Quantifying Privacy in Smart Meter Data: A Comparative Analysis of Aggregation and AI-Generated Synthetic Data

John Corsten

Wolfson College

University of Oxford

A dissertation submitted for the degree of

MSc in Energy Systems

2024

Word count: 14,926

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Grünewald, for his invaluable guidance, insightful feedback, and continuous support throughout the course of my research. His passion, expertise, and patience were essential in the development of my own passion for the field and a topic whose outcome yields true impact.

I would also like to thank both Dr. Wallom and Dr. Sparrow, the course director and deputy course director for the MSc in Energy Systems at the University of Oxford, for their exceptional leadership, curation of an outstanding program, and their assembly of a cohort of peers whose diverse perspectives and collaborative spirit greatly enriched my academic experience.

My sincere thanks to the Energy Demand Observatory and Laboratory for their academic support and expertise. Mr. Chai and our research partners at the Octopus Centre for Net Zero also provided valuable feedback and contributed to the development of the ultimate direction of this dissertation.

To my parents, Katherine Corsten and Martin Corsten, thank you for instilling in me a passion for education and a reverence for science, your unwavering support and encouragement have been the foundation of my academic success, and I am forever grateful.

This dissertation would not have been possible without the contributions and support of all these individuals and organizations. Thank you to all.

Abstract

This dissertation addresses the critical challenge of balancing privacy and utility in smart meter data, with a focus on the UK's Smart Meter Implementation Program (SMIP). Smart meters provide essential data for optimising energy efficiency, grid management, and supporting sustainable energy transitions, but they also pose significant privacy risks by revealing sensitive consumer information. Currently, data aggregation, or averaging multiple profiles into one representative profile, is the primary method for reducing identification risks for privacy. However, the lack of a standardised, quantitative definition of "risk of identification" in policy and literature makes it difficult to assess or replace this approach. As advanced data analytics and AI become more integral to the energy sector, the demand for high-fidelity data increases, along with the potential for privacy breaches. Aggregation has proven vulnerable to certain identification attacks and often compromises data utility, especially in time series data like smart meter readings.

In response to these challenges, this research explores AI-powered synthetic data generation as an alternative to aggregation and introduces a novel, scientific framework for assessing privacy. This framework, adaptable to any anonymised time series model, uses random forest classification and Bayesian inference to evaluate identification risks based on a model output and the input data being anonymised. By implementing a rudimentary generative AI for smart meter data and systematically evaluating both this approach and traditional aggregation methods in terms of utility and privacy, this study demonstrates that synthetic data can enhance privacy protection while effectively preserving data utility. These findings have significant implications for the future management of smart meter data and could be applied to other sensitive time series data, as aggregation is among the most popular methods for anonymising distributed sensitive data worldwide. Adopting synthetic data could improve privacy protections in smart grid initiatives while maintaining the necessary data utility for decision-making and innovation. The novel framework this study develops also provides a robust tool for evaluating and comparing privacy-preserving data release models, offering valuable insights for future research and policy development. This work fills critical gaps in literature and sets the stage for more secure and efficient management of smart meter data in the UK and beyond.

Contents

1

Introduction

1.1 The Promise and Challenge of Smart Meters

The deployment of smart meters revolutionise the energy sector by enabling real-time monitoring of energy consumption, enhancing grid reliability, and supporting the transition to sustainable energy systems. The United Kingdom's Smart Meter Implementation Program (SMIP) represents one of the most ambitious smart grid projects globally, aiming to install over 53 million smart meters in homes and small businesses by the end of the decade. These devices, by providing detailed data on electricity and gas usage, hold significant promise for optimising energy efficiency, reducing costs, and enabling consumers to make informed decisions about their energy use. However, with this wealth of data comes a critical challenge: ensuring the privacy of consumers while maintaining the utility of the data for various stakeholders, including energy providers, researchers, and policymakers.

Smart meter data, despite its potential benefits, is inherently sensitive. Detailed consumption patterns can reveal intimate information about individuals' daily routines, their

presence or absence from home, and even personal habits. Furthermore, even data which is ostensibly benign to some, like energy consumption patterns, can be combined with other information to deduce highly sensitive data. This sensitivity classifies smart meter data as personal data under current data protection regulations in the UK, and necessitates robust privacy-preserving mechanisms. The fundamental challenge lies in striking a balance between maximising the utility of the data for legitimate purposes and obscuring the data such that the privacy of individuals is protected.

Smart meter data includes detailed information on energy usage, collected every 30 minutes for electricity and daily for gas, which is then transmitted to energy suppliers via a secure wireless network [1]. Over half of UK households are now equipped with smart meters, representing an approximately £13.5 billion endeavor [2], marking the largest engineering project ever undertaken in Europe [3]. This program's primary motivation is to lay foundation for a *smart grid*, a digitised electricity grid that dynamically shapes demand and tailors resource allocation for an intermittent supply and volatile load profile [4]. The smart grid also provides invaluable insights into how the energy system can be efficiently improved for the benefit of consumers and the environment. However, integral to this mission is widespread access to geographically diverse, metadata-rich, granular smart meter data [5]. Unfortunately, users of smart meter data encounter significant legal obstacles in obtaining it due to privacy regulation, often relying on inaccessible representations or heavily aggregated data that lose much of their real value.

1.2 Access to Smart Meter Data

The complex privacy policy and resulting difficulty in accessing relevant data serves as a major roadblock for research into smart meter data privacy and utility. However, a significant aspect of this dissertation's novelty is its unfettered access to a large, metadata-rich database of smart meter profiles provided by the Energy Demand Observatory and Laboratory (EDOL). More detail about the data resources is available at [6]. This access enables a comprehensive analysis that would otherwise be impossible, allowing for the development and testing of novel techniques without the constraints typically imposed by using randomly generated or small datasets.

This dissertation has approval to conduct tests on all contributions to the EDOL dataset, and permission from one participant to release their personal data. Therefore, while the entire dataset is used for research in this dissertation, all visual representations and aggregations of data presented here are specifically derived from the contributions of the one consenting participant.

1.3 The Tension Between Privacy and Utility

The current approach to privacy in the UK's smart meter data distribution primarily relies on aggregation, whereby multiple profiles are averaged to create a representative dataset which is assumed to minimise the risk of identification. However, this approach has significant limitations; aggregation diminishes data utility, particularly for applications that require high-fidelity data, such as detailed energy consumption analysis and load forecasting. Moreover, recent studies highlight the vulnerabilities of aggregation to various attacks, including differential attacks, where adversaries can use additional information to potentially identify individuals within the aggregated data.

