosen to retain the closed-form formulations for this case study.

In addition, as we showed in Section 6.3.4, correlations between privacy parameters and value have an impact on performance. We therefore consider four privacy scenarios of how the total DP noise, $\sum_{i \in M} {\sigma_i^{DP}}^2$, is distributed among consumers:

1. Correlated (corr): Noise is allocated in proportion to the proportion of schedulable load, l_i^s, of each consumers. As a result, consumers with a higher value have higher privacy preferences.

2. Inverse (inv): Noise is allocated inversely proportionally to the proportion of schedulable load. Scenario where, consumers with a higher value have lower privacy preferences.

3. Uniform (uni): Noise is allocated uniformly among consumers, indicating they all have the same privacy preferences.

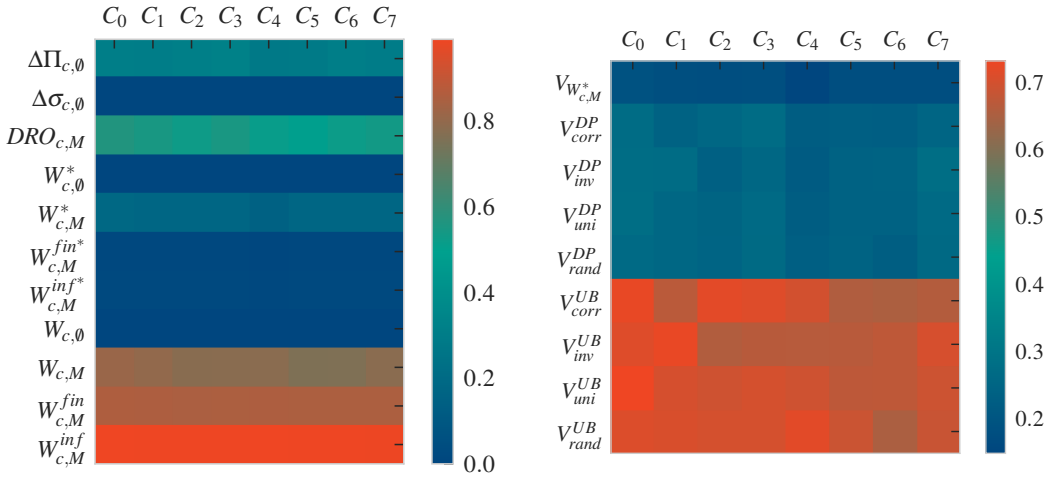
4. Random (rand): Noise allocated randomly indicating random privacy preferences.

As in [39], we assume market prices are, $\lambda^r = 0.1$£/kWh, $\lambda^w = 0.06$£/kWh, $\lambda^- = 0.16$£/kWh and, $\lambda^+ = 0.03$£/kWh[17]. We also consider there are 8 consumers and use the same procedure to generate the demand forecasts over 48 HH time steps. However, we use the modified smart meter dataset described in Section 5.1.5. For the differential privacy scenarios we first assume $\gamma = 0.5$.

Results

Calibration As discussed in Section 7.3.3, the looseness of the Lipschitz bound for the newsvendor problem can result in overly conservative value estimates. We observe this phenomenon in this case study. Figure 7.6a shows a heatmap of the proportion of non-positive coalition values for each consumer. Broadly, we see that the proportions are consistent across consumers. However, we see that the Wasserstein valuation mechanism ($W_{c,M}, W_{c,M}^{fin}, W_{c,M}^{inf}$) result in a high proportions, above 0.7. As a result, we do not gain any information from over 70% of the valuations. Indeed, we see that the resulting Shapley allocations give all consumers the same payoffs (not shown). Instead, we can calibrate the constant using the reference data: $K^* = \frac{\Pi(q_M) - \Pi(q_\emptyset)}{W(D_\emptyset, D_M)}$.

[17]The prices have been chosen to satisfy the Assumption 7.1 and are purely illustrative. As with other case studies presented in this thesis (see Section 5.1) we are focusing on the dynamics of data value rather than the actual monetary value. In order to obtain an estimate of real-world value of smart meter data or forecasts future work could use the presented framework with the appropriate market prices.



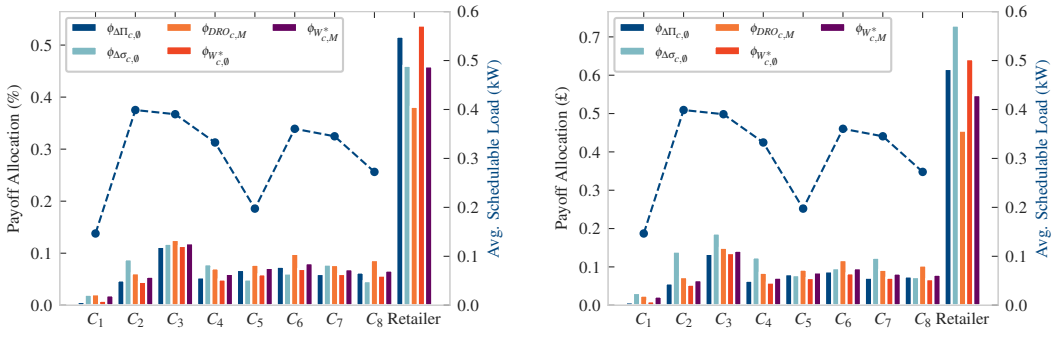
(a) Lipschitz Calibration

(b) Effect of DP

Figure 7.6: Proportion of Non-Positive Coalition Values

Following calibration, there is a significant decrease in the proportion of non-positive coalitions ($W^*_{c,M}, W^{fin^*}_{c,M}, W^{inf^*}_{c,M}$) and resulting more information that can be used for Shapley calculation. Again, this is broadly consistent across consumers.

Forecast Valuation Figure 7.7 shows the Shapley allocations determined using the different valuations methods. We see that they are broadly in line with each other and are also correlated to the average scheduleable load of each consumer, as was identified in [39]. To evaluate the performance of the different valuation methods we compare them against the ground truth, the actual profit and the associated the Shapley allocations. We summarise the key performance metrics in Table 7.1.



(a) Allocation %

(b) Allocation in £

Figure 7.7: Shapley Values under Different Valuation Metrics

First we see that the calibrated WD based approaches and the DRO method are highly correlated with the actual profit, with our proposed method giving a correlation

coefficient of $\rho(\Delta\Pi, W^*_{c,M}) = 0.986$. The data value rate ($\Delta\sigma$), however, has a correlation coefficient of 0.46. This is because it only considers changes in the standard deviation of the forecast whereas the WD as it captures both mean and standard deviation differences. If the scheduled load, l^s_i had been equal to the mean, μ^s_i, the data rate value and reference based Wasserstein approach, $W^*_{c,\emptyset}$ would be equivalent up to a constant factor. These methods also ensure there are no negative coalition values. However, these methods do not ensure budget balance ($\sum \phi - \sum \phi_{\Delta\Pi}$), where as the other methods using the WD w.r.t. the target distribution do ensure budget balance. The infinite Hoeffding approximation leads to a budget feasible but not balanced allocation.

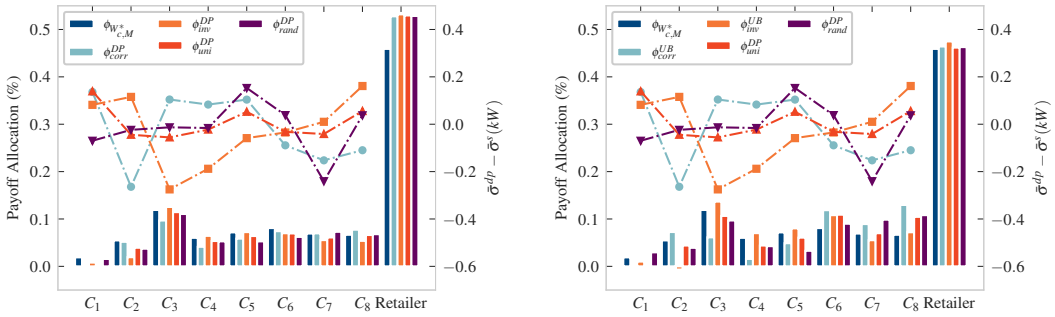
Table 7.1: Performance Metrics across Valuation Metrics

	$\Delta\sigma_{c,\emptyset}$	$DRO_{c,M}$	$W_{c,\emptyset}$		$W_{c,M}$		$W^{fin}_{c,M}$		$W^{inf}_{c,M}$	
			K	K^*	K	K^*	K	K^*	K	K^*
$\rho(\Delta\Pi, *)$	0.46	0.94	0.76	0.96	0.88	**0.99**	0.59	0.83	0.29	0.86
$\%\leq 0$	**0.00**	0.58	**0.00**	**0.00**	0.82	0.20	0.90	0.04	0.99	0.05
$\sum \phi - \sum \phi_{\Delta\Pi}$	0.39	**0.00**	3.63	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	-1.16	-0.53
% Retailer	0.46	0.38	0.55	0.54	0.30	0.46	0.21	0.48	0.30	0.57
$\sum\lvert\phi_{\Delta\Pi} - \phi\rvert$(£)	1.42	0.36	4.28	0.25	0.66	**0.24**	1.37	0.83	1.24	0.84
$\sum\lvert\phi_{\Delta\Pi} - \phi\rvert$(%)	0.02	0.03	0.01	**0.01**	0.05	**0.01**	0.07	0.02	0.05	0.02

Interestingly, although the calibration results in a reduction in the conservatism of the valuation, it leads to the retailer being given a higher proportion of the value under the Shapley allocations. This is likely due to the un-calibrated values providing no way to distinguish between consumers or the retailer. We also see that our proposed valuation mechanism, $W^*_{c,M}$, results in the lowest mis-allocation, both in terms of £ and %, compared to using the actual profit difference, $\Delta\Pi$. Overall, we see that the proposed valuation method performs well and maintains theoretical properties such as budget balance. Although the DRO method also performs well, it requires the calculation of the profit of each coalition and the WD between the coalition and the grand coalition. The Hoeffding bound approximations provide further computational improvements but do reduce performance in terms of mis-allocations.

Effect of Differential Privacy We now consider the addition of DP. The noise introduced into the forecast leads to increased costs compared to the non-DP scenario. As a result, the total payoffs in £ will be less. As we define $\sum_{i\in M} {\sigma^{DP}_i}^2 = \gamma \sum_{i\in M} {\sigma^s_i}^2$, the overall value

reduction is the same in each privacy setting, only the distribution across consumers varies. When γ=0.5 the profit reductions observed are ~32% and ~60% using the exact and upper bound formulations, respectively.



(a) Exact DP Formulation

(b) Upper Bound

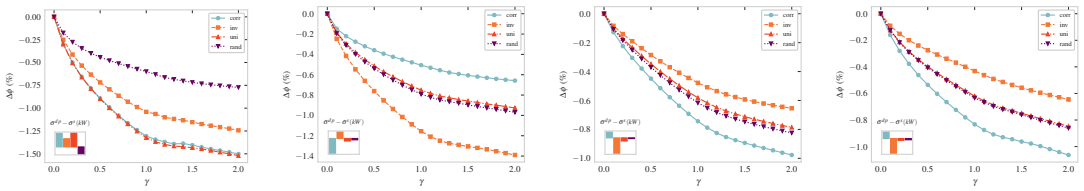
Figure 7.8: Shapley Values with Differential Privacy

Figure 7.8 shows the Shapley allocations using the Wasserstein valuation mechanism under different DP formulations. We also plot the average differences between the scheduling uncertainty, σ^s and the DP noise, σ^{dp}. We see that with the exact formulation the main driver remains the average schedulable load, as the allocations are in line with the non-DP scenario ($\phi_{W^*_{c,M}}$). The relative rankings of consumers are preserved across the DP correlation scenarios (e.g. C_3 has the highest allocation in the non-DP case as well as in all the DP scenarios). The impact of DP noise is more prominent when using upper bound formulation. This results in zero or negative allocations for some consumers (e.g. C_2 in the random scenario)[18], as their contribution to improving the mean forecast is less than the introduced noise. Indeed, the upper bound formulation results in an increased proportion of non-positive coalition values, as shown in Figure 7.6b.

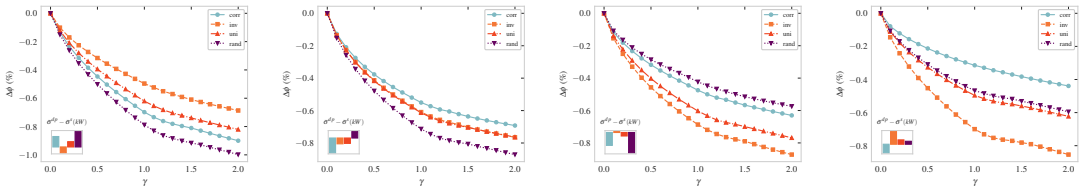
We see that changes in allocations across the different correlation scenarios are complex and depend on multiple factors. For example, C_3, exhibits a reduction in allocation in the correlated DP scenario and an increase in allocation in the inverse case, whereas, for C_6, the allocations decrease for all scenarios. This is despite the two consumers being the highest and second highest valued in the non-DP case. The changes depend on the average schedulable load, the difference in the scheduling uncertainty, σ^s and the DP noise, σ^{dp}, as well as the proportion of positive coalition values.

To further investigate the effect of DP, we vary the noise multiplier, γ. Figure 7.9 shows the percentage change in the Shapely allocations (in £) for each consumer as a function

[18]Although we use the zero-Shapley value to avoid negative coalition values occurring. This does not ensure monotonicity of the value function, hence it is still possible to generate negative allocations.



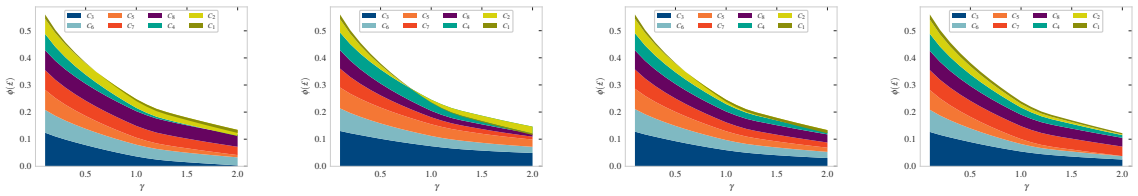
(a) C_1 (b) C_2 (c) C_3 (d) C_4



(e) C_5 (f) C_6 (g) C_7 (h) C_8

Figure 7.9: Percentage Change in Shapley Allocations with Exact DP Formulation

of γ, when using the exact formulation. The inset in the bottom left of each plot shows $\sigma^s - \sigma^{dp}$ for each consumer. This further elucidates the role $\sigma^s - \sigma^{dp}$ plays in determining the changes in allocations, in the presence of DP noise. For example, we see that for C_1, its' $\sigma^s - \sigma^{dp}$ is lowest in the random correlation scenario, followed by the inverse scenario, with the uniform and correlated scenario resulting in similar values. We see this ordering (see inset bars in Figure 7.9) reflected directly in the change in payoffs as γ increases. Effectively we see that the higher $\sigma^s - \sigma^{dp}$ is, the higher the reduction in payoffs.



(a) Correlated (b) Inverse (c) Uniform (d) Random

Figure 7.10: Change in Shapley Allocations with DP

The previous figures showed relative changes in allocations for each consumer. We know look at the absolute changes as a function of γ, as shown in the stacked area plots in Figure 7.10. In general across correlation scenarios we observe a decrease in value as we increase γ, and as a result the DP noise added. We see that value reductions follow the expected DP correlation scenarios. Specifically, value reduction is most prominent for highest value consumers (C_3) in correlated case, with the opposite observed in the inverse case consumers (C_1 & C_2). In the uniform and random case, value reduction is

more evenly spread. We also see that in the inverse case, in particular, as we increase γ past 1 the value of lowest valued consumers becomes negative.

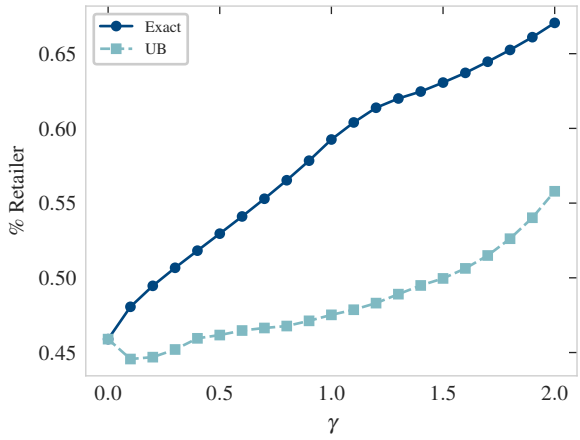


Figure 7.11: Proportion of Value Captured by Retailer

Alongside the overall reduction in value, we also observe that as privacy concerns increase (i.e. as γ increases), the retailer captures more of the value that remains, as shown in Figure 7.11. This is due to the retailers pivotal role in producing positive coalition values, which becomes more pronounced as overall value decreases. Effectively, the contribution from consumers data is diminished due to the increased noise, however the retailers contribution remains fixed. We see that this effect is more prominent for the exact formulation. As mentioned above, this is likely due to the inability to distinguish contributions across consumers and the retailer when there is a large proportion of zero-valued coalitions (which is higher for the upper bound formulation).

7.4.2 Case Study: Procuring Smart Meter Data

The previous case study modelled the effect of schedulable load information on demand forecast uncertainty. As such, it valued the demand forecast, rather than smart meter data. In this case study we apply the integrated forecasting and procurement framework, for the retailers' energy procurement problem, to value real smart meter data directly.

Experimental Setup

We assume a scenario where a retailer has a set of customers, $\mathcal{N}$, each with their own private smart meter data, X_i. As discussed in Chapters 2 to 4, Privacy by Design principles strongly indicate that full data privacy should be the default option for consumers, i.e. any sharing or purchasing of data be done on an opt-in basis. As such, the retailer needs to be actively incentivising its customers to share their smart meter data. The retailer

offers customers an annual contract consisting of a fixed annual retail tariff, λ^r, and an additional annual payment, t_i, for data sharing, i.e. a discount on their energy bill for providing data. To determine the discount for each customer, historical data stored on the smart meter or the training set (X_i^{tr}) is used. A holdout validation set (X_i^{va}) is used to better estimate out-of-sample performance. The true out-of-sample performance is characterised using a test set (X_i^{te}), which immediately follows the validation set. The target demand distribution in this scenario is, $X_T = \frac{1}{N}\sum_{i\in\mathcal{N}} X_i$. The retailer also has access to a publicly available reference dataset, X_R, which may have very different load dynamics to the retailer's actual customer base. In the absence of it's customers smart meter data, it is assumed the retailer will make decisions solely using X_R.

We again use the modified Irish CER smart meter data described in Section 5.1.5. The dataset of HH consumption data for 6010 smart meters spans 75 weeks. We split the data into a training, validation and test set with a 70-10-20 split, (see Figure 7.12). The training set covers a full year from August 2009 to August 2010 to account for seasonal effects. We assume the reference dataset is the GB national demand, as the equivalent Irish data was not available[19]. We re-scale the national demand to ensure its maxima coincides with that of the average smart meter, X_T.

We assume the retailer has 8 customers, which we generate by clustering the smart meters by their average DLCs using KMeans clustering. Each cluster is considered a single customer resulting in more distinct load dynamics in each cluster, than if we had selected 8 smart meters at random. Concurrently, we can leverage load aggregation effects, resulting in more predictable load profiles allowing us to use simpler forecasting models[320]. For this case study, we use a simple ANN, with no exogenous variables, as our primary focus is on relative performance differences rather than producing the optimal forecasting model. We use the ANN defined in (7.4) with the following inputs:

$$\mathbf{x} = [X_{t-h}, X_{t-h-1}, X_{t-2h}, X_{t-2h-1}, X_{t-3h}, X_{t-3h-1}] \tag{7.30}$$

where, X_{t-*} are the lagged/historical load values, h is the number of periods in the day (48).

The ANN forecasting models are implemented in Python using the skforecast package [321], with the ANN defined using the Scikit-Learn wrapper for Keras[322]. The hyper-parameters of the ANN (epochs, learning rate, number of hidden layers, and optimiser)

[19]Other potential references include; available academic smart meter datasets, load profile classes developed by settlement body, or generating synthetic load data through bottom-up demand models e.g. the CREST Demand Model [145]

are tuned using the mean forecaster with the full training dataset using a backtesting-based grid search. The resulting parameters (20 epochs, 0.001 learning, 3 neurons in hidden layer, ADAM optimiser) are then used to produce all the forecasts in the case study. To ensure the ANNs are Lipschitz we include weight clipping constraints which ensure the weights of in each layer do not exceed 1. As a result, the Lipschitz constant of the forecasting model, $K_f = 1$.

First, we explore how the retailer's bidding strategy, q, affects the retailer's profit, Π. We consider the following strategies:

- MSE: The retailer bids the mean day-ahead demand forecast,

$$q = \underset{q(\mathbf{\Phi}, \mathbf{x}_t^{tr})}{\arg\min} \frac{1}{N^{tr}} \sum_{i}^{N^{tr}} \left(y_t - q_t(\mathbf{\Phi}, \mathbf{x}_t^{tr})\right)^2 \tag{7.31}$$

- MAE: The retailer bids the median day-ahead demand forecast

$$q = \underset{q(\mathbf{\Phi}, \mathbf{x}_t^{tr})}{\arg\min} \frac{1}{N^{tr}} \sum_{i}^{N^{tr}} \left|y_t - q_t(\mathbf{\Phi}, \mathbf{x}_t^{tr})\right| \tag{7.32}$$

- NV: The retailer forecasts and bids the critical quantile directly,

$$q = \underset{q(\mathbf{\Phi}, \mathbf{x}_t^{tr})}{\arg\min} \frac{1}{N^{tr}} \sum_{i}^{N^{tr}} \lambda^u \left[y_t - q_t(\mathbf{\Phi}, \mathbf{x}_t^{tr})\right]^+ + \lambda^o \left[q_t(\mathbf{\Phi}, \mathbf{x}_t^{tr}) - y_t\right]^+ \tag{7.33}$$

- SAA: The retailer bids the mean day-ahead demand forecast plus the critical quantile of the empirical forecast error distribution.

$$q = (7.31) + F^{-1}_{\epsilon^{va}}(\tau) \tag{7.34}$$

  where, ϵ^{va}, is the empirical error of the validation set and $F^{-1}_{\epsilon^{va}}(\tau)$ is the inverse CDF of the error distribution.

Next, we also consider how the choice of data valuation metric affects data procurement and compensation. Here, we consider five metrics:

- $\Delta\Pi$: The actual difference in profit.

- ΔMPL: The MPL. As shown in Section 7.2.2, this to be equivalent to the profit difference up to a constant factor.

- $\Delta RMSE$: A standard assumption in extant literature on data valuation is to use the MSE or RMSE. This is then related to monetary losses via the buyer's WTP[37],

[38], [40]. In [38], [40], which focus on purchasing features in regression markets, the WTP is assumed to be a known quantity, estimated by the buyer, a priori. In [37], the £ per % improvement in normalised RMSE is calculated using holdout cross-validation. We use the latter method to calculate the buyer's average WTP using the validation set.

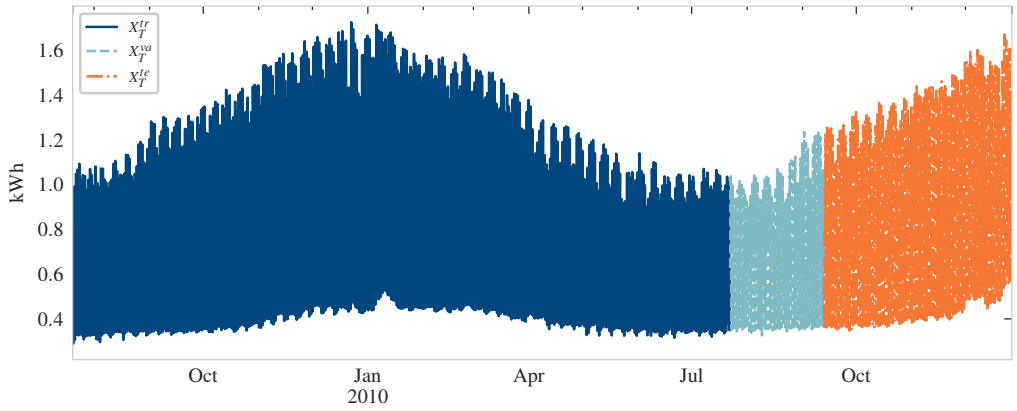
- ΔMAE: The same method as above with the MAE.
- $K \cdot W$: The proposed WD based data valuation mechanism. We also consider the calibrated version, $\hat{K} \cdot W$, again based on reference data, X_R.

For both the bidding strategies and the valuation metrics we compare their Shapley based payoff allocations. As in the case study in Section 5.2.3, we assume the coalition values are defined as follows: $V(X_c) = \max_{c \in C}(L(X_c)) - L(X_c)$ or $V(X_c) = \Pi(X_c) - \min_{c \in C}(\Pi(X_c))$.

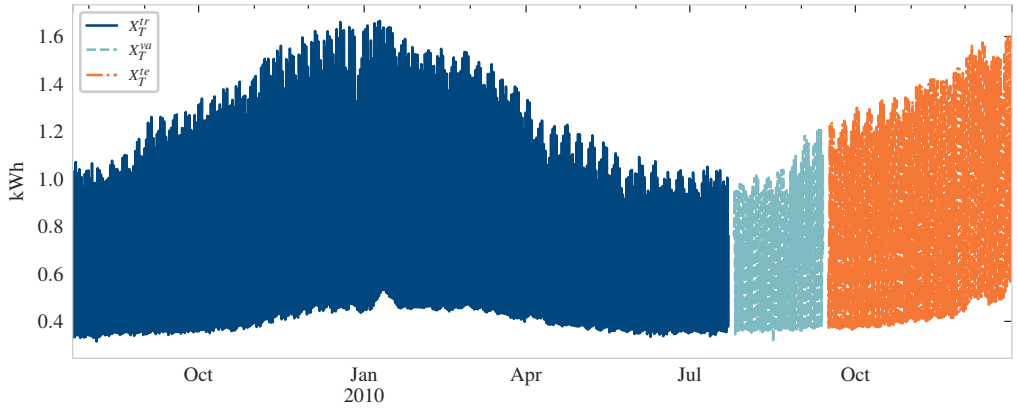
We then examine the impact of model mis-specification by altering the lag structure. The lag structure in (7.30) was selected given the cyclical diurnal dynamics of electricity load. However, the performance of a given forecasting model will vary according to the lag structure, input feature transformations and any exogenous inputs that are included, as well as the data used to develop the features. As such, data procurement based on actual performance for a particular model specification can be manipulated to depress apparent value. The buyer can then use the purchased data and apply an improved model specification, fully capturing any additional improvements. To illustrate this, we also include a mis-specified model with the following lag structure, which does not account for the diurnal dynamics (while maintaining the same number of input features):

$$\mathbf{x} = [X_{t-h+6}, X_{t-h+10}, X_{t-2h+6}, X_{t-2h+10}, X_{t-3h-10}, X_{t-3h-4}] \tag{7.35}$$

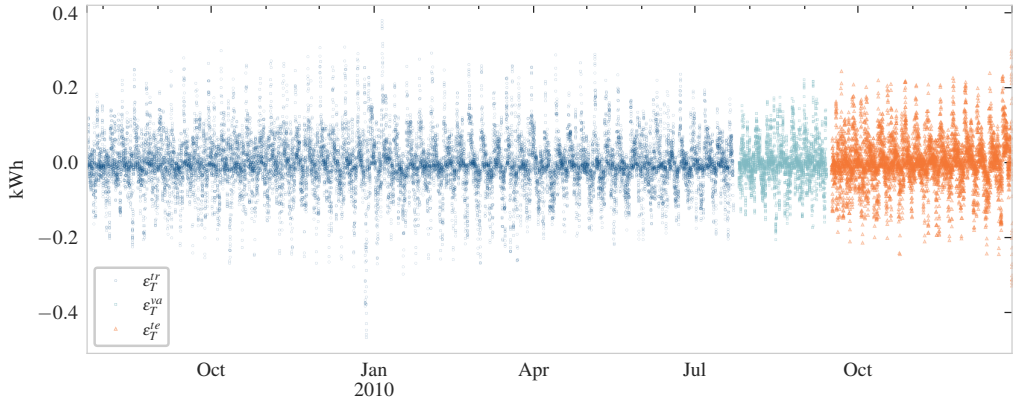
Finally, we apply the joint optimisation mechanism to procure data and determine the payments made, t. The procurement mechanism is formulated at the MISOCP in (6.22) and (6.18) for the finite and infinite formulation, respectively. We initially assume the reference budget is $B(X_R) = \Pi^{va}(X_T) - \Pi^{va}(X_R)$ and the Lipschitz constant is $K = 2 \cdot \max(\lambda^u, \lambda^o)$. We also investigate the effect of changing these two parameters, as well as the confidence level, δ, of the Hoeffding bound. In addition, we generate uniformly distributed reserve prices, θ, this time correlated with the validation set profit differences, $\Delta\Pi^{va}$. The full experimental parameters are outlined in Table 7.2



(a) Actual



(b) Forecast



(c) Residuals

Figure 7.12: Target Load Profile and Forecast

Table 7.2: Experimental Parameters for Data Procurement

Parameter	Value/Range
N	8
Trials	50
Valuation Metric	RMSE, MAE, Π, $K \cdot W$, $K^* \cdot W$
Reserve Prices	$\theta \sim U(0, \bar{\theta})$, where, $\bar{\theta} \in \left[0, 0.1\frac{B(X_R)}{N_d}, \ldots, 4\frac{B(X_R)}{N_d}\right]$
Correlation	$\rho(\Delta\Pi, \theta) \in -1, 0, 1$
Confidence Level	$\delta \in (0, 1)$
K	$K \in (0, 2\max(\lambda^u, \lambda^o)]$
Budget	$B(X_R) \in [0, KW(X_R, X_T)]$

Results

Effect of Bidding Strategy In order to asses the performance of the different bidding strategies considered, we start by evaluating the achieved quantile, $\bar{\tau}$ [20], of the forecast. From (7.3) we know that the optimal bid is the critical quantile, τ. Both MSE and MAE result in underestimation as they aim to forecast the mean and median, respectively, instead of the critical quantile which in this case is 0.769. The achieved quantiles for SAA and the integrated approach, NV, are similar across in-sample (train, validation) and out-of-sample (test) datasets, and are close to τ.

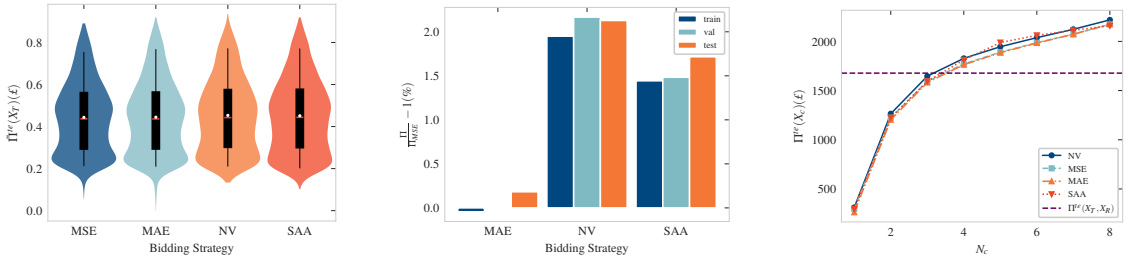
Table 7.3: Error in Achieved Quantile ($\bar{\tau} - \tau$)

	MSE	MAE	NV	SAA
X^{tr}	-0.207	-0.253	0.030	**-0.020**
X^{va}	-0.222	-0.274	-0.025	**0.000**
X^{te}	-0.259	-0.306	**0.010**	-0.031

As shown in Table 7.3, SAA results in the closest quantile for the training and validation set (with the latter result being inherent to the method). However, we see that NV performs better for the test set. This is potentially because SAA generates quantile based on the in-sample (historical) errors alone. On the other hand, NV explicitly captures the effect of inputs on the critical quantile, and, is therefore, able to incorporate heteroscedasticity[310].

[20]The empirical quantile achieved across the dataset $\frac{1}{N_x}\sum \mathbf{1}_{\hat{y} \leq y}$, where, $\mathbf{1}$ is the indicator function, and N_x is the number of samples in the dataset

We note that the differences between SAA and NV on the one hand, and MSE on the other, are driven by market prices and the resulting τ. Although we assume they are fixed for this case study, in practice they also need to be forecast. As a result, one advantage of SAA over NV, is that it only requires training one model, namely the mean forecast. Conversely, NV would have to be trained for each price scenario.



(a) Profit Distribution (b) Profit Improvement (c) Profit by Coalition Size

Figure 7.13: Effect of Bidding Strategy on Profits

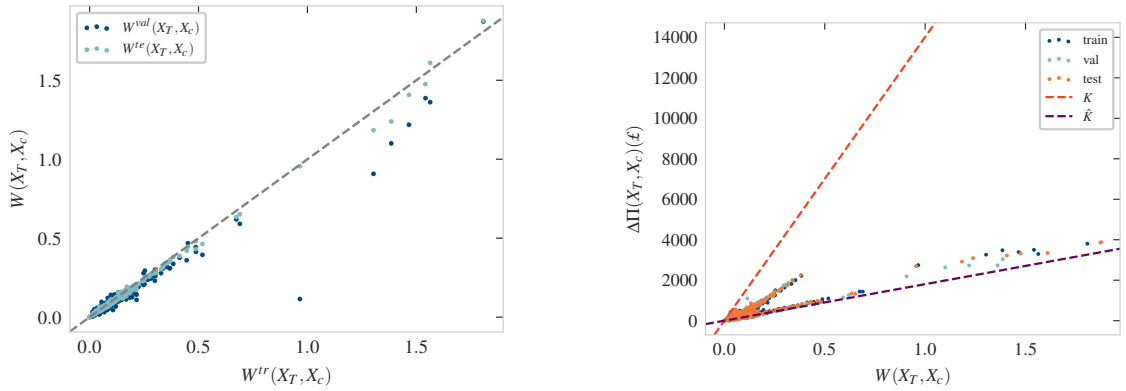
Overall, as shown in Figure 7.13a, the different bidding strategies do not result in large differences in the out-of-sample profit distribution ($\bar{\Pi}^{te}(X_T)$ is the profit per time step). However, we do see differences in the overall profit. The best performance is achieved using NV, with total profits are ~2% higher compared to MSE. SAA also results in increased profits of around ~1.5%, with the results being consistent across datasets, as shown in Figure 7.13b. Interestingly, we see in Figure 7.13c that, on average, SAA performs better than NV for coalition sizes between 5 and 6. We note that both methods are data-driven, and as such, do not require any distributional assumptions on the forecast errors. However, the relative performance of the approaches depends on the underlying forecasting model, in this case an ANN, and the resulting bias-variance trade-offs[310].

Data Valuation Metric In Chapter 5 we observed that correlations between the true value function (the profit difference, $\Delta\Pi$) and the value functions generated using alternative valuation metrics is a good indicator of performance. The correlations with $\Delta\Pi$ across the three dataset splits (X^{tr}, X^{va} and X^{te}), are summarised in Table 7.4. Naturally, $\rho(\Delta\Pi, \Delta MPL) = 1$, as they are equivalent up to a constant factor. We observe that the $\Delta RMSE$, ΔMAE and $K \cdot W$ exhibit similar correlation coefficients between 0.82 and 0.87. This is consistent across the data splits. However, by calibrating the Lipschitz constant we can significantly increase the correlations to between 0.88 and 0.92. This suggests that the WD is better able to capture the value dynamics of our integrated framework than the difference in RMSE or MAE.

Table 7.4: Valuation Metric Correlations

	$\Delta RMSE$	ΔMAE	ΔMPL	$K \cdot W$	$\hat{K} \cdot W$
X^{tr}	0.860	0.869	**1.000**	0.837	0.920
X^{va}	0.826	0.834	**1.000**	0.829	0.880
X^{te}	0.855	0.864	**1.000**	0.840	0.910

The proposed framework assumes the WDs calculated using the in-sample data (training set) are reflective of the out-of-sample distribution (Assumption 7.2). Figure 7.14a shows a scatterplot of the training set WDs of each coalition against the validation and test set distances, We see that the values are broadly consistent, indicating the validity of our assumption. Figure 7.14b plots the WDs against the profit difference with the target distribution. Like with the Gaussian data in the previous studies, we see that the Lipschitz bound is loose when using the global Lipschitz constant, K, due to the asymmetric nature of the newsvendor profit.