Furthermore, the concept of privacy itself, as it pertains to smart meter data, lacks a clear, universally accepted quantitative definition. The UK's Office of Gas and Electricity Markets (Ofgem) mandate that data should be anonymised to the point where the risk of identification is "remote," [7] yet this standard is not quantitatively defined, leading to inconsistencies in the implementation and enforced minimum number of profiles aggregated per representation. Some distributors recognise 3 as a standard, UK Power Networks enforces 5 [7], and the prominent Smart Energy Research Lab (SERL) [8] uses 10 as the minimum aggregation size, however these policies are arbitrary as the respective privacy trade-offs of different parameters have no quantified nor scientific bases; this method is used because of its ubiquity across many industries and a qualitative, intuitive assessment of it's sufficiency. The absence of a rigorous scientific basis for these privacy-preserving techniques exacerbates the tension between the need for data utility and the imperative to protect consumer privacy as we have no mechanism for iterating on or replacing current practices while definitively not reducing its protection.

1.4 Objectives of the Dissertation

This dissertation seeks to tackle the challenges of balancing privacy and utility in smart meter data by developing a robust scientific framework for assessing privacy in data release models. With this framework, we evaluate the effectiveness of the existing aggregation model against an alternative approach: generative AI-powered synthetic data. Synthetic data generation involves creating artificial datasets that replicate the statistical properties of real data without releasing any individual's data directly. This method offers a promising solution to the privacy-utility trade-off by potentially providing high utility while maintaining stringent privacy standards. The proposed scientific and quantitative assessment framework enables a direct and unprecedented comparison between the incumbent aggregation model and this innovative synthetic approach. This dissertation lays the groundwork for future research in the nascent and "not yet properly researched" [9] field of privacy-preserving time series data release models.

1.5 Review of Literature and Methodological Approach

This study begins by reviewing the existing literature on smart meter data privacy and utility, focusing on the legal and technical frameworks that govern data protection. The literature review identifies significant gaps in current privacy-preserving methods and the need for a more nuanced understanding of privacy risks. It also examines advancements in other sectors, such as the differential privacy techniques used by the U.S. Census Bureau, and their potential applicability to smart meter data. Building on these insights, this dissertation develops a novel methodology for quantifying privacy in smart meter data, emphasising the concept of "identifiability", or the likelihood that an individual's data can be isolated within a dataset. By integrating techniques like random forest classification and Bayesian inference, this research proposes a quantitative framework for assessing privacy risks across different data release models. This framework is then applied to both the traditional aggregation model and a synthetic data generation approach, offering a comparative analysis of their effectiveness in balancing privacy and utility.

1.6 Significance and Implications

The findings of this research have significant implications for the future of smart meter data management in the UK. By providing a rigorous, quantitative assessment of privacy and utility, the study offers valuable insights into how privacy-preserving data release models can be improved and adapted to meet the needs of all stakeholders. This dissertation concludes by discussing the potential for synthetic data to replace aggregation as the standard for smart meter data distribution, thus contributing to the development of more secure and efficient energy systems.

This research not only fills a critical gap in literature but also provides a foundation for policymakers and energy providers to re-evaluate current data privacy strategies. As the adoption of smart meters continues to grow, ensuring the privacy of consumer data while maximising its utility will remain a paramount concern. This dissertation's contribution lies in its quantification of privacy and its demonstration of a framework for optimising the utility-privacy trade-off, offering a path forward for the future of smart meter data privacy in the UK.

2

Literature Review

This chapter reviews the literature relevant to assessing privacy in smart meter data release models, which forms the foundation for this dissertation's analysis. We begin with an overview of the utility of smart meter data and the associated privacy risks when this data is made publicly accessible. We examine the current privacy policies in place, highlighting their strengths, limitations, and areas for improvement within the context of the UK's regulatory framework.

We then delve into the concept of privacy as it pertains to smart meter data, considering how existing methods for quantifying privacy can be applied to evaluate data aggregation and other distribution models. We compare privacy practices in the UK's Smart Meter Implementation Program (SMIP) with those used by the United States Census Bureau, to understand why certain strategies have not been widely adopted for smart meter data. Additionally, alternative models to data aggregation are explored, assessing their potential to enhance privacy while maintaining data utility. The chapter concludes by identifying gaps in the current literature and outlining how this dissertation aims to address these issues, contributing to the ongoing

development of privacy-preserving practices for the UK's smart grid. Figure 2.1 visualises the high-level logical structure and objectives of this review of literature.

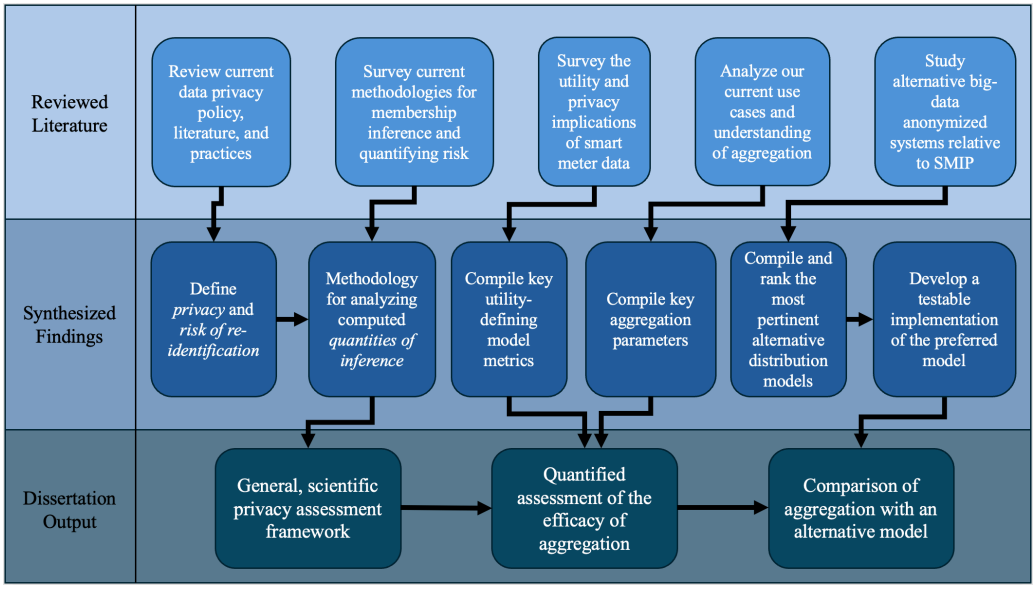

Figure 2.1: Depiction of the primary topics being surveyed in this literature review, our synthesised findings, and how these outputs relate to the ultimate products of this dissertation.

2.1 Smart Meter Data Privacy and Utility

The problems of increasing the utility of publicly available consumption data or increasing the anonymity of release models are trivial in isolation, however a central theme of this dissertation is the tension that has emerged between these mutually vital imperatives. In this section we review the nature of smart meter data, its utility, and the consumer privacy implications of publishing it today and in the future.

2.1.1 Balancing Privacy and Utility

Utility

At present, smart meter data is primarily consumed by three stakeholders: utilities and energy providers, researchers and academics, and regulators and policymakers [10].

Utilities enhance grid reliability and efficiency by leveraging real-time and historical data from consumers' smart meters alongside public databases. This data is utilised to plan trans-

former upgrades, analyse customer behavior for strategic planning, and identify opportunities for storage facility placement, thereby reducing renewable energy waste [1] [11]. Researchers typically access this data through utility partnerships, consumer studies, or public repositories, applying it to benchmark energy systems, develop new system innovations [12], model technology adoption, and study energy equity among many other applications [13]. Policymakers rely on anonymised data from mandatory utility reporting to plan infrastructure and develop policies that meet future demands and integrate new technologies [14], however this data is also aggregated before distribution [15]. In the context of this dissertation, "anonymity" refers to preservation of user data in public representations which can be identified, not the simple removal of associated metadata as can be the case.

Privacy Implications

While one's smart meter consumption data may seem benign to some, it is legally personal data and the preservation of customer privacy in the management of it is essential [10]. Whether the risk of unauthorised public access to consumption data appears severe or not, it is imperative that all personal data be protected to the highest standards. Even ostensibly benign data, like energy consumption patterns, can be combined with other information to deduce sensitive details about individuals' daily lives [16], such as their presence or absence from home, daily routines, and lifestyle habits [17]. This type of data access can lead to serious security risks and privacy breaches if misused [18].

In the UK, smart meter adoption is stagnating due to an increased pubic skepticism and waning public trust [19]. Uniform protection of all personal data is crucial to prevent exploitation and to maintain consumer trust in digital system [20, 21] as trust can be undermined by the inconsistent interpretation and application of legal mechanisms for data sharing systems [22]. The integrity of privacy frameworks hinges on consistently rigorous data protection measures, therefore the current outdated UK smart meter data protection schemes threaten to undermine the largest engineering project in the history of Europe.

2.2 Privacy Legislation and Smart Meter Data

2.2.1 Legislative Definition of Privacy

Personal data like residential smart meter readings must be anonimised before being publicly shared to maintain user privacy [10]. It is important to note that policies pertaining to the preservation of data privacy do not provide a quantitative definition of privacy or any risk thresholds. From reviewing the current policy as articulated by the UK Office of Gas and Electricity Markets (Ofgem), we see that privacy is considered a spectrum, and that *privacy* is conflated with a minimal *risk of identification* [23]. Data which is anonimised to the point where there is no risk has no "useful purpose" [23] from a network management or research point of view, but the ICO and Ofgem mandate that the risk of identification be mitigated until "it is remote" [7]. In this context, the risk of identification can be thought of as the likelihood that a malicious party could identify an individual's load profile as having contributed to a public data release.

2.2.2 Aggregation to Preserve Privacy

Researchers at Columbia University note that 'identifiability' is often mentioned in data distribution discussions but rarely quantitatively defined [24]. The SMIP reflects this, lacking quantitative definitions or security guarantees for 'risk of identification.' The standard practice in the UK, endorsed by Ofgem, is to aggregate multiple profiles into one, assuming this technique sufficiently preserves privacy by making it 'generally difficult' to identify individuals, provided small aggregations are avoided [7] as aggregation size is assumed to be a prominent driver of identifiability [25]. All data reported at the feeder level may only be used if it feeds a minimum of 5 individual residences, meaning its output is effectively a 5-house aggregation, and individual meter readings, obtained with consumer consent, may only be distributed and used if they are similarly aggregated with most applications enforcing 10-house aggregations [8]. For example, if an academic repository, like the Smart Energy Research Lab (SERL), were to release data from users who have installed heat pumps, they may average 10 profiles from their database of users with heat pumps and release this single representation. In some sense, the UK defines

its threshold for the risk of identifying public smart meter data as the result of its preferred method to enforce it, creating a circular definition and one that is inherently rigid.

2.2.3 Criticisms of Aggregation

Privacy Vulnerabilities

Aggregation is believed to be particularly vulnerable to identification when the number of aggregated profiles is small, when adversaries have access to external information that can infer individual contributions [26], or when malicious actors employ differential attacks, which compare aggregated data over time or against other datasets to identify patterns that could reveal individual profiles [27]. Aggregation is especially ineffective when input profiles have distinct, large features as these features often remain recognisable, though at a reduced amplitude [28] (see Figure 2.2). This particular vulnerability of aggregation is a critical concern specific to domestic consumption data as it is often characterised by distinct, large spikes in consumption at specific times of day such as mornings or early evenings [29]. Furthermore, regardless of aggregation size, the method is highly susceptible to reconstruction through elimination. For instance, if an aggregation of 1,000 households is re-released after a contributor revokes their consent, it is possible to then reconstruct the individual's profile by analysing the delta between the two releases. [30]

Utility Loss

High-fidelity data is essential for smart meter applications like research, network management, and policy-making because it ensures accurate analysis, reliable decisions, and effective energy system optimisation, ultimately supporting the development of smarter, more efficient infrastructures [31]. It is expected that these analyses will be increasingly dominated by machine learning models with AI-powered management systems seeming inevitable [32], however these applications require a level of fidelity and nuance aggregation has been found to obscure [33, 34]. Most impacted are forecasting models, which are expected to be integral to our efficient integration of renewable resources [35], as their accuracy diminishes significantly with

aggregated data due to the loss in granularity [36]. Figure 2.2 visualises an aggregate profile for aggregation sizes 1 through 10.