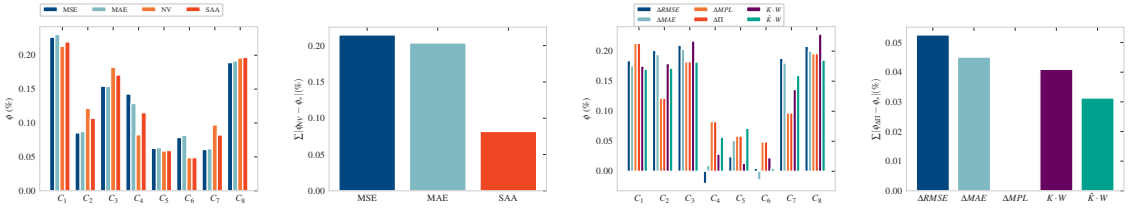
(a) Data Split Correlations

(b) Lipschitz Bound Tightness

Figure 7.14: Wasserstein Approximation

Next, we turn to the Shapley allocations under the different bidding strategies and valuation metrics considered. Figure 7.15a shows the Shapley allocation proportions for the test set. The methods estimating the critical quantile, NV and SAA, result in a slightly different distribution revealing the differences in contribution of consumers to the mean and the critical quantile. For example, C_1 and C_6 are less valuable for estimating the critical quantile than the mean, resulting in lower allocations for SAA and NV. Conversely, the allocations for C_2, C_3, and C_7 are higher for NV and SAA, indicating they are more valuable. Considering that NV provides the highest profits, we evaluate

the mis-allocation proportions of the other bidding strategies relative to it. We see that SAA is closest on average in terms of allocation proportions. Overall, we observe that the bidding strategy affects both the total profit and the profit distribution. Figure 7.15c



(a) Bid Allocation (b) Bid Mis-allocation (c) Metric Allocation (d) Metric Mis-allocation

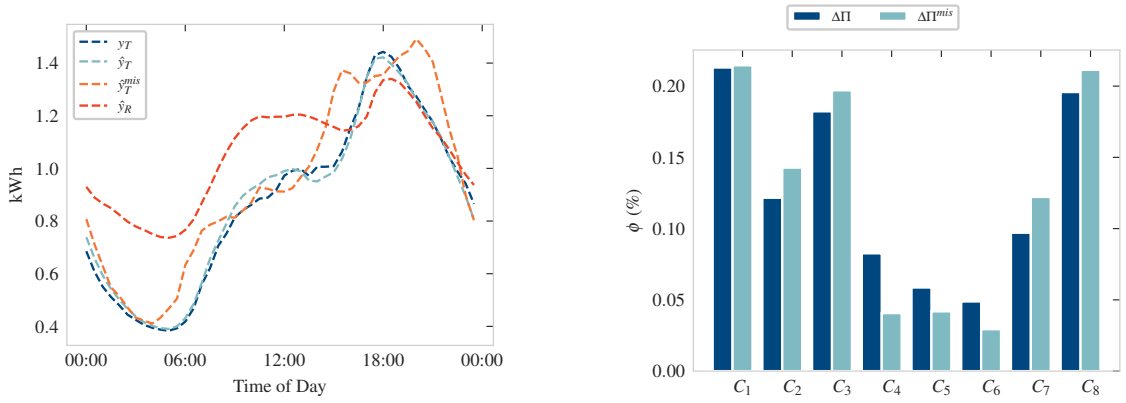
Figure 7.15: Shapley Allocation Performance

and 7.15d show the Shapley allocations and mis-allocations across the valuation metrics considered for the test set. We see that the choice of valuation metric has a much bigger impact on the allocation distribution than the bidding strategies. For example, using $\Delta RMSE$ and ΔMAE, lead to negative allocations and changes the relative rankings of consumers. Wasserstein valuation also causes a change in the allocations, however the average mis-allocation are lower, especially for the calibrated metric, $\hat{K} \cdot W$.

Mis-specification and Reference Data We now investigate the effect of model mis-specification and using reference data, on valuation. Figure 7.16a shows the true load, y_T, and various forecasts produced for 31st October 2009. We see that the model with a principled lag structure, $\hat{y}_T$, is able to capture the diurnal load dynamics correctly. On the other hand, the mis-specified model results in a forecast, $\hat{y}_T^{mis}$, with two afternoon/evening peaks as opposed to one. Similarly, using the reference data (national demand in this case), with the correct model, results in a forecast, $\hat{y}_R$, with different intra-day dynamics and scaling.

Both model mis-specification and incorrect data can lead to significant profit reductions. In this example, using national demand leads to ~25% reduction in profit and the mis-specified model results in a ~5% reduction[21]. In addition, we see that mis-specification also affects allocation percentages, as shown in Figure 7.16b. In a centralised valuation mechanism, such as a cooperative game, we assume the market platform or

[21] Although this may seem minor given that the Wasserstein approach also introduces errors, we stress that we could produce examples of mis-specified models, for example, with fewer features, with sub-optimal hyperparameters, or use less complex (linear) models, which would have a significantly larger impact on profit.



(a) Example Daily Profile

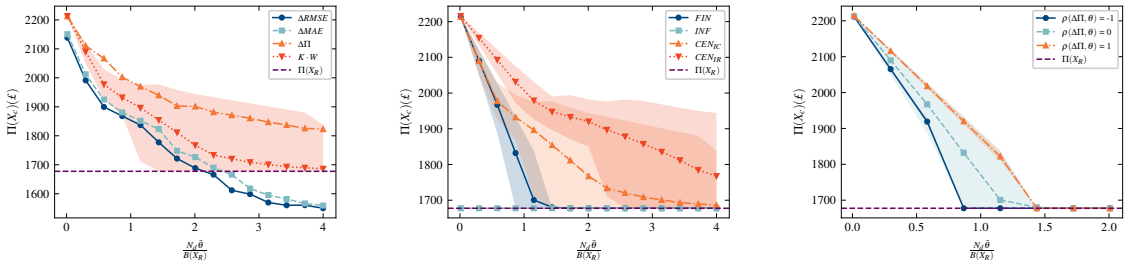
(b) Shapley Allocation

Figure 7.16: Effect of Model Mis-specification and Reference Data

buyer will run the correct model. However, the buyer could purposely provide a sub-optimal model to depress value during the valuation process and then run an improved model after purchasing data. This is especially relevant in our scenario, where there is a single buyer (the retailer), and assumptions of a competitive environment or a trusted intermediary do not apply. This highlights an advantage of our mechanism which relies on differences in inputs (i.e. WD between inputs).

Retailer Profit We now evaluate the procurement mechanism. Figure 7.17a, shows the profit achieved by the retailer using the optimal bidding strategy (NV), assuming they have access to all coalition values, across different valuation metrics. The figure plots the profit as the upper bound of the uniformly distributed reserve prices, $\bar{\theta}$, increases. $\Delta\Pi$ provides a benchmark of the best-case performance if the retailer knew the profit achieved by each coalition. Conversely, $\Pi(X_R)$ provides the worst-case profit the retailer achieves if they used the reference data. We see that the Wasserstein approach, $K \cdot W$, performs similar to $\Delta\Pi$ for lower reserve prices and remains better than the other potential metrics for larger reserve prices, selecting higher valued coalitions. Importantly, the Wasserstein approach maintains budget feasibility, i.e. $\Pi(X_c) \geq \Pi(X_R)$ (when K is the global Lipschitz constant and $B(X_R) = \Pi(X_T) - \Pi(X_R)$), as shown by the shaded region representing the profit range across the 50 reserve price trials. On the other hand, neither $\Delta RMSE$ nor ΔMAE maintain budget balance.

Figure 7.17b shows the performance of our proposed procurement mechanisms, FIN and INF. We see that the finite formulation procures data for $\frac{N_d\,\bar{\theta}}{B(X_R)} \leq 1.5$ whereas the infinite formulation is too conservative in this case. One reason for this conservatism



(a) Centralised Comparison vs. Other Metrics ($\rho(\Delta\Pi, \theta) = 0$),

(b) Proposed vs Centralised ($\rho(\Delta\Pi, \theta) = 0$)

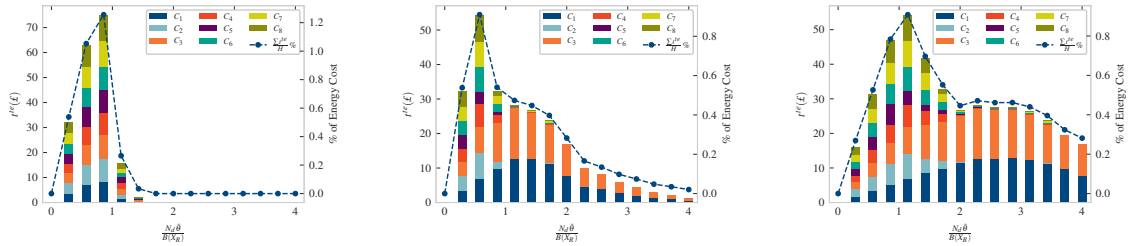
(c) Effect of Correlations

Figure 7.17: Data Procurement Performance

is the Hoeffding bound. Indeed, the centralised mechanisms (CEN_{IC},CEN_{IR}) perform much better offering increased profits across the reserve price range considered. The effect of reserve price-value correlations on the finite mechanism is shown in Figure 7.17c. We observe similar dynamics as in the other case studies with synthetic data in Chapter 6. When reserve prices are positively correlated with the profit difference, $\Delta\Pi$, the mechanism procures more data, resulting in higher profits. A possible reason for the limited difference across correlation scenarios is that the procurement decision is mainly limited by the looseness of the Lipschitz and Hoeffding bounds.

Data Payments Next, we turn to the data payments made to consumers. Figure 7.18, shows the average annualised data payments made to the eight consumers for the FIN, CEN_{IC} and CEN_{IR} procurement mechanisms in the uncorrelated scenario. We omit INF, as it does not procure any data under the base experiment parameters (see Figure 7.17b). We see that as consumers' reserve prices increase, their payments initially increase before reducing. This is because in the initial phase the budget constraints are on average not affecting the procurement decisions. As a result, the retailer can continue to procure data even as the compensation demanded by consumers increases. However, at a certain point the budget constraints start to limit the feasible coalitions that can be purchased. This is particularly visible for CEN_{IC} and CEN_{IR}, where we see that for higher reserve prices the mechanism chooses smaller coalitions (in particular, C_3 and C_1). We note that although the actual payments, for a given trial, increase as reserve prices increase the average payments decrease due to the increasing proportion of trials where no or less data is procured.

The annual average total data payments range between £2 and £75 for the FIN, which represent up to 1.25% of the total energy costs (H). This is similar to the proportions



(a) FIN (b) CEN_{IC} (c) CEN_{IR}

Figure 7.18: Expected Annual Data Payments to Consumers ($\rho = 0$)

observed in other similar studies [39] but is on the lower end of the willingness to accept analysis estimated in Chapter 3. However, as shown in Table 7.5, individual proportions can be much higher. For example, C_5 and C_7 receive a 16.24% and a 5.79% discount, respectively, for sharing data under particular runs of FIN.

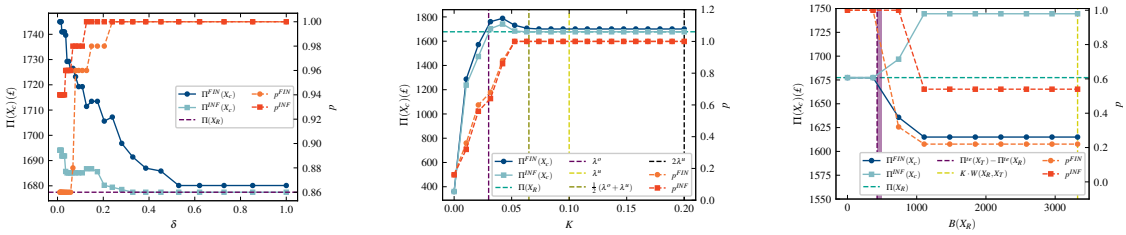
Table 7.5: Maximum Data Payments for each Consumer using FIN (% of Energy Cost)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
FIN	2.83	3.76	3.49	3.41	**16.24**	3.44	5.79	2.16
CEN_{IC}	4.23	2.58	4.22	2.57	**6.92**	2.94	4.62	1.63
CEN_{IR}	4.28	2.68	4.13	2.55	**7.28**	3.41	4.58	1.62

In this case study we have assumed a single annual contract between the retailer and consumers. However, in reality, data value evolves over time and also depends on changes in the reference used. For example, once the retailer purchases data from consumers it can update its reference distribution. Over time this accumulation of data from customers will degrade data value. This could be addressed by, for example, placing time or usage limits on data, forcing the retailer to purchase data annually. This would also address issues around data hoarding highlighted in [7]. In addition, if load dynamics evolve over time, data value may be preserved, as these changes will results in a new target distribution. See for example the online regression market studies explored in [38], where parameters evolve over time.

Risk and Calibration We now investigate the effect of risk adjustment and calibration on the retailer's profit. First, we vary the confidence level, δ, of the Hoeffding bound. As shown in Figure 7.19a, reducing δ, results in increased profits in the uncorrelated scenario when $\delta < 0.5$ for both FIN and INF. However, this comes with the increased probability

of budget infeasibility, as discussed in Chapter 6. The observed probability of budget feasibility, p, is plotted on the right-hand axis. For example, we see that by reducing δ to 0.06, we see an average profit increase of 2.73% but also a reduction in p to 0.86. We note, however, that there is a significant range ($\delta = 1$ to 0.24) for which budget feasibility is not affected, suggesting our underlying valuation is conservative.



(a) δ ($\frac{N_d\ \bar{\theta}}{B(X_R)} = 1.44$, $\rho = 0$) (b) K ($\frac{N_d\ \bar{\theta}}{B(X_R)} = 1.15$, $\rho = 0$) (c) $B(X_R)$ ($\frac{N_d\ \bar{\theta}}{B(X_R)} = 1.44$, $\rho = 0$)

Figure 7.19: Effect of Calibration and Risk Adjustment on Retailer Profits

Although the Hoeffding bound adjustment allows the retailer to explore the trade-off between risk and profit, the main driver of the conservatism of our mechanism is the looseness of the Lipschitz bound. As discussed in Section 7.3.3, we can relax the Lipschitz constant, K, to improve our approximation. In Figure 7.19b we vary K from 0 to $2\lambda^u$, the global Lipschitz constant. We see that if K is too small, the valuation mechanism over-estimates the potential improvement, $B(X_R) - \Pi(X_T) - \Pi(X_c)$, and purchases too much data. As a result, for both FIN and INF, the resulting profits, $\Pi(X_c)$, are lower than the reference profit, $\Pi(X_R)$. However, if K is between λ^o and $\frac{1}{2}(\lambda^o + \lambda^u)$, we see an improvement in the resulting profit[22]. We again see that this comes with risk of budget in feasibility, with p between 0.6 and 1. As K is increased beyond $\frac{1}{2}(\lambda^o + \lambda^u)$, mechanism no longer procures any data for INF and reverts to the subset procured when using the global Lipschitz constant under FIN. We see that budget feasibility is also maintained ($p = 1$) for $K > 0.05$.

Finally, we look at the effect of the reference budget $B(X_R)$. Figure 7.19c shows the change in profit for the INF and FIN mechanisms as we increase the budget over the range specified in (7.18). In this example, for the base case ($B(X_R) = \Pi(X_T) - \Pi(X_R)$), the mechanism does not procure any data as the estimated improvement minus payments is less than the budget. We note that the in-sample budget values ($\Pi^{tr}(X_T) - \Pi^{tr}(X_R)$ or $\Pi^{va}(X_T) - \Pi^{va}(X_R)$) are close to the out-of-sample values as shown by the purple shaded

[22]The range in which we observe improved profits depends on the particular values, and, as such, is only illustrative.

area in the figure. This is in line with Assumption 7.2. While $B(X_R) \leq \Pi^{te}(X_T) - \Pi^{te}(X_R)$, budget feasibility is maintained. Increasing $B(X_R)$ beyond this can lead to either an increase in profits as is the case with INF or a decrease in overall profits in the case of FIN. When the base case procurement decision was overly conservative (INF), increasing the budget reference allows more data, on average, to be purchased while potentially maintaining budget feasibility. We see a 4% improvement in profit but this comes at the cost of p reducing to 0.54. Conversely, when using a less conservative valuation approach (FIN), increasing the budget leads to significant over-procurement, on average leading to reduced profits and reduced p.

7.5 Discussion

This chapter applied the proposed data valuation and procurement mechanism to the retailers' day-ahead energy procurement problem. We developed the notion of a joint energy and data market, by outlining an integrated forecasting and optimisation model. This differentiates our work from existing data market mechanisms that have investigated this problem, by establishing a direct connection between raw smart meter data and retailer profits, rather than valuing forecasts. We use this framework to determine our mechanism parameters. Specifically, we determined the Lipschitz constant assuming a constrained ANN-based forecasting model, the benchmark budget as a function of the retailer's profit using a reference dataset, and the appropriate data distributions and resulting WDs to use.

In Chapter 6, we identified that quantile estimation and the associated MPL function result in a loose Lipschitz bound. This affects the performance of the mechanism and can lead to overly conservative data procurement decisions, specifically, under-estimation of potential improvement offered by data. We showed that the energy procurement problem is equivalent to a linear newsvendor problem and therefore to a quantile regression. As such, it suffers from the same conservatism. We explore the underlying reasons for the conservatism, namely the asymmetry introduced by the underage and overage costs. We explore ways to deal with this by adapting the Lipschitz constant used. First, we show how a transfer function, i.e. a linear relationship, can be used to estimate the average Lipschitz constant, using either reference or synthetic data. Next, we describe a more principled method of altering the Lipschitz constant by introducing the notion of local Lipschitzness. This allows us to bound the Lipschitz constant of a loss function within a given input range. We show, with the example of the Gaussian newsvendor problem, how

this notion together with additional assumptions on distributional spread and a bound on the difference in bidding quantities, we can produce tighter Lipschitz bounds. An additional method to deal with the conservatism of the Lipschitz constant is motivated by the results from Chapter 5 in Figure 5.11. Specifically, we note that the average Lipschitz constant is significantly lower for the MPL, suggests a probabilistic approach could improve average performance, consistent with the ex-interim budget feasibility constraints of our mechanisms.

We then present two case studies to evaluate the performance of our proposed joint energy and data market framework. The first focuses on procuring forecasts, which allows us to compare our Wasserstein based valuation mechanism against existing techniques. The second implements our full framework, including the procurement mechanism to value smart meter data directly.

In the forecast procurement case study we apply the modelling framework proposed in [39]. We show that, with appropriate calibration of the Lipschitz constant, our method performs well in terms of correlations, and Shapely mis-allocations compared to using the true profit. We then extend the framework to incorporate DP. We explore both the closed forms and upper bounds on the WD developed in Section 5.3.2. We show that data value is linked to the difference between forecast uncertainty reduction in terms of the schedulable load and the noise introduced by DP and depends on privacy-value correlation among consumers. Of particular note is that, as privacy concerns increase, not only does data value decrease, but the retailer captures an increasing proportion of the remaining value. This is especially prominent in the exact formulation.

The second case study implements the full integrated forecasting, optimisation and procurement mechanism using real smart meter data. Although the main motivation for outlining the integrated approach was to establish a connection between raw smart meter data and the retailers' profit, we show that it results in higher profits than using the mean forecast or SAA. The case studies in Chapter 6 and the forecast procurement study assumed Gaussian data. We therefore again evaluated the performance of our data valuation metric with real smart meter data. We find that without calibration, our method is similar to other proposed metrics (e.g. RMSE) in terms of correlations w.r.t. the true profit. However, Shapley mis-allocations are less using our metric, and improve further with calibration of Lipschitz constant.

We show how model mis-specification can cause value depression. In a cooperative game framework, with an un-trusted platform or buyer this can be used to manipulate the

market. Aside from its ability to capture value and private computation properties, the choice of the WD also means the metric is based on inputs alone. As a result, it is immune to model mis-specification by the buyer/platform. Our mechanism still requires buyer to provide certain inputs, namely, the Lipschitz constant and reference budget, however, over or under estimation of these quantities will result either in increased costs or risk of budget infeasibility.

We then explored the mechanism's distribution of profits, both in terms of the retailer's profits and consumers' data payments. We found that data payments were, on average at the lower end of WTA values estimated in Chapter 3. However, there were instances where much higher payments were made, in line with the WTA estimate. Finally, we explore the effect of the conservatism of our approach. The implications of risk adjustment, through the Hoeffding bound confidence level, the Lipschitz constant and the reference budget were explored. The implications and trade-offs, in terms increased average profits and probability of budget infeasibility, were discussed. We note that although we consider a risk neutral (in terms of energy procurement) retailer our mechanism could be extended to incorporate risk aversion through the CVaR[23]. The next chapter concludes the work and discusses broader future research directions.

[23] A closed-form solution for the risk averse newsvendor problem can be found in [323]

CHAPTER 8
Conclusions and Future Research Directions

This thesis tackles the dilemma of balancing access to smart meter data and protecting privacy in the context of the UK's Smart Meter Implementation Programme. To this end, we proposed the development of a privacy-preserving market mechanism to procure smart meter data. We identified four key challenges in developing such a mechanism, namely:

1. The ability to obtain truly informed consent is hindered by the complexity of potential privacy risks and the wide range of potential uses for smart meter data.

2. Given the variety of PPTs, identifying the most suitable PPT for smart meter data and its associated costs and benefits.

3. The valuations of smart meter data need to be assessed, both from the perspective of data users (e.g. retailers) and data owners (e.g. consumers).

4. A market design approach with the desired properties needs to be identified and tested for its viability.

In order to address these challenges, we employed an interdisciplinary approach consisting of two distinct, but interdependent research directions:

- **Determining the design criteria of a data market for smart meter data.** This involved developing a mapping of the dependence of benefits as well as potential privacy risks on smart meter data characteristics, investigating consumers' privacy concerns and valuations of smart meter data, and establishing the suitability of different PPTs for smart meter data.

- **Developing a novel data market framework.** This consists of a data valuation mechanism, which embodies the drivers of smart meter data value, including the effect of PPTs, and a procurement mechanism which incorporates the buyer and seller preferences identified in the first.

Pursuant to these research paths this thesis makes the following contributions:

- A critical assessment of the existing SMIP against Privacy by Design principles, and a mapping of the data dependence of the benefits as well as the privacy risks of sharing smart meter data, providing a basis for obtaining informed consent.
- A comprehensive review of suitability of different privacy-preserving techniques for smart meter data, with specific focus on the requirements of consumers and the existing SMIP infrastructure, adding to the evidence base for policymakers.
- A novel survey and discrete choice experiment of a sample of nationally-representative GB bill payers (n=686) quantifying their WTA and WTP for smart meter data anonymisation and investigating the effect of information asymmetries by means of a randomised control trial. This broadened empirical evidence on data valuations, substantiated and quantified behavioural economics effects, specifically, the endowment effect and bounded rationality, in the context of smart meter data privacy valuations.
- A novel data valuation mechanism and framework based on the WD. This included a method to endogenously incorporate the effect of DP on data utility, a theoretically grounded approach to ensuring performance guarantees in the output space using Lipschitz bounds and a linear-time approximation scheme for calculating the combinatorial value using the Hoeffding bound.
- A novel set of procurement mechanisms based on incentive mechanism design theory to accommodate task-agnostic budget feasible procurement, task-specific value maximising procurement and task-specific profit maximising procurement. The latter two are able to capture decision-dependent structure. The task-specific profit maximising procurement mechanism was used to develop a joint energy and data market.

In this chapter we highlight the relevance of the key findings and contributions of this thesis. We start by briefly summarising the findings of each chapter in Section 8.1. We then synthesise their connections through the four challenges addressed in the thesis in

Sections 8.2.1 to 8.2.4. We then move on to make a number of policy recommendations based on the research findings in Section 8.3. Finally, in Section 8.4 we suggest potential avenues for future research.

8.1 Chapter Summaries and Key Results

8.1.1 Smart Meter Data Privacy in the UK

Chapter 2 established a mapping between the benefits as well as the privacy risks associated with smart meter data and the form in which the smart meter data is shared. First, the existing data sharing framework, the DAPF, under the SMIP was assessed against Privacy by Design principles. We found it to be heavily reliant on consumer consent and permissions controls. The introduction of CADs and the move towards an opt-out model as part of the MHHS create a complicated landscape, without strong privacy defaults or preventative measures, for consumers to navigate.

Next, the dependence of different use cases for smart meter data on the data's temporal and spatial resolution were mapped. We found that, of the benefits (e.g. automated readings, energy savings) that have been quantified under BEIS's cost-benefit analyses of the SMIP, many (£7.58 bln or 39%) are independent of sharing high temporal resolution (e.g. half-hourly) smart meter data. However, a majority (£11.3 bln or 58%) of the benefits would be enhanced if such data were available to energy sector actors. Similarly, most benefits can be realised with spatially aggregated smart meter data and do not require access to individual level data. Although, access to individual level data could enhance these benefits. We also found that many innovative uses for smart meter data lie beyond the energy sector, for example, in healthcare for monitoring assisted living facilities. As such, wider access would allow the full potential and value of smart meter data to be realised but will also require an increasing number of entities, who may or may not have direct contractual relationships with consumers, to access the data.

Finally, the potential privacy risks and infringements associated with sharing smart meter data were mapped. Smart meter data has a vast amount of personal information embedded within it. This includes energy usage related information (e.g.the use of certain appliances), derived information (e.g. daily routines and home occupancy), as well as socio-demographic information (e.g. age and income). The ability to extract such information and the accuracy with which it can be done depends on the technical tools used, but notably, also on the form of the smart meter data. Specifically, whether

the temporal resolution of the smart meter data and whether data is aggregated have a significant impact on inferrable information. Smaller appliances and activities can only be identified at sub-HH resolutions (with up to 83% accuracy with sub-1 second data), whereas larger appliances and demographic information can be extracted even at daily or monthly resolutions (e.g. 82% accuracy for heating/cooling appliances with hourly data). Similarly, for aggregated high-resolution data larger appliances are still identifiable (e.g. 74% accuracy for EVs). In addition, given the quickly evolving nature of machine learning and the increasing availability of linked datasets it is difficult to properly define and communicate the potential implications of data sharing. Overall, we see that obtaining informed consent in the case of high-resolution smart meter data, and specifically for uses outside of regulated activities outlined in the DAPF, is difficult, as the extent of potential privacy infringements are not known.

8.1.2 Consumer Privacy Concerns and Valuations

Chapter 3 detailed the novel survey of a nationally representative sample of GB bill payers (n = 686). It provided the first estimate of consumers' WTP/A for privacy and anonymisation in the context of smart meter data. It also investigated the effect of information asymmetries, in terms of the implications of data sharing through an embedded randomised control trial. The sample was split between a control group (n = 337), who received educational material on their data sharing options (the data sharing frequency and anonymisation), and the treatment group (n = 349), who were also told specifically what personal information would be shared under each option. Respondents' willingness-to-pay/accept for anonymisation and the different data sharing frequencies were estimated through a discrete choice experiment and mixed logit models. In addition to quantifying the monetary aspects of privacy, the survey also considered respondents' willingness-to-share their smart meter data.

We found that the general willingness-to-share half-hourly smart meter data was high at 62% across both the control and treatment group. In addition, 41% were more likely to share their half-hourly data if it were anonymised after receiving the information on their data sharing options, suggesting that there is demand for protecting data. This also translated into respondents' WTP/A. For the complete sample, on average, respondents were willing to pay up to 8.80% of their monthly electricity bill to share daily data instead of real-time data and would require a discount of at least 15.93% to share their real-time data instead of daily data. No significant difference was observed between half-hourly

and real-time data, likely due to the similar type of information that could be inferred from them. We found that if data was anonymised, respondents no longer differentiated between different data sharing frequencies. The WTP and WTA for anonymised data, as opposed to non-anonymised data, ranged from 4.63% to 7.93% and 8.53% to 14.50%, respectively.

The difference between the WTP and WTA is significant, which implications from a policy perspective. Given that Privacy by Design advocates strong privacy defaults the WTA would be the more appropriate figure. On average we observed a WTA/WTP ratio of 1.83 confirming the presence of endowment effects. Estimated market shares based on the underlying mixed logit model for different pricing scenarios showed that framing effects and the option to anonymise have a significant impact of the availability of high-resolution data. For example, a 20% discount would incentivise up to 80% of consumers to provide either half-hourly or real-time data instead of daily data, with this proportion increasing to 97% if data were anonymised. If instead consumers had to pay a fee to avoid sharing real-time data, a 20% fee would result in 99.9% of consumers sharing high-resolution data.

We observed significant heterogeneity among consumers with many respondents having no privacy concerns and therefore no value for anonymisation or lower resolutions and others with very high privacy concerns and high valuations. The drivers of this heterogeneity were explored using a multinomial logit model with interactions. The following covariates were considered based on previous studies: socio-demographics (age,gender and socio-economic grouping), general attitudes to data sharing (whether respondents share data with third parties), supply characteristics (smart meter ownership, tariff type and IHD engagement), and their level of understanding of the educational material (self-reported feedback and manipulation checks). Older respondents, women, those who did not share data with third parties, non-smart meter owners, those on time-varying tariffs, those who engaged with their IHD more than once a week and those with a higher level of understanding of the educational material were found to have significantly higher WTP/A for anonymisation compared to the corresponding reference groups.

The presence of information asymmetries was explored through the WTP/A analysis, respondents WTS and their knowledge of their existing options where applicable. We found that over 30% of smart meter owners did not know their current data sharing options. In addition, 24% of smart meter owners did not or did not know if they owned an IHD. This is indicative of a high level of apathy in line with general attitudes observed

with regards to energy related issues. The effect of the treatment, additional information on the personal information embedded in smart meter data for each sharing data option, the above-mentioned multinomial logit model with interactions was used. The treatment resulted in 2.45% increase in WTA compared to those in the control group and was significantly higher at 5.41% among those who did not regularly share data with third parties.

Although we observed significant value and thus demand for anonymisation, a subset of the sample was convinced by the option of anonymisation. Of those who were initially unwilling to share their half-hourly smart meter data only 24% in the control group were more likely to share if anonymised. This was even more pronounced in the treatment group with only 6% being attracted by anonymisation. Although this is a small proportion of the overall sample (14%), it highlights the need to consider other factors such as the role of trust, a sentiment expressed through respondents open-ended responses, in conjunction with technical solutions such as data anonymisation.

8.1.3 Privacy-Preserving Techniques

Chapter 4 reviewed several PPTs focusing on their suitability for smart meter data and the SMIP. Many PPTs have been proposed for smart meter data ranging from general techniques, such as DP, to the domain specific user demand shaping. PPTs vary in their definitions of privacy (e.g. anonymity, limiting inference), which may be continuous (e.g. the privacy budget in DP) or discrete (e.g. homomorphic encryption), and hence so do the privacy infringements (e.g. membership inference, non-intrusive load monitoring, linking attacks or data breaches) they protect against. Therefore, we developed a set of privacy guarantees and desirable properties a suitable PPT should possess. We found that DP is a prime candidate to achieve this given that it can be implemented either centrally or in a distributed manner, provides provable, tuneable and future proof guarantees of privacy protection and has been implemented in other sectors. It has already been widely implemented in practice by both private entities such as Apple and Google and public entities such as the US Census Bureau.

Although it only allows for access to aggregate data, many of the applications for smart meter data can be performed using aggregated data with marginal improvements provided by access to individual level data. By way of example, we use the US Census's recent incorporation of DP for the 2020 census to explore the challenges of implementing DP and lessons for the SMIP. Overall, we see that smart meter data under the existing

SMIP has similarly conceivable disclosure risks to those faced by the US Census data. Based on extant academic literature aggregate load statistics could be used to reconstruct individual load profiles with up to 90% accuracy. Socio-demographic data could then be extracted from these individual load profiles and matched to external databases, which could result in up to 99% of individuals being identified. The US Census Bureau faced significant challenges in deciding the appropriate privacy-utility trade-off, specifically, how to determine the privacy budget, ϵ, and how to split it across the multiple statistics they release as part of the census. This would also be an issue with regards to smart meter data and would require the development of a data catalogue which maps out different levels of aggregation and types of groupings that one would want to publish. From a technical perspective differentially-private noise can be incorporated at the meter given the randomisation and two-way communications capabilities defined under the SMETS2 specifications. In addition, we identified two unique challenges for smart meter data: who should implement a differentially-private smart metering system and how to account for the dynamic nature of smart meter data.

8.1.4 Valuation of Differentially-Private Data

Chapter 5 proposed a data valuation framework based on the WD with a focus on applications for differentially private smart meter data. First, we identified the drivers of value for differentially private data through a forecasting and energy procurement framework. We showed that value is highly contextual, depending on the availability of reference data, risk appetites and the level and distribution of privacy concerns as well as energy market conditions.

We then presented the WD, a statistical distance, as a theoretically grounded data valuation metric. As a measure of difference between data distributions in the input space, it is a task agnostic valuation metric. As such, it induces the inevitable trade-off between accuracy for a particular task and applicability, as a measure of value, across multiple tasks. We provided justification for choosing the WD over other potential statistical distances based on its statistical properties (being a metric, non-saturating, defined on disjoint supports, and defined by CDFs). We also showed, through simulations with synthetic data, that the WD performs better, in terms of correlations (between 0.7 and 1.0) and Shapley allocations, than the other distances considered across a number of tasks (e.g. mean, quantile estimation) and distributions (Gaussian, uniform, exponential).

Conventional data valuation mechanisms use performance improvement for a partic-

ular task as a indication of value. We show that the WD can also be used in this context as it provides a natural upper bound for tasks with Lipschitz loss functions. We then also tackled the combinatorial nature of data value by presenting a linear approximation scheme for the WD using the Hoeffding bound. We showed that these approximations do degrade performance. However, we saw that Shapley allocations for a range of tasks are broadly consistent using these approximations versus using the actual task performance.

Finally, we provided novel closed-forms for the WD for location-scale distributions, which are used to endogenously model DP in the WD. We also showed how the WD can be computed in a privacy-preserving manner, using existing MPC techniques, both for location-scale distributions defined by their distributional parameters and for empirical data.

8.1.5 Market for Differentially-Private Data

Chapter 6 proposed three data procurement mechanisms, based on Bayesian incentive mechanism design, which build upon the proposed WD based data valuation mechanism. The first was an exogenous budget mechanism which allows for task-agnostic procurement while capturing both the non-I.I.D. nature of data and the effect of privacy preferences. The other mechanisms, namely the endogenous budget and joint optimisation mechanisms, allow for task-specific procurement and capture the decision-dependent nature of data procurement in this scenario. We reformulated these mechanisms as MISOCP, which can be solved using commercial solvers. Although these are still computationally intensive, the use of the WD decouples the complexity from the specific task. It also ensures that the data valuation process is immune to model mis-specification. This contrasts with the existing state-of-the-art cooperative game approach.

We perform extensive case studies using synthetic data to highlight mechanism performance and dynamics. We show that our proposed exogenous budget mechanism results in more stable performance than existing budget feasible mechanisms, with the finite formulation outperforming existing mechanism when the valuation metric and reserve prices are uncorrelated or positively correlated. In addition, we showed that by using the WD to capture both the effect of DP and the non-I.I.D. nature of data, our proposed valuation metric performs better than considering only DP or the non-I.I.D. setting. The case studies showed that the performance of the endogenous budget and joint optimisation mechanism is heavily influenced by the tightness of the Lipschitz bound. We showed the dynamics of the mechanisms and explored the effect of risk adjustment on procurement

performance and budget feasibility. The level of conservatism and the bias introduced by the Hoeffding bound affect performance differently depending again on the tightness of the Lipschitz bound and the correlations between valuation metrics and reserve prices.