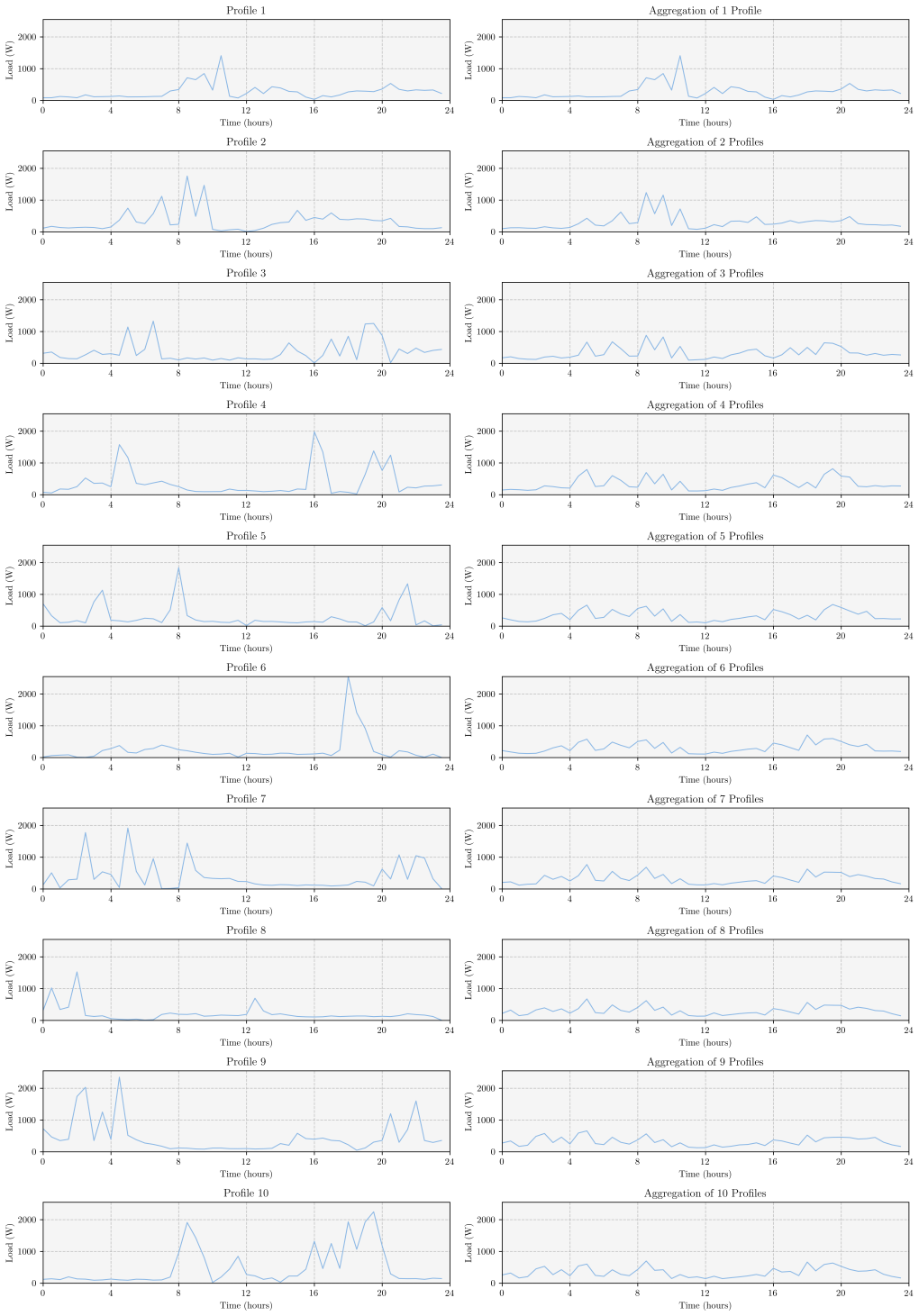

Figure 2.2: A visual depiction of the impact of aggregation at various aggregation sizes; each aggregation of profiles, n, represents the aggregation of profiles 1 through n. All profiles aggregated and plotted here are derived from a user's annual profile which this dissertation has permission to release.

From Figure 2.2, we can see that the qualitative shape of major features can be preserved in small aggregations. As aggregation size increases, the aggregation process tends to attenuate extreme values and distinctive spikes from individual profiles, smoothing out the composite profile as the sum of divergent time series tends to cancel out extreme fluctuations, resulting in a more uniform and less pronounced overall signal.

Lack of Scientific Basis

Critically, aggregation has little scientific basis; its effectiveness likely depends on dataset size, data characteristics, and cross-referencing capabilities [7], but because its design is not in service of some set metric and there exists no standard framework to assess its adherence to policy or ability to minimize identifiability, we have no general quantitative descriptions of these relationships. A palpable tension has emerged in the industry and in literature; the conversation is dominated by those who advocate for a more secure anonymisation policy and those who criticize the unnecessary rigidity and anonymity of aggregation, but the tension is perpetuated by our inability to compare the relative privacy of adaptations to aggregation nor the merits of competitive models. There are several examples in literature of studies which attempt to quantify aggregation's efficacy in terms of anonimising specific attributes, like appliance identifiability [37, 38], or the relationship between utility loss and aggregation parameters [39], but none attempt to generalise an identifiability assessment. This means literature is sparse with direct, objective comparisons of aggregation to competitive models, nor evaluations of its success at enforcing UK privacy policy.

2.3 Quantifying Identifiability

In our pursuit of a method by which aggregation and some competitive model may be assessed, with this section we survey literature on current approaches to defining *identifiability* objectively and how they may translate to a general privacy-quantifying model.

2.3.1 What is Identifiability?

The pursuit of a universal metric which completely represents the identifiability of a data release model is improper framing of the problem, as a model's security can vary significantly across different datasets due to variations in quality, noise, and data volume [40]. For example, let us imagine a published aggregation of smart meter data for a small neighborhood where one resident, a werewolf, consumes an inordinate amount of energy following every full moon. This unique pattern makes the resident highly identifiable within the aggregation, unless all residents exhibit similar behavior. If we model a malicious actor's process for identifying users as cross-referencing each profile in a data pool with some public representation, known as *membership inference*, then identifiability depends on the nature of the input profiles used to generate the representation, the remaining non-input profiles being considered, and the privacy model itself. Researchers at Columbia University frame this concept by arguing that identifiability should be defined as a function of model + data, not just the model [24].

It is additionally misguided to view identifiability as a binary concept when it is better described in "shades of gray" [24], with a continuous distribution. [24] discusses a potential Bayesian approach to this problem; by defining a *quantity of inference* (QOI) or a score which indicates the amount of an individual's information identifiable within a data representation, they suggest constructing a probability distribution of QOIs and measuring the identifiability of input profiles as the distance between their posterior and prior probabilities relative to the others in the set of potential contributors. This structure lends itself to the formulation of ranking the identifiability of a release model as a continuous distribution which depends on the model output and data being anonymised.

2.3.2 Measuring Identifiability in Sensitive Time Series

Security of major systems requires constant testing and research to keep pace with evolving threats. Therefore, generalised privacy evaluations should not aim to replicate every possible attack or cover all potential vulnerabilities, but serve as a benchmark from which more targeted, context-specific research and model development can be performed [41].

[9] recognises that identifiability assessments and data membership inference within sensitive time series is a nascent field which has not yet been properly researched, but proposes a generalised identifiability quantification through random forest Membership Inference Attacks (MIAs). MIAs are a key method for this kind of [generalised] evaluation [9] wherein random forest classifiers are trained on specific critical features of the protected data and output a score which represents the similarity between model outputs and potential contributors. Therefore, rather than rigidly testing how conducive a model is to some specific attack, MIAs quantify the intact input information within model outputs which would be available for identification to any attack vector.

From analysing sensitive time series medical patient data, [9] found that trend and seasonality were the essential features of time series data and most identifiable to MIAs. Training on these features, [9] produced quantified scores with which the relative identifiability of patient data, anonymised using various models, was ranked and their results demonstrate that training on these features enhance the efficiacy MIAs for time series. Such an MIA score would represent a suitable *quantity of inference* to support the hypothetical Bayesian structure proposed by [24], and collectively these findings provide a foundation to our pursuit of an unbiased assessor of the relative identifiability afforded by aggregation and other competitive models.

2.4 Privacy-Preserving Techniques: Case Study of the US Census Bureau

Maintaining big data privacy is more crucial than ever due to the increasing capabilities of data analytics and AI, which can easily exploit vulnerabilities in poorly protected datasets, leading to severe breaches of personal information [11, 42]. Additionally, the rise in regulatory frameworks like GDPR in the EU underscore the growing legal and ethical demands for robust data protection practices to prevent misuse and ensure public trust [43, 44]. In this section, we examine the U.S. Census, one of the world's largest modern privacy-preserving data release programs, and their strategies for ensuring security. We then compare their approach with the SMIP to highlight its inadequacies relative to international standards and identify potential areas for improvement.

The US Census Bureau (USCB) releases detailed demographic data to support policy-making, economic planning, and research with a mandate to ensure privacy. Using 2010 census data which implemented aggregation as the standard privacy-preservation technique, an internal experiment showed 46% of the population could be identified with 100% accuracy using third-party data [45]. To address this, the USCB conducted thorough tests of a variety of models and ultimately adopted differential privacy for the 2020 census, balancing privacy and accuracy by adding controlled noise according to an allocated privacy budgets across geographic levels. This approach has not only made recent censuses quantifiably more secure with defined guarantees, but more statistically reliable for data users. Table 2.1 contrasts the respective development stages of the USCB and the SMIP privacy strategies throughout their evolutions.

Development Stage	**US Census Bureau (USCB)**	**UK Smart Meter Implementation Program (SMIP)**
Initial Privacy Measures	No formal privacy protection (pre-1840)	No formal privacy protection (Pre-2012)
Early Enhancements	Established response confidentiality as a legal requirement (1840-1910)	Developed a data access and privacy framework (2010-2015)
Implementation of Privacy-Preservation Model	Introduced statistical disclosure avoidance measures like aggregation (1920-1970)	Implemented aggregation as the primary privacy-preservation measure (2018-2020)
Advanced Privacy Techniques	Adopted advanced statistical techniques such as data swapping and table suppression (1970-2010)	No significant advancements; continues default data distribution without enhanced privacy guarantees (2020-Present)
Recent Innovations	Introduced differential privacy with quantified guarantees to counter identification risks (2020-Present)	N/A
Future Directions	Continues to evolve privacy measures with a focus on innovation and adaptation	No planned evolution or enhancement of privacy policies

Table 2.1: Comparison of the evolution of privacy measures between the US Census Bureau (USCB) and the UK Smart Meter Implementation Program (SMIP).

From Table 2.1, we can see the USCB approach to privacy has been that of constant iteration and quantification which is in stark contrast to the SMIP. The USCB's formerly unquantified approach exposed the program to unforeseen vulnerabilities as their aggregation approach was not based on a guarantee nor an adherence to some specific threshold. Their defined thresholds and "privacy budgets" [45] provide concrete security guarantees and allow

for tailored privacy policies to meet specific needs. However, without an assessment framework, aggregation cannot be similarly adapted while maintaining its privacy standards.

[46] discusses how longer data sequences can lead to more effective anonymisation and reduced risk of identification through aggregation due to increased data complexity. This approach could, in theory, be used to adapt SMIP's aggregation mandate by proportionally reducing aggregation size as the length of input profiles increases, thereby enabling the targeting of specific context-dependant needs. However, we have no standardised mechanism for quantifying this phenomenon relative to current practices.

The USCB's discovery of aggregation vulnerabilities could have spurred advancements within the SMIP, but momentum remains lacking both academically and politically. Table 2.1 does not address the technical and logistical challenges of implementing advanced privacy-preserving techniques. These methods demand significant infrastructure changes, extensive testing, and validation to ensure that data utility and system functionality are not compromised, which can deter government agencies from adoption [47]. Moreover, time series privacy-preserving release models remain underdeveloped academically [9]; while general data privacy is well-researched, the specific challenges of smart meter data and the limitations of aggregation receive less attention, resulting in limited academic advocacy and sparse policy recommendations driving change [48]. Overcoming these barriers requires heightened public awareness, increased academic focus on smart meter data privacy, and the development of frameworks to test and compare existing systems against new approaches efficiently.

2.5 Alternative Privacy Preservation Techniques

From a survey of literature, four alternative privacy-preserving data release models emerge as most pertinent to the conversation of replacing aggregation: (1) differential privacy, (2) k-anonymity (and l-diversity), (3) federated learning, and (4) synthetic data. Differential Privacy (DP) adds controlled noise to data, providing strong privacy guarantees, but it is unpopular for smart meter data due to its high sensitivity to added noise, compromising utility [49]. k-

Anonymity and l-Diversity anonymise data by grouping records by similarity and verifying a minimum number of records share any individual's traits, but are vulnerable to pattern analysis [50] and can reduce data utility through generalisation in time series [51]. Federated Learning (FL) trains models locally and shares only updates, enhancing privacy but is not designed to anonymise individual records, making it better suited for system optimisations without direct human interaction [52]. An approach gaining traction in the field is to generate synthetic datasets that mimic the statistical properties of real smart meter data without containing any actual personal information using a generative AI model trained on historic data. This approach ensures high privacy and utility, making it ideal for publicly sharing data while avoiding privacy risks and maintaining the detailed insights necessary for analysis and decision-making [53, 54]. Table 2.2 summarises the relative strengths and trade-offs of each model described in literature.

Privacy Strategy	Privacy Level	Data Utility	Complexity	Scalability	Adaptability	Rank
Synthetic Data Generation	↑	↑	●	↑	↑	**1**
k-Anonymity and l-Diversity	●	●	↓	↑	●	**2**
Federated Learning	↑	↑	↑	●	↑	**3**
Differential Privacy	↑	↓	↑	●	↑	**4**

Table 2.2: Comparison of the relative performance of privacy-preservation techniques for big-data distribution as surveyed from literature.

Synthetic data emerges as the preferred choice; with this, we will review current implementations of synthetic data and how it may be implemented to replace aggregation as a superior model.