8.1.6 A Joint Energy and Data Market

Chapter 7 applied the proposed data valuation and procurement mechanism to the retailers' day-ahead energy procurement problem. We developed the notion of a joint energy and data market by outlining an integrated forecasting and optimisation model. This differentiates our work from existing data market mechanisms that have investigated this problem, by establishing a direct connection between raw smart meter data and retailer profits, rather than valuing forecasts.

The tightness of the Lipschitz bound affects the performance of the mechanism and can lead to overly conservative data procurement decisions, specifically, under-estimation of potential improvement offered by data. We showed that the energy procurement problem is equivalent to a linear newsvendor problem and therefore to a quantile regression. As such, it suffers from the same conservatism. We explored the underlying reasons for the conservatism, namely the asymmetry introduced by the underage and overage costs. We explored ways to deal with this by adapting the Lipschitz constant used, including the use of transfer functions estimated from data as well as a more principled method of altering the Lipschitz constant by introducing the notion of local Lipschitzness. This allowed us to bound the Lipschitz constant of a loss function within a given input range.

We then presented two case studies to evaluate the performance of our proposed joint energy and data market framework. The first focused on procuring forecasts of schedulable load, which allowed us to compare our Wasserstein based valuation mechanism against existing techniques. The second implemented our full framework, including the procurement mechanism to value smart meter data directly. In the forecast procurement case study (proposed in [39]) we showed that with appropriate calibration of the Lipschitz constant our method performs well in terms of correlations (0.83 to 0.99), and Shapely mis-allocations compared to using the true profit. We then extended the framework to incorporate DP and show that in this case data value is linked to the difference between forecast uncertainty reduction in terms of the schedulable load and the noise introduced by DP and depends on privacy-value correlation among consumers. Of note is that, as privacy concerns increase, not only does data value decrease, but the retailer captures an increasing proportion of the remaining data value under a Shapley allocation policy.

The second case study implemented the full integrated forecasting, optimisation and procurement mechanism using real smart meter data. We found that without calibration, our method is similar to other proposed metrics (e.g. RMSE) in terms of correlations w.r.t. the true profit. However, Shapley mis-allocations are less using our metric, and improve further with calibration of Lipschitz constant. We then explored the mechanism's distribution of profits, both in terms of the retailers' profits and consumers data payments. We found that data payments were, on average, at the lower end of WTA values estimated in Chapter 3 (up to ~1.2%), there were instances where much higher payments (up to 16.2%) were made. Finally, we explored the effect of the conservatism of our approach. The implications of risk adjustment, through the Hoeffding bound confidence level, the Lipschitz constant and the reference budget were explored and the implications and trade-offs in terms increased average profits and probability of budget infeasibility were discussed.

8.2 Synthesis: Balancing Privacy and Access to Smart Meter Data

In this section we synthesise our findings and conclusions in relation to the four challenges posed in this thesis.

8.2.1 The Benefits and Privacy Risks of Smart Meter Data

The first challenge we identified concerns the ability to obtain truly informed consent, given the complexity of potential privacy risks and the wide range of potential uses/abuses for smart meter data.

In Chapter 2, we conducted a comprehensive review of the existing data sharing framework under the SMIP, the contemporary as well as potential future benefits of sharing smart meter data, and the state-of-the-art literature on information extraction techniques for smart meter data. We find that the use cases and resulting benefits of smart meter data also vary depending on data resolution. Many of the benefits of smart metering are independent of sharing high-resolution smart meter data. However, a large majority of the benefits would be enhanced if such data were accessible to energy sector actors. Significantly, many of the operational benefits of smart meter data sharing can be achieved with aggregate data. We also find that many innovative uses for smart meter data lie beyond the operational use cases e.g. evaluating policy interventions or monitoring assisted living facilities.

Smart meter data has a vast amount of personal information embedded within it,

including demographic, financial and medical information. The ability to extract such information and the accuracy with which it can be done is driven largely by the spatial and temporal resolution of the data. Specifically, highly aggregated data (> 100 households) and low temporal resolution (> Daily) limit the inferences and embedded information that can extracted. However, as detailed in Section 4.2.2, the release of low temporal resolution data together with high temporal resolution aggregated data leaves it vulnerable to database reconstruction.

Extant surveys, as outlined in Chapter 3, suggest consumers consider smart meter to be less sensitive than other forms of personal data. Our mapping of the personal information and data dependencies run counter to these studies. This incongruity highlights the presence of information asymmetry in consumers' current decision making around data sharing choices. In addition, we found a significant proportion of consumers with smart meters are unaware of their current data sharing options. Our survey, and particularly the RCT element, further suggests that existing privacy concerns are depressed due to information asymmetries, manifesting in lower WTP/A.

Although we provide a detailed mapping of the data dependence of privacy risks, the quickly evolving nature of machine learning and the increasing availability of linked datasets, make it difficult to convey the full extent of potential future privacy risks. As such, obtaining informed consent remains challenging, motivating the need for additional forms of privacy protection.

8.2.2 An Assessment of the Suitability of PPTs

This led us to the second challenge addressed in this thesis which was the identification of the most suitable PPT for smart meter data. We developed a set of privacy guarantees, and desirable properties a suitable PPT should possess. These included technical considerations such as the ability to provide anonymity and preserving data utility as well as incorporating the implications our survey results from Chapter 3. Specifically, accommodating preference heterogeneity and removing the need for a trusted third party. Chapter 4 critically assessed a number of techniques: data obfuscation techniques like pseudonymisation, aggregation and DP, cryptographic techniques like homomorphic encryption and multi-party computation, and the domain specific user demand shaping, against these criteria.

Overall, we found that DP, is a prime candidate given that it can be implemented either centrally or in a distributed manner, provides provable, tuneable and future proof

guarantees of privacy protection and has been implemented in other sectors. To further illustrate this, in Section 4.4, we detailed the implementation of DP for the US Census, highlighting the challenges of determining the appropriate privacy-utility trade-off. We then proceeded to concretise the drivers of value for differentially private smart meter data in Section 5.1, through addressing the energy retailer procurement problem. This led to a novel framework to quantify the privacy-utility trade-off induced by DP. We then showed that privacy protected data can still provide significant value, while noting that the amount of value is heavily influenced by contextual factors.

8.2.3 The Value of Smart Meter Data

The third challenge we tackled is the investigation of valuations of smart meter data. We considered this question from the perspectives of both data users and consumers.

Conventional data valuation mechanisms, for data users, use performance improvement for a particular task as an indication of value. Indeed, this provides an accurate assessment of value in that particular setting. However, given the multiple uses for smart meter data, detailed in Section 2.3, an intrinsic assessment of data value is required. In Section 5.2, we showed that statistical distances, specifically the WD, provides such as measure. We outlined its superior statistical properties and showed, through case studies, across Chapters 5, 6, and 7, with synthetic as well as real smart meter data, that it performs better than other alternatives considered across the different tasks and distributions. The WD largely results in higher correlations with actual task performance and more consistent Shapley allocations.

Given the significance of also considering consumers' valuations of smart meter data, in Chapter 3 we quantified their WTA and WTP for smart meter data anonymisation. We observed a significant endowment effect, with WTA being greater than WTP. From a regulatory perspective, this has significant implications. Indeed, under a Privacy by Design approach, establishing strong privacy defaults in the form of opt-ins rather than opt-outs would make the WTA the more appropriate valuation. We found that many consumers would require compensation before sharing their data irrespective of the privacy implication, while others were willing to pay for their privacy with compensation depending on the level of privacy afforded.

In Section 5.1.2 we endogenised the effect of DP in the WD through improved bounds and novel closed-forms. This facilitates the modelling of compensation dependent privacy preferences and removes the need for Monte-Carlo simulations, and as such, also

provides computational advantages. We outlined how the WD can be computed in a privacy-preserving manner using existing multi-party computation techniques. Overall, we developed a novel valuation framework with a unified metric of intrinsic data value.

8.2.4 A Privacy-Preserving Data Market

Finally, to implement the design criteria established above and assess its viability we tackled the fourth challenge which was the development of a privacy-preserving data market mechanism.

We chose to pursue an incentive mechanism design approach since in Chapter 3, we had identified that consumers value their smart meter data, implying they have non-zero reserve prices. The data procurement mechanisms, proposed in Chapter 6 built upon the data valuation mechanism proposed in Chapter 5, specifically, to minimise the WD of the data procured. Given the many potential users and use cases for smart meter data identified in Chapter 2, we proposed three types of mechanism: an exogenous budget mechanism with applications in task-agnostic data procurement, an endogenous budget mechanism which has task-specific applications where the aim is data value/welfare maximisation while ensuring welfare improvements are commensurate with procurement, costs and a joint optimisation mechanism, which can be used in settings where the data buyer aims to co-optimise data value and procurement costs. The latter two mechanisms capture the decision-dependent nature of data procurement highlighted in Section 5.1.

By incorporating the Hoeffding bound to approximate the combinatorial nature of data value, an endogenous modelling of the effect of DP, and the Lipschitz bound to incorporate task-specific dynamics we overcame the shortcomings of existing data procurement mechanisms outlined in Chapter 6. Specifically, in contrast to cooperative game based approaches, our mechanism decouples data valuation from model complexity improving computational tractability, ensures valuation is immune to model manipulation, and lends itself to extension to compensation-dependent privacy settings as in [52]. Similarly, compared to existing incentive mechanism design approaches our mechanism links task performance to payments and thus captures the decision-dependent nature of data procurement. At the same time, our approach also offers privacy to both the buyer, as they do not need to share their model/task, and sellers, as the WD is calculated in a privacy-preserving manner thus removing the need for a TTP.

To concretise further in Chapter 7, we applied the proposed mechanism to our moti-

vating example, namely, the retailers' day-ahead energy procurement problem to develop a novel joint energy and data market framework. We then presented two case studies to evaluate the performance of our proposed framework. The first focused on procuring forecasts, which allows us to compare our WD based valuation mechanism against existing techniques. The second implemented our full framework, including the procurement mechanism to value data directly, using actual smart meter data. In the forecast procurement case study, we showed that data value is linked to the difference between forecast uncertainty reduction and the noise introduced by DP and depends on privacy-value correlations among consumers. Of note is that, as privacy concerns increase, not only does data value decrease, but the retailer captures an increasing proportion of the remaining value.

In the second case study we explored the mechanism's distribution of profits, in terms of data payments to consumers. We found that data payments were, on average, at the lower end of WTA values estimated in Chapter 3. However, there were instances where much higher payments, in line with the WTA estimates, were made, confirming the viability of such a data market to resolve the dilemma of balancing access and privacy to smart meter data.

8.3 Recommendations for Policymakers

This thesis developed a means of widening data access while simultaneously improving privacy protections for consumers, with a focus on the GB's existing smart metering landscape and the specific concerns of GB consumers. To advance the smart meter data sharing landscape forward, our research findings provide policymakers with several insights which we develop into recommendations.

8.3.1 Fostering Informed Consent

Existing material provided by suppliers, the government, through Smart Energy GB and the data guide developed by Citizens Advice[73], do not communicate the privacy implications of data sharing adequately. This incongruity confirms the presence of information asymmetries which is likely to result in significant underestimations of consumers' privacy concerns. A centralised dashboard encompassing all energy related data flows, including smart appliances, demand response and CADs, would provide consumers with clarity on their data sharing options and choices. Our mapping of the data dependence of privacy risks provides a rigorous assessment of the privacy implications of data

sharing choices under the DAPF. This is essential to foster substantive informed consent. The proposed smart meter data guides could be expanded to include such information while acknowledging that, given the evolving nature of smart meter data analytics, it is not possible to detail the full extent of privacy risk in the future.

8.3.2 Transparency around Benefits and Usage

The move to an opt-out model for sharing HH smart meter data as part of the MHHS represents a departure from the Privacy by Design approach. It was justified based on the notion that data access for suppliers is required to ensure the accuracy of the settlement process and to achieve the benefits of the MHHS. However, the role of suppliers (and their data aggregation agents), in relation to smart meter data collection and its use for settlement purposes, were discussed at the early stages of the SMIP and operating models which did not require the suppliers to handle smart meter data were envisioned. Consequently, it is important for regulators, suppliers and third parties to consider how much data and what level of granularity is required for different uses. Transparency around how smart meter data will be used, the benefits of data sharing and how these benefits will be distributed should be communicated to consumers. As we laid out, the benefits of providing more granular data can, for most regulated activities, be quantified and should also be clearly set out. Limiting data access to only what is necessary and where benefits have been clearly identified prevents data hoarding and limits potential privacy infringements, in adherence with Privacy by Design principles. Here our mapping of the dependence of benefits on smart meter data characteristics provides a basis to develop these limits.

8.3.3 Proactive and Preventative Risk Management

Permissions controls and consent provide a level of privacy protection, but these place the onus on consumers and still leave their data vulnerable to privacy infringements. This is especially relevant given the issues we identified around obtaining informed consent. In addition, with the widening of data access, sharing outside of regulated entities exposes consumers to risks of misuse. Privacy by Design advocates for preventative and proactive management of such risks. Going beyond permissions control, the use of PPTs can ensure data is not easily re-identifiable and that even in the case of a data breach any inferences cannot be linked back to individual consumers. Importantly, this can be achieved while preserving data utility, allowing a wide range of benefits which hinge

on access to high resolution data, to be realised. Privacy-preserving techniques can help balance the legitimate privacy concerns of consumers and the value and benefits of wider access to high resolution smart meter data.

8.3.4 A Blueprint for Implementation

Differential Privacy has already been widely implemented in practice by both private entities such as Apple and Google and public entities such as the US Census Bureau. The US Census Bureau provides a detailed framework for implementing DP, which could be built upon and integrated within the UK's existing SMIP. There are many similarities between the challenges faced by the US Census Bureau and the SMIP. Specifically, the move towards presumed open data will result in similar disclosure risks and threats to those faced by the US census. Although the US Census already has a fixed catalogue of what data will be released a similar framework could be developed for smart meter data accounting for all stakeholders needs. Our mapping of benefits provides a basis upon which policymakers can build. There are several challenges which remain and require further investigation in order to fully implement differential privacy. These include how to determine the appropriate privacy-utility trade-off, how to account for the dynamic nature of smart meter data and who should implement the system.

8.3.5 Leveraging Heterogeneity

Privacy concerns regarding sharing of smart meter data vary significantly and depend on the intended use. Some consumers are unwilling to share high-resolution data whereas others are happy to share data at any resolution. Anonymising the data prior to sharing also increases consumers' willingness to share. In addition, many potential uses of smart meter data are not confined to regulated activities such as electricity settlement and billing. A transition towards a more dynamic and active domestic electricity sector will introduce new market actors, such as aggregators and local energy system operators, as well as new uses for smart meter data such as personalised tariffs and product recommendations. The current DAPF options do not accommodate this heterogeneity in preferences. A market mechanism would provide a means of leveraging the heterogeneity in privacy concerns and valuations. Extant cooperative game frameworks could be used, however these require a trusted third party to run the market. Given issues around trust, our proposed method provides an alternative which focuses on ensuring privacy preservation, not requiring a trust third party, while maintaining an acceptable level of performance.

8.4 Future Research Directions

The results and findings of this thesis motivate several future research directions pertaining to potential privacy risks, data valuation and procurement, and eliciting consumer preferences. We discussed the limitations and potential incremental improvements in the respective chapter summaries. We now focus on substantive new research directions based on our findings.

8.4.1 Concretising Privacy Risks

Our first contribution was mapping the data dependence of the potential privacy risks associated with sharing smart meter data. This informed much of the subsequent work contained within this thesis. In discussions with stakeholders like Sustainability First and the Energy Systems Catapult, following the publication of [Paper B], the theoretical nature of the privacy risks outlined was raised. Specifically, there was a demand to see real-world examples of privacy infringements, for example linking attacks, using smart meter data.

To be able to convince stakeholders of the need for additional privacy protections an important area of extension is to concretise these privacy risks by way of example. Indeed, the US Census Bureau only pursued DP following an internal experiment using their own datasets[219]. Although academic literature has shown that there are similar disclosure risks associated with smart meter data, these have used datasets which are already linked i.e. the dataset of personal information and the dataset of smart meter consumption are known to contain the same individuals. Developing examples, which link disjoint datasets and then attempt to infer personal information using a separate training dataset is needed, to determine the true out-of-sample performance and generalisability.

We see competitions which aim to bridge the gap between academic and industry practice, as a promising venue to develop these examples. This could be structured in a similar vein as other competitions in the field, such as the longstanding Global Energy Forecasting Competition[324], the recently announced Hybrid Energy Forecasting and Trading Competition[325], and of particular note, SYN-MAD 2022, a competition on morphing attack detection using privacy-aware data[326]. Our mapping of data dependencies in Chapter 2 and assessment of PPTs in Chapter 4 provide a basis for designing the parameters and assessment criteria. A particular focus should be placed on the risks of re-identification and reconstruction as many other potential infringements

stem from access to individual high-resolution data.

In addition, the effect of different PPTs could be studied to quantify the protection offered on a comparable basis. For example, DP applies a worst-case framing, protecting individual contributions to an aggregate even against an attacker with knowledge of all other individuals in the aggregation. Although this approach provides robust privacy guarantees, different assumptions on background knowledge, based on practical considerations, could ensure guarantees with less utility degradation[327]. Overall, this would provide a mechanism for translating the theoretical models of privacy risk into demonstrable infringements and better communicate privacy risks to policymakers and consumers alike.

8.4.2 Investigating Synergies among Privacy-Preserving Techniques

In this thesis we assessed different PPTs based on their individual properties and applied them (DP and MPC) for different purposes. However, as discussed in Chapter 4, the protections provided by different PPTs are not mutually exclusive. For example, calculating aggregates with MPC, as we propose for the WD calculations, would enable us to employ centralised DP, which requires much less noise, while not requiring a trusted third party. The example in [216], which proposes a privacy-preserving protocol for financial applications, takes this compounding of PPTs one step further by also including federated learning. As such, individual's data are hidden from the central server via MPC, protected from reconstruction/linking attacks by DP and also minimise the impact of data breaches, as there are no transfers of data under federated learning.

The synergies among PPTs warrant further investigation given the possibility to reduce the induced privacy-utility trade-offs and the different notions of privacy protection provided. A particularly crucial area of investigation, in the context of smart meter data is dealing with the dynamic nature of the database[170]. Given the temporal correlations within smart meter data the sequential composition under DP applies, resulting in ever increasing privacy losses[229]. Although we highlighted a temporal discounting based solution[170], the extent to which consumers exhibit such preferences requires further research. As such, combining different PPTs could provide alternative methods to overcome this challenge.

Federated learning models also face the sequential composition issue as successive model updates are generated from the same dataset[229]. Our work focused on data procurement rather than information or model procurement, motivated by the diverse

range of uses for smart meter data and the existing computational capabilities of smart meters. However, in the future we are likely to see increased computational capacity at the nodes allowing for more advanced distributed computing capabilities. Dedicated, task-specific federated learning/data processing techniques would therefore be more attractive solutions but would need to be complemented with appropriate privacy protection likely through a combination of PPTs.

8.4.3 Combinatorial Accuracy, and Computational Efficiency

Our data market mechanism employs the Hoeffding bound to incorporate the combinatorial effect of data value for aggregated data. This method was chosen to maintain computational tractability, as we only require the individual WD rather than each combination. As such, we assume individuals' values are fixed and independent, allowing us to employ a direct incentive mechanism. However, the computational advantages come at the cost of approximation error and a bias (see Figure 5.15a). In order to improve accuracy, we could apply sampling techniques, similar to those developed for the Shapley value[295], [296]. This approach would also require the re-design of our incentive mechanism. Alternatively, the use distributed computation techniques such as federated learning could alleviate computational issues by shifting computation to the smart meter which would retain accuracy. However, this would require smart meters with significantly more computational power and/or CADs with appropriate secure communications infrastructure.

The use of WD offers the potential for a unique approach to modelling combinatorial effects. Namely, using the geometric properties of the WD, specifically, Wasserstein barycenters, an alternative to Euclidean aggregation of distributions[277], and geodesics (the shortest path) in the Wasserstein space, [328]. The Wasserstein barycenter of a convex combination of distributions lies on a geodesic between them allowing for exact interpolation[329]. In addition, we note that location-scale distributions are closed for barycenters, i.e. if the distributions under consideration all belong to the same location-scale distribution then the associated barycenter also belongs to the same location-scale distribution [330]. These observations suggest the possibility to develop a computationally efficient, yet accurate, method to calculate WDs of subsets of aggregated distributions, and therefore to calculate data value. This could involve graph theoretic representations, which open up possibilities for efficient combinatorial incentive mechanisms (combinatorial reverse auctions)[331].

8.4.4 The Value of Interdisciplinarity

Finally, our research has highlighted the significant value of interdisciplinarity. More interdisciplinary research needs to be done using social science methodologies such as focus groups, in-depth qualitative interviews and household surveys which are longitudinal, with disaggregated data to account for heterogeneity in relation to smart meter data sharing. Studies on other aspects of smart technologies using such methodologies have provided interesting insights. For example, [168] consider perceptions and experiences of the attempt to roll out smart meters amongst low-income householders, drawing on longitudinal qualitative interviews of 24 residents from a Welsh Valleys community case site. Despite a small sample it brings out the importance of disaggregated data on smart technologies assumed to reduce fuel poverty showing the different needs and concerns of the elderly and people from low-income backgrounds. They stress the importance of policies to be sensitive to vulnerable consumers. Similar points are made by [332] on the elderly using a combination of sensors etc and interviews with elderly respondents on behaviours regarding energy and technology use.

To move forward towards a decarbonised future, for which smart meters and the data they gather are key enablers, it is important to adopt an interdisciplinary approach to balance the interests and concerns of both consumers and suppliers. This needs to be done within a perspective that values the principles of trust, transparency and respect for consumers preferences, as much as the requirements of data for innovative technical and market solutions which are efficient and sustainable. This also requires close collaboration between policy makers, academia and citizens, and more investigation and assessment of chosen policy instruments.

Bibliography

[1] BEIS, "Energy white paper," BEIS, 2020, [Online]. Available: https://www.gov.uk/government/publications/energy-white-paper-powering-our-net-zero-future.

[2] R. Hledik, P. Bagci, and S. Chhachhi, "Two paths for advancing great britain's smart metering programme," 2018, [Online]. Available: https://www.brattle.com/wp-content/uploads/2021/05/15050_advancing_gbs_smart_metering_programme_-_discussion_paper_final.pdf.

[3] BEIS, *Smart meter statistics in great britain: Quarterly report to end june 2023*, Aug. 2023, [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1181440/Q2_2023_Smart_Meters_Statistics_Report.pdf.

[4] B. K. Sovacool, P. Kivimaa, S. Hielscher, *et al.*, "Vulnerability and resistance in the united kingdom's smart meter transition," *Energy Policy*, vol. 109, pp. 767–781, Oct. 2017, ISSN: 03014215, DOI: 10.1016/j.enpol.2017.07.037.

[5] Y. Wang, Q. Chen, D. Gan, *et al.*, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Transactions on Smart Grid*, vol. 10, pp. 2593–2602, 3 May 2019, ISSN: 1949-3053, DOI: 10.1109/TSG.2018.2805723.

[6] L. Stankovic, V. Stankovic, J. Liao, *et al.*, "Measuring the energy intensity of domestic activities from smart meter data," *Applied Energy*, vol. 183, pp. 1565–1580, Dec. 2016, ISSN: 03062619, DOI: 10.1016/j.apenergy.2016.09.087.

[7] C. Véliz and P. Grunewald, "Protecting data privacy is key to a smart energy future," *Nature Energy*, vol. 3, pp. 702–704, 9 Sep. 2018, ISSN: 2058-7546, DOI: 10.1038/s41560-018-0203-3.

[8] Data Privacy Manager, "20 biggest gdpr fines so far [2019, 2020 & 2021]," Aug. 2021, [Online]. Available: https://dataprivacymanager.net/5-biggest-gdpr-fines-so-far-2020/.

[9] R. Cellan-Jones, "British airways faces record £183m fine for data breach," *BBC News*, Jul. 2019, [Online]. Available: https://www.bbc.co.uk/news/business-48905907.

[10] D. Lu, "The social dilemma review: How big tech companies use us for profit," *New Scientist*, Sep. 2020, [Online]. Available: https://institutions.newscientist.com/article/2255588-the-social-dilemma-review-how-big-tech-companies-use-us-for-profit/.

[11] BEIS, "Smart metering implementation programme: Review of the data access and privacy framework," 2018, [Online]. Available: https://www.gov.uk/government/publications/smart-metering-implementation-.

[12] J. Banks and K. Mcglinchey, "Access to half-hourly electricity data for settlement purposes: A data protection impact assessment," OFGEM, 2019, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2019/06/data_protection_impact_assessment_v2_june_2019.pdf.

[13] OFGEM, *Consultation on access to half-hourly electricity data for settlement purposes: Ofgem decision and response to stakeholder feedback.* Jun. 2019, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2019/06/access_to_data_consultation_ofgem_response_0.pdf.

[14] A. Acquisti, C. Taylor, and L. Wagman, "The economics of privacy," *Journal of Economic Literature*, vol. 54, pp. 442–492, 2 Jun. 2016, ISSN: 0022-0515, DOI: 10.1257/jel.54.2.442.

[15] S. A. Azcoitia and N. Laoutaris, "A survey of data marketplaces and their business models," *ACM SIGMOD Record*, vol. 51, pp. 18–29, 3 Nov. 2022, ISSN: 0163-5808, DOI: 10.1145/3572751.3572755.

[16] J. Wang, F. Gao, Y. Zhou, *et al.*, "Data sharing in energy systems," *Advances in Applied Energy*, vol. 10, p. 100 132, Jun. 2023, ISSN: 26667924, DOI: 10.1016/j.adapen.2023.100132.

[17] BEIS, "Digitalising our energy system for net zero strategy and action plan 2021," Jul. 2021, [Online]. Available: https://assets.publishing.service

.gov.uk/government/uploads/system/uploads/attachment_data/file/1004011/energy-digitalisation-strategy.pdf.

[18] Sustainability First and CSE, "Smart meter energy data: Public interest advisory group (piag). final report - phase 2.," Sustainability First & CSE., Apr. 2021, [Online]. Available: https://www.sustainabilityfirst.org.uk/images/publications/piag/PIAG-phase-2-final-report.pdf%20-.

[19] BEIS, "Smart meter roll-out cost benefit analysis 2019," BEIS, Sep. 2019, [Online]. Available: https://assets.publishing.service.gov.uk/media/5d7f54c4e5274a27c2c6d53a/smart-meter-roll-out-cost-benefit-analysis-2019.pdf.

[20] Y. Wang, Q. Chen, T. Hong, *et al.*, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, pp. 3125–3148, 3 May 2019, ISSN: 1949-3053, DOI: 10.1109/TSG.2018.2818167.

[21] B. Zhao, L. Stankovic, and V. Stankovic, "Electricity usage profile disaggregation of hourly smart meter data," *4th International Workshop on Non-Intrusive Load Monitoring*, 2018, [Online]. Available: https://strathprints.strath.ac.uk/63692/.

[22] J. L. Ramírez-Mendiola, P. Grünewald, and N. Eyre, "Linking intra-day variations in residential electricity demand loads to consumers' activities: What's missing?" *Energy and Buildings*, vol. 161, pp. 63–71, Feb. 2018, ISSN: 03787788, DOI: 10.1016/j.enbuild.2017.12.012.

[23] C. Shin, S. Rho, H. Lee, *et al.*, "Data requirements for applying machine learning to energy disaggregation," *Energies*, vol. 12, p. 1696, 9 May 2019, ISSN: 1996-1073, DOI: 10.3390/en12091696.

[24] A. Knight, "Consumer views on sharing half-hourly settlement data," OFGEM, Tech. Rep., Jul. 2018, [Online]. Available: https://www.ofgem.gov.uk/publications-and-updates/consumer-research-datasets.

[25] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Security and Privacy Magazine*, vol. 3, pp. 26–33, 1 Jan. 2005, ISSN: 1540-7993, DOI: 10.1109/MSP.2005.22.

[26] A. Skatova, R. L. McDonald, S. Ma, *et al.*, "Unpacking privacy: Willingness to pay to protect personal data," *PsyArXiv*, 2019, DOI: 10.31234/osf.io/ahwe4.

[27] L.-L. Richter and M. G. Pollitt, "Which smart electricity service contracts will consumers accept? the demand for compensation in a platform market," *Energy Economics*, vol. 72, pp. 436–450, May 2018, ISSN: 01409883, DOI: 10.1016/j.eneco.2018.04.004.

[28] A. G. Winegar and C. R. Sunstein, "How much is data privacy worth? a preliminary investigation," *SSRN Electronic Journal*, Jul. 2019, ISSN: 1556-5068, DOI: 10.2139/ssrn.3413277.

[29] G. Eibl and D. Engel, "Differential privacy for real smart metering data," *Computer Science - Research and Development*, vol. 32, pp. 173–182, 1-2 Mar. 2017, ISSN: 1865-2034, DOI: 10.1007/s00450-016-0310-y.

[30] G. Giaconi, D. Gunduz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 129–142, 1 Jan. 2018, ISSN: 1556-6013, DOI: 10.1109/TIFS.2017.2744601.

[31] M. Jawurek, F. Kerschbaum, and G. Danezis, "Privacy technologies for smart grids - a survey of options," Tech. Rep. MSR-TR-2012-119, Nov. 2012, [Online]. Available: https://www.microsoft.com/en-us/research/publication/privacy-technologies-for-smart-grids-a-survey-of-options/.

[32] M. R. Asghar, G. Dan, D. Miorandi, *et al.*, "Smart meter data privacy: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, pp. 2820–2835, 4 2017, ISSN: 1553-877X, DOI: 10.1109/COMST.2017.2720195.

[33] G. Giaconi, D. Gunduz, and H. V. Poor, "Privacy-aware smart metering: Progress and challenges," *IEEE Signal Processing Magazine*, vol. 35, pp. 59–78, 6 Nov. 2018, ISSN: 1053-5888, DOI: 10.1109/MSP.2018.2841410.

[34] S. Desai, R. Alhadad, N. Chilamkurti, *et al.*, "A survey of privacy preserving schemes in ioe enabled smart grid advanced metering infrastructure," *Cluster Computing*, vol. 22, pp. 43–69, 1 Mar. 2019, ISSN: 15737543, DOI: 10.1007/s10586-018-2820-9.

[35] S. Finster and I. Baumgart, "Privacy-aware smart metering: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 1088–1101, 2 2015, ISSN: 1553-877X, DOI: 10.1109/COMST.2015.2425958.

[36] F. Farokhi, "Review of results on smart-meter privacy by data manipulation, demand shaping, and load scheduling," *IET Smart Grid*, vol. 3, pp. 605–613, 5 Oct. 2020, ISSN: 2515-2947, DOI: 10.1049/iet-stg.2020.0129.

[37] C. Gonçalves, P. Pinson, and R. J. Bessa, "Towards Data Markets in Renewable Energy Forecasting," *IEEE Transactions on Sustainable Energy*, p. 1, 2020, ISSN: 1949-3037 VO -, DOI: 10.1109/TSTE.2020.3009615.

[38] P. Pinson, L. Han, and J. Kazempour, "Regression markets and application to energy forecasting," *arxiv*, Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.03633.

[39] L. Han, J. Kazempour, and P. Pinson, "Monetizing customer load data for an energy retailer: A cooperative game approach," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.05519.

[40] L. Han, P. Pinson, and J. Kazempour, "Trading data for wind power forecasting: A regression market with lasso regularization," *arxiv*, Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.07432.

[41] Z. Sun, L. V. Krannichfeldt, and Y. Wang, "Trading and valuation of day-ahead load forecasts in an ensemble model," *IEEE Transactions on Industry Applications*, vol. 59, pp. 2686–2695, 3 May 2023, ISSN: 0093-9994, DOI: 10.1109/TIA.2023.3244171.

[42] A. Ghorbani, M. P. Kim, and J. Zou, "A distributional framework for data valuation," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.12334.

[43] K. Ren, "Differentially private auction for federated learning with non-iid data," vol. 2022-May, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 305–312, ISBN: 9781665498616, DOI: 10.1109/ICSS55994.2022.00054.

[44] Julien, D. C. Emiliano, U. E. B. Igor, *et al.*, "What's the gist? privacy-preserving aggregation of user profiles," in *Computer Security - ESORICS 2014*, J. K. Mirosław and Vaidya, Eds., Springer International Publishing, 2014, pp. 128–145, ISBN: 978-3-319-11212-1, DOI: 10.1007/978-3-319-11212-1_8.

[45] Y. Zhao, M. Li, L. Lai, *et al.*, "Federated learning with non-iid data," Jun. 2018, DOI: 10.48550/arXiv.1806.00582.

[46] S. Fan, H. Zhang, Y. Zeng, *et al.*, "Hybrid blockchain-based resource trading system for federated learning in edge computing," *IEEE Internet of Things Journal*, vol. 8, pp. 2252–2264, 4 Feb. 2021, ISSN: 2327-4662, DOI: 10.1109/JIOT.2020.3028101.

[47] M. Jia, Y. Wang, C. Shen, *et al.*, "Privacy-preserving distributed clustering for electrical load profiling," *IEEE Transactions on Smart Grid*, vol. 12, pp. 1429–1444, 2 Mar. 2021, ISSN: 1949-3053, DOI: 10.1109/TSG.2020.3031007.

[48] N. Ding, Z. Fang, and J. Huang, "Optimal contract design for efficient federated learning with multi-dimensional private information," *IEEE Journal on Selected Areas in Communications*, vol. 39, pp. 186–200, 1 Jan. 2021, ISSN: 0733-8716, DOI: 10.1109/JSAC.2020.3036944.

[49] J. Weng, J. Weng, H. Huang, *et al.*, "Fedserving: A federated prediction serving framework based on incentive mechanism," IEEE, May 2021, pp. 1–10, ISBN: 978-1-6654-0325-2, DOI: 10.1109/INFOCOM42981.2021.9488807.

[50] A. Ghosh and A. Roth, "Selling privacy at auction," *Games and Economic Behavior*, 2015, ISSN: 10902473, DOI: 10.1016/j.geb.2013.06.013.

[51] M. Zhang, F. Beltran, and J. Liu, "Selling data at an auction under privacy constraints," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, J. Peters and D. Sontag, Eds., vol. 124, PMLR, Dec. 2020, pp. 669–678, [Online]. Available: https://proceedings.mlr.press/v124/zhang20b.html.

[52] A. Fallah, A. Makhdoumi, A. Malekian, *et al.*, "Optimal and differentially private data acquisition: Central and local mechanisms," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2201.03968.

[53] DECC, "Smart metering implementation programme: Information leaflet," Department of Energy & Climate Change, 2013, [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/245736/smart_meters_domestic_leaflet.pdf.

[54] ICO, *Data protection by design and default*, May 2023, [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-gove

rnance/accountability-and-governance/data-protection-by-design-and-default/.

[55] S. Kingsmill and A. Cavoukian, "Privacy by design setting a new standard for privacy certification," 2015, [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/risk/ca-en-ers-privacy-by-design-brochure.PDF.

[56] I. Brown, "Britain's smart meter programme: A case study in privacy by design," *https://doi.org/10.1080/13600869.2013.801580*, vol. 28, pp. 172–184, 2 2014, ISSN: 13646885, DOI: 10.1080/13600869.2013.801580.

[57] Information Commissioner's Office, "Response to OFGEM's open letter of 17 december 2015: Half-hourly settlement (hhs): The way forward.," Information Commissioner's Office, Feb. 2016, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2016/03/information_commissioner_response_-_dec_15_open_letter.pdf.

[58] N. C. S. Centre, "The smart security behind the gb smart metering system," National Cyber Security Centre, Tech. Rep., Apr. 2016, [Online]. Available: https://www.ncsc.gov.uk/information/the-smart-security-behind-the-gb-smart-metering-system.

[59] C. Cuijpers and B.-J. Koops, "Smart metering and privacy in europe: Lessons from the dutch case," in Springer Netherlands, 2013, pp. 269–293, ISBN: 9789400751705, DOI: 10.1007/978-94-007-5170-5_12.

[60] Smart Energy GB, *Smart meter installation process*, 2021, [Online]. Available: https://www.smartenergygb.org/en/get-a-smart-meter/the-installation-process.

[61] Citizen's Advice, "Summary report on energy suppliers' communication with consumers regarding smart meter data," Citizen's Advice, 2016, [Online]. Available: http://www.energy-uk.org.uk/files/docs/Policies/Smart%20Meter%20policies%20consultation%20r.

[62] BEIS, "Smart metering equipment technical specifications version 2.0," BEIS, Jan. 2013, [Online]. Available: https://www.smartme.co.uk/documents/smart_meters_equipment_technical_spec_version_2.pdf.

[63] BEIS, "Consultation on smart metering system proportional load control functionality," BEIS, 2019, pp. 1–23, [Online]. Available: https://smartenergycodecompany.co.uk/download/17882/.

[64] Citizen's Advice, "Clear and in control," 2019, [Online]. Available: https://www.citizensadvice.org.uk/about-us/our-work/policy/policy-research-topics/energy-policy-research-and-consultation-responses/energy-policy-research/clear-and-in-control/.

[65] OFGEM, "Electricity retail market-wide half-hourly settlement: Decision document," Ofgem, Apr. 2021, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2021/04/mhhs_draft_ia_consultation_decision_document_final_version_for_publication_20.04.21.pdf.

[66] OFGEM, "Statutory consultation on proposals to modify electricity supply licence condition 47: "smart metering-matters relating to obtaining and using consumption data"," May 2022, [Online]. Available: www.ofgem.gov.uk..

[67] Sustainability First and CSE, "Smart meter energy data: Public interest advisory group (paig). annex 1 - working paper on dno privacy plans.," Sustainability First and CSE, Apr. 2021, [Online]. Available: https://www.cse.org.uk/downloads/file/PIAG-phase-2-privacy-plans-annex.pdf.

[68] D. for Digital Culture Media & Sport, "Data: A new direction," Sep. 2021, [Online]. Available: https://www.gov.uk/government/consultations/data-a-new-direction.

[69] OFGEM, "Electricity retail market-wide half-hourly settlement: Consultation," Ofgem, 2020, [Online]. Available: https://www.ofgem.gov.uk/system/files/docs/2020/05/mhhs_draft_impact_assessment_consultation.pdf.

[70] Sustainability First and CSE, "Smart meter energy data: Public interest advisory group (paig). stimulus paper 2 - international experience-smart meter data access. maxine frerk, sustainability first," Maxine Frerk, Sustainability First, Aug. 2018, [Online]. Available: https://www.smartenergydatapiag.org.uk/.

[71] BEIS, "PAS 1878:2021 energy smart appliances –system functionality and architecture – specification," BEIS, May 2021, [Online]. Available: https://www.bsigroup.com/en-GB/about-bsi/uk-national-standards-body/about-

standards/Innovation/energy-smart-appliances-programme/pas-1878/.

[72] BEIS, "PAS 1879:2021 energy smart appliances-demand side response operation-code of practice," BEIS, May 2021, [Online]. Available: https://www.bsigroup.com/globalassets/localfiles/en-gb/energy-smart-appliances-programme/pas1879.pdf.

[73] C. Advice, *Smart metering data dashboard*, 2018, [Online]. Available: https://www.citizensadvice.org.uk/Global/Public/Corporate%20content/Publications/Personal%2.

[74] Energy Digitalisation Taskforce, "Delivering a digitalised energy system," Energy System Catapult, Jan. 2022, [Online]. Available: https://es.catapult.org.uk/report/delivering-a-digitalised-energy-system/.

[75] R. Carmichael, R. Gross, and A. Rhodes, "Unlocking the potential of residential electricity consumer engagement with demand response," Energy Futures Lab, Imperial College London, 2018, [Online]. Available: https://www.imperial.ac.uk/energy-futures-lab/policy/briefing-papers/paper-3/.

[76] R. Carmichael, A. Rhodes, R. Hanna, *et al.*, "Smart and flexible electric heat an energy futures lab briefing paper," 2020, [Online]. Available: http://imperial.ac.uk/energy-futures-lab.

[77] A. Rhodes, "Digitalisation of energy an energy futures lab briefing paper," 2020, [Online]. Available: http://imperial.ac.uk/energy-futures-lab.

[78] Elexon, *Settlement & invoicing*, [Online]. Available: https://www.elexon.co.uk/settlement/.

[79] BEIS, "Review of the average annual domestic gas and electricity consumption levels: Methodology note," BEIS, Tech. Rep., May 2020, [Online]. Available: https://assets.publishing.service.gov.uk/media/5ec54d0ad3bf7f45fb3213b7/annual-domestic-gas-electricity-consumption-levels-review-methodology-note.pdf.

[80] Elexon, "Load profiles and their use in electricity settlement," Elexon, 2018, [Online]. Available: https://www.elexon.co.uk/documents/training-guidance/bsc-guidance-notes/load-profiles.

[81] Cornwall Insights, "Market wide half-hourly settlement: Half-way home or just the first steps on the journey to a smart, flexible, energy system?," Jun. 2019, [Online]. Available: http://www.cornwall-insight.com/market-wide-half-hourly-settlement-half-way-home-or-just-the-first-steps-on-the-journey-to-a-smart-flexible-energy-system/.

[82] OFGEM, "Market-wide half-hourly settlement: Final impact assessment," Apr. 2021, [Online]. Available: https://www.ofgem.gov.uk/energy-policy-and-regulation/policy-and-regulatory-programmes/electricity-settlement-reform.

[83] C. Dromaque, R. Grigoriou, T. Mikkelsen, *et al.*, "The role of data for consumer centric energy markets and solutions," 2018, [Online]. Available: https://www.esmig.eu/wp-content/uploads/2021/10/The-Role-of-Data-for-Consumer-Centric-Energy-Markets-and-Solutions.pdf.

[84] Carbon Trust and Imperial College London, "An analysis of electricity system flexibility for great britain," Nov. 2016, [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/568982/An_analysis_of_electricity_flexibility_for_Great_Britain.pdf.

[85] P. Steele, "Octopus energy: Agile pricing explained," 2019, [Online]. Available: https://octopus.energy/blog/agile-pricing-explained/.

[86] R. Hledik, W. Gorman, M. Fell, *et al.*, "The value of tou tariffs in great britain: Insights for decision-makers volume I: Final report," The Brattle Group, 2017, [Online]. Available: https://www.citizensadvice.org.uk/Global/CitizensAdvice/Energy/The%5C%20Value%5C%20of%5C%20TOU%5C%20Tariffs%5C%20in%5C%20GB%5C%20-%5C%20Volume%5C%20I.pdf.

[87] Energy Systems Catapult, "Smart systems and heat," 2020, [Online]. Available: https://es.catapult.org.uk/case-studies/smart-systems-and-heat/.

[88] Energy Systems Catapult, "Enabling smart local energy systems: The value of digitalisation and data best practice," Energy Systems Catapult, 2021, [Online]. Available: https://es.catapult.org.uk/reports/enabling-smart-local-energy-systems-the-value-of-digitalisation-and-data-best-practice/.

[89] C. Vigurs, C. Maidment, M. Fell, *et al.*, "Customer privacy concerns as a barrier to sharing data about energy use in smart local energy systems: A rapid realist review," *Energies*, vol. 14, p. 1285, 5 Feb. 2021, ISSN: 1996-1073, DOI: 10.3390/en14051285.

[90] M. Duesterberg and L. Mirviss, "Reinventing residential demand response breaking through barriers with gamification and devices," California Energy Commission, Mar. 2021, [Online]. Available: https://www.energy.ca.gov/sites/default/files/2021-05/CEC-500-2021-019.pdf.

[91] C. Maidment, C. Vigurs, M. J. Fell, *et al.*, "Privacy and data sharing in smart local energy systems: Insights and recommendations," Tech. Rep., 2020, [Online]. Available: www.energyrev.org.uk.

[92] R. Hledik, A. Faruqui, J. Weiss, *et al.*, "The tariff transition considerations for domestic distribution tariff redesign in great britain volume I: Final report," 2016, [Online]. Available: https://www.brattle.com/wp-content/uploads/2017/10/7286_the_tariff_transition_-_considerations_for_domestic_distribution_tariff_redesign_in_great_britain.pdf.

[93] National Grid ESO, "Virtual energy system," 2021, [Online]. Available: https://www.nationalgrideso.com/virtual-energy-system.

[94] P. Grunewald, "How has behaviour changed under the covid-19 lockdown?" *Joju Solar*, May 2020, [Online]. Available: https://www.jojusolar.co.uk/opinion/how-has-behaviour-changed-under-covid-19-lockdown/.

[95] E. Webborn, S. Elam, E. Mckenna, *et al.*, "Utilising smart meter data for research and innovation in the uk," *ECEEE SUMMER STUDY PROCEEDINGS*, 2019, [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/10075955/1/8-113-19_Webborn%5C%20Final_Published.pdf.

[96] DECC, "Smart meters, smart data, smart growth," DECC, Jan. 2015, [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/591322/09022017_-_Smart_Meters__Data__Growth_DR_-_updated.pdf.

[97] M. Fell, H. Kennard, G. Huebner, *et al.*, "Energising health: A review of the health and care applications of smart meter data," UCL Energy Institute, May 2017, [Online]. Available: https://www.smartenergygb.org/media/bpcbgvid/energising-health-final-report.pdf.

[98] C. Chalmers, P. Fergus, C. A. C. Montanez, *et al.*, "Detecting activities of daily living and routine behaviours in dementia patients living alone using smart meter load disaggregation," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2020, ISSN: 2168-6750, DOI: 10.1109/TETC.2020.2993177.

[99] J. Paxman, M. James, E. Costanza, *et al.*, "Smart future of healthcare," 20/20health, Nov. 2020, [Online]. Available: https://www.smartenergygb.org/media/ueybkg5v/2020health.pdf.

[100] C. Dinesh, S. Welikala, Y. Liyanage, *et al.*, "Non-intrusive load monitoring under residential solar power influx," *Applied Energy*, vol. 205, pp. 1068–1080, Nov. 2017, ISSN: 03062619, DOI: 10.1016/j.apenergy.2017.08.094.

[101] A. Cominola, M. Giuliani, D. Piga, *et al.*, "A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring," *Applied Energy*, vol. 185, pp. 331–344, Jan. 2017, ISSN: 03062619, DOI: 10.1016/j.apenergy.2016.10.040.

[102] V. Singhal, J. Maggu, and A. Majumdar, "Simultaneous detection of multiple appliances from smart-meter measurements via multi-label consistent deep dictionary learning and deep transform learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2969–2978, 2019, DOI: 10.1109/TSG.2018.2815763.

[103] H. Kim, M. Marwah, M. Arlitt, *et al.*, "Unsupervised disaggregation of low frequency power measurements," in *Proceedings of the 2011 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Apr. 2011, pp. 747–758, ISBN: 978-0-89871-992-5, DOI: 10.1137/1.9781611972818.64.

[104] J. M. Abreu, F. C. Pereira, and P. Ferrão, "Using pattern recognition to identify habitual behavior in residential electricity consumption," *Energy and Buildings*, vol. 49, pp. 479–487, Jun. 2012, ISSN: 03787788, DOI: 10.1016/j.enbuild.2012.02.044.

[105] J. Munkhammar, J. D. Bishop, J. J. Sarralde, *et al.*, "Household electricity use, electric vehicle home-charging and distributed photovoltaic power production in the city of westminster," *Energy and Buildings*, vol. 86, pp. 439–448, Jan. 2015, ISSN: 0378-7788, DOI: 10.1016/J.ENBUILD.2014.10.006.

[106] M. Parti and C. Parti, "The total and appliance-specific conditional demand for electricity in the household sector," *The Bell Journal of Economics*, vol. 11, p. 309, 1 1980, ISSN: 0361915X, DOI: 10.2307/3003415.

[107] Z. Zhang, J. H. Son, Y. Li, *et al.*, "Training-free non-intrusive load monitoring of electric vehicle charging with low sampling rate," in *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, Oct. 2014, pp. 5419–5425, ISBN: 978-1-4799-4032-5, DOI: 10.1109/IECON.2014.7049328.

[108] K. S. Cetin, M. Siemann, and C. Sloop, "Disaggregation and future prediction of monthly residential building energy use data using localized weather data network," in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2016, pp. 1–12, [Online]. Available: https://www.aceee.org/files/proceedings/2016/data/papers/12_410.pdf.

[109] C. M. R. do Carmo and T. H. Christensen, "Cluster analysis of residential heat load profiles and the role of technical and household characteristics," *Energy and Buildings*, vol. 125, pp. 171–180, Aug. 2016, ISSN: 03787788, DOI: 10.1016/j.enbuild.2016.04.079.

[110] B. Liu, W. Luan, and Y. Yu, "Dynamic time warping based non-intrusive load transient identification," *Applied Energy*, vol. 195, pp. 634–645, Jun. 2017, ISSN: 03062619, DOI: 10.1016/j.apenergy.2017.03.010.

[111] V. Hoffmann, B. I. Fesche, K. Ingebrigtsen, *et al.*, "Automated detection of electric vehicles in hourly smart meter data," in *CIRED 2019 Conference*, AIM, 2019, p. 1531, DOI: http://dx.doi.org/10.34890/666.

[112] S. Welikala, N. Thelasingha, M. Akram, *et al.*, "Implementation of a robust real-time non-intrusive load monitoring solution," *Applied Energy*, vol. 238, pp. 1519–1529, Mar. 2019, ISSN: 03062619, DOI: 10.1016/j.apenergy.2019.01.167.

[113] K. Li, F. Wang, Z. Mi, *et al.*, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," *Applied Energy*, vol. 253, Nov. 2019, DOI: 10.1016/J.APENERGY.2019.113595.

[114] P. Zhang, C. Zhou, B. G. Stewart, *et al.*, "An improved non-intrusive load monitoring method for recognition of electric vehicle battery charging load,"

Energy Procedia, vol. 12, pp. 104–112, 2011, ISSN: 18766102, DOI: 10.1016/j.egypro.2011.10.015.

[115] C. M. Cheung, S. R. Kuppannagari, R. Kannan, *et al.*, "Disaggregation of behind-the-meter solar generation in presence of energy storage resources," *2020 IEEE Conference on Technologies for Sustainability, SusTech 2020*, Apr. 2020, DOI: 10.1109/SUSTECH47890.2020.9150506.

[116] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, pp. 184–194, Jun. 2013, ISSN: 03605442, DOI: 10.1016/j.energy.2013.03.086.

[117] N. Buescher, S. Boukoros, S. Bauregger, *et al.*, "Two is not enough: Privacy assessment of aggregation schemes in smart metering," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, pp. 198–214, 4 Oct. 2017, ISSN: 2299-0984, DOI: 10.1515/popets-2017-0045.

[118] S. Wang, R. Li, A. Evans, *et al.*, "Regional nonintrusive load monitoring for low voltage substations and distributed energy resources," *Applied Energy*, vol. 260, p. 114 225, Feb. 2020, ISSN: 03062619, DOI: 10.1016/j.apenergy.2019.114225.

[119] G. S. Ledva, L. Balzano, and J. L. Mathieu, "Real-time energy disaggregation of a distribution feeder's demand using online learning," *IEEE Transactions on Power Systems*, vol. 33, pp. 4730–4740, 5 Sep. 2018, ISSN: 0885-8950, DOI: 10.1109/TPWRS.2018.2800535.

[120] G. S. Ledva and J. L. Mathieu, "Separating feeder demand into components using substation, feeder, and smart meter measurements," *IEEE Transactions on Smart Grid*, vol. 11, pp. 3280–3290, 4 Jul. 2020, ISSN: 1949-3053, DOI: 10.1109/TSG.2020.2967220.

[121] E. C. Kara, C. M. Roberts, M. Tabone, *et al.*, "Disaggregating solar generation from feeder-level measurements," *Sustainable Energy, Grids and Networks*, vol. 13, pp. 112–121, Mar. 2018, ISSN: 23524677, DOI: 10.1016/j.segan.2017.11.001.

[122] M. Aydinalp, V. I. Ugursal, and A. S. Fung, "Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks," *Applied Energy*, vol. 79, pp. 159–178, 2 Oct. 2004, ISSN: 03062619, DOI: 10.1016/j.apenergy.2003.12.006.

[123] M. Aydinalp-Koksal and V. I. Ugursal, "Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector," *Applied Energy*, vol. 85, pp. 271–296, 4 Apr. 2008, ISSN: 03062619, DOI: 10.1016/j.apenergy.2006.09.012.

[124] C. Beckel, L. Sadamori, and S. Santini, "Automatic socio-economic classification of households using electricity consumption data," in *Proceedings of the the fourth international conference on Future energy systems - e-Energy '13*, ACM Press, 2013, p. 75, ISBN: 9781450320528, DOI: 10.1145/2487166.2487175.

[125] C. Beckel, L. Sadamori, T. Staake, *et al.*, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, Dec. 2014, ISSN: 03605442, DOI: 10.1016/j.energy.2014.10.025.

[126] S. Williams and K. Gask, "Modelling sample data from smart-type electricity meters to assess potential within official statistics," *The Office for National Statistics, UK*, 2015.

[127] S. Hattori and Y. Shinohara, "Actual consumption estimation algorithm for occupancy detection using low resolution smart meter data," SCITEPRESS - Science and Technology Publications, 2017, pp. 39–48, ISBN: 978-989-758-211-0, DOI: 10.5220/0006129400390048.

[128] X. Tong, R. Li, F. Li, *et al.*, "Cross-domain feature selection and coding for household energy behavior," *Energy*, vol. 107, pp. 9–16, Jul. 2016, ISSN: 03605442, DOI: 10.1016/j.energy.2016.03.135.

[129] P. Grünewald and M. Diakonova, "Societal differences, activities, and performance: Examining the role of gender in electricity demand in the united kingdom," *Energy Research & Social Science*, vol. 69, p. 101 719, Nov. 2020, ISSN: 22146296, DOI: 10.1016/j.erss.2020.101719.

[130] A. Satre-Meloy, M. Diakonova, and P. Grunewald, "Daily life and demand: New data on behavioral drivers of residential electricity use patterns," *2018 ACEEE Summer Study on Energy Efficiency in Buildings*, 2018, [Online]. Avail-

able: https://aceee.org/files/proceedings/2018/index.html#/paper/event-data/p264.

[131] P. Grunewald and T. Reisch, "The trust gap - privacy perceptions of location data for energy services in the uk," *Energy Research and Social Science*, vol. 68, p. 101 534, Oct. 2020, ISSN: 22146296, DOI: 10.1016/j.erss.2020.101534.

[132] U. Greveler, B. Justus, and D. Loehr, "Multimedia content identification through smart meter power usage profiles," *IN COMPUTERS, PRIVACY AND DATA PROTECTION (CPDP)*, 2012, DOI: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.727.4674.

[133] S. Schafer, "With capital in panic, pizza deliveries soar," *Washington Post*, Dec. 1998, [Online]. Available: https://www.washingtonpost.com/wp-srv/politics/special/clinton/stories/pizza121998.htm.

[134] Scottish & Southern Electricity Networks, "Smart meter data privacy plan (access to household electricity consumption data)," Scottish & Southern Electricity Networks, Apr. 2020, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2020/05/ssen_smart_meter_data_privacy_plan_redacted_final.pdf.

[135] "Privacy, technology, and norms: The case of smart meters," *Social Science Research*, vol. 51, pp. 64–76, May 2015, ISSN: 0049089X, DOI: 10.1016/j.ssresearch.2014.12.003.

[136] T. Jakobi, S. Patil, D. Randall, *et al.*, "It is about what they could do with the data: A user perspective on privacy in smart metering," *ACM Transactions on Computer-Human Interaction*, vol. 26, 1 Jan. 2019, ISSN: 15577325, DOI: 10.1145/3281444.

[137] A. Dickman and A. P. Aslaksen, "Consumer attitudes to dno access to half hourly electricity consumption data," Ipsos Mori, 2017, [Online]. Available: https://www.ipsos.com/ipsos-mori/en-uk/data-privacy-and-smart-meters.

[138] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Anomalies: The endowment effect, loss aversion, and status quo bias," *Journal of Economic Perspectives*, vol. 5, pp. 193–206, 1 Feb. 1991, ISSN: 0895-3309, DOI: 10.1257/jep.5.1.193.

[139] H. A. Simon, "Bounded rationality," in Palgrave Macmillan UK, 1990, pp. 15–18, DOI: 10.1007/978-1-349-20568-4_5.

[140] S. Sundt and K. Rehdanz, "Consumers' willingness to pay for green electricity: A meta-analysis of the literature," *Energy Economics*, vol. 51, pp. 1–8, Sep. 2015, ISSN: 01409883, DOI: 10.1016/j.eneco.2015.06.005.

[141] G. R. Parsons, M. K. Hidrue, W. Kempton, *et al.*, "Willingness to pay for vehicle-to-grid (v2g) electric vehicles and their contract terms," *Energy Economics*, vol. 42, pp. 313–324, Mar. 2014, ISSN: 01409883, DOI: 10.1016/j.eneco.2013.12.018.

[142] E. R. Frederiks, K. Stenner, E. V. Hobman, *et al.*, "Evaluating energy behavior change programs using randomized controlled trials: Best practice guidelines for policymakers," *Energy Research & Social Science*, vol. 22, pp. 147–164, Dec. 2016, ISSN: 22146296, DOI: 10.1016/j.erss.2016.08.020.

[143] M. Palinski, "Paying with your data. privacy tradeoffs in ride-hailing services," *Applied Economics Letters*, pp. 1–7, Aug. 2021, ISSN: 1350-4851, DOI: 10.1080/13504851.2021.1959891.

[144] W. Kuhfeld, "%Choiceff macro," *Marketing Research Methods in SAS*, pp. 806–955, 2010, [Online]. Available: http://support.sas.com/techsup/technote/mr2010choiceff.pdf.

[145] E. McKenna, M. Thomson, and J. Barton, *Crest demand model*, 2015, DOI: https://doi.org/10.17028/rd.lboro.2001129.v8.

[146] Office for National Statistics, *Ons population projections*, Oct. 2019, [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopulationprojections/2018based.

[147] UK Government, *Uk government ethnicity facts and figures*, 2011, [Online]. Available: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest.

[148] National Readership Survey, *National readership survey social grades*, 2016, [Online]. Available: http://www.nrs.co.uk/nrs-print/lifestyle-and-classification-data/social-grade/.

[149] BEIS, "Smart meter statistics in great britain: Quarterly report to end december 2020," Department for Business, Energy & Industrial Strategy, Mar. 2021, [Online]. Available: https://assets.publishing.service.gov.uk/govern

ment/uploads/system/uploads/attachment_data/file/968356/Q4_2020_Smart_Meters_Statistics_Reportv2.pdf.

[150] D. for Energy Security and N. Zero, *Subnational estimates of domestic properties not on the gas grid, great britain, 2015 - 2022*, Jan. 2024, [Online]. Available: https://assets.publishing.service.gov.uk/media/65b0be04f2718c0014fb1bdb/Subnational_estimates_of_properties_not_connected_to_the_gas_network_2015-2022.xlsx.

[151] OFGEM, *Standard variable tariff indicators – previous updates: Svt non-ppm customer accounts for the 11 larger suppliers - july 2021*, Jul. 2021, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/2022-06/svt%20non-ppm%20customer%20accounts%20for%20the%2011%20largest%20suppliers%20-%20July%202021.xlsx.

[152] OFGEM, *Vulnerable consumers in the energy market: 2019*, 2019, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2019/09/vulnerable_consumers_in_the_energy_market_2019_final.pdf.

[153] K. E. Train, *Discrete Choice Methods with Simulation*. Cambridge University Press, Jan. 2001, ISBN: 9780521766555, DOI: 10.1017/CBO9780511805271.

[154] A. Daly, S. Hess, and K. Train, "Assuring finite moments for willingness to pay in random coefficient models," *Transportation*, vol. 39, pp. 19–31, 1 Jan. 2012, ISSN: 0049-4488, DOI: 10.1007/s11116-011-9331-3.

[155] S. Hess and D. Palma, "Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application," *Journal of Choice Modelling*, vol. 32, p. 100 170, Sep. 2019, ISSN: 1755-5345, DOI: 10.1016/J.JOCM.2019.100170.

[156] B. Lanz, A. Provins, I. J. Bateman, *et al.*, "Investigating willingness to pay–willingness to accept asymmetry in choice experiments," in *Choice Modelling: The State-of-the-art and The State-of-practice*, Emerald Group Publishing Limited, Jan. 2010, pp. 517–541, ISBN: 978-1-84950-772-1, DOI: 10.1108/9781849507738-024.

[157] K. Train and M. Weeks, "Discrete choice models in preference space and willingness-to-pay space," *Applications of Simulation Methods in Environmental and Resource Economics*, pp. 1–16, Dec. 2005, DOI: 10.1007/1-4020-3684-1_1.

[158] M. C. Bliemer and J. M. Rose, "Confidence intervals of willingness-to-pay for random coefficient logit models," *Transportation Research Part B: Methodological*, vol. 58, pp. 199–214, Dec. 2013, ISSN: 01912615, DOI: 10.1016/j.trb.2013.09.010.

[159] I. Krinsky and A. L. Robb, "On approximating the statistical properties of elasticities," *The Review of Economics and Statistics*, vol. 68, p. 715, 4 Nov. 1986, ISSN: 00346535, DOI: 10.2307/1924536.

[160] S. Hess, K. E. Train, and J. W. Polak, "On the use of a modified latin hypercube sampling (mlhs) method in the estimation of a mixed logit model for vehicle choice," *Transportation Research Part B: Methodological*, vol. 40, pp. 147–163, 2 Feb. 2006, ISSN: 01912615, DOI: 10.1016/j.trb.2004.10.005.

[161] M. Sobolewski, "Measuring consumer well-being from using zero price digital services. the case of navigation apps and location-based services," European Commission, Tech. Rep., 2021, [Online]. Available: https://ec.europa.eu/jrc.

[162] D. S. Bunch, D. M. Gay, and R. E. Welsch, "Algorithm 717: Subroutines for maximum likelihood and quasi-likelihood estimation of parameters in non-linear regression models," *ACM Transactions on Mathematical Software*, vol. 19, pp. 109–130, 1 Mar. 1993, ISSN: 0098-3500, DOI: 10.1145/151271.151279.

[163] G. Glasgow, S. Butler, and S. Iyengar, "Survey response bias and the 'privacy paradox': Evidence from a discrete choice experiment," *Applied Economics Letters*, vol. 28, pp. 625–629, 8 May 2021, ISSN: 1350-4851, DOI: 10.1080/13504851.2020.1770183.

[164] D. Rigby and M. Burton, "Modeling disinterest and dislike: A bounded bayesian mixed logit model of the uk market for gm food," *Environmental & Resource Economics*, vol. 33, pp. 485–509, 4 Apr. 2006, ISSN: 0924-6460, DOI: 10.1007/s10640-005-4995-9.

[165] M. Fosgerau and S. L. Mabit, "Easy and flexible mixture distributions," *Economics Letters*, vol. 120, pp. 206–210, 2 Aug. 2013, ISSN: 0165-1765, DOI: 10.1016/J.ECONLET.2013.03.050.

[166] Information Commisioner's Office, "Guide to the general data protection regulation (gdpr)," Information Commissioners Office, Jan. 2021, [Online]. Available: https://ico.org.uk/media/for-organisations/guide-to-data-pr

otection/guide-to-the-general-data-protection-regulation-gdpr-1-1.pdf.

[167] Which? "Control, alt or delete? the future of consumer data," Which? Tech. Rep., 2018, pp. 1–48, [Online]. Available: https://consumerinsight.which.co.uk/articles/consumer-data-summary.

[168] F. Shirani, C. Groves, K. Henwood, *et al.*, "'i'm the smart meter': Perceptions of smart technology amongst vulnerable consumers.," *Energy Policy*, vol. 144, p. 111 637, Sep. 2020, ISSN: 03014215, DOI: 10.1016/j.enpol.2020.111637.

[169] P. J. van de Waerdt, "Information asymmetries: Recognizing the limits of the gdpr on the data-driven market," *Computer Law & Security Review*, vol. 38, p. 105 436, Sep. 2020, ISSN: 02673649, DOI: 10.1016/j.clsr.2020.105436.

[170] F. Farokhi, "Temporally discounted differential privacy for evolving datasets on an infinite horizon," in *Proceedings - 2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems, ICCPS 2020*, 2020, ISBN: 9781728155012, DOI: 10.1109/ICCPS48487.2020.00008.

[171] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, *et al.*, "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics," ENISA, 2015, DOI: 10.2824/641480.

[172] C. Zimmerman, *Part 3: Privacy-preserving technologies*, Jun. 2022, [Online]. Available: https://the-privacy-blog.eu/2022/06/03/part-3-privacy-preserving-technologies/.

[173] Article 29 Data Protection Working Party, "Opinion 05/2014 on anonymisation techniques," 2014, [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[174] Information Commisioner's Office, "Anonymisation: Managing data protection risk code of practice," Information Commissioner's Office, 2012, [Online]. Available: https://ico.org.uk/media/1061/anonymisation-code.pdf.

[175] M. Elliot, E. Mackey, and K. O'Hara, *Anonymisation Decision Making Framework: European Practitioners' Guide*, 2nd. UK Anonymisation Network, 2020, [Online]. Available: https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf.

[176] B. LLP, "Access to data arrangements: Evaluation," Jun. 2018, [Online]. Available: https://www.ofgem.gov.uk/sites/default/files/docs/2018/07/baringa_report_for_ofgem_enhanced_privacy_evaluation_for_hhs_published_version_2.0_0.pdf.

[177] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications*, vol. 10, p. 3069, 1 Dec. 2019, ISSN: 2041-1723, DOI: 10.1038/s41467-019-10933-3.

[178] S. Cleemput, M. A. Mustafa, E. Marin, *et al.*, "De-pseudonymization of smart metering data: Analysis and countermeasures," in *2018 Global Internet of Things Summit (GIoTS)*, IEEE, Jun. 2018, pp. 1–6, ISBN: 978-1-5386-6451-3, DOI: 10.1109/GIOTS.2018.8534430.

[179] L. Sweeney, "Simple demographics often identify people uniquely," 2000, [Online]. Available: http://dataprivacylab.org/projects/identifiability/.

[180] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *arXiv preprint*, Oct. 2006, [Online]. Available: http://arxiv.org/abs/cs/0610105.

[181] T. Q. Hoan. "A study on privacy level in publishing data of smart tap network." (Jun. 2014), [Online]. Available: https://www.slideshare.net/Eniod/zoro-iece.

[182] California Public Utilities Commission, "The california public utility commission decision 14-05-016," California Public Utilities Commission, 2014, [Online]. Available: https://docs.cpuc.ca.gov/publisheddocs/published/g000/m090/k845/90845985.pdf.

[183] G. Danezis, "Smart meter aggregation assessment: Review of the evidence, prepared for citizen's advice," Aug. 2015.

[184] N. Sheikh, Z. Lu, H. Asghar, *et al.*, "Trace recovery: Inferring fine-grained trace of energy data from aggregates," in *Proceedings of the 18th International Conference on Security and Cryptography*, SCITEPRESS - Science and Technology Publications, 2021, pp. 283–294, ISBN: 978-989-758-524-1, DOI: 10.5220/0010560302830294.

[185] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, pp. 211–407, 3-4 Aug. 2013, ISSN: 1551-305X, DOI: 10.1561/0400000042.

[186] G. Eibl, K. Bao, P.-W. Grassal, *et al.*, "The influence of differential privacy on short term electric load forecasting," *Energy Informatics*, 2018, ISSN: 2520-8942, DOI: 10.1186/s42162-018-0025-3.

[187] D. Desfontaines and B. Pejó, "Sok: Differential privacies," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, pp. 288–313, 2 Apr. 2020, ISSN: 2299-0984, DOI: 10.2478/popets-2020-0028.

[188] J. Hsu, M. Gaboardi, A. Haeberlen, *et al.*, "Differential privacy: An economic method for choosing epsilon," in *2014 IEEE 27th Computer Security Foundations Symposium*, IEEE, Jul. 2014, pp. 398–410, ISBN: 978-1-4799-4290-9, DOI: 10.1109/CSF.2014.35.

[189] Apple Inc., "Learning with privacy at scale," Apple Inc., 2017, [Online]. Available: https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

[190] Google, "Covid-19 community mobility reports," 2021, [Online]. Available: https://www.google.com/covid19/mobility/.

[191] US Census Bureau, "A history of census privacy protections," 2019, [Online]. Available: https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html.

[192] Recurve, "Real world use-cases for energy differential privacy: Using edp to track covid impacts.," *Recurve Blog*, 2021, [Online]. Available: https://www.recurve.com/blog/traditional-approaches-to-protecting-energy-data-dont-work-heres-what-to-do-instead-part-3-of-3.

[193] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? personalized differential privacy," in *2015 IEEE 31st International Conference on Data Engineering*, IEEE, Apr. 2015, pp. 1023–1034, ISBN: 978-1-4799-7964-6, DOI: 10.1109/ICDE.2015.7113353.

[194] G. Ács and C. Castelluccia, "I have a dream! (differentially private smart metering)," in *Lecture Notes in Computer Science*. 2011, pp. 118–132, ISBN: 9783642241772, DOI: 10.1007/978-3-642-24178-9_9.

[195] Smart Energy Code Company, *The smart energy code: Schedule 9*, 2021, [Online]. Available: https://smartenergycodecompany.co.uk/the-smart-energy-code-2/.

[196] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *2010 First IEEE International Conference on Smart Grid Communications*, IEEE, Oct. 2010, pp. 327–332, ISBN: 978-1-4244-6510-1, DOI: 10.1109/SMARTGRID.2010.5622064.

[197] K. Xue, B. Zhu, Q. Yang, *et al.*, "An efficient and robust data aggregation scheme without a trusted authority for smart grid," *IEEE Internet of Things Journal*, vol. 7, pp. 1949–1959, 3 Mar. 2020, ISSN: 2327-4662, DOI: 10.1109/JIOT.2019.2961966.

[198] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Springer Berlin Heidelberg, 1999, pp. 223–238, DOI: 10.1007/3-540-48910-X_16.

[199] C. Gentry, "A fully homomorphic encryption scheme," AAI3382729, Ph.D. dissertation, Stanford, CA, USA, 2009, ISBN: 9781109444506, [Online]. Available: https://crypto.stanford.edu/craig/craig-thesis.pdf.

[200] G. Danezis, C. Fournet, M. Kohlweiss, *et al.*, "Smart meter aggregation via secret-sharing," ACM, Nov. 2013, pp. 75–80, ISBN: 9781450324922, DOI: 10.1145/2516930.2516944.

[201] A. Blanco-Justicia and J. Domingo-Ferrer, "Privacy-preserving computation of the earth mover's distance," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12472 LNCS, pp. 409–423, 2020, ISSN: 16113349, DOI: 10.1007/978-3-030-62974-8_23/FIGURES/10.

[202] IEEE, *Applications of multiparty computation*, [Online]. Available: https://digitalprivacy.ieee.org/publications/topics/applications-of-multiparty-computation#:~:text=As%20stated%20before%2C%20multiparty%20computation,to%20gain%20conclusive%20industry%20insights..

[203] M. A. Mustafa, S. Cleemput, A. Aly, *et al.*, "A secure and privacy-preserving protocol for smart metering operational data collection," *IEEE Transactions on Smart Grid*, vol. 10, pp. 6481–6490, 6 Nov. 2019, ISSN: 19493061, DOI: 10.1109/TSG.2019.2906016.

[204] *Boston women's workforce council report 2016*, Jan. 2017, [Online]. Available: https://www.boston.gov/sites/default/files/document-file-09-2017/bwwcr-2016-new-report.pdf.

[205] G. Kalogridis, C. Efthymiou, S. Z. Denic, *et al.*, "Privacy for smart meters: Towards undetectable appliance load signatures," in *2010 First IEEE International Conference on Smart Grid Communications*, IEEE, Oct. 2010, pp. 232–237, ISBN: 978-1-4244-6510-1, DOI: 10.1109/SMARTGRID.2010.5622047.