2.6 Synthetic Data

2.6.1 Framing Aggregation as a Synthetic Model

Aggregation is essentially a method for producing synthetic data; it takes real data with common features and creates artificial representations which preserve the statistical characteristics of those features, much like any other generative model. While aggregation may provide sufficient utility for some applications of anonymised time series data, most applications of published smart meter data for which individual house-level, high fidelity data is required, aggregation is misplaced. The persistence of aggregation as the enforced model for distributed smart meter data in the UK is largely a result of legislation defining it as a method to achieve an ultimately undefined goal. This dissertation argues for the scrutiny of aggregation as yet another synthetic profile generation method and for its direct, levelised comparison with more sophisticated synthetic generation models.

2.6.2 Comparing Synthetic Models with Aggregation

The most prominent big-data synthetic generation frameworks available depend on AI algorithms which train on source datasets and make predictions on future data given some constraints [55], and while it remains an underdeveloped field, time series analysis models, which are tailored to generate such data from time series input, are increasing in availability [55]. The granularity and fidelity achieved through synthetic data and the fact that it is often framed as containing "no information" [56] about real people is driving a rapidly growing popularity within the field of private data distribution [57]; however, the privacy afforded by synthetic data and how it should be treated legislatively is emerging as a point of controversy.

[58] discusses the legal issues and opportunities of using synthetic data in release frameworks. Current data protection laws often do not fully address synthetic data; for example, synthetic datasets are usually exempt from GDPR, the EU's data protection regulation, unless there are identification risks, which leads to legal uncertainty [58] as "identification risk" is an ambiguously

defined stipulation. Some advocates argue synthetic models conatin no *real* data and hence should be exempt from scrutiny. Alternatively, while the current state of research and the existing legal frameworks in the UK are insufficient to rollout synthetic data as an immediate replacement to the existing aggregation model, perhaps the problem should be approached differently. Rather than taking the current popular approach of asserting a binary, that is, synthetic data should either be exempt from privacy scrutiny or not and until it is proven one way or the other assume it is unqualified to displace aggregation [59], let us instead imagine a levelised metric upon which synthetic smart meter data can be compared with aggregation as if it were any other synthetic data model. The problem then becomes demonstrating a superior average utility and privacy preservation in synthetic data across a representative set of smart meter data relative to aggregation. In this case, policy may be more susceptible to a 'nudge' away from aggregation towards synthetic solutions without requiring a paradigm shift, as advocated for in [59]. The increased adoption of synthetic data in smart meter data applications could potentially drive changes in current data protection orthodoxy, moving beyond the binary classification of personal and non-personal data. Such a shift could lead to improved standards of protection overall, benefiting both data utility and privacy across all data distribution.

2.7 Implementing a Competitive Synthetic Model

[60] demonstrates a promising Python-based model called 'Faraday' for synthetic data generation using a Variational Auto-encoder (VAE) combined with a Gaussian Mixture Model (GMM); the model is trained on millions of real world smart meter readings with detailed metadata. The results of extensive comparisons between real and synthetic data indicate the generated data closely resembles real-world substation readings, yielding a high fidelity and utility. These results are not, however, framed in the context of aggregation.

[60] discusses aspects of smart meter data which are both critical for analysis and vulnerable to being identified, which include peak load and consumption variability. Additionally, [60] stresses that analysis of real-time smart meter data is crucial in the investigation of privacy-

preserving data release frameworks for identifying and mitigating privacy risks, which cannot be fully replicated with synthetic data or simulated environments. Therefore integrating peak load and variability features with the MIA assessment framework discussed in 2.3.2, tested on real consumer data, presents as an efficient, scientifically-derived methodology to quantify the relative privacy of two competitive smart meter data distribution models.

2.8 Identified Literary Gaps and Resulting Research Aims

This chapter highlights significant gaps in the current literature that sustain the unresolved tension between smart meter data utility and privacy in the UK. Ofgem's UK privacy policy mandates that smart meter time series data be anonymised to ensure a "remote" risk of identification, yet it lacks a clear method for quantifying this risk or establishing a quantitative threshold. Aggregation, the standard anonymisation method in the Smart Meter Implementation Program (SMIP), lacks a scientific foundation; its effectiveness across different aggregation sizes and data types is presumed but remains unquantified and untested. While the literature suggests that aggregation size significantly impacts privacy [25], there is no standardised method to quantify the risk of identification. Without addressing both utility and privacy quantification, there's no basis for adapting or validating alternative approaches.

The US Census Bureau's revision of its aggregation model improves data utility while guaranteeing privacy, potentially serving as a model for SMIP. Researchers often use Membership Inference Attacks to measure the preservation of private information, but these methods fall short as universal benchmarks due to the interplay between the data, the model, and the nature of the anonymised data.

Our literature review suggests that using a random forest classifier to perform membership inference attacks, focusing on key smart meter data features like trend, seasonality, peak load, and variability, could provide an objective measure of identification risk. Analysing these scores with Bayesian inference could offer a continuous, relative metric, or the *probability of identification*, to assess the performance of aggregation and other distribution methods.

The evolution of SMIP's privacy standards and the security of the UK's smart grid investment require an objective, comparable privacy metric. This metric would help set minimum thresholds, facilitate comparisons, and drive the development of alternative models that balance utility and privacy. AI-generated synthetic data, often cited as a promising alternative to aggregation, requires a direct, quantified comparison against the current model. This approach could establish a new methodology for evaluating privacy-preserving data release models and potentially prompt a reevaluation of aggregation's effectiveness. This dissertation seeks to answer the question:

Can synthetic data provide a better balance between consumer privacy and data utility than the current model, and can we create a quantifiable, universal metric to objectively evaluate and compare these and other data release models?

3

Methodology

The methodological design of this dissertation aims to define and test a quantitative privacy assessment framework for smart meter data release models. A generative AI model is developed and assessed as an alternative to aggregation, and multiple aggregation implementations are evaluated for their relative security.

The current aggregation model, constrained by a lack of quantitative foundation, reveals significant shortcomings in the SMIP data privacy policy, contributing to the tension between privacy protection and data utility in the literature. Existing approaches to assessing a model's privacy, such as reconstruction attacks or adversarial testing [61], simulate potential attacks to measure the effectiveness of anonymization techniques. While these methods are effective for identifying specific vulnerabilities, they are often limited by their specificity, assessing vulnerability to certain attack types without generalizing across datasets or anonymization models. Statistical disclosure control methods, like k-anonymity, l-diversity, or differential privacy, offer structured, quantifiable approaches but often at the cost of significant data utility

loss for time series data, with privacy metrics that are inherent to the model and not comparable across different model types. Despite their merits in providing concrete privacy guarantees, these methods can be too rigid and context-specific for broader applications.

To address these limitations, the assessment of aggregation and alternative models necessitates a levelized metric that is model-independent and does not rely on knowledge of specific attack vectors. The framework presented in this dissertation measures the information preserved in the output that would be available to any attack vector for reconstruction, offering a more generic, objective approach.