[206] P. Vepakomma, O. Gupta, T. Swedish, *et al.*, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint*, Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.00564.

[207] Q. Yang, Y. Liu, T. Chen, *et al.*, "Federated machine learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–19, 2 Feb. 2019, ISSN: 2157-6904, DOI: 10.1145/3298981.

[208] E. Bagdasaryan, A. Veit, Y. Hua, *et al.*, "How to backdoor federated learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2938–2948, [Online]. Available: http://proceedings.mlr.press/v108/bagdasaryan20a.html.

[209] C. Zhang, S. Li, J. Xia, *et al.*, "Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning," in *USENIX Annual Technical Conference*, USENIX Association, Jul. 2020, pp. 493–506, ISBN: 978-1-939133-14-4, [Online]. Available: https://www.usenix.org/conference/atc20/presentation/zhang-chengliang.

[210] C. Gonçalves, R. J. Bessa, and P. Pinson, "A critical overview of privacy-preserving approaches for collaborative forecasting," *International Journal of Forecasting*, vol. 37, pp. 322–342, 1 Jan. 2021, ISSN: 01692070, DOI: 10.1016/j.ijforecast.2020.06.003.

[211] T. Yang, G. Andrew, H. Eichner, *et al.*, "Applied federated learning: Improving google keyboard query suggestions," *arXiv preprint*, Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.02903.

[212] M. J. Sheller, G. A. Reina, B. Edwards, *et al.*, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic*

Brain Injuries, Springer International Publishing, 2019, pp. 92–104, DOI: 10.1007/978-3-030-11723-8_9.

[213] MELLODDY, "Machine learning ledger orchestration for drug discovery," 2019, [Online]. Available: https://www.melloddy.eu/objectives.

[214] A. Taik and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, IEEE, Jun. 2020, pp. 1–6, ISBN: 978-1-7281-5089-5, DOI: 10.1109/ICC40277.2020.9148937.

[215] Y. Wang, I. L. Bennani, X. Liu, *et al.*, "Electricity consumer characteristics identification: A federated learning approach," *IEEE Transactions on Smart Grid*, vol. 12, pp. 3637–3647, 4 Jul. 2021, ISSN: 1949-3053, DOI: 10.1109/TSG.2021.3066577.

[216] D. Byrd and A. Polychroniadou, "Differentially private secure multi-party computation for federated learning in financial applications," Association for Computing Machinery, 2021, ISBN: 9781450375849, DOI: 10.1145/3383455.3422562.

[217] M. B. Hawes, "Implementing differential privacy: Seven lessons from the 2020 united states census," *Harvard Data Science Review*, vol. 2, 2 Apr. 2020, DOI: 10.1162/99608f92.353c6f99.

[218] US Census Bureau, "Comparing differential privacy with older disclosure avoidance methods," 2021, [Online]. Available: https://www.census.gov/library/fact-sheets/2021/comparing-differential-privacy-with-older-disclosure-avoidance-methods.html.

[219] M. B. Hawes, "Differential privacy and the 2020 census: Modernizing disclosure avoidance at scale to mitigate growing privacy threats," *Zenodo*, Jan. 2020, DOI: 10.5281/ZENODO.4122103.

[220] M. Hawes, *Understanding the 2020 census disclosure avoidance system: Simulated reconstruction-abetted re-identification attack on the 2010 census*, May 2021.

[221] Government Statistical Service, "Privacy and data confidentiality methods: A data and analysis method review," Dec. 2018, [Online]. Available: https://gss.civilservice.gov.uk/policy-store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr/.

[222] S. L. Garfinkel, J. M. Abowd, and S. Powazek, "Issues encountered deploying differential privacy," in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, ACM, Jan. 2018, pp. 133–137, ISBN: 9781450359894, DOI: 10.1145/3267323.3268949.

[223] J. M. Abowd, "Disclosure avoidance for block level data and protection of confidentiality in public tabulations," 2018, [Online]. Available: https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf.

[224] US Census Bureau, "Key parameters set to protect privacy in 2020 census results," Jun. 2021, [Online]. Available: https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html.

[225] J. Mervis, "Can a set of equations keep u.s. census data private?" *Science*, Jan. 2019, ISSN: 0036-8075, DOI: 10.1126/science.aaw5470.

[226] Brennan Center for Justice, "Court rejects alabama challenge to census plans for redistricting and privacy," 2021, [Online]. Available: https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy.

[227] S. Petti and A. Flaxman, "Differential privacy in the 2020 us census: What will it do? quantifying the accuracy/privacy tradeoff," *Gates Open Research*, vol. 3, p. 1722, Apr. 2020, ISSN: 2572-4754, DOI: 10.12688/gatesopenres.13089.2.

[228] Sustainability First and CSE, "Smart meter energy data: Public interest advisory group (paig). stimulus paper 7 - possible routes to the data for a public interest.," Sustainability First and CSE, Jan. 2019, [Online]. Available: https://www.smartenergydatapiag.org.uk/.

[229] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, *et al.*, "A critical review on the use (and misuse) of differential privacy in machine learning," *ACM Computing Surveys*, vol. 55, pp. 1–16, 8 Aug. 2023, ISSN: 0360-0300, DOI: 10.1145/3547139.

[230] J. Tang, A. Korolova, X. Bai, *et al.*, "Privacy loss in apple's implementation of differential privacy on macos 10.12," *arXiv preprint*, 2019, DOI: 10.48550/arXiv.1709.02753.

[231] US Census Bureau, "Github 2020 census repository," 2019, [Online]. Available: https://github.com/uscensusbureau/census2020-das-2010ddp.

[232] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!" *Journal of Privacy and Confidentiality*, vol. 9, 2 Oct. 2019, ISSN: 2575-8527, DOI: 10.29012/jpc.689.

[233] US Census Bureau, "The "72-year rule"," 2020, [Online]. Available: https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html.

[234] J. Liu, J. Lou, J. Liu, *et al.*, "Dealer," *Proceedings of the VLDB Endowment*, vol. 14, pp. 957–969, 6 Feb. 2021, ISSN: 2150-8097, DOI: 10.14778/3447689.3447700.

[235] S. W. Driessen, G. Monsieur, and W.-J. V. D. Heuvel, "Data market design: A systematic literature review," *IEEE Access*, vol. 10, pp. 33 123–33 153, 2022, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2022.3161478.

[236] C. Feng, Y. Wang, K. Zheng, *et al.*, "Smart Meter Data-Driven Customizing Price Design for Retailers," *IEEE Transactions on Smart Grid*, 2020, ISSN: 19493061, DOI: 10.1109/TSG.2019.2946341.

[237] J. Schofield, R. Carmichael, S. Tindemans, *et al.*, *Report a3 low carbon london learning lab - residential consumer responsiveness to time-varying pricing*, 2014, DOI: 10.13140/RG.2.1.3155.0887.

[238] M. Sun, Y. Wang, F. Teng, *et al.*, "Clustering-based residential baseline estimation: A probabilistic perspective," *IEEE Transactions on Smart Grid*, vol. 10, pp. 6014–6028, 6 Nov. 2019, ISSN: 1949-3053, DOI: 10.1109/TSG.2019.2895333.

[239] Elexon, "Assurance in a smart metered world," Elexon, Tech. Rep., 2011, [Online]. Available: https://www.elexon.co.uk/wp-content/uploads/2011/10/Assurance_Smart_Metered_World.pdf.

[240] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential Privacy Techniques for Cyber Physical Systems: A Survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 746–789, 2020, ISSN: 1553-877X, DOI: 10.1109/COMST.2019.2944748.

[241] S. Thorve, L. Kotut, and M. Semaan, "Privacy Preserving Smart Meter Data," *The 7th International Workshop on Urban Computing (UrbComp)*, 2018, [Online]. Available: https://faculty.washington.edu/kotut/papers/UrbComp18-privacy-preserving-smart-meter-data.pdf.

[242] B. Stölb and J. Domingo-Ferrer, "Protecting Consumer Privacy in Smart Metering by Randomized Response," in *Conference of European Statisticians*, 2019, [Online]. Available: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S4_URV_Protecting_Consumer_Privacy_AD.pdf.

[243] H. Wang and C. Wu, "Understanding Differential Privacy in Non-Intrusive Load Monitoring," in *e-Energy 2020 - Proceedings of the 11th ACM International Conference on Future Energy Systems*, New York, NY, USA: Association for Computing Machinery, Inc, Jun. 2020, pp. 401–403, ISBN: 9781450380096, DOI: 10.1145/3396851.3403508.

[244] M. Abadi, A. Chu, I. Goodfellow, *et al.*, "Deep learning with differential privacy," ACM, Oct. 2016, pp. 308–318, ISBN: 9781450341394, DOI: 10.1145/2976749.2978318.

[245] V. Dvorkin, F. Fioretto, P. V. Hentenryck, *et al.*, "Differentially private optimal power flow for distribution grids," *IEEE Transactions on Power Systems*, vol. 36, pp. 2186–2196, 3 May 2021, ISSN: 0885-8950, DOI: 10.1109/TPWRS.2020.3031314.

[246] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, "The limits of differential privacy (and its misuse in data release and machine learning)," *Communications of the ACM*, vol. 64, pp. 33–35, 7 Jul. 2021, ISSN: 0001-0782, DOI: 10.1145/3433638.

[247] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, IEEE, 2007, pp. 94–103, DOI: 10.1109/FOCS.2007.66.

[248] J. W. Kim, B. Jang, and H. Yoo, "Privacy-preserving aggregation of personal health data streams," *PLoS ONE*, 2018, ISSN: 19326203, DOI: 10.1371/journal.pone.0207639.

[249] M. Alaggan, S. Gambs, and A.-M. Kermarrec, "Heterogeneous Differential Privacy," *Journal of Privacy and Confidentiality*, 2017, DOI: 10.29012/jpc.v7i2.652.

[250] Y. Wang, Q. Chen, M. Sun, *et al.*, "An ensemble forecasting method for the aggregated load with subprofiles," *IEEE Transactions on Smart Grid*, vol. 9, pp. 3906–3908, 4 Jul. 2018, ISSN: 1949-3053, DOI: 10.1109/TSG.2018.2807985.

[251] S. I. Vagropoulos, E. G. Kardakos, C. K. Simoglou, *et al.*, "Artificial neural network-based methodology for short-term electric load scenario generation," in *2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*, 2015, pp. 1–6, DOI: 10.1109/ISAP.2015.7325540.

[252] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine Learning in {P}ython," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, DOI: 10.48550/arXiv.1201.0490.

[253] M. Song and M. Amelin, "Price-Maker Bidding in Day-Ahead Electricity Market for a Retailer With Flexible Demands," *IEEE Trans. on Power Syst.*, vol. 33, no. 2, pp. 1948–1958, 2018, ISSN: 1558-0679, DOI: 10.1109/TPWRS.2017.2741000.

[254] Elexon, *Balancing mechanism reporting service (BMRS)*, [Online]. Available: https://www.bmreports.com.

[255] H. P. Williams, "Building integer programming models I," in *Model Building in Mathematical Programming*. Wiley, 2013, ISBN: 9781118506189, [Online]. Available: https://www.wiley.com/en-in/Model+Building+in+Mathematical+Programming%2C+5th+Edition-p-9781118506172.

[256] S. Roberts and W. Penny, "Variational bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002, DOI: 10.1109/TSP.2002.801921.

[257] Commision for Energy Regulation, *CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010.* Irish Social Science Data Archive, 2012, [Online]. Available: www.ucd.ie/issda/CER-electricity.

[258] S. Pfenninger and I. Staffell, "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data," *Energy*, vol. 114, pp. 1251–1265, 2016, ISSN: 0360-5442, DOI: https://doi.org/10.1016/j.energy.2016.08.060.

[259] T. Dodson and S. Slater, "Electric Vehicle Charging Behaviour Study, Final Report for National Grid ESO," Element Energy, Tech. Rep., 2019, [Online]. Available: https://www.element-energy.co.uk/wordpress/wp-content/uploads/2019/04/20190329-NG-EV-CHARGING-BEHAVIOUR-STUDY-FINAL-REPORT-V1-EXTERNAL.pdf.

[260] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation*, New York, NY, USA: Association for Computing Machinery, Inc, Jun. 2019, pp. 701–726, ISBN: 9781450367929, DOI: 10.1145/3328526.3329589.

[261] J. Liu, "Dealer: End-to-End Data Marketplace with Model-based Pricing," *arXiv e-prints*, Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.13103.

[262] C. Goncalves, R. J. Bessa, and P. Pinson, "Privacy-preserving distributed learning for renewable energy forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, pp. 1777–1787, 3 Jul. 2021, ISSN: 1949-3029, DOI: 10.1109/TSTE.2021.3065117.

[263] T. Falconer, J. Kazempour, and P. Pinson, "Bayesian regression markets," Oct. 2023, DOI: 10.48550/arXiv.2310.14992.

[264] A. M. Piotrowska and M. Klonowski, "Some remarks and ideas about monetization of sensitive data," in Guillermo, A. Alessandro, M. Fabio, *et al.*, Eds. Springer International Publishing, 2016, pp. 118–133, ISBN: 978-3-319-29883-2, DOI: 10.1007/978-3-319-29883-2_8.

[265] Y. Jiao, P. Wang, D. Niyato, *et al.*, "Toward an automated auction framework for wireless federated learning services market," Dec. 2019, DOI: 10.1109/TMC.2020.2994639.

[266] R. Mieth, J. M. Morales, and H. V. Poor, "Data valuation from data-driven optimization," May 2023, [Online]. Available: http://arxiv.org/abs/2305.01775.

[267] L. Chen, Z. Wu, J. Wang, *et al.*, "Toward future information market: An information valuation paradigm," in *2021 IEEE Power & Energy Society General Meeting (PESGM)*, IEEE, Jul. 2021, pp. 1–5, ISBN: 978-1-6654-0507-2, DOI: 10.1109/PESGM46819.2021.9638205.

[268] B. Wang, Q. Guo, and Y. Yu, "Mechanism design for data sharing: An electricity retail perspective," *Applied Energy*, vol. 314, p. 118 871, May 2022, ISSN: 03062619, DOI: 10.1016/j.apenergy.2022.118871.

[269] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, pp. 419–435, 3 Dec. 2002, ISSN: 0306-7734, DOI: 10.1111/j.1751-5823.2002.tb00178.x.

[270] J. Cho and C. Suh, "Wasserstein gan can perform pca," IEEE, Sep. 2019, pp. 895–901, ISBN: 978-1-7281-3151-1, DOI: 10.1109/ALLERTON.2019.8919827.

[271] Y. Cai and L. .-. Lim, "Distances between probability distributions of different dimensions," *IEEE Transactions on Information Theory*, vol. 68, pp. 4020–4031, 6 2022, ISSN: 1557-9654, DOI: 10.1109/TIT.2022.3148923.

[272] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., vol. 70, PMLR, Mar. 2017, pp. 214–223, [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html.

[273] A. T. Lopez and V. Jog, "Generalization error bounds using wasserstein distances," in *2018 IEEE Information Theory Workshop (ITW)*, IEEE, Nov. 2018, pp. 1–5, ISBN: 978-1-5386-3599-5, DOI: 10.1109/ITW.2018.8613445.

[274] S. A. Hosseini, J.-F. Toubeau, N. Amjady, *et al.*, "Day-ahead wind power temporal distribution forecasting with high resolution," *IEEE Transactions on Power Systems*, pp. 1–11, 2023, ISSN: 0885-8950, DOI: 10.1109/TPWRS.2023.3295915.

[275] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual Review of Statistics and Its Application*, 2019, DOI: 10.1146/annurev-statistics.

[276] O. Besbes, W. Ma, and O. Mouchtaki, "Beyond iid: Data-driven decision-making in heterogeneous environments," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.09642.

[277] M. Agueh and G. Carlier, "Barycenters in the wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, pp. 904–924, 2 Jan. 2011, ISSN: 0036-1410, DOI: 10.1137/100805741.

[278] H. Gouk, E. Frank, B. Pfahringer, *et al.*, "Regularisation of neural networks by enforcing lipschitz continuity," *Machine Learning*, vol. 110, pp. 393–416, 2 Feb. 2021, ISSN: 15730565, DOI: 10.1007/S10994-020-05929-W/FIGURES/4.

[279] T. T.-K. Lau and H. Liu, "Wasserstein distributionally robust optimization with wasserstein barycenters," Mar. 2022, DOI: 10.48550/arXiv.2203.12136.

[280] Y. Yan, L. J. Chen, and Z. Zhang, "Error-bounded sampling for analytics on big sparse data," *Proceedings of the VLDB Endowment*, vol. 7, pp. 1508–1519, 13 Aug. 2014, ISSN: 2150-8097, DOI: 10.14778/2733004.2733022.

[281] C. Villani, "The wasserstein distances," in *Optimal Transport: Old and New*, Springer Berlin Heidelberg, 2009, pp. 93–111, ISBN: 978-3-540-71050-9, DOI: 10.1007/978-3-540-71050-9_6.

[282] M. D. Angelis and A. Gray, "Why the 1-wasserstein distance is the area between the two marginal cdfs," *arXiv*, Nov. 2021, DOI: 10.48550/arxiv.2111.03570.

[283] R. Flamary, N. Courty, A. Gramfort, *et al.*, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021, [Online]. Available: http://jmlr.org/papers/v22/20-451.html.

[284] M. Tsagris, C. Beneki, and H. Hassani, "On the folded normal distribution," *Mathematics 2014, Vol. 2, Pages 12-28*, vol. 2, pp. 12–28, 1 Feb. 2014, ISSN: 2227-7390, DOI: 10.3390/MATH2010012.

[285] F. Farokhi, "Distributionally-robust machine learning using locally differentially-private data," *Optimization Letters*, vol. 16, pp. 1167–1179, 4 May 2022, ISSN: 18624480, DOI: 10.1007/S11590-021-01765-6/FIGURES/4.

[286] Y. Liu and T. J. Kozubowski, "A folded laplace distribution," *Journal of Statistical Distributions and Applications*, vol. 2, pp. 1–17, 1 Dec. 2015, ISSN: 21955832, DOI: 10.1186/S40488-015-0033-9/FIGURES/3.

[287] W. J. Reed, "The normal-laplace distribution and its relatives," in *Advances in Distribution Theory, Order Statistics, and Inference*, Birkhäuser Boston, 2006, pp. 61–74, DOI: 10.1007/0-8176-4487-3_4.

[288] D. Morales, I. Agudo, and J. Lopez, *Private set intersection: A systematic literature review*, Aug. 2023, DOI: 10.1016/j.cosrev.2023.100567.

[289] A. A. Jolfaei, H. Mala, and M. Zarezadeh, "Eo-psi-ca: Efficient outsourced private set intersection cardinality," *Journal of Information Security and Applications*, vol. 65, p. 102 996, Mar. 2022, ISSN: 2214-2126, DOI: 10.1016/J.JISA.2021.102996.

[290] R. B. Christensen, S. R. Pandey, and P. Popovski, "Semi-private computation of data similarity with applications to data valuation and pricing," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1978–1988, 2023, ISSN: 1556-6013, DOI: 10.1109/TIFS.2023.3259879.

[291] M. Purcell, Y. Li, and K. S. Ng, "Split, count, and share: A differentially private set intersection cardinality estimation protocol," in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, R. J. Evans and I. Shpitser, Eds., ser. Proceedings of Machine Learning Research, vol. 216, PMLR, Aug. 2023, pp. 1684–1694, [Online]. Available: https://proceedings.mlr.press/v216/purcell23a.html.

[292] O. Nohadani and K. Sharma, "Optimization under decision-dependent uncertainty," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1773–1795, Jun. 2018, ISSN: 10526234, DOI: 10.1137/17M1110560.

[293] R. Xie, P. Pinson, and Y. Chen, "Robust scheduling with improved uncertainty sets via purchase of distributed predictive information," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.00291.

[294] A. M. Kharman, C. Jursitzky, Q. Zhou, *et al.*, "An adversarially robust data-market for spatial, crowd-sourced data," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.06299.

[295] L. Watson, R. Andreeva, H.-T. Yang, *et al.*, "Differentially private shapley values for data evaluation," Jun. 2022, DOI: 10.48550/arXiv.2206.00511.

[296] R. Jia, D. Dao, B. Wang, *et al.*, "Towards efficient data valuation based on the shapley value," K. Chaudhuri and M. Sugiyama, Eds., vol. 89, PMLR, Nov. 2019, pp. 1167–1176, [Online]. Available: https://proceedings.mlr.press/v89/jia19a.html.

[297] B. Wang, Q. Guo, T. Yang, *et al.*, "Data valuation for decision-making with uncertainty in energy transactions: A case of the two-settlement market system," *Applied Energy*, vol. 288, p. 116 643, Apr. 2021, ISSN: 03062619, DOI: 10.1016/j.apenergy.2021.116643.

[298] L. Ensthaler and T. Giebe, "Bayesian optimal knapsack procurement," *European Journal of Operational Research*, vol. 234, pp. 774–779, 3 May 2014, ISSN: 03772217, DOI: 10.1016/j.ejor.2013.09.031.

[299] A. A. Raja, P. Pinson, J. Kazempour, *et al.*, "A market for trading forecasts: A wagering mechanism," *International Journal of Forecasting*, Feb. 2023, ISSN: 01692070, DOI: 10.1016/j.ijforecast.2023.01.007.

[300] F. Moret and P. Pinson, "Energy collectives: A community and fairness based approach to future electricity markets," *IEEE Transactions on Power Systems*, vol. 34, pp. 3994–4004, 5 Sep. 2019, ISSN: 0885-8950, DOI: 10.1109/TPWRS.2018.2808961.

[301] F. Jarman and V. Meisner, "Ex-post optimal knapsack procurement," *Journal of Economic Theory*, vol. 171, pp. 35–63, Sep. 2017, ISSN: 00220531, DOI: 10.1016/j.jet.2017.06.001.

[302] R. B. Myerson, "Optimal auction design," *Mathematics of Operations Research*, vol. 6, pp. 58–73, 1 Feb. 1981, ISSN: 0364-765X, DOI: 10.1287/moor.6.1.58.

[303] F. Glover, "Improved linear integer programming formulations of nonlinear integer problems," *Management Science*, vol. 22, pp. 455–460, 4 Dec. 1975, ISSN: 0025-1909, DOI: 10.1287/mnsc.22.4.455.

[304] D. Pisinger, "A fast algorithm for strongly correlated knapsack problems," *Discrete Applied Mathematics*, vol. 89, pp. 197–212, 1-3 Dec. 1998, ISSN: 0166218X, DOI: 10.1016/S0166-218X(98)00127-9.

[305] K. Bhawalkar and T. Roughgarden, "Welfare guarantees for combinatorial auctions with item bidding," Society for Industrial and Applied Mathematics, Jan. 2011, pp. 700–709, ISBN: 978-0-89871-993-2, DOI: 10.1137/1.9781611973082.55.

[306] D. Acemoglu, A. Makhdoumi, A. Malekian, *et al.*, "Too Much Data: Prices and Inefficiencies in Data Markets," National Bureau of Economic Research, Cambridge, MA, Tech. Rep., Sep. 2019, DOI: 10.3386/w26296.

[307] S. R. Pandey, P. Pinson, and P. Popovski, "Privacy-aware data acquisition under data similarity in regression markets," Dec. 2023.

[308] M. Yan and F. Teng, "Towards joint electricity and data trading: A scalable cooperative game theoretic approach," Oct. 2022, DOI: 10.48550/arxiv.2210.04051.

[309] S. Lee, H. Kim, and I. Moon, "A data-driven distributionally robust newsvendor model with a wasserstein ambiguity set," *Journal of the Operational Research Society*, vol. 72, pp. 1879–1897, 8 Aug. 2021, ISSN: 0160-5682, DOI: 10.1080/01605682.2020.1746203.

[310] J. Huber, S. Müller, M. Fleischmann, *et al.*, "A data-driven newsvendor problem: From data to decision," *European Journal of Operational Research*, vol. 278, pp. 904–915, 3 Nov. 2019, ISSN: 03772217, DOI: 10.1016/j.ejor.2019.04.043.

[311] R. Levi, G. Perakis, and J. Uichanco, "The data-driven newsvendor problem: New bounds and insights," *Operations Research*, vol. 63, pp. 1294–1306, 6 Dec. 2015, ISSN: 0030-364X, DOI: 10.1287/opre.2015.1422.

[312] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, pp. 914–938, 3 Jul. 2016, ISSN: 01692070, DOI: 10.1016/j.ijforecast.2015.11.011.

[313] G.-Y. Ban and C. Rudin, "The big data newsvendor: Practical insights from machine learning," *Operations Research*, vol. 67, pp. 90–108, 1 Jan. 2019, ISSN: 0030-364X, DOI: 10.1287/opre.2018.1757.

[314] C. Liu, A. N. Letchford, and I. Svetunkov, "Newsvendor problems: An integrated method for estimation and optimisation," *European Journal of Operational Research*, vol. 300, pp. 590–601, 2 Jul. 2022, ISSN: 03772217, DOI: 10.1016/j.ejor.2021.08.013.

[315] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, pp. 707–738, 3-4 Aug. 2015, ISSN: 0178-8051, DOI: 10.1007/s00440-014-0583-7.

[316] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, pp. 115–166, 1-2 Sep. 2018, ISSN: 14364646, DOI: 10.1007/s10107-017-1172-1.

[317] J.-y. Gotoh, M. J. Kim, and A. E. B. Lim, "A data-driven approach to beating saa out-of-sample," May 2021, [Online]. Available: http://arxiv.org/abs/2105.12342.

[318] A. F. Siegel and M. R. Wagner, "Profit estimation error in the newsvendor model under a parametric demand distribution," *Management Science*, vol. 67, pp. 4863–4879, 8 Aug. 2021, ISSN: 0025-1909, DOI: 10.1287/mnsc.2020.3766.

[319] J. Liu, "Absolute shapley value," vol. 12, xxxx–yyyy, Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.10076.

[320] Y. Peng, Y. Wang, X. Lu, *et al.*, "Short-term load forecasting at different aggregation levels with predictability analysis," IEEE, May 2019, pp. 3385–3390, ISBN: 978-1-7281-3520-5, DOI: 10.1109/ISGT-Asia.2019.8881343.

[321] J. Amat Rodrigo and J. Escobar Ortiz, *Skforecast*, version 0.11.0, Nov. 2023, DOI: 10.5281/zenodo.8382788.

[322] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[323] J. ya Gotoh and Y. Takano, "Newsvendor solutions via conditional value-at-risk minimization," *European Journal of Operational Research*, vol. 179, pp. 80–96, 1 May 2007, ISSN: 03772217, DOI: 10.1016/j.ejor.2006.03.022.

[324] T. Hong, J. Xie, and J. Black, "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting," *International Journal of Forecasting*, vol. 35, pp. 1389–1399, 4 Oct. 2019, ISSN: 01692070, DOI: 10.1016/j.ijforecast.2019.02.006.

[325] J. Browell, S. Haglund, H. Kälvegren, *et al.*, *Hybrid energy forecasting and trading competition*, 2023, DOI: 10.21227/5hn0-8091.

[326] M. Huber, F. Boutros, A. T. Luu, *et al.*, "Syn-mad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data," IEEE, Oct. 2022, pp. 1–10, ISBN: 978-1-6654-6394-2, DOI: 10.1109/IJCB54206.2022.10007950.

[327] K. Grining and M. Klonowski, "Towards extending noiseless privacy-dependent data and more practical approach," 2017, ISBN: 9781450349444, DOI: 10.1145/3052973.3052992.

[328] L. Ambrosio, N. Gigli, and G. Savare, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, 1st ed. Birkhäuser-Verlag, 2005, ISBN: 3-7643-2428-7, DOI: 10.1007/b137080.

[329] J. Zhu, J. Qiu, A. Guha, *et al.*, "Interpolation for robust learning: Data augmentation on wasserstein geodesics," Feb. 2023.

[330] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, *et al.*, "Wide consensus aggregation in the wasserstein space. application to location-scatter families," *Bernoulli*, vol. 24, 4A Nov. 2018, ISSN: 1350-7265, DOI: 10.3150/17-BEJ957.

[331] A. Eden, M. Feldman, O. Friedler, *et al.*, "A simple and approximately optimal mechanism for a buyer with complements," *Operations Research*, vol. 69, pp. 188–206, 1 Jan. 2021, ISSN: 0030-364X, DOI: 10.1287/opre.2020.2039.

[332] G. Barnicoat and M. Danson, "The ageing population and smart metering: A field study of householders' attitudes and behaviours towards energy use in scotland," *Energy Research & Social Science*, vol. 9, pp. 107–115, Sep. 2015, ISSN: 22146296, DOI: 10.1016/j.erss.2015.08.020.

[333] V. Bianco, O. Manca, and S. Nardini, "Electricity consumption forecasting in italy using linear regression models," *Energy*, vol. 34, pp. 1413–1421, 9 Sep. 2009, ISSN: 03605442, DOI: 10.1016/j.energy.2009.06.034.

[334] M. Salani, M. Derboni, D. Rivola, *et al.*, "Non intrusive load monitoring for demand side management," *Energy Informatics*, vol. 3, p. 25, S1 Oct. 2020, ISSN: 2520-8942, DOI: 10.1186/s42162-020-00128-2.

[335] K. Cooray, S. Gunasekera, and M. M. A. Ananda, "The folded logistic distribution," *Communications in Statistics - Theory and Methods*, vol. 35, pp. 385–393, 3 Apr. 2006, ISSN: 0361-0926, DOI: 10.1080/03610920500476234.

[336] S. Psarakis and J. Panaretoes, "The folded t distribution," *Communications in Statistics - Theory and Methods*, vol. 19, pp. 2717–2734, 7 Jan. 1990, ISSN: 0361-0926, DOI: 10.1080/03610929008830342.

[337] D. Chafaï, *Wasserstein distance between two gaussians*, Apr. 2010, [Online]. Available: https://djalil.chafai.net/blog/2010/04/30/wasserstein-distance-between-two-gaussians/.

[338] M. Fréchet, "Sur la distance de deux lois de probabilité," *Annales de l'ISUP*, vol. VI, no. 3, pp. 183–198, 1957, [Online]. Available: https://hal.science/hal-04093677.

[339] D. J. Sutherland, *Earth mover's distance (emd) between two gaussians*, Cross Validated, (version: 2019-03-26), eprint: https://stats.stackexchange.com/q/144896, [Online]. Available: https://stats.stackexchange.com/q/144896.

[340] D. Chafaï and F. Malrieu, "On fine properties of mixtures with respect to concentration of measure and sobolev type inequalities," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 46, pp. 72–96, 1 Feb. 2010, DOI: 10.1214/08-AIHP309.

[341] L. An, A. A. Li, B. Moseley, *et al.*, "The nonstationary newsvendor with (and without) predictions," May 2023, DOI: 10.48550/arXiv.2305.07993.

APPENDIX A
Supplementary Tables for Benefit and Privacy Risk Mapping

Table A.1: Identifiable Socio-Demographic Information at Different Data Resolutions.

Type	< 1 hour	Daily	Monthly	Yearly
No. of Residents	[5], [124], [125]			[122], [123], [126], [333]
Residents Age	[5], [124], [125]			
Marital Status	[5], [124], [125]			
Employment Status	[5], [124], [125], [128]			
Long-term Illness	[5], [124], [125]			
Household Income	[5], [124], [125]			
Children and Pets	[5], [124], [125]			[122], [123], [126], [333]
Occupancy	[125], [127]			[122], [123], [126], [333]
House Type	[5], [124], [125]			[122], [123]
No. of Rooms	[5], [124], [125]			
Size of House	[5], [124], [125]			
House Location				
House Ownership		[116]		[122], [123]

Note: Considered identifiable if accuracy is greater 50%. Summarised from [5], [116], [122]–[128].

Table A.2: Breakdown of Benefits Dependence on High-Resolution Data.

Benefit	Function	BEIS Estimate	High-Resolution Dependence			Description
			Dependent	Enhanced	Independent	
Consumer		**7,623**	**0**	**6,247**	**1,376**	
Energy Savings	Energy Savings	6,247	0	6,247	0	Real-time feedback can be provided via IHD without sharing. Personalised and detailed recommendation, which may require sharing, could increase engagement and persistence, enhancing savings.
Time Savings	Automated Readings	1,376	0	0	1,376	Time savings relate to remote and automated readings which could be sent at monthly or lower resolution basis.
Supplier		**8,069**	**260**	**1,701**	**6,108**	
Avoided Site Visits	Automated Readings	2,327	0	0	2,327	Automated readings can be taken on a monthly or less granular resolution as up to 13 months of data is stored on the smart meter.
Customer Switching	Automated Readings	1,249	0	0	1,249	Account closure can be achieved using a single reading and switching can be done remotely.

continued . . .

. . . continued

Benefit	Function	BEIS Estimate	High-Resolution Dependence			Description
			Dependent	Enhanced	Independent	
Customer Calls	Automated Readings	1,242	0	0	1,242	Assumed to accrue from accurate billing resulting in less customer calls. This can be achieved using low resolution data.
Avoided PPM Premium	Automated Readings	1,119	0	0	1,119	Benefits based on more payment options available for pre-paid customers. This is independent of data sharing.
Debt Handling	Automated Readings	1,051	0	1,051	0	Projected to improve through more frequent billing, earlier identification of debt build-up and providing faster follow-up action to help consumers. Could be enhanced with access to higher resolution data.
Reduced Theft and Losses	Network Operations	910	260	650	0	Access to granular data allows the identification of patterns of behaviour that may indicate theft (£260 mln). High-resolution could also enhance operational efficiency (settlement costs through better forecasting) and reduce losses (due to demand shifting) (£650 mln).

continued . . .

. . . continued

Benefit	Function	BEIS Estimate	High-Resolution Dependence			Description
			Dependent	Enhanced	Independent	
Remote Change of Tariff	Automated Readings	171	0	0	171	Smart meters have tariff registers that can be updated remotely.
System		**1,738**	**375**	**1,363**	**0**	
Demand Shifting	Demand Shifting	1,363	0	1,363	0	Time-of-use tariffs do not require sharing high resolution data as smart meters have tariff registers. More complex demand response schemes may require access to high resolution data to verify demand changes or for personalised tariffs.
Investment Decisions	Network Operations	209	209	0	0	High-resolution data would be required to better determine intra-day fluctuations in load and decide on reinforcement requirements.
Outage Management	Network Operations	166	166	0	0	High-resolution data allows DNOs to better identify faults in the network, restore electricity supply more quickly when outages occur.

continued . . .

. . . continued

Benefit	Function	BEIS Estimate	High-Resolution Dependence			Description
			Dependent	Enhanced	Independent	
Environmental		**2,027**	**0**	**2,027**	**0**	
Reduced Emissions	Energy Savings	1,633	0	1,633	0	Carbon emissions reductions relate directly to the expected reduction in usage from informational feedback.
Air Quality	Energy Savings	394	0	394	0	Air quality improvements relate directly to the expected reduction in usage from informational feedback.

Note: BEIS estimates from [19, Table 11 p. 63]. All values represent net present value as of 2019 and are reported in £ millions.

Table A.3: Maximum Reported Accuracy of NILM Algorithms for Different Temporal and Spatial Resolutions.

Type	Temporal Resolution						Spatial Resolution		
	< 1 sec	1 sec	1 min	15-30 min	1 hour	Daily	~10 Houses	~100 Houses	~1000 Houses
Small Appliances		83%[100], [112]	70%[110]	20%[110]					
Cooking Appliances	97% [110]	83%[100], [112]	95%[101]				0.65[117]		
Heating & Cooling	95%[110]	83%[100], [112]	95%[101]	80% [104]	60%[21]	82%[108]	89% [334], 0.7[117]		[119], [120]
Electric Vehicles		94.5%[114]	92.2%[107]		75%[111]		0.735[117]	[118]	
Distributed Generation	96%[109]			70%[115]	38%[113]			80.2%[118]	94% [121]

Note: Maximum reported accuracy and source for given data resolution of smart meter data used. Values are given in percentages for accuracy of actual consumption or as a fraction for average correct identification rate when consumption accuracy is not available. When neither of these figures is available only the citation is provided indicating the appliance is identifiable at the given data resolution. Temporal resolution accuracy based on review of [21], [100], [101], [104], [106]–[116]. Spatial resolution indicates order of magnitude (e.g. an aggregation of 200 houses is ~ 100 houses). Spatial resolution accuracy based on review of [117]–[121], [334].

APPENDIX B

Supplementary Survey Information

Table B.1: Full Sample Statistics

		Control (n = 477)		Treatment (n = 488)		GB
		n	%	n	%	%
Age[1]	18-34	123	25.8	134	27.5	28
	35-54	157	32.9	172	35.2	34
	55-64	75	15.7	76	15.6	15
	65+	119	24.9	105	21.5	23
	Refused	3	0.6	1	0.2	
Gender[1]	Male	232	48.6	237	48.6	49
	Female	241	50.5	248	50.8	51
	Refused	4	0.8	3	0.6	
Ethnicity[2]	White	421	88.3	425	87.1	86
	Asian	30	6.3	20	6.8	8
	Black	13	2.7	9	2.3	3
	Mixed	9	1.9	9	2.9	2
	Other	4	0.8	3	0.8	1
	Refused	0	0.0	0	0.0	
Socio-Economic Group[3]	AB	120	25.2	116	23.8	27
	C1	127	26.6	131	26.8	28
	C2	90	18.9	110	22.5	20
	DE	135	28.3	123	25.2	25
	Refused	5	1.0	8	1.6	
Region[1]	England	410	86.0	388	79.5	87
	Wales	20	4.2	27	5.5	5
	Scotland	32	6.7	42	8.6	8
	Refused	0	0.0	0	0.0	
Smart Meter Ownership[4]	Yes	248	42.2	272	52.0	44
	No	226	47.4	206	42.2	56
	Don't Know	3	0.6	10	2.0	

Note: Percentages may not add up due to rounding. [1]2018 ONS Population Projections[146]. [2]2011 UK Government Ethnicity Facts and Figures[147]. [3]National Readership Survey Social Grades[148]. [4]December 2020 Quarterly Smart Meter Statistics[149, Table 5a]. Includes meters in smart and traditional mode.

Table B.2: Full Sample Electricity Supply Characteristics

		Control (n = 477)		Treatment (n = 488)		GB[1]
		n	%	n	%	%
Fuel Type	Dual Fuel	357	74.8	325	66.6	84.9
	Electricity Only	102	21.4	136	27.9	15.1
	Don't Know	18	3.8	27	5.5	
Tariff	Standard Variable	106	22.2	108	22.1	36.4
	Fixed	221	46.3	205	42.0	28.7
	Pre-Payment	51	10.7	71	14.5	16.7[2]
	Time-of-Use	19	4.0	17	3.5	18.3[3]
	Economy 7/10	24	5.0	34	7.0	
	Don't Know	56	11.7	53	10.9	

Note: Percentages may not add up due to rounding.

[1] Fuel type data based on proportion of domestic customers connected to the gas grid in 2021[150]. Tariff data from OFGEM for April 2021 [151, Tab 1].

[2] Pre-payment meters estimated at 4.4 million based on latest available OFGEM data [152, p. 49].

[3] Category 'Other non-standard variable tariffs'. Split between Time-of-Use and Economy 7/10 not available.

Table B.3: Multinomial Logit Models with Full Sample

	Base	w/o TR	with TR	with TR x SH
Fee(%)	-0.057 (0.005)***	-0.058 (0.005)***	-0.058 (0.005)***	-0.058 (0.005)***
Discount(%)	0.031 (0.003)***	0.032 (0.003)***	0.032 (0.003)***	0.032 (0.003)***
Half-Hourly	-0.059 (0.107)	-0.081 (0.109)	-0.081 (0.109)	-0.077 (0.109)
Daily	0.427 (0.089)***	0.431 (0.090)***	0.430 (0.090)***	0.430 (0.090)***
Anon	0.089 (0.112)	-0.566 (0.137)***	-0.610 (0.140)***	-0.272 (0.144)+
Anon X HH	0.263 (0.211)	0.304 (0.214)	0.303 (0.214)	0.295 (0.214)
Anon X Daily	-0.507 (0.159)**	-0.499 (0.160)**	-0.498 (0.160)**	-0.503 (0.161)**
Anon X AGE 55+		0.167 (0.052)**	0.168 (0.052)**	0.170 (0.052)**
Anon X Female		0.305 (0.050)***	0.306 (0.050)***	0.297 (0.050)***
Anon X SEG C1C2		-0.020 (0.059)	-0.023 (0.059)	-0.014 (0.059)
Anon X SEG AB		0.148 (0.069)*	0.148 (0.069)*	0.147 (0.069)*
Anon X No SM		0.189 (0.058)**	0.192 (0.058)***	0.185 (0.058)**
Anon X TV Tariff		0.375 (0.083)***	0.374 (0.083)***	0.373 (0.083)***
Anon X IHD		0.072 (0.067)	0.069 (0.067)	0.067 (0.067)
Anon X FEED		0.043 (0.055)	0.045 (0.055)	0.053 (0.055)
Anon X MANIP		0.074 (0.052)	0.076 (0.052)	0.082 (0.052)
Anon X MR		0.447 (0.067)***	0.452 (0.067)***	
Anon X BA		0.406 (0.057)***	0.408 (0.057)***	
Anon X TR			0.081 (0.049)+	
Anon X TR X TP				-0.335 (0.069)***
Anon X C X TP				-0.249 (0.070)***
Anon X TR X BA+MR				0.259 (0.070)***
n Ind	965	965	965	965
n Obs	7720	7720	7720	7720
AIC	9915	9773	9772	9760
BIC	9964	9898	9904	9892
LL	-4951	-4868	-4867	-4861
pseudo-R2	0.075	0.090	0.090	0.092
Adj pseudo-R2	0.074	0.087	0.087	0.088

Note: Standard Errors shown in Parentheses.+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.4: Full Sample Structured Feedback

		Strongly Disagree	Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree	Don't Know	% Agree
I was able to understand the choices	Control	15	27	96	169	163	7	69.6
	Treatment	17	29	113	173	149	8	66.0
	Total	31	56	210	341	312	15	67.8
I found the options realistic	Control	22	38	123	188	94	12	59.1
	Treatment	14	37	121	172	130	14	61.9
	Total	36	75	243	360	224	27	60.5
I found it easy to choose between the options	Control	25	38	110	164	133	7	62.3
	Treatment	25	53	109	151	140	10	59.6
	Total	50	91	219	314	273	18	60.9

Note: Control (n= 477), Treatment (n=488), Total (n=965).

Table B.5: Full Sample Manipulation Checks

			True	False	% Incorrect
1	A negative expected change in your monthly bill (shown in green) indicates a reduction in your electricity bill.	Control	373	104	21.8
		Treatment	384	104	21.3
		Total	757	208	21.6
2	A half-hourly frequency means that your total electricity consumption for each half-hour is sent to your electricity supplier every 30 minutes.	Control	407	70	14.7
		Treatment	405	83	17.0
		Total	812	153	15.9
3	Anonymisation of your consumption data ensures your information can be linked back to you in the event of a data breach	Control	204	273	42.8
		Treatment	217	271	44.5
		Total	421	544	43.6

Note: Control (n= 477), Treatment (n=488), Total (n=965).

B.1 Survey Quenstionnaire and DCE Screenshots

B.1.1 Questionnaire

3470 Imperial College WTP Questionnaire

1.1 Introduction

Thank you very much for agreeing to complete this online survey which is being conducted by Accent on behalf of Imperial College London.

We are interested in understanding attitudes to Smart Meters. The survey should take about 15-20 minutes to complete.

Any answer you give will be treated in confidence in accordance with the Code of Conduct of the Market Research Society. If you would like to confirm Accent's credentials type Accent in the search box at: https://www.mrs.org.uk/researchbuyersguide.

IF MOBILE DEVICE SHOW: **This survey is best undertaken on a tablet or a PC. If you do use a smartphone you can switch between desktop mode and mobile mode at any time by clicking the button at the bottom of the screen.**

We will just ask you a couple of questions to check that you are eligible to take part in this research. For convenience you can stop and return to complete the questionnaire as many times as you wish, although once submitted you will not be able to enter again. The date for completion is 31st March 2021.

Q1.1 Are you responsible, **either fully or jointly**, for paying the electricity bills in your household?

1. Yes
2. No

DP Thank and close if Q1.1 == NO

Q1.2 Please select your gender:

1. Female
2. Male
3. Other
4. Prefer not to say

Q1.3 Please select your age bracket:

5. 18 - 24
6. 25 - 34
7. 35 - 44
8. 45 - 54
9. 55 - 64
10. 65 - 74
11. 75 - 84
12. 85 or older
13. Prefer not to say

Q1.4 Choose one option that best describes your ethnic group or background:

A - White

1. British/English/Scottish/Welsh/Northern Irish
2. Irish
3. Gypsy or Irish Traveller
4. Any other White background

B – Mixed/Multiple ethnic groups

1. White and Black Caribbean
2. White and Black African
3. White and Asian
4. Any other mixed background

C – Asian/Asian British

1. Indian
2. Pakistani
3. Bangladeshi
4. Chinese
5. Any other Asian background

D - Black/African/Caribbean

1. Caribbean
2. African
3. Any other Black background

E - Other ethnic group

1. Arab
2. Any other ethnic group

F - Prefer not to say

Q1.5 How would you describe the occupation of the chief income earner in your household?

1. Senior managerial or professional
2. Intermediate managerial, administrative or professional
3. Supervisor; clerical; junior managerial, administrative or professional
4. Manual worker (with industry qualifications)
5. Manual worker (with no qualifications)
6. Unemployed due to ill health
7. Unemployed for another reason
8. Retired
9. Student
10. Prefer not to say

Q1.6: **IF Q1.5=8 (RETIRED), ASK, ELSE SKIP** Does the chief income earner have a state pension, a private pension or both?

1. State only
2. Private only
3. Both

IFQ1.6=1 (STATE ONLY), GO TO Q1.8

Q1.7 **IF Q1.6 = PRIVATE OR BOTH, ASK, ELSE SKIP**: How would you describe the chief income earner's occupation before retirement?

1. Senior managerial or professional
2. Intermediate managerial, administrative or professional
3. Supervisor; clerical; junior managerial, administrative or professional
4. Manual worker (with industry qualifications)
5. Manual worker (with no qualifications)
6. None of these

SEG: CODE AS FOLLOWS:

IF Q1.4= 1 or 2; SEG = AB

IF Q1.4= 3; SEG = C1

IF Q1.4= 4; SEG = C2

IF Q1.4= 5; SEG = DE

IF Q1.4= 6; SEG = DE

IF Q1.4= 7; SEG = DE

IF Q1.4= 9; SEG = C1

IF Q1.4= 8 and Q1.5 = State only; SEG = DE

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 1; SEG = AB

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 2; SEG = AB

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 3; SEG = C1

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 4; SEG = C2

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 5; SEG = DE

IF Q1.4= 8 and Q1.5= Private only OR Both and Q1.6= 6; SEG = DE

IF Q1.4= 10; SEG = Not stated

Q1.7 Please indicate the **first part of your postcode** (e.g. for SW7 2AZ enter SW7). Select the letter prefix (e.g. SW) from the dropdown menu and enter the number (e.g. 7) in the box next to it. If you prefer not to say please leave blank.

DP INCLUDE VALIDATION CHECK OF POSTCODE, IF INVALID POSTCODE IS ENTERED PROMPT TO CHECK POSTCODE. IF STILL INVALID, THANK AND CLOSE.

Q1.9 Do you have a smart meter (an example is pictured above, however each supplier provides a slightly different device)?

- Yes
- No
- Don't know

DP information bubble: A smart meter is the next generation of a gas and electricity meter. They were first introduced in 2012 in the UK. Smart meters measure how much gas and electricity you're using, as well as what it's costing you and display this on an in-home display.

Q1.10 As part of survey you will be asked about preferences for different electricity supply options. In order to accurately reflect the financial benefits/costs of these options we'd like to understand how much you pay for **your household's electricity**. If you know the amount (excluding gas) please enter **either** the monthly or yearly amount below. A copy of your latest electricity bill should include an estimate of your yearly bill (broken down by gas and electricity if you have both).

1. _____ Monthly
2. _____ Yearly
3. Don't know

DP PLEASE PROVIDE BOXES TO ENTER IN THE AMOUNT EITHER PER MONTH OR PER YEAR. IF IT IS ENTERED PER YEAR, DIVIDE BY 12 TO PRODUCE MONTHLY AMOUNT. IF DON'T KNOW PLEASE USE £57/MONTH FOR CHOICE TASK INPUT.

Q1.11 **IF Q1.10 (MONTHLY)>300 OR (YEARLY)>3600, ASK, ELSE SKIP** The amount you have entered for your household electricity bill is more than 5 times the national average. You said your (yearly/monthly) electricity bill is £XXX. Is this correct? If not please enter the correct monthly or yearly amount below.

1. My (monthly/yearly) electricity bill is £XXX
2. _____ Monthly
3. _____ Yearly

DP THANK AND CLOSE IF Q1.11 == 1 OR (MONTHLY)>300 OR (YEARLY)>3600

1.2 Participant Information Sheet

Study Title: Smart Meter Data Sharing

Principal Investigator and Co-investigators: Dr. Fei Teng and Mr. Saurab Chhachhi

Introduction

You are being invited to take part in a research study which aims to understand people's attitudes towards smart meters and sharing their smart meter data. You can participate even if you don't own a smart meter at the moment.

The results will be published as part of the Imperial College London Energy Futures Lab Briefing Paper series. In addition, detailed results, methodology and anonymised response data will be submitted to an academic journal.

Imperial College London is the study sponsor and aims to include a 1000 participants across Great Britain. This study is supported by funding from Research England through the Strategic Priorities Fund - Evidence-based policy making. This study was given ethical approval by the Head of Department and the Research Governance and Integrity Team (RGIT).

Information about your involvement in the study

You will be asked to complete a survey which should last approximately 15-20 minutes. First, we will ask about your general attitudes towards data sharing and preferences on several data sharing options for your electricity consumption data. This is followed by a series of questions about your current electricity usage. Finally, we will ask you for some additional demographic information.

Possible Disadvantages and Risks of Taking Part

There are no risk or disadvantages to taking part in this online survey. You will not be asked for any identifying data (name, phone number, email etc.) or other sensitive questions. All questions include a 'Prefer not to say/Don't know' option if you are not comfortable with answering a particular

question or do not know the answer. All data collected will be anonymous. Anonymised data collected in the study is to be published with any resulting academic publications.

Participation

Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. It is up to you to decide whether or not to take part. By completing the survey you will have given implied consent. If you decide to take part you are still free to withdraw at any time while completing the survey and without giving a reason. We will keep all data collected.

Contact

If you have questions at any time about the study or the procedures please contact the researcher, Saurab Chhachhi by email at [redacted]. If you wish to complain or have any concerns about any aspect this study then please contact the Investigator Dr. Fei Teng at [redacted]

If you do decide to take part, you can print a copy of this information sheet to keep. Thank you for reading this.

Q2.1 Please confirm whether you consent to the following statements (in order to participate you will need to consent to all):

I confirm that I have read and understand the participant information for this study and have had the opportunity to ask questions which have been answered fully.	**o**
I understand that my participation is voluntary, and I am free to withdraw at any time before submitting the survey, without giving any reason and without my legal rights being affected.	o
I give permission for Imperial College London to access my research records that are relevant to this research.	o
I give consent for anonymised data collected in the study to be published in any resulting academic publications.	o
I give consent for information collected about me to be used to support other research in the future, including those outside of the European Economic Area (EEA).	o
I consent to take part in this study.	o

DP MUST SELECT ALL IN ORDER TO CONTINUE

1.3 Data Sharing Preferences

We would like to understand your views on sharing your data with different types of organizations

Q3.1 Typically, when you share your data in order to use a service or product, for example when you shop online, use an energy company or book a holiday online, which level of data sharing do you sign up to? **Please tick all that apply**.

1. Sharing the basic information the company requires to provide me with the service they offer
2. Allowing the company to use my information for marketing, research, forecasting etc.
3. Allowing my data and information to be passed to third parties.

1.4 Introduction to Choice Task

Smart meters are the new generation of electricity meters being rolled out across Great Britain. They show you how much energy you are consuming, in real-time, in pounds and pence.

Your electricity consumption data can also be shared with your energy supplier which can help them operate more efficiently and pass on savings to you through reduced electricity bills.

In Great Britain, if you choose to install a smart meter, you have the option to choose how your electricity consumption data is shared and who can access it. By default electricity consumption data is only sent on a daily basis, similar to the way traditional electricity metering works.

However, to achieve some of the operational benefits, your electricity supplier may need access to more detailed data, for example, half-hourly or minute-by-minute readings.
DP INCLUDE TIMING FOR ABOVE SECTION (Introduction to Choice Task)

Q4.1 How willing would you be to share your **half-hourly** electricity consumption data with your energy supplier?

1. Very Willing
2. Quite Willing
3. Neither willing nor unwilling
4. Not very willing
5. Not at all willing

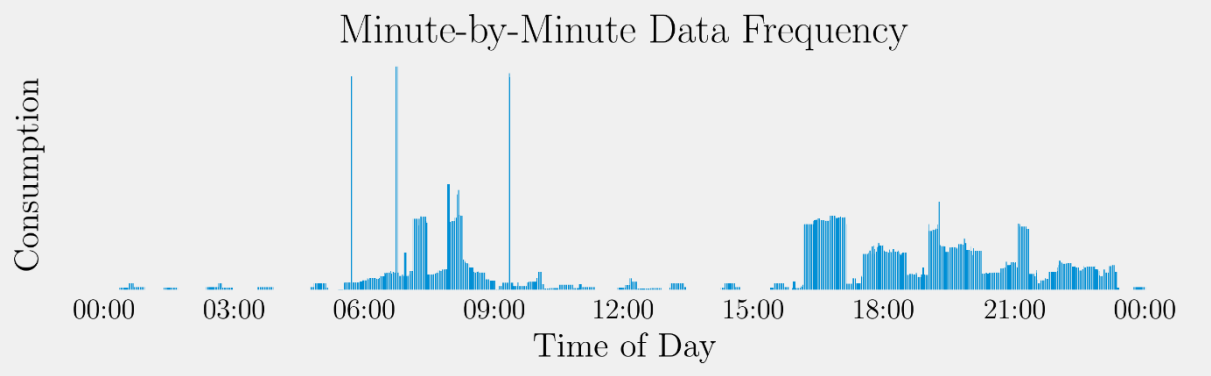
1.5 Data Sharing Options

You will be shown a series of scenarios and asked about your preferences for different electricity consumption data sharing options. The following section describes these data sharing options and the personal information that you would be providing to your supplier. Please take your time to read the descriptions and understand the options.

DP underlined text only shown to Group B

1. Frequency of Data Sharing

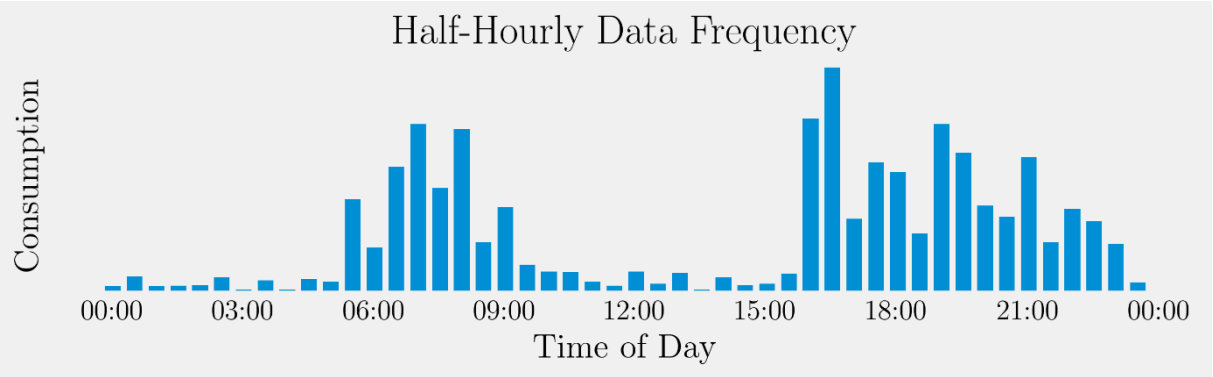
Minute-by-minute – your electricity consumption data for **every minute** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a minute-by-minute basis. An illustrative example is shown below:



1. Frequency of Data Sharing

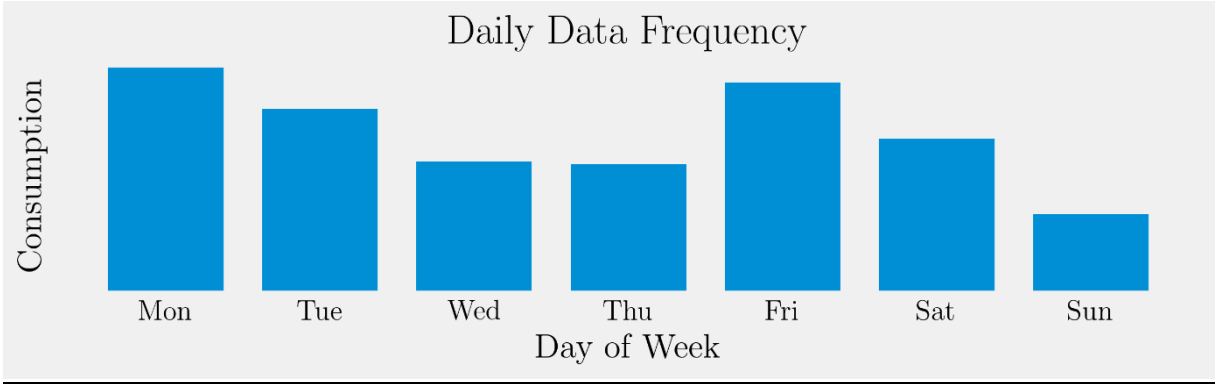
Half-hourly – your electricity consumption data for **every 30 minutes** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy

consumption on a half-hourly basis. The same example data as used above is now shown at half-hourly frequency:



1. Frequency of Data Sharing

Daily – your electricity consumption data for **every day** is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a daily basis. The same example data as used above is now shown at daily frequency:



2. Anonymisation

Anonymised – Anyone with access is able to extract insights and patterns from your electricity consumption data, but these **cannot be linked to you personally**. This ensures that even in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **cannot** use it to identify or build a profile of you.
None – Anyone with access is able to extract insights and patterns from your electricity consumption data and these are **linked to you personally**. This means that in the event of a data breach, for

example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **can** use it to identify or build a profile of you.

1.6 Personal Information in your Electricity Consumption Data

The data sharing options (frequency and anonymisation) determine the type, amount and accuracy of personal information, **about you,** that can be extracted from your electricity consumption. The higher the frequency of data sharing the more details and greater the accuracy.

The following section will describe a number of different personal characteristics which your supplier and any third parties who may access the data, may be able to determine from electricity consumption data. Please take your time to read the descriptions and understand the options.

Appliance Ownership – identifying which electrical appliances you own. Large appliances such as ovens, microwaves, toasters, kettles, fridges, solar panels, electric vehicles and battery storage as well as small appliances such as TVs, laptops and lighting.

Appliance Usage and Routines – how you are using your appliances and what your daily routine is. For example, when you turn on the kettle or have a shower as well as what temperature your thermostat is set to.

Occupancy – if anyone is in the house as well as how many people are there at a particular point in time.

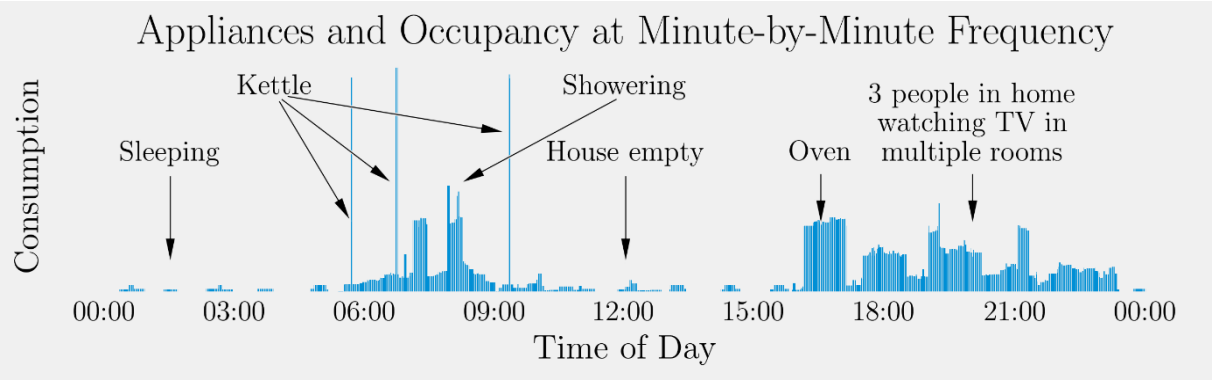
Household Details – the number of people ordinarily living in your house and their approximate ages as well as the presence of children and pets.

Socio-Demographic Details - marital status, whether there is someone with a long-term illness and employment status.

Economic Details - approximate household income level and socio-economic class

Housing Details – location, type of housing (e.g. detached house or flat), size, ownership (rented or owned) and age of dwelling.

An illustrative example showing appliances usage and occupancy at minute-by-minute frequency is shown below:



The frequency at which you choose to share your consumption data determines what personal information can be extracted and how accurate this is. Below is a table showing what personal information can be extracted that the different data sharing frequencies (minute-by-minute, half-hourly and daily).

Please take the time to carefully read the table below.

Frequency	Minute-by-Minute	Half-hourly	Daily
Household Details	X	X	X
Income Level	X	X	X
Marital & Employment Status	X	X	.
Housing Details	X	X	X
Large Appliance Ownership	X	X	.
Small Appliance Ownership	X	.	.
Appliance Usage and Routines	For every minute	For every half-hour	.
Occupancy	For every minute	For every half-hour	For every day

For example at a minute-by-minute frequency it is possible to determine which appliances you own, how you use them across the day and what you are doing at a particular point in time.

At a half-hourly frequency one can determine the use of large appliances such as ovens and if they have been used within a specific half hour interval but with a lower accuracy than if data was shared minute-by-minute.

At daily frequency it is not possible to determine which appliances you own or when they are being used.

DP TEXT HIGHLIGHTED IN YELLOW ONLY FOR GROUP B

Q4.2 Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **not anonymised** before being shared?

1. More likely
2. It makes no difference
3. Less likely

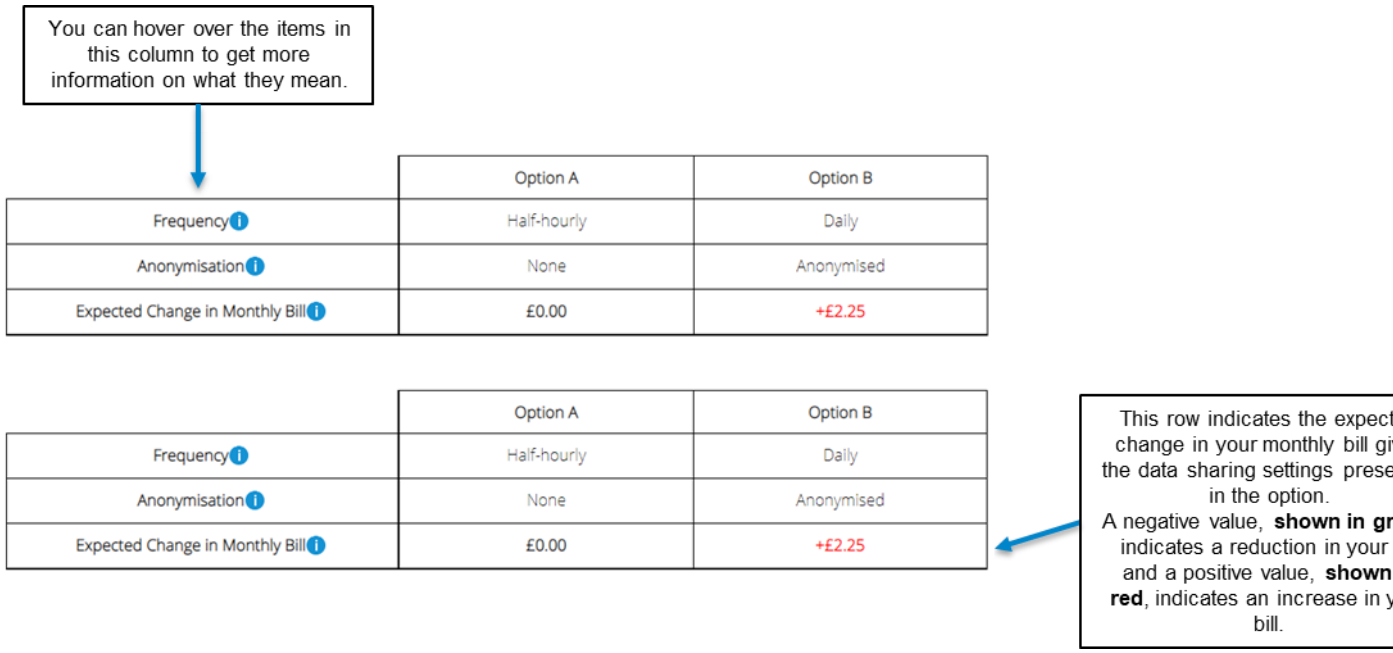
Q4.3 Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **anonymised** before being shared?

1. More likely
2. It makes no difference
3. Less likely

You will now be shown an example of the scenarios. Each scenario will consist of two options (A,B) defined by the data sharing settings described (frequency and anonymisation) and what personal information you share in each option.

1.7 Introduction to the Scenarios – Group A

Below is an example of the scenarios you will be presented: **DP Please refer to the latest powerpoint examples**

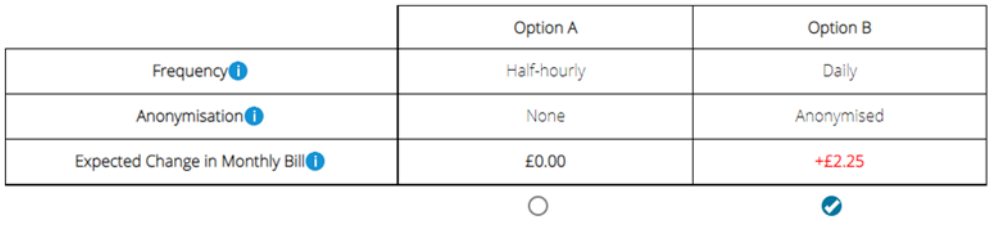


You will now be presented with eight scenarios. Please select your preferred option in each scenario.

DP INCLUDE TIMING FOR ABOVE SECTION (Introduction to Scenarios)

1.8 Scenario 1 (Group A)

Q7.1 Which option (A or B) would you prefer?. Base your choice on the options on this page only.



1. Option A
2. Option B

DP INCLUDE TIMING FOR EACH SCENARIO (1-8)

1.9 Introduction to Choice Task – Group B

Below is an example of the scenarios you will be presented: **DP Please refer to the latest powerpoint examples**

You can hover over the items in this column to get more information on what they mean.

	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.25
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.

This row indicates the expected change in your monthly bill given the data sharing settings presented in the option.
A negative value, **shown in green**, indicates a reduction in your bill and a positive value, **shown in red**, indicates an increase in your bill.

	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.25
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.

	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.25
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.

If the type of data in this row is identifiable and can be linked to you, the box will have an **X**. If not, it will be blank.
For example, in Option A it is possible to identify and link your income level but not small appliances such as TVs.

	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.25
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.

If the occupancy of your home or your usage of appliances can be detected the frequency at which this is possible will be shown. If not, it will be blank.

You will now be presented with eight scenarios. Please select your preferred option in each scenario.

Q6 Timing

DP INCLUDE TIMING FOR ABOVE SECTION (Introduction to Scenarios, SUM of the four screens)

1.10 Scenario 1 – Group B

Q7.1 Which option (A or B) would you prefer? Base your choice on the options on this page only.

	Option A	Option B	Option C
Anonymisation	None	None	None
Aggregation	None	Aggregate	None
Frequency	Minute-by-minute	Daily	Daily
Expected Change in Monthly Bill	-£2.85	+£8.55	£0.00
Household Details	X	X	X
Income Level	X	.	X
Marital & Employment Status	X	.	.
Housing Details	X	X	X
Large Appliance Ownership	X	.	.
Small Appliance Ownership	X	.	.
Appliance Usage and Routines	For every minute	.	.
Occupancy	For every minute	For every day	For every day
	○	●	○

1. Option A
2. Option B

DP INCLUDE TIMING FOR EACH SCENARIO (1-8)

Q8-14 Other choice cards (8 in total)

1.11 Choice Task Evaluation

We would now like to ask you a few questions about the choices you have just made.

Q15.1 Thinking about the choices you have just made, would you opt to have a smart meter with one of the data sharing options you selected or choose not to have a smart meter?

1. Have a smart meter with one of the selected data sharing options
2. Not have a smart meter

Q15.2 Thinking about the descriptions of the different data sharing options you have just read, please answer the following questions about the text and scenarios:

	True	False
A **negative** expected change in your monthly bill (shown in green) indicates a **reduction** in your electricity bill.		
A **half-hourly** frequency means that your total electricity consumption for each half hour is sent to your supplier every 30 minutes.		
Anonymisation of your consumption data ensures your information **can be** linked back to you in the event of data breach.		
If you share your electricity consumption data at a **half-hourly** frequency your supplier **will** know how many people were in your house between, for example, 17:00 and 17:30.		
If you share your electricity consumption data at a **daily** frequency your supplier **will know** that, for example, you put the kettle on at 08:00.		

DP OPTIONS HIGHLIGHTED IN YELLOW ONLY SHOWN TO GROUP B

Q15.3 How strongly do you agree or disagree with the following statements about the choices you have just made?

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree	Don't Know
I was able to understand the choices						
I found the options realistic						
I found it easy to choose between the options						

Q15.4 Do you have any comments or feedback about the choice task? Were any of the data sharing options (frequency, anonymisation) difficult to understand? Were the differences between the options in each scenario clear?

1.12 Additional Information on Current Electricity Supply

We would now like to understand more about your current electricity supply. To answer the following set of questions we recommend having a copy of your latest electricity bill at hand.

Q16.1 Does your home have a dual fuel (both electricity and gas) connection or electricity only connection?

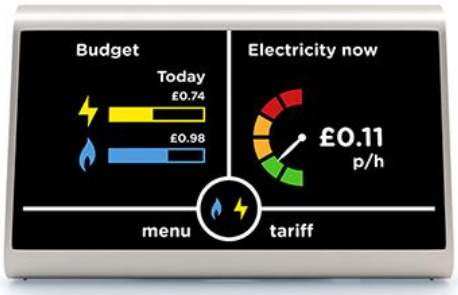
1. Dual Fuel (if you have a gas boiler for heating and/or hot water and/or a gas stove)
2. Electricity Only
3. Don't know

Q16.2 What type of electricity tariff are you currently on?

1. Standard Variable Tariff (default tariff if you made no active tariff selection)
2. Fixed Rate (fixed rate guaranteed for specified time e.g. 12 months)
3. Pre-payment (if you have a pre-payment meter which you have to top up)
4. Time-of-Use Tariff (rate varies depending on the time of day e.g. Octopus Agile)
5. Economy 7 or 10 (you pay a lower rate for 7 or 10 hours of day, usually at night and have two separate electricity meters)
6. Don't know

Q16.3 **ALL WHO SAID YES AT Q1.9** When you had your smart meter installed you should have been asked what frequency (half-hourly, daily or monthly) you want your smart meter to be sent to your supplier. Thinking about when you had your smart meter installed, which frequency did you select?

1. Half-hourly
2. Daily
3. Monthly
4. Don’t know



Q16.4 **ALL WHO SAID YES AT Q1.9** Do you own an In-Home Energy Display (an example is pictured above, however each supplier provides a slightly different device)?

1. Yes
2. No
3. Don't know

DP information bubble: An In-Home Energy Display is an electronic device that is connected to your smart meter. It allows you to see, in real-time, how much gas and electricity you're using, as well as what it's costing you.