3.1 About the Dataset

Access to the unprecedented EDOL Smart Meter Dataset and associated metadata contributes to the novelty of this dissertation, and is integral to the quantitative analyses conducted in this chapter. The dataset contains more than 80 individual residential load profiles over 365 days in 2020, or more than 1.4 million smart meter readings. Despite the more than 70 survey questions answered by the study's participants, this study considers (1) total income, (2) house size, and (3) number of residents to be its primary characteristics as a more detailed analysis of this metadata would be beyond the set scope. This dissertation develops a Python-based processing pipeline which divides all year-long profiles into segments of some specified length, parses all survey answers, and couples the household survey answers with each associated profile as metadata.

3.2 Quantifying Identifiability

Foundational to the analyses this dissertation conducts is a method by which anonymising release models for smart meter data can be quantitatively assessed for their privacy-preservation, and compared to competitive models. Our methodology frames this metric as the *probability of identification*. This metric is agnostic to the identification attack vector, therefore it should not necessarily be perceived as the chance of a malicious actor reconstructing sensitive data,

but as a statistic which indicates the identifiability of *true input* data in an anonymised output profile relative to the total pool of profiles being considered for membership.

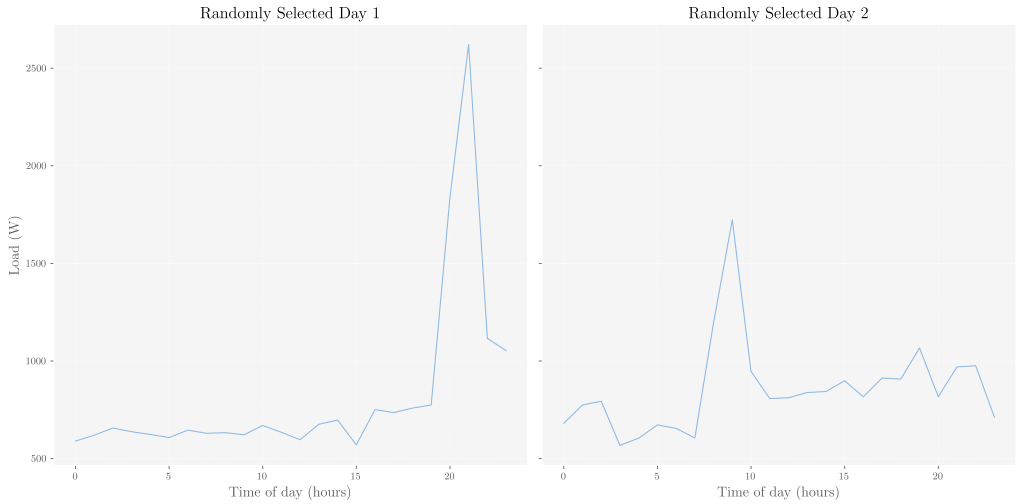

Figure 3.1: Plots of consumption profiles for two randomly selected days from the participant whose data release permission has been granted for this dissertation.

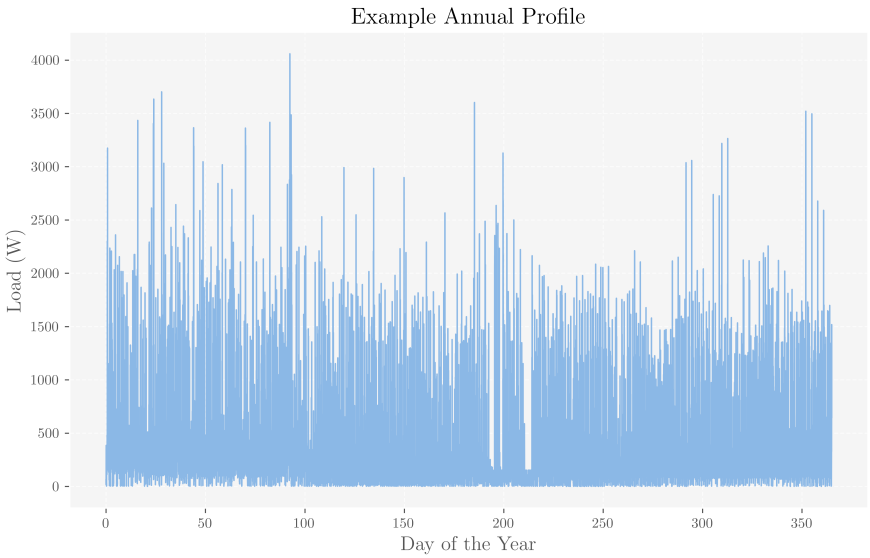

Figure 3.2: Plot of the full annual consumption profile corresponding to the participant whose data release permission has been granted for this dissertation.

Figure 3.3 graphically depicts the process for computing this probability metric, Table 3.1 describes the literature basis to each step in the process, and in this section we elaborate on the mechanics of each of these steps.

Step	High-Level Explanation	Basis
1	From the total pool of EDOL profiles, a subset is selected and used as input to some data release model.	While the selected subset interacts with the model, the remaining pool is relevant to this process as the identifiability of the input profiles is only substantial if framed relative to non-input profiles.
2	The anonymisation model outputs a profile representation of the input.	This methodology is general to any anonymising model which takes real data as input and outputs a representative profile intended for anonymous distribution, such as aggregation or a synthetic model.
3	Using a set of literature-based critical classification features, we use a random forest classification to score all profiles in the EDOL dataset in terms of their information's identifiability within the model output.	[9] presents random forest classification as an elegant approach to an unbiased assessment of the relative identifiability of information within some aggregate representation. The classification computes critical profile features to guide the classification; [9] shows that trend and seasonality are most pertinent to time series data, and [60] demonstrate that incorporating peak load and variability information is critical for evaluating smart meter data representations specifically.
4	Every EDOL profile now has an associated *identifiability* score, which is independent of the other profiles in the pool.	N/A
5	A probability density function is constructed using the identifiability scores.	The output scores are independent of the other assigned scores in the set. The PDF contextualizes the scores relative to the other profiles being considered for membership, as [24] posited that an assessment of identifiability for some model should be a function of both the dataset being anonymised and the model.
6	Using a Bayesian Inference structure, each profile is assigned a *probability of identification* based on the statistical significance of their scores relative to the set.	[24] recommends the Bayesian structure, assuming some quantity of inference, as the correct formulation as it frames identifiability as a continuous, or in "shades of grey" rather than as a binary.
7	The probabilities associated with the *true input* profiles are separated from the set.	N/A
8	Based on the probabilities associated with the *true input* profiles, a final *identifiability* statistic is assigned to the model/dataset combination.	The final assessment is a function of both the model and the studied dataset.