Q16.5 **ALL WHO SAID YES AT Q16.4**How often do you check your In-Home Energy Display?

1. Daily
2. 2-3 times a week
3. Once a week
4. Once a month
5. Less than once a month
6. Never

1.13 Demographics Information

Q17.1 Which of the following best describes your household?

1. I/we own my own home (mortgage or outright) **CODE AS HOMEOWNER**
2. I/we own my own home (through a shared ownership or Keyworker scheme) **CODE AS HOMEOWNER**
3. I/we rent from a private landlord **CODE AS TENANT**
4. I/we live in Student Accommodation **CODE AS TENANT**
5. I/we rent from a Housing Association/Council **CODE AS TENANT**
6. I/we live with my parents **CODE AS TENANT**
7. Prefer not to say

Q17.2 Please select your **pre-tax** household income bracket:

1. Less than £20,000
2. Between £20,000 and £39,999
3. Between £40,000 and £59,999
4. Between £60,000 and £79,999
5. Between £80,000 and £99,999
6. More than £100,000
7. Prefer not to say

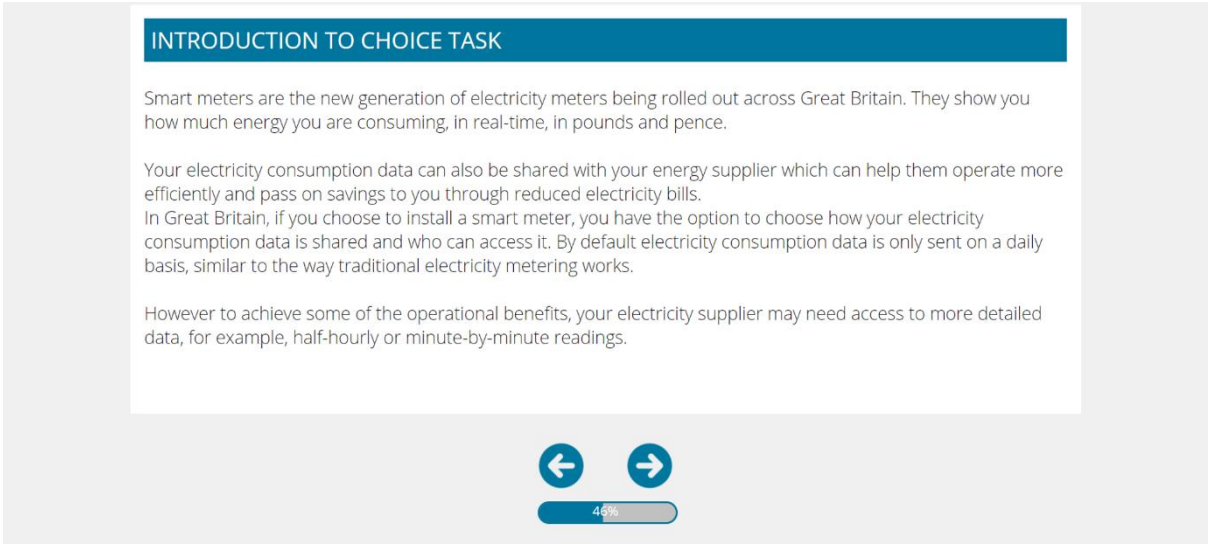
Q17.3 Since the Covid-19 pandemic started, how much time have you spent at home compared to before the pandemic:

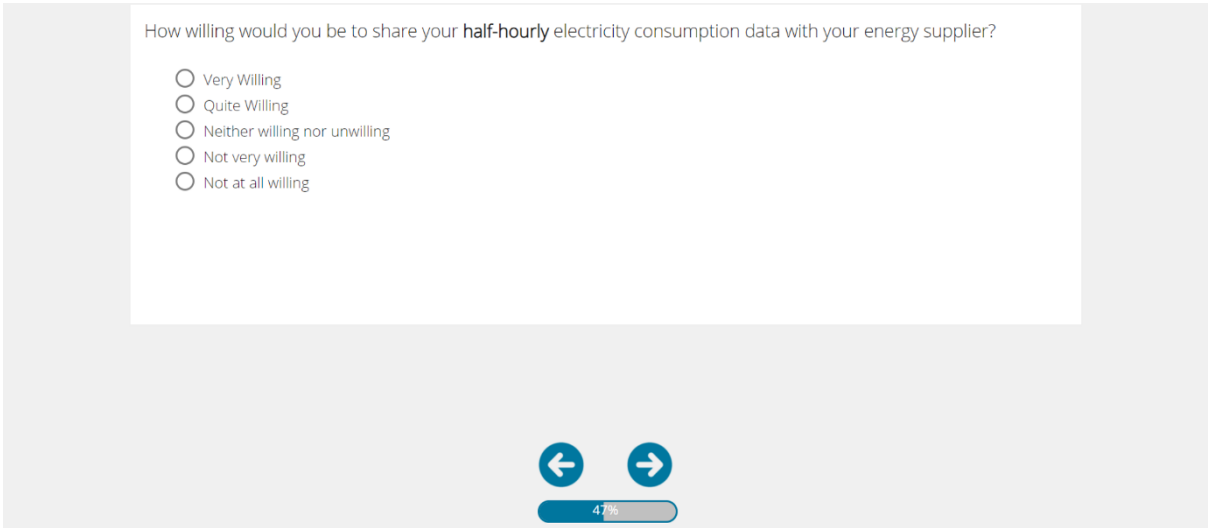
1. A lot more
2. Slightly more
3. About the same
4. Slightly less
5. A lot less
6. Prefer not to say

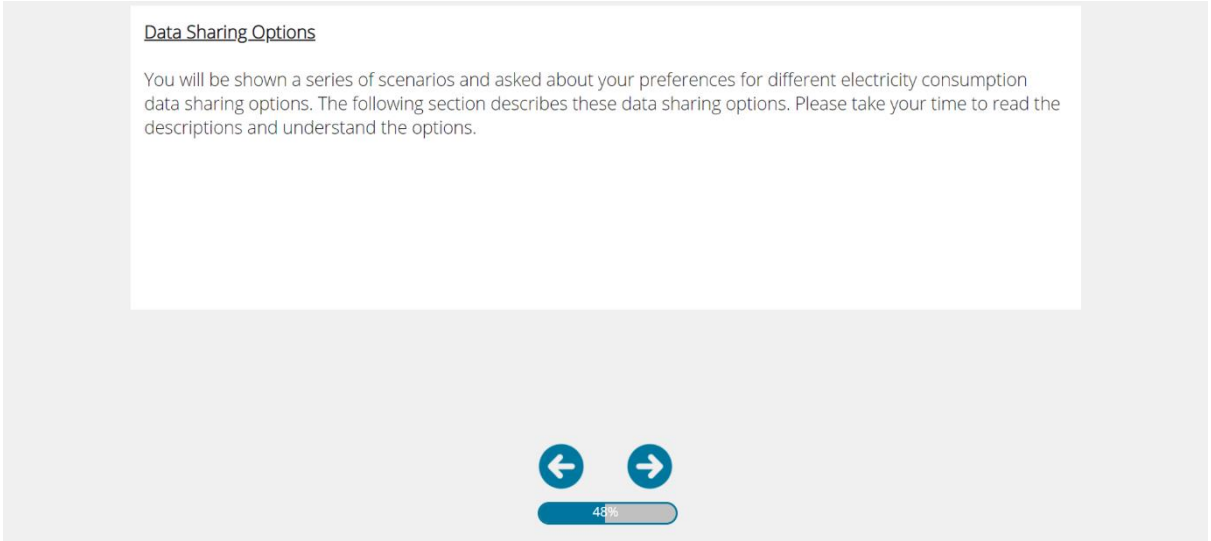
DP INCLUDE MINIMUM TIMING FOR FULL SURVEY OF 5 MINUTES

B.1.2 Discrete Choice Experiment

Group A – Control

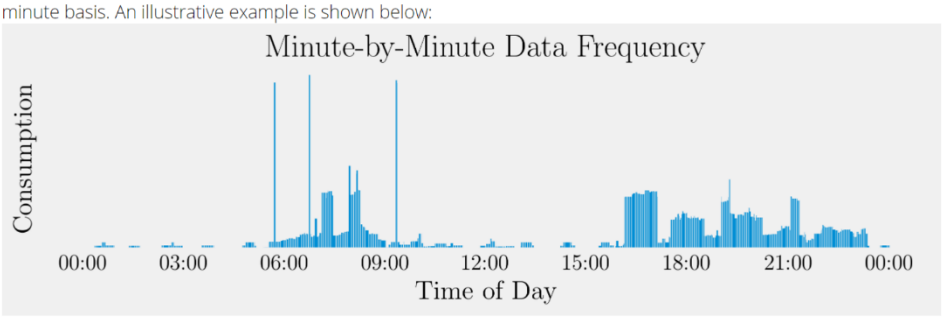






1. Frequency of Data Sharing

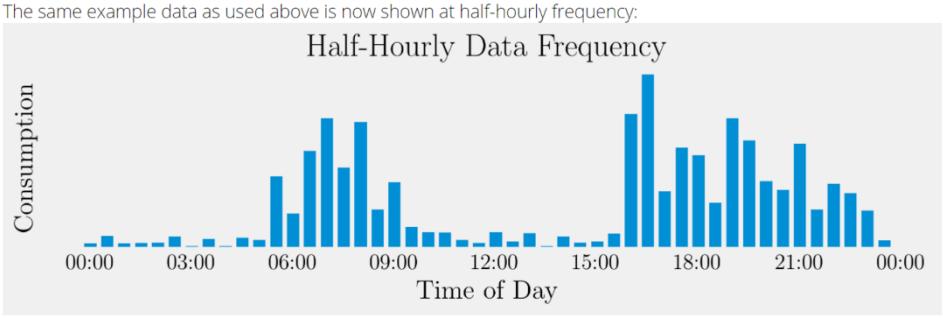
Minute-by-minute - your electricity consumption data for **every minute** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a minute-by-minute basis. An illustrative example is shown below:



49%

1. Frequency of Data Sharing

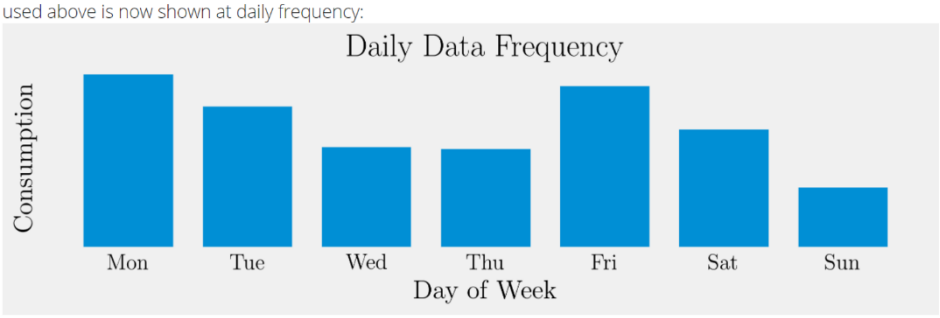
Half-hourly - your electricity consumption data for **every 30 minutes** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a half-hourly basis. The same example data as used above is now shown at half-hourly frequency:



50%

1. Frequency of Data Sharing

Daily - your electricity consumption data for **every day** is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a daily basis. The same example data as used above is now shown at daily frequency:



51%

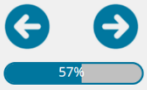
2. Anonymisation

Anonymised - Anyone with access is able to extract insights and patterns from your electricity consumption data, but these **cannot be linked to you personally**. This ensures that even in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **cannot** use it to identify or build a profile of you.

None - Anyone with access is able to extract insights and patterns from your electricity consumption data and these are **linked to you personally**. This means that in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **can** use it to identify or build a profile of you.

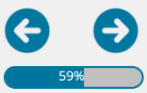
Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **not anonymised** before being shared?

- ○ More likely
- ◉ It makes no difference
- ○ Less likely



Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **anonymised** before being shared?

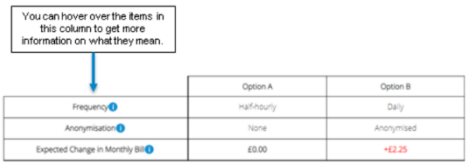
- ◉ More likely
- ○ It makes no difference
- ○ Less likely



You will now be shown an example of the scenarios. Each scenario will consist of two options (A,B) defined by the data sharing settings described (frequency and anonymisation). Please take your time to read the descriptions and understand the options.

60%

Below is an example of the scenarios you will be presented:



61%

You will now be presented with eight scenarios. Please select your preferred option in each scenario.

64%

Which option (A or B) would you prefer? Base your choice on the options on this page only.

	Option A	Option B
Frequency	Minute-by-minute	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	-£1.50	+£1.00
	○	○

65%

We would now like to ask you a few questions about the choices you have just made.

Thinking about the choices you have just made, would you opt to have a smart meter with one of the data sharing options you selected or choose not to have a smart meter?

- ○ Have a smart meter with one of the selected data sharing options
- ● Not have a smart meter

74%

Choice Task Evaluation

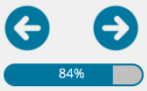
Thinking about the descriptions of the different data sharing options you have just read, please answer the following questions about the text and scenarios:

	True	False
A **negative** expected change in your monthly bill (shown in green) indicates a **reduction** in your electricity bill.	○	○
A **half-hourly** frequency means that your total electricity consumption for each half hour is sent to your supplier every 30 minutes.	○	○
Anonymisation of your consumption data ensures your information **can be** linked back to you in the event of data breach.	○	○

77%

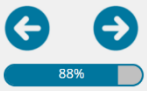
How strongly to do you agree or disagree with the following statements about the choices you have just made?

	Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	Don't know
I was able to understand the choices	○	○	○	○	○	○
I found the options realistic	○	○	○	○	○	○
I found it easy to choose between the options	○	○	○	○	○	○

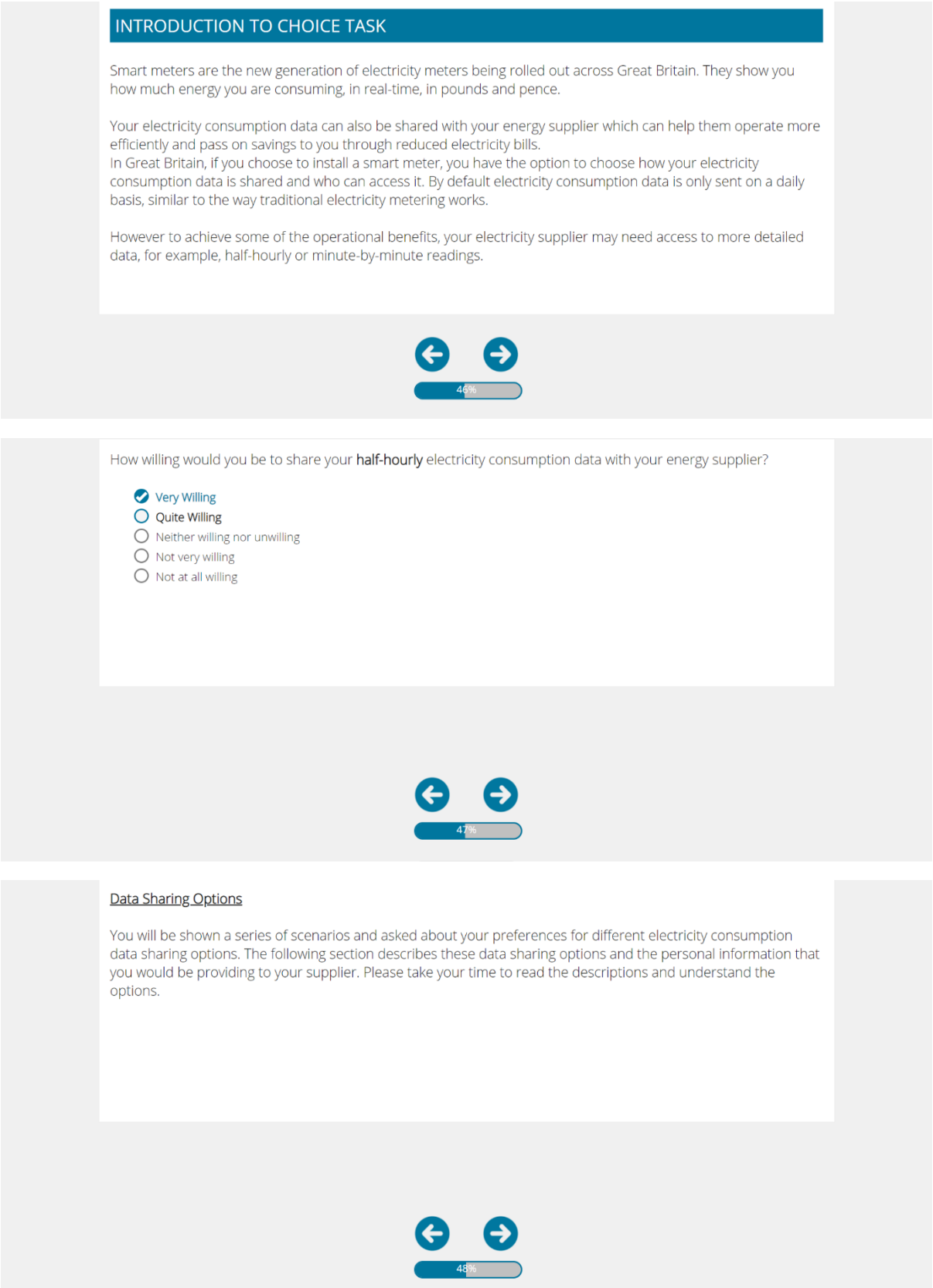


Do you have any comments or feedback about the choice task? Were any of the data sharing options (frequency, anonymisation) difficult to understand? Were the differences between the options in each scenario clear?

☐ No comments

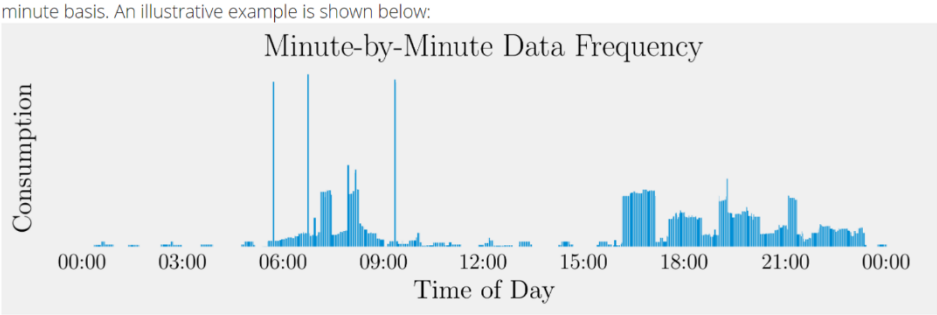


Group B – Treatment



1. Frequency of Data Sharing

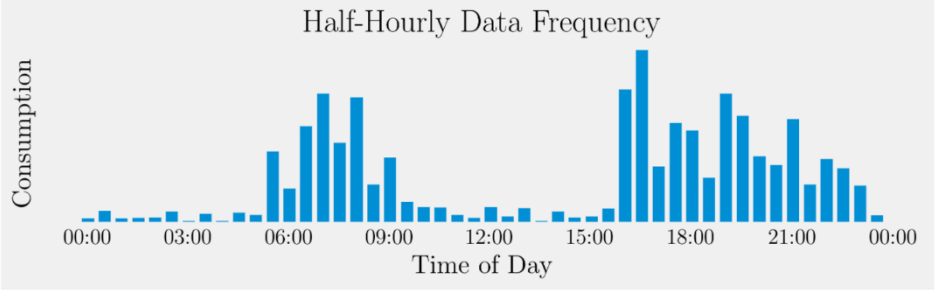
Minute-by-minute - your electricity consumption data for **every minute** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a minute-by-minute basis. An illustrative example is shown below:



49%

1. Frequency of Data Sharing

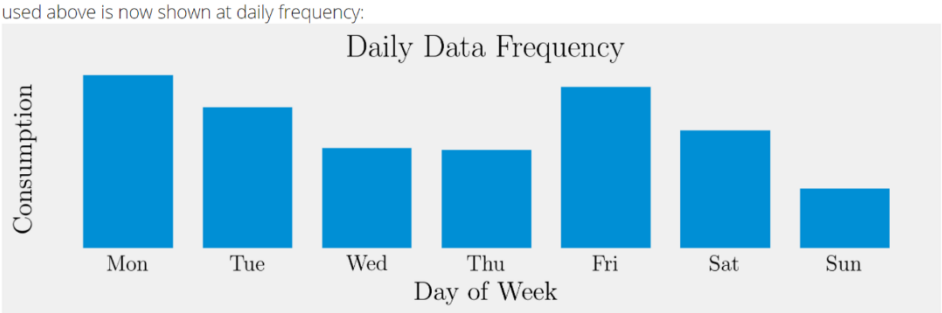
Half-hourly - your electricity consumption data for **every 30 minutes** of the day is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a half-hourly basis. The same example data as used above is now shown at half-hourly frequency:



50%

1. Frequency of Data Sharing

Daily - your electricity consumption data for **every day** is sent to your supplier. Your supplier and any other authorised entities can see the changes in your energy consumption on a daily basis. The same example data as used above is now shown at daily frequency:



51%

2. Anonymisation

Anonymised - Anyone with access is able to extract insights and patterns from your electricity consumption data, but these **cannot be linked to you personally**. This ensures that even in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **cannot** use it to identify or build a profile of you.

None - Anyone with access is able to extract insights and patterns from your electricity consumption data and these are **linked to you personally**. This means that in the event of a data breach, for example if your supplier was hacked or if data were misused, whoever has access to your electricity consumption data **can** use it to identify or build a profile of you.

Personal Information in your Electricity Consumption Data

The data sharing options (frequency and anonymisation) determine the type, amount and accuracy of personal information, **about you**, that can be extracted from your electricity consumption. The higher the frequency of data sharing the more details and greater the accuracy.

The following section will describe a number of different personal characteristics which your supplier and any third parties who may access the data, may be able to determine from electricity consumption data. Please take your time to read the descriptions and understand the options.

Appliance Ownership - identifying which electrical appliances you own. Large appliances such as ovens, microwaves, toasters, kettles, fridges, solar panels, electric vehicles and battery storage as well as small appliances such as TVs, laptops and lighting.

Appliance Usage and Routines - how you are using your appliances and what your daily routine is. For example, when you turn on the kettle or have a shower as well as what temperature your thermostat is set to.

Occupancy - if anyone is in the house as well as how many people are there at a particular point in time.

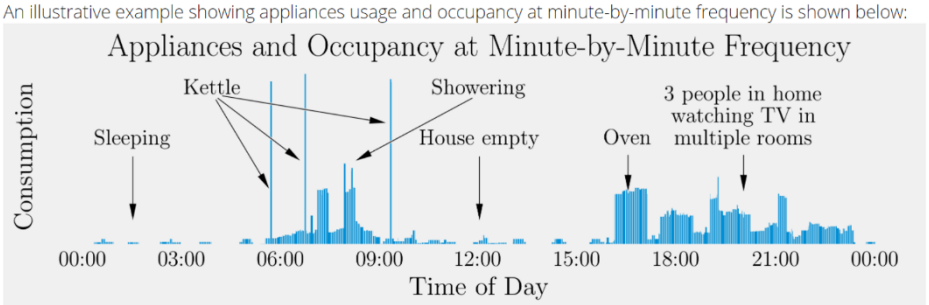
Household Details - the number of people ordinarily living in your house and their approximate ages as well as the presence of children and pets.

Socio-Demographic Details - marital status, whether there is someone with a long-term illness and employment status.

Economic Details - approximate household income level and socio-economic class

Housing Details - location, type of housing (e.g. detached house or flat), size, ownership (rented or owned) and age of dwelling.

An illustrative example showing appliances usage and occupancy at minute-by-minute frequency is shown below:



55%

The frequency at which you choose to share your consumption data determines what personal information can be extracted and how accurate this is. Below is a table showing what personal information can be extracted that the different data sharing frequencies (minute-by-minute, half-hourly and daily).

Please take the time to carefully read the table below.

Frequency	Minute-by-Minute	Half-hourly	Daily
Household Details	X	X	X
Income Level	X	X	X
Marital & Employment Status	X	X	.
Housing Details	X	X	X
Large Appliance Ownership	X	X	.
Small Appliance Ownership	X	.	.
Appliance Usage and Routines	For every minute	For every half-hour	.
Occupancy	For every minute	For every half-hour	For every day

For example at a minute-by-minute frequency it is possible to determine which appliances you own, how you use them across the day and what you are doing at a particular point in time.

At a half-hourly frequency one can determine the use of large appliances such as ovens and if they have been used within a specific half hour interval but with a lower accuracy than if data was shared minute-by-minute.

At daily frequency it is not possible to determine which appliances you own or when they are being used.

Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **not anonymised** before being shared?

- ○ More likely
- ○ It makes no difference
- ○ Less likely

57%

Considering the information you have just read, would you be more or less likely to share your **half-hourly** electricity consumption data if it was **anonymised** before being shared?

- ○ More likely
- ○ It makes no difference
- ○ Less likely

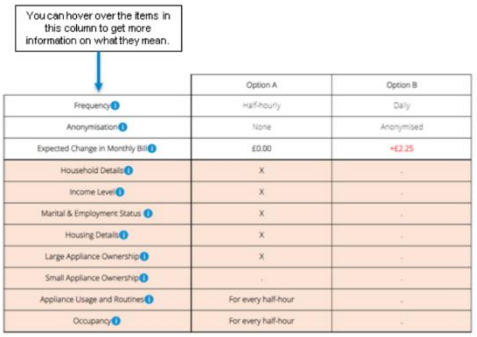
59%

You will now be shown an example of the scenarios. Each scenario will consist of two options (A,B) defined by the data sharing settings described (frequency and anonymisation) and what personal information you share in each option. Please take your time to read the descriptions and understand the options.

[No Title]

60%

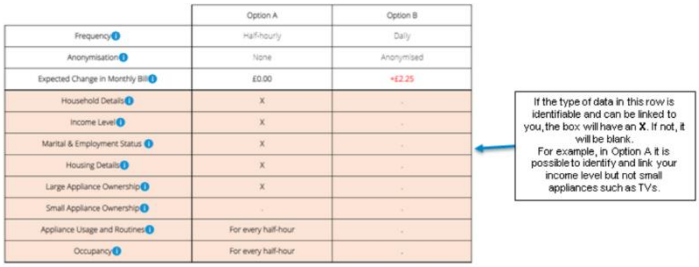
Below is an example of the scenarios you will be presented:



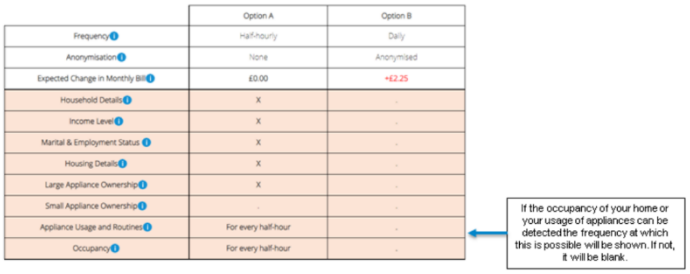
	Option A	Option B
Frequency	Half-hourly	Daily
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	+£2.25
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.

61%

62%



63%



You will now be presented with eight scenarios. Please select your preferred option in each scenario.

64%

Which option (A or B) would you prefer? Base your choice on the options on this page only.

	Option A	Option B
Frequency	Half-hourly	Minute-by-minute
Anonymisation	None	Anonymised
Expected Change in Monthly Bill	£0.00	-£1.00
Household Details	X	.
Income Level	X	.
Marital & Employment Status	X	.
Housing Details	X	.
Large Appliance Ownership	X	.
Small Appliance Ownership	.	.
Appliance Usage and Routines	For every half-hour	.
Occupancy	For every half-hour	.
	○	○

We would now like to ask you a few questions about the choices you have just made.

Thinking about the choices you have just made, would you opt to have a smart meter with one of the data sharing options you selected or choose not to have a smart meter?

- ○ Have a smart meter with one of the selected data sharing options
- ✓ Not have a smart meter

74%

Choice Task Evaluation

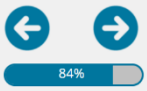
Thinking about the descriptions of the different data sharing options you have just read, please answer the following questions about the text and scenarios:

	True	False
A **negative** expected change in your monthly bill (shown in green) indicates a **reduction** in your electricity bill.	○	○
A **half-hourly** frequency means that your total electricity consumption for each half hour is sent to your supplier every 30 minutes.	○	○
Anonymisation of your consumption data ensures your information **can be** linked back to you in the event of data breach.	○	○
If you share your electricity consumption data at a **half-hourly** frequency your supplier **will** know how many people were in your house between, for example, 17:00 and 17:30.	○	○
If you share your electricity consumption data at a **daily** frequency your supplier **will know** that, for example, you put the kettle on at 08:00.	○	○

77%

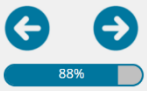
How strongly to do you agree or disagree with the following statements about the choices you have just made?

	Strongly disagree	Disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	Don't know
I was able to understand the choices	○	○	○	○	○	○
I found the options realistic	○	○	○	○	○	○
I found it easy to choose between the options	○	○	○	○	○	○



Do you have any comments or feedback about the choice task? Were any of the data sharing options (frequency, anonymisation) difficult to understand? Were the differences between the options in each scenario clear?

☐ No comments



B.2 Study Block Design

Table B.6: SAS Output

	Value
Design	56
Choice Sets	96
Alternatives	2
Parameters	11
Maximum Parameters	96
D-Efficiency	4.737
Relative D-Eff	17.4
D-Error	0.211
1/Choice Sets	0.0104

Listing B.1: SAS Macro for Experimental Design

```
1  title Willingness-to-Pay for Privacy of Smart Meter Data;
2  %let m = 2; /* number of alternatives */
3  %let n = 8; /* number of choice sets per person */
4  %let blocks = 12; /* number of blocks */
5  %let seed = 42; /* random seed for reproducibility */
6  %let iters = 100; /* max number of iterations */
7  %let n_freq = 3; /* levels of frequency */
8  %let n_anon = 2; /* levels of anonymisation */
9  %let n_bill = 9; /* levels of expected bill changes */
10
11 /* generate full-factorial candidate list */
12 %mktruns(&n_freq &n_anon &n_bill);
13 %mktex(&n_freq &n_anon &n_bill, /* all attrs of all alternatives
       */
14                 n=%eval(&n_freq*&n_anon*&n_bill), /* number of
                      choice sets */
15                 seed=&seed) /* random number seed */
16 %mktlab(data=design,
17                 int=f1-f2,
18                 out=design_out)
19 proc sort data=design_out out=design_sorted;
20         by x1 x2 x3;
21 run;
22 /* reduce candidate list */
23 data cand_list;
24         set design_sorted;
25         if  (x1 = 3 & x2 = 1 & x3 ^= 5) /* remove invalid status
               quo options */
26                 |(x1 = 3 & x2 = 2 & x3 < 5) /* anonymised daily
                      results in bill increase */
27                 |(x1 < 3 & x2 = 1 & x3 > 5) /* non-anonymised
                      granualar data results in bill decrease */
28                 then do; f1 = 0; f2 = 0; end;
```

```
29                  else do; f1 = 1; f2 = 1; end;
30          /*if (x3=1)
31                  then do; x3=-0.2; end;
32          if (x3=2)
33                  then do; x3=-0.15; end;
34          if (x3=3)
35                  then do; x3=-0.1; end;
36          if (x3=4)
37                  then do; x3=-0.05; end;
38          if (x3=5)
39                  then do; x3=0; end;
40          if (x3=6)
41                  then do; x3=0.05; end;
42          if (x3=7)
43                  then do; x3=0.1; end;
44          if (x3=8)
45                  then do; x3=0.15; end;
46          if (x3=9)
47                  then do; x3=0.2; end;*/
48          run;
49  proc print data=cand_list; run;
50  %macro res;
51  do i = 1 to nalts;
52          do k = i + 1 to nalts;
53                  if (x[i,1] > x[k,1] & x[i,2] >= x[k,2] & x[i,3]
                       < x[k,3]) /* less granular is more expensive
                       a.e.e. */
54                          then bad = bad + 1;
55                  if (x[i,1] < x[k,1] & x[i,2] <= x[k,2] & x[i,3]
                       > x[k,3]) /* less granular is more expensive
                       a.e.e. */
56                          then bad = bad + 1;
57                  if (x[i,2] > x[k,2] & x[i,1] >= x[k,1] & x[i,3]
                       < x[k,3]) /* anonymised is more expensive
```

```
                    than non-anonmymised a.e.e */
58                      then bad = bad + 1;
59                  if (x[i,2] < x[k,2] & x[i,1] <= x[k,1] & x[i,3]
                    > x[k,3]) /* anonymised is more expensive
                    than non-anonmymised a.e.e */
60                      then bad = bad + 1;
61                  if (x[i,2]=2 & x[k,2]=2)/* both alternatives
                    cannot be anonymised */
62                      then bad = bad + 1;
63                  if (x[i,1:2] = x[k,1:2]) /* exclude dominated
                    alternatives and duplicates */
64                      then bad = bad + 1;
65                  end;
66          end;
67  %mend;
68  %choiceff(data=cand_list, /* candidate set of alternatives */
69                  model=class(x1-x3/ zero=first), /* model with
                    alternative specific constant */
70                  nsets=%eval(&n*&blocks), /* number of choice
                    sets */
71                  flags= f1-f2, /* number of alternatives per
                    choice set */
72                  restrictions=res, /* name of the restrictions
                    macro */
73                  resvars=x1-x3, /* vars used in defining
                    restrictions */
74                  rscale = alt,
75                  seed=&seed, /* random number seed */
76                  maxiter=&iters, /* maximum number of designs to
                    make */
77                  options=relative coded, /* display relative D-
                    efficiency */
78                  beta= 0.28166 0.29662 0.22407 -0.15215 -0.12801
                    -0.33625 -0.57899 -0.97788 -1.76238 -1.67874
```

```
                    -1.86093) /* assumed beta vector, Ho: b=0 */
79 proc print data=bestcov label;
80 title 'Variance-Covariance Matrix';
81 id __label;
82 label __label = '00'x;
83 run;
84 title 'Blocked design';
85 %mktblock(data=best, /* D-effcient design */
86                   out=blocked, /* output */
87                   nalts = &m, /* number of alternatives */
88                   nblocks=&blocks, /* number of blocks */
89                   factors = x1-x3, /* design variables */
90                   iter=&iters, /* number of trys */
91           seed=&seed) /* random seed */
92 proc format;
93 value af 1 = 'None' 2 = 'Anonymised';
94 value tf 1 = 'Minute-by-minute' 2 = 'Half-hourly' 3 = 'Daily';
95 value bf 1 = -0.2 2 = -0.15 3 = -0.1 4 = -0.05 5 = 0 6 = 0.05 7
     = 0.1 8 = 0.15 9 = 0.2;
96 run;
97 data ChoiceDesign;
98 set blocked;
99 format x1 tf. x2 af. x3 bf.;
100 label x1 = 'Frequency' x2 = 'Anonymisation' x3 = 'Expected Bill
     Change';
101 run;
102 proc print label; var x1-x3; id block set alt; run;
103
104 proc export
105   data=ChoiceDesign
106   dbms=xlsx
107   outfile="final_design.xlsx"
108   replace;
109 run;
```

APPENDIX C
Closed-Form Expressions for Location-Scale Distributions

C.1 1-Wasserstein Distance between Selected Distributions

Theorem 5.3 provides a convenient method for determining the 1-Wasserstein distance between two distributions within a location-scale family. For distributions with non-negative support (e.g. exponential, Weibull) the 1-Wasserstein is simply the mean of such a distribution (Y) with a location parameter of $\alpha_y = (\alpha_1 - \alpha_2)$ and scale parameter of $\beta_y = (\beta_1 - \beta_2)$. For distributions with real support ($x \in \mathbb{R}$) (e.g. Gaussians) or bounded but both positive and negative support (e.g. $U(-1, 1)$) the theorem requires an additional step to obtain a closed-form/analytical solution. The 1-Wasserstein distance is the mean of the absolute value (folded) of the distribution. In many cases explicit formulae for the folded distribution are readily available. Table C.1 summarises the 1-Wasserstein distance for widely used location-scale distributions.

C.1.1 Closed-form Bounds

There are a number of well established bounds on the 1-Wasserstein distances for distributions with finite first and second moments. We note that these bounds are conventionally presented in terms of means and standard deviations (μ, σ) rather than location and scale (α, β).

Existing Bounds

Below we present the tightest existing bounds for location-scale distributions before providing a new upper bound for univarite location-scale distributions and discussing the

conditions under which this new bound is tighter than the extant literature.

Upper Bound

Lemma C.1. *Given two univariate independent distributions $X_1 \sim (\mu_1, \sigma_1)$ and $X_2 \sim (\mu_2, \sigma_2)$ within a location-scale family, the 1-Wasserstein distance between them is upper bounded by:*

$$W_1^{UB2}(X_1, X_2) = \sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2} \tag{C.1}$$

Proof. Given two spherical distributions $X_1 \sim (\mu_1, \Sigma_1)$ and $X_2 \sim (\mu_2, \Sigma_2)$, where the marginal distributions are orthogonal (i.e. $\Sigma_i = \sigma_i^2 I_d$) the 2-Wasserstein distance admits a closed-form:

$$W_2^2(X_1, X_2) = (\mu_1 - \mu_2)^2 + d\,(\sigma_1 - \sigma_2)^2 \tag{C.2}$$

This is also known as the Frechet distance [338]. Next, note that $W_p \leq W_q$ for $p \leq q$ by Hölder's inequality[281, Remark 6.6] meaning:

$$W_1(X_1, X_2) \leq \sqrt{W_2^2(X_1, X_2)} \tag{C.3}$$

$$= \sqrt{(\mu_1 - \mu_2)^2 + d\,(\sigma_1 - \sigma_2)^2} \tag{C.4}$$

In the univariate case $d = 1$, which concludes the proof. □

Lower Bound

Lemma C.2. *Given two distributions X_1 and X_2 with $E[X_1] = \mu_1$ and $E[X_2] = \mu_2$ the 1-Wasserstein distance between them is lower bounded by [339]:*

$$W_1^{LB}(X_1, X_2) = |\mu_1 - \mu_2| \tag{C.5}$$

This result also agrees with analysis in [282, Proposition 3.2] where when one distribution dominates the other ($X_1 > X_2$), for example if $\sigma_1 = \sigma_2$ and $\mu_1 > \mu_2$ then the 1-Wasserstein distance is:

$$W_1(X_1, X_2) = \mathbb{E}[X_1] - \mathbb{E}[X_2] \tag{C.6}$$

$$= W_1^{LB}(X_1, X_2) \tag{C.7}$$

New Linear Upper Bound

We now present a new upper bound on the 1-Wasserstein distance specific to location-scale distributions.

Theorem C.3. *Given two univariate independent random variables $X_1 = \alpha_1 + \beta_1 Z$ and $X_2 = \alpha_2 + \beta_2 Z$ the 1-Wasserstein distance between them is upper bounded by:*

$$W_1^{UB}(X_1, X_2) = |\alpha_1 - \alpha_2| + \mathbb{E}[|Z|]|\beta_1 - \beta_2| \tag{C.8}$$

where, $Z \sim (0, 1)$ is a standard distribution within the location-scale family.

Proof.

$$W_1(X_1, X_2) = \int_0^1 |F_1^{-1}(q) - F_2^{-1}(q)| dq \tag{C.9}$$

$$\begin{aligned} &= \int_0^1 |\Big(\alpha_1 + \beta_1 \Phi_Z^{-1}(q)\Big) \\ &\qquad - \Big(\alpha_2 + \beta_2 \Phi_Z^{-1}(q)\Big)| dq \end{aligned} \tag{C.10}$$

by the triangle inequality ($|x + y| \leq |x| + |y|$):

$$\begin{aligned} W_1(X_1, X_2) \leq &|\alpha_1 - \alpha_2| \\ &+ |\beta_1 - \beta_2| \int_0^1 |\Phi_Z^{-1}(q)| dq \end{aligned} \tag{C.11}$$

then note that $\int_0^1 |\Phi_Z^{-1}(q)| dq = \mathbb{E}[|Z|]$. □

The new upper bound in Theorem C.3 is linear in distributional parameters, as opposed to the existing bound in Lemma C.1. As X_1 and X_2 are from the same location-scale family we can reframe the expressions such that $\mu_i = \alpha_i$ and $\sigma_i = \beta_i$. We note that when $\alpha_1 = \alpha_2 / \mu_1 = \mu_2$ then (C.1) reduces to $|\sigma_1 - \sigma_2|$ and (C.8) reduces to $\mathbb{E}[|Z|]|\beta_1 - \beta_2|$. Therefore if $\mathbb{E}[|Z|] \leq 1$, then (C.8) will provide a tighter upper bound.

C.1.2 Gaussians Random Variables

In this section we study in greater detail the 1-Wasserstein distance between independent univariate Gaussians using the closed-form bounds and exact expressions developed in the previous sections. We note here that for Gaussians $\mu = \alpha$ and $\sigma = \beta$ so we parameterise them using μ and σ throughout.

Improved Upper Bound

In Section C.1.1 we showed a generic upper bound based on the 2-Wasserstein distance. However, a specific upper bound for 1-Wasserstein distance between Gaussian distributions, based on distributional parameters, has also been developed[340].

Lemma C.4. *Given two independent multivariate Gaussians $X_1 \sim N(\mu_1, \Sigma_1)$ and $X_2 \sim N(\mu_2, \Sigma_2)$ an upper bound for the d-dimensional 1-Wasserstein distance is[340, Lemma 2.4]:*

$$\begin{aligned} W_1(X_1, X_2) \leq & |\mu_1 - \mu_2| \\ & + \left(\sum_{j=1}^{d} \left\{ \left(\sqrt{\lambda_{1,j}} - \sqrt{\lambda_{2,j}} \right)^2 \right. \right. \\ & \left. \left. + 2\sqrt{\lambda_{1,j}\lambda_{2,j}} \left(1 - v_{1,j} \cdot v_{2,j}\right) \right\} \right)^{1/2} \end{aligned} \tag{C.12}$$

where, μ_i are the means, $\lambda_{i,j}$ is the ordered spectrum of the d-dimensional Gaussians and $v_{i,j}$ is the associated orthonormal basis of the eigenvectors.

In the univaiate case (C.12) simplifies to (denoted $W_1^{UB1}(X_1, X_2)$):

$$W_1(X_1, X_2) \leq |\mu_1 - \mu_2| + |\sigma_1 - \sigma_2| \tag{C.13}$$

We can recover (C.13) which by applying the Cauchy-Schwartz inequality to (C.1) $\left(\sum_i x_i^2 \leq (\sum_i x_i)^2\right)$:

$$W_1(X_1, X_2) \leq |\mu_1 - \mu_2| + \sqrt{d}|\sigma_1 - \sigma_2| \tag{C.14}$$

$$= |\mu_1 - \mu_2| + |\sigma_1 - \sigma_2| \tag{C.15}$$

We show that our new bound in Theorem C.3 is tighter bound than either $W_1^{UB1}(X_1, X_2)$ or $W_1^{UB2}(X_1, X_2)$ under certain conditions. Although $W_1^{UB2}(X_1, X_2)$ provides a tighter bound in general, it is possible to obtain a linear bound that is always tighter than $W_1^{UB1}(X_1, X_2)$ and also tighter than $W_1^{UB2}(X_1, X_2)$ when $\mu_1 = \mu_2$. The new linear upper bound using Theorem C.3 is detailed in the corollary below:

Corollary C.5. *Given two univariate independent Gaussians $X_1 = N(\mu_1, \sigma_1^2)$ and $X_2 = N(\mu_2, \sigma_2^2)$ the 1-Wasserstein distance is upper bounded by:*

$$W_1(X_1, X_2) \leq |\mu_1 - \mu_2| + \sqrt{\frac{2}{\pi}}|\sigma_1 - \sigma_2| \tag{C.16}$$

Exact Analytical Expression

Using Theorem 5.3, the 1-Wasserstein distance between two univarite Gaussians is the mean of a folded Gaussian[284, Equation 7].

Corollary C.6. *Given two univariate independent Gaussians $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ the 1-Wasserstein distance is equal to the mean of a folded Gaussian $E[|Y|]$ where, $Y \sim N(\mu_y = \mu_1 - \mu_2, \sigma_y^2 = (\sigma_1 - \sigma_2)^2)$:*

$$\begin{aligned} W_1(X_1, X_2) = |\mu_y| \left[1 - 2\Phi_N\left(-\frac{|\mu_y|}{|\sigma_y|}\right)\right] \\ + |\sigma_y|\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_y^2}{2\sigma_y^2}\right) \end{aligned} \tag{C.17}$$

As shown above the 1-Wasserstein distance can be expressed as a function of distributional parameters and the standard normal CDF, $\Phi_N(x)$.

Asymptotic Bounds

Given the exact analytical representation of the 1-Wasserstein distance in terms of the distributional parameters we determine the tightness of the closed-form upper and lower bounds and establish asymptotic bounds. By taking limits over the distributional parameters (μ_y, σ_y) we produce the following proposition.

Proposition C.7. *Given two univariate independent Gaussians $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ the 1-Wasserstein distance between them converges asymptotically to:*

$$\lim_{\sigma_y \to 0 | \mu_y \to \infty/-\infty} W_1(X_1, X_2) = W_1^{LB}(X_1, X_2) \tag{C.18}$$

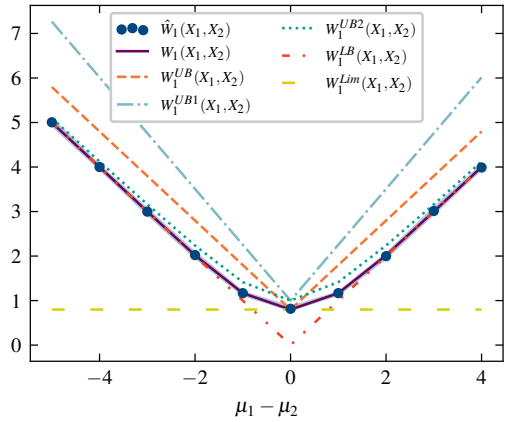
$$\lim_{\sigma_y \to \infty | \mu_y \to 0} W_1(X_1, X_2) = W_1^{Lim}(X_1, X_2) \tag{C.19}$$

where, $W_1^{Lim}(X_1, X_2) = W_1^{UB}(X_1, X_2) - |\mu_1 - \mu_2| = \sqrt{\frac{2}{\pi}}|\sigma_1 - \sigma_2|$.

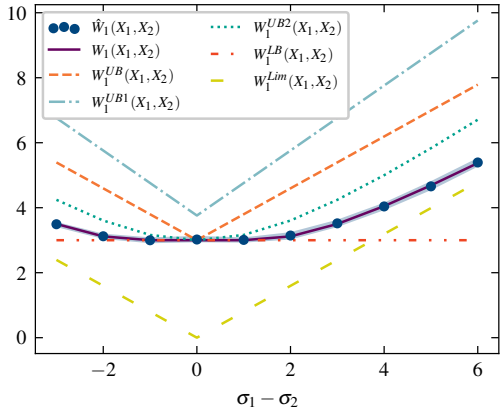
Proof.

$$\begin{aligned} \lim_{|\sigma_y| \to 0} W_1(X_1, X_2) &= |\mu_y| \left[1 - 2\Phi\left(-\infty\right)\right] \\ &\quad + (0)\sqrt{\frac{2}{\pi}} \exp\left(-\infty\right) && \text{(C.20)} \\ &= |\mu_y| \left[1 - 2(0)\right] && \text{(C.21)} \\ &= |\mu_1 - \mu_2| && \text{(C.22)} \\ &= W_1^{LB}(X_1, X_2) && \text{(C.23)} \end{aligned}$$

□



(a) $X_1 \sim N(\mu_1, 4), X_2 \sim N(5, 9)$



(b) $X_1 \sim N(2, \sigma_1^2), X_2 \sim N(5, 9)$

Figure C.1: 1-Wasserstein distance and bounds for univariate independent Gaussians

The proofs for the remaining limits ($\sigma_y \to \infty, \mu_y \to 0, \mu_y \to \infty/-\infty$) are similar and therefore omitted. We note that the results in the degenerate cases have been reported in [340, Example 2.5 & 2.6].

Improved Lower Bound

Based on the asymptotic analysis in Proposition C.7 we can see that a tighter lower bound can be obtained for univariate Gaussians.

Proposition C.8. *Given two univariate independent Gaussians* $X_1 = N(\mu_1, \sigma_1^2)$ *and* $X_2 = N(\mu_2, \sigma_2^2)$ *the 1-Wasserstein distance is lower bounded by:*

$$W_1(X_1, X_2) \geq \max\left(|\sigma_y|\sqrt{\frac{2}{\pi}}, |\mu_y|\right) \tag{C.24}$$

Proof. Staring from (5.34) we see that each component is positive. As a result:

$$W_1(X_1, X_2) \geq |\sigma_y|\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_y^2}{2\sigma_y^2}\right) \tag{C.25}$$

$$\geq W_1^{Lim}(X_1, X_2) \tag{C.26}$$

$$= |\sigma_y|\sqrt{\frac{2}{\pi}} \tag{C.27}$$

Additionally, we know from Lemma C.2 we have the lower bound $W_1^{LB}(X_1, X_2) = |\mu_y|$. □

Figure C.1 illustrates the improved bounds and exact expressions for the 1-Wasserstein distance between univariate Gaussians. $\hat{W}_1(X_1, X_2)$ is the empirical 1-Wasserstein distance

averaged over $N_r = 10^2$ simulations with $N_s = 10^4$ samples in each simulation. The shaded area indicates the 95% confidence interval.

Although $W_1^{UB2}(X_1, X_2)$, defined in (C.1), is much tighter in general, the linear upper bound $W_1^{UB}(X_1, X_2)$ defined in (C.16) is better than the existing linear upper bound $W_1^{UB1}(X_1, X_2)$ defined in (C.13). We see that when either $\sigma_y \to 0$ or $\mu_y \to \infty/-\infty$ the lower bound $W_1^{LB}(X_1, X_2)$ is tight. However, when either $\sigma_y \to \infty$ or $\mu_y \to 0$ the asymptotics do not converge to the upper bound but rather an intermediate value $W_1^{Lim}(X_1, X_2)$, as discussed in Proposition C.8. As a result the upper bound $W_1^{UB}(X_1, X_2)$ is only tight when $\sigma_y = 0$ resulting in $W_1^{UB}(X_1, X_2) = W_1^{LB}(X_1, X_2)$. This is clearly visible in Figure C.1a.

C.1.3 Uniformly Distributed Random Variables

Although uniformly distributed random variables are part of a location-scale family, a closed-form expression for the 1-Wasserstein distance between them does not have a single expression in the general case. Instead, there are two distinct cases as the mean of the resulting folded distribution is different depending on α_y and β_y.

Proposition C.9. *Given two univariate uniformly distributed random variables $X_1 = \alpha_1 + \beta_1 Z$ and $X_2 = \alpha_2 + \beta_2 Z$. The 1-Wasserstein distance between them is:*

$$W_1(X_1, X_2) = \frac{1}{2}\left(|a_y| + |b_y|\right), \quad \textit{if } a_y, b_y \geq 0 \textit{ or } \leq 0 \tag{C.28}$$

$$W_1(X_1, X_2) = \frac{1}{2}\left(\frac{a_y^2 + b_y^2}{b_y - a_y}\right), \quad \textit{otherwise} \tag{C.29}$$

Proof. By Theorem 5.3 the underlying random variable $Y \sim (\alpha_1 - \alpha_2) + (\beta_1 - \beta 2)Z$. This is equivalent to a uniform random variable $Y \sim U(a_y = min((\alpha_1 - \alpha_2), (\beta_1 - \beta 2) + (\alpha_1 - \alpha_2), b_y = max((\alpha_1 - \alpha_2), (\beta_1 - \beta 2) + (\alpha_1 - \alpha_2))$. Below we derive the two cases depending on the resulting random variable $Y \sim U(a_y, b_y)$; (1) when Y is non-negative or non-positive (i.e. $a_y, b_y \geq 0$ or $a_y, b_y \leq 0$) or (2) when Y spans the origin (i.e. $a_y < 0, b_y > 0$).

(1) $a_y, b_y \geq 0$ **or** $a_y, b_y \leq 0$ In this case $|Y| \sim U(|a_y|, |b_y|)$ which means that $\mathbb{E}[|Y|] = |\mathbb{E}[Y]|$, resulting in a straight forward closed-form expression:

$$W_1(X_1, X_2) = \frac{1}{2}\left(|a_y| + |b_y|\right) \tag{C.30}$$

(2) $a_y < 0, b_y > 0$ In this case $|Y|$ will not have a uniform distribution, instead the negative support ($a_y \leq x \leq 0$) of Y will be folded over into the positive domain. The

expected value will then be:

$$E[|Y|] = \int_0^{a_y} 2Cx\,dx + \int_{a_y}^{b_y} Cx\,dx \tag{C.31}$$

$$= \frac{C}{2}\left(a_y^2 + b_y^2\right) \tag{C.32}$$

where C is the normalising constant for Y:

$$2Ca_y + C(b_y - a_y) = 1 \tag{C.33}$$

$$C = \frac{1}{b_y - a_y} \tag{C.34}$$

The 1-Wasserstein distance is then:

$$W_1(X_1, X_2) = \frac{1}{2}\left(\frac{a_y^2 + b_y^2}{b_y - a_y}\right) \tag{C.35}$$

□

Table C.1: 1-Wasserstein Distance for Selected Location-Scale Distributions

Distribution	$W_1(X_1, X_2)/\mathbb{E}[\lvert Y\rvert]$ [1]	Source
Uniform[2]	$\frac{1}{2}\left(\lvert a_y\rvert + \lvert b_y\rvert\right), a_y \,\&\, b_y \geq 0 \text{ or } \leq 0$ $\frac{1}{2}\left(\frac{a_y^2 + b_y^2}{b_y - a_y}\right), \quad \text{otherwise}$	
Gaussian	$\lvert\alpha_y\rvert\left[1 - 2\Phi_N\left(-\frac{\lvert\alpha_y\rvert}{\lvert\beta_y\rvert}\right)\right] + \lvert\beta_y\rvert\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\alpha_y^2}{2\left(\beta_y\right)^2}\right)$	[284]
Laplace	$\lvert\alpha_y\rvert + \lvert\beta_y\rvert\exp\left(-\frac{\lvert\alpha_y\rvert}{\lvert\beta_y\rvert}\right)$	[286]
Logistic	$\lvert\alpha_y\rvert + 2\lvert\beta_y\rvert\ln\left(1 + \exp\left(-\frac{\lvert\alpha_y\rvert}{\lvert\beta_y\rvert}\right)\right)$	[335]
Gamma [3]	$\alpha_y + k\beta_y, \quad \alpha_y, \beta_y \geq 0$	
Weibull [4]	$\alpha_y + \beta_y\Gamma(1 + 1/k), \quad \alpha_y, \beta_y \geq 0$	
Exponential[5]	$\alpha_y + \beta_y, \quad \alpha_y, \beta_y \geq 0$	
Rayleigh[6]	$\alpha_y + \lvert\beta_y\rvert\sqrt{\frac{\pi}{2}}, \quad \alpha_y, \beta_y \geq 0$	
Student's t [7]	$2\lvert\beta_y\rvert\sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\nu-1)}$	[336]

[1] $Y \sim$ (location: $\alpha_y = (\alpha_1 - \alpha_2)$, scale: $\beta_y = (\beta_1 - \beta_2)$).

[2] $Y \sim \alpha_y + \beta_y Z_U = U(a_y, a_y + b_y)$. The conventional upper and lower bounds of the uniform distribution Y are $a_y = \min\left((\alpha_1 - \alpha_2), (\beta_1 - \beta_2) + (\alpha_1 - \alpha_2)\right), b_y = \max\left((\alpha_1 - \alpha_2), (\beta_1 - \beta_2) + (\alpha_1 - \alpha_2)\right)$.

[3] The Gamma distribution is a location-scale distribution for any given k. It is non-negative when $\alpha_y, \beta_y \geq 0$ meaning $E[|Y|] = E[Y]$.

[4] The Weibull distribution is a location-scale distribution for any given k. It is non-negative when $\alpha_y, \beta_y \geq 0$ meaning $E[|Y|] = E[Y]$.

[5] $\beta_y = \frac{\lambda_1 - \lambda_2}{\lambda_1, \lambda_2}$ where λ_i is the conventional inverse scale parameter. Equivalent to $Y \sim Weibull(\beta_y, k = 1)$. Also reported independently in [337] by evaluating (5.23) directly.

[6] Equivalent to $Y \sim Weibull(\sqrt{2}\beta_y, k = 2)$.

[7] Only applies for $\alpha_i = 0$ and $\nu > 1$, where ν is the degrees of freedom.

APPENDIX D
Supplementary Proofs and Reformulations

D.1 Proof of Myerson's Lemma

Assume a seller has two potential reserve prices $0 \leq \theta_i \leq \theta_i + h$. If the sellers actual price is θ_i then the IC constraints dictate that:

$$T_i(\theta_i) - \theta_i Q_i(\theta_i) \geq T_i(\theta_i + h) - \theta_i Q_i(\theta_i + h) \tag{D.1}$$

If instead the seller's actual prices is $\theta_i + h$ then:

$$T_i(\theta_i + h) - (\theta_i + h)Q_i(\theta_i + h) \geq T_i(\theta_i) - (\theta_i + h)Q_i(\theta_i) \tag{D.2}$$

Combining the above two inequalities gives us the following 'payment sandwich':

$$(\theta_i + h)[Q_i(\theta_i + h) - Q_i(\theta_i)] \leq T_i(\theta_i + h) - T_i(\theta_i) \leq (\theta_i)[Q_i(\theta_i + h) - Q_i(\theta_i)] \tag{D.3}$$

To ensure IC then:

$$Q_i(\theta_i + h)[(\theta_i + h) - \theta_i] \leq Q_i(\theta_i)[(\theta_i + h) - \theta_i] \tag{D.4}$$

$$Q_i(\theta_i + h) \leq Q_i(\theta_i) \tag{D.5}$$

The allocation $Q(\theta)$ must therefore be monotonically decreasing.

Now to determine the payment rule:

$$\lim_{h \to 0^+} \frac{(\theta_i + h)[Q_i(\theta_i + h) - Q_i(\theta_i)]}{h} \leq \frac{T_i(\theta_i + h) - T_i(\theta_i)}{h} \leq \frac{(\theta_i)[Q_i(\theta_i + h) - Q_i(\theta_i)]}{h} \tag{D.6}$$

$$\frac{dT_i(\theta_i)}{d\theta_i} = \theta_i \frac{dQ_i(\theta_i)}{d\theta_i} \tag{D.7}$$

The payment for seller i can then be determined:

$$T_i(\theta_i) = \int_{\underline{\theta}_i}^{\bar{\theta}_i} -\theta_i \frac{dQ_i(\theta_i)}{d\theta_i} d\theta_i \tag{D.8}$$

$$= -\bar{\theta}_i Q_i(\theta_i) + \theta_i Q_i(\theta_i) + \int_{\theta_i}^{\bar{\theta}_i} Q_i(\theta_i) d\theta_i \tag{D.9}$$

$$= \theta_i Q_i(\theta_i) + \int_{\theta_i}^{\bar{\theta}_i} Q_i(\theta_i) d\theta_i \tag{D.10}$$

Using the payment rule established above we can reformulate the platform's problem in terms of selection probabilities and virtual costs which are defined as:

$$\psi_i(\theta_i) = \theta_i + \frac{F_i(\theta_i)}{f_i(\theta_i)}, \quad \forall i \in \mathcal{N}, \theta_i \in \mathbb{R}_+ \tag{D.11}$$

The payment terms are reformulated by applying the payment rule and integrating by parts:

$$\int_\Theta \sum_{i\in\mathcal{N}} t_i(\theta) f(\theta) d\theta = \sum_{i\in\mathcal{N}} \int_{\underline{\theta}_i}^{\bar{\theta}} T_i(\theta_i) f_i(\theta_i) d\theta \tag{D.12}$$

$$= \sum_{i\in\mathcal{N}} \int_{\underline{\theta}_i}^{\bar{\theta}} \left[\theta_i Q_i(\theta_i) + \int_{\theta_i}^{\bar{\theta}} Q_i(s) ds \right] f_i(\theta_i) d\theta \tag{D.13}$$

$$= \sum_{i\in\mathcal{N}} \int_{\underline{\theta}_i}^{\bar{\theta}} Q_i(\theta_i) \psi_i(\theta_i) f_i(\theta_i) d\theta \tag{D.14}$$

$$= \mathbb{E}_\Theta \left[\sum_{i\in\mathcal{N}} q_i(\theta_i) \psi_i(\theta_i) \right] \tag{D.15}$$

D.2 Proof of Monotonicity

To ensure feasibility of the mechanism has been maintained we need to show that the allocation q is monotonically decreasing in θ, the data owners' reserve prices. As the MISOCP reformulation is exact we can analyse the original problem in (6.17). Let $q = [q_0, \ldots, q_N]$ be the optimal solution of the (6.17) for the reserve price vector $\theta = [\theta_1, \ldots, \theta_N]$. Now, following the approach in [52], suppose we increase one of the θ_i's, which, without loss of generality, we assume is the first one. Let $\theta_1' > \theta_1$ and $\theta_i' = \theta_i \ \forall i > 1$ and suppose $q' = [q_0', \ldots, q_N']$ is the corresponding optimal solution of (6.17). The optimality condition implies that, if the true reserve price vector is θ:

$$q_0 B_{\mathcal{M}}(X_R) + C^{inf} \frac{\|q_i W_i\|}{\sum_{i\in\mathcal{N}} q_i} + \sum_{i\in\mathcal{N}} q_i \psi_i(\theta_i) \leq q_0' B_{\mathcal{M}}(X_R) + C^{inf} \frac{\|q_i' W_i\|}{\sum_{i\in\mathcal{N}} q_i'} + \sum_{i\in\mathcal{N}} q_i' \psi_i(\theta_i) \tag{D.16}$$

and if the true reserve price vector is θ':

$$q_0' B_{\mathcal{M}}(X_R) + C^{inf} \frac{\|q_i' W_i\|}{\sum_{i\in\mathcal{N}} q_i'} + \sum_{i\in\mathcal{N}} q_i' \psi_i(\theta_i') \leq q_0 B_{\mathcal{M}}(X_R) + C^{inf} \frac{\|q_i W_i\|}{\sum_{i\in\mathcal{N}} q_i} + \sum_{i\in\mathcal{N}} q_i \psi_i(\theta_i') \tag{D.17}$$

Then taking the summation of both sides of the inequalities above and using the fact that $\theta_i' = \theta_i \ \ \forall i > 1$ we obtain:

$$\left(q_1 - q_1'\right)\left(\psi_1(\theta_1) - \psi_1(\theta_1')\right) \leq 0 \tag{D.18}$$

Assumption 6.1 and the above inequality show that the point-wise optimisation problem (6.17) is monotonically decreasing in θ (i.e. $\left(q_1 - q_1'\right) \geq 0$).

We note that the monotonicity of the formulations can be shown in a similar manner to the joint optimisation mechanism using the Lagrangian relaxation (of the budget constraint) of the problem.

D.3 MISOCP Formulations for Budgeted Mechanisms

D.3.1 Infinite Population

The MISOCP formulations for the infinite exogenous and endogenous budget mechanisms are:

$$\min_{q,s,z} \quad s \tag{D.19}$$

$$\text{s.t.} \sum_{i\in\mathcal{N}} q_i \psi_i \leq B_{\mathcal{M}}(X_R) - C^{inf} s \text{ or } \sum_{i\in\mathcal{N}} q_i \psi_i \leq B \tag{D.19a}$$

$$\|q_i W_i\| \leq \sum_{i\in\mathcal{N}} z_i \quad \Longleftrightarrow \quad \left(\sum_{i\in\mathcal{N}} z_i, q^T W\right) \in \mathcal{Q}^{N+1} \tag{D.19b}$$

$$0 \leq z_i \leq M q_i, \quad \forall i \in \mathcal{N} \tag{D.19c}$$

$$0 \leq s - z_i \leq M(1 - q_i), \quad \forall i \in \mathcal{N} \tag{D.19d}$$

$$1 \leq \sum_{i\in\mathcal{N}} q_i \leq N \tag{D.19e}$$

$$q \in \{0,1\}^N, s \in \mathbb{R}_+, z \in \mathbb{R}_+^N \tag{D.19f}$$

D.3.2 Finite Population

The MISOCP formulations for the finite exogenous and endogenous budget mechanisms are:

$$\min_{q,r,s,z} \quad s \tag{D.20}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{N}} q_i\psi_i(\theta_i) \le B_{\mathcal{M}}(X_R) - C^{fin}s \text{ or } \sum_{i\in\mathcal{N}} q_i\psi_i \le B \tag{D.20a}$$

$$\|Wr\| \le \sum_{i\in\mathcal{N}} z_i \quad \Longleftrightarrow \quad \left(\sum_{i\in\mathcal{N}} z_i, r^T W\right) \in \mathcal{Q}^{N+1} \tag{D.20b}$$

$$r_{i,j} \le q_i, \quad \forall i \in \mathcal{N} \tag{D.20c}$$

$$r_{i,j} \le 1 - q_j, \quad \forall j \in \mathcal{N} \tag{D.20d}$$

$$r_{i,j} \ge q_i - q_j, \quad \forall i \in \mathcal{N}, j \in \mathcal{N}/i \tag{D.20e}$$

$$0 \le z_i \le Mq_i, \quad \forall i \in \mathcal{N} \tag{D.20f}$$

$$0 \le s - z_i \le M(1 - q_i), \quad \forall i \in \mathcal{N} \tag{D.20g}$$

$$1 \le \sum_{i\in\mathcal{N}} q_i \le N \tag{D.20h}$$

$$q \in \{0,1\}^N, r \in \{0,1\}^{N^2-N}, s \in \mathbb{R}_+, z \in \mathbb{R}_+^N \tag{D.20i}$$

D.4 Joint Energy and Data Market Proofs

D.4.1 Equivalence of the Data-Driven Newsvendor and Quantile Regression.

This is adapted from similar results for a variant of the newsvendor problem in [314, Appendix A] and [310, Equation 8]. We start by reformulating the retailer's profit function into the canonical form for newsvendor problems, namely, in terms of underage and overage costs:

$$\Pi(q, D) = \lambda^r D - \lambda^w q - \lambda^- [D - q]^+ + \lambda^+ [q - D]^+ \tag{D.21}$$

$$= \lambda^r D - \lambda^w D + \lambda^w D - \lambda^w q - \lambda^- [D - q]^+ + \lambda^+ [q - D]^+ \tag{D.22}$$

$$= (\lambda^r - \lambda^w) D + \lambda^w \left([D - q]^+ - [q - D]^+\right) - \lambda^- [q - D]^+ + \lambda^+ [D - q]^+ \tag{D.23}$$

$$= (\lambda^r - \lambda^w) D + (\lambda^w - \lambda^-) [D - q]^+ - \left(\lambda^w - \lambda^+\right) [q - D]^+ \tag{D.24}$$

$$= (\lambda^r - \lambda^w) D - \lambda^u [D - q]^+ - \lambda^o [q - D]^+ \tag{D.25}$$

where, $\lambda^u = \lambda^- - \lambda^w$ and $\lambda^o = \lambda^w - \lambda^+$ are the overage and underage costs, respectively. Now, we incorporate the forecasting model to develop the link between the optimal bidding quantity, $q(\mathbf{\Phi}, \mathbf{x})$, and a set of historical features, $\mathbf{x}$. The retailer's profit maximisation problem is then:

$$\max_{\mathbf{\Phi}} \mathbb{E}_D[\Pi(q(\mathbf{\Phi}, \mathbf{x}), D)] = \frac{1}{N_s} \sum_{i}^{N_s} (\lambda^r - \lambda^w)\, d_i - \lambda^u \left[d_i - q_i(\mathbf{\Phi}, \mathbf{x})\right]^+ - \lambda^o \left[q_i(\mathbf{\Phi}, \mathbf{x}) - d_i\right]^+ \tag{D.26}$$

Notice that as the actual demand, D, is fixed so we can equivalently write the retailer's profit maximisation problem as a cost minimisation problem:

$$\min_{\mathbf{\Phi}} \mathbb{E}_D[C(q(\mathbf{\Phi}, \mathbf{x}))] = \min_{\mathbf{\Phi}} \frac{1}{N_s} \sum_{i}^{N_s} \lambda^u \left[d_i - q_i(\mathbf{\Phi}, \mathbf{x})\right]^+ + \lambda^o \left[q_i(\mathbf{\Phi}, \mathbf{x}) - d_i\right]^+ \tag{D.27}$$

$$= (\lambda^u + \lambda^o) \min_{\mathbf{\Phi}} \left[\frac{\lambda^u}{\lambda^u + \lambda^o} \left[d_i - q_i(\mathbf{\Phi}, \mathbf{x})\right]^+ + \frac{\lambda^o}{\lambda^u + \lambda^o} \left[q_i(\mathbf{\Phi}, \mathbf{x}) - d_i\right]^+ \right] \tag{D.28}$$

$$= \min_{\mathbf{\Phi}} \tau \left[d_i - q_i(\mathbf{\Phi}, \mathbf{x})\right]^+ + (1 - \tau) \left[q_i(\mathbf{\Phi}, \mathbf{x}) - d_i\right]^+ \tag{D.29}$$

where, $\tau = \frac{\lambda^u}{\lambda^u + \lambda^o}$, is the critical fractile, as discussed above. The retailer energy procurement problem is thus equivalent to a quantile regression problem, with the objective function differing by a constant factor, $\lambda^u + \lambda^o$.

D.4.2 Lipschitz Constant for Integrated Forecasting and Optimisation Problem

In order to obtain the Lipschitz constant for the entire framework we need to consider the Lipschitz constants of each stage. We start by noting that the newsvendor cost, as defined in (D.27), is Lipschitz with constant $K_{nv} = \max(\lambda^u, \lambda^o)$[276]. Next, we show that the expected profit difference between the optimal profit, achieved at the optimal order quantity of q^* when the retailer has access to the true demand distribution D, and the expected profit when ordering a different quantity $\hat{q}^*$ based on access to a different

demand distribution $\hat{D}$, can be upper bounded as follows[1]:

$$\Pi(q^*, D) - \Pi(\hat{q}^*, D) = C(\hat{q}^*, D) - C(q^*, D) \tag{D.30}$$

$$= C(\hat{q}^*, D) - C(q^*, \hat{D}) + C(q^*, \hat{D}) - C(q^*, D) \tag{D.31}$$

$$\overset{(a)}{\leq} C(\hat{q}^*, D) - C(q^*, \hat{D}) + \max(\lambda^u, \lambda^o) W(D, \hat{D}) \tag{D.32}$$

$$\overset{(b)}{\leq} C(\hat{q}^*, D) - C(\hat{q}^*, \hat{D}) + \max(\lambda^u, \lambda^o) W(D, \hat{D}) \tag{D.33}$$

$$\overset{(c)}{\leq} 2 \max(\lambda^u, \lambda^o) W(D, \hat{D}) \tag{D.34}$$

where, (a) and (c) result from the Lipschitzness of the newsvendor cost function and (b) follows from the fact that $C(\hat{q}^*, \hat{D}) \leq C(q^*, \hat{D})$. The Lipschitz constant is therefore $2 \max(\lambda^u, \lambda^o)$.

Next, to incorporate the forecasting model we leverage the composition properties of Lipschitz continuity. We assume the ANN forecasting model has appropriately clipped/bounded weight matrices, resulting in a Lipschitz constant, K_f. The resulting Lipschitz constant for the full framework is:

$$K = 2K_f \max(\lambda^u, \lambda^o) \tag{D.35}$$

[1]Similar results are presented in [276], [341], however they provide bounds on the DRO newsvendor problem using the WD and bounds on the classical newsvendor problem using the distribution means, respectively.