Table 3.1: Explanation and justification for the identifiability test steps depicted in Figure 3.3.

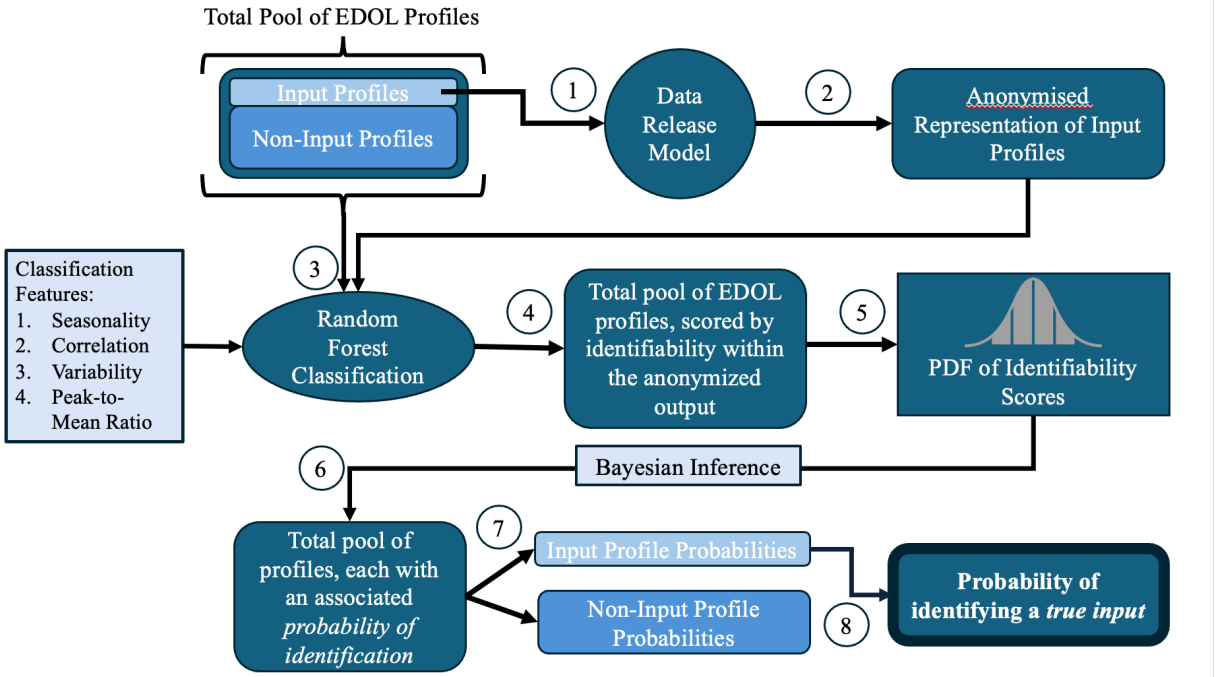

Figure 3.3: Step-by-step process for quantifying the probability of identification. See Table 3.1 for detailed explanation of steps.

3.2.1 Random Forest Classification

We train our random forest classification on the complete EDOL-provided database of profiles with all associated metadata and creates an ensemble of decision trees, where each tree is a simple model which classifies data based on profile features. Each tree is trained on a different random subset of the data, and at each split, it randomly selects which features to consider. The final prediction is made by aggregating the outcomes of all the trees to produce an estimate of the information in common between the anonymised output and each profile, indicating how identifiable a profile is within the anonymised representation. We implement this common form of trained classification in Python using the `RandomForestClassifier` class from the `sklearn.ensemble` module as follows:

1. The classification model in this study utilizes critical features from the time series data, specifically focusing on seasonality, trend, peak-to-mean ratio, and variability. The seasonality component is captured using a Fourier Transform, which decomposes the time series

into its constituent frequencies, thereby highlighting periodic patterns. Mathematically, this is represented as:

$$X(f) = \sum_{t=0}^{N-1} x_t e^{-i2\pi ft/N} \tag{3.1}$$

where $X(f)$ denotes the Fourier coefficients at frequency f, N is the total number of discrete time points, and x_t is the value of the time series at time t. This transform enables the identification of dominant frequencies within the time series, which are indicative of its seasonal patterns. The extracted Fourier coefficients serve as training parameters for the model, capturing the essential periodic features of the time series data.

The trend is represented by the observed Pearson correlation between the model output and input data, computed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

Additionally, the peak-to-mean ratio (PMR) and standard deviation are computed to encapsulate the peak load behaviour and variability, respectively. PMR is defined as

$$\text{PMR} = \frac{\max(x_t)}{\frac{1}{N}\sum_{t=0}^{N-1} x_t} \tag{3.3}$$

and variability, or the standard deviation as

$$\text{Variability} = \sqrt{\frac{1}{N}\sum_{t=0}^{N-1}(x_t - \mu)^2} \tag{3.4}$$

where μ is the mean of the series.

2. These features form the input vectors for training the random forest classifier, $\mathcal{F}$, comprising multiple decision trees $\{T_1, T_2, \ldots, T_m\}$. The classifier predicts the likelihood, or a score out of 100, of a profile's information's inclusion in the model output by averaging the predictions of these trees,

$$\mathcal{F}(\mathbf{f}_i) = \frac{1}{m} \sum_{j=1}^{m} T_j(\mathbf{f}_i) \tag{3.5}$$

3. We assess the identifiability of synthetic profiles using the output of the RandomForestClassification Python class which are normalised scores representing the likelihood that each individual tested profile was a true input

$$\text{Score}(\mathbf{f}_{\text{synthetic}}) = \mathcal{F}(\mathbf{f}_{\text{synthetic}}) \tag{3.6}$$

3.2.2 Probability of Identification

Given our methodology for scoring a profile's identifiability within a model output, this section places those scores in a statistical context by comparing them to the rest of the dataset. We construct a probability density function (PDF) using the identifiability scores of all profiles in the dataset. This PDF helps us determine how common or rare a particular score is, which is used to assess the likelihood of a profile being a true input. Using Bayes' Theorem (Eq. 3.7), we calculate the posterior probability that any individual profile, p_i, within a pool S, contributed to the models output or would be identified as a contributor.

$$P(\mathcal{I} \mid \text{Score}(p_i)) = \frac{P(\text{Score}(p_i) \mid \mathcal{I}) \cdot P(\mathcal{I})}{P(\text{Score}(p_i))} \tag{3.7}$$

where:

- $P(\mathcal{I} \mid \text{Score}(p_i))$ is the **posterior probability**: the probability that profile p_i is a true input given its identifiability score.
- $P(\text{Score}(p_i) \mid \mathcal{I})$ is the **likelihood**: the probability of obtaining a specific identifiability score if p_i is indeed a true input.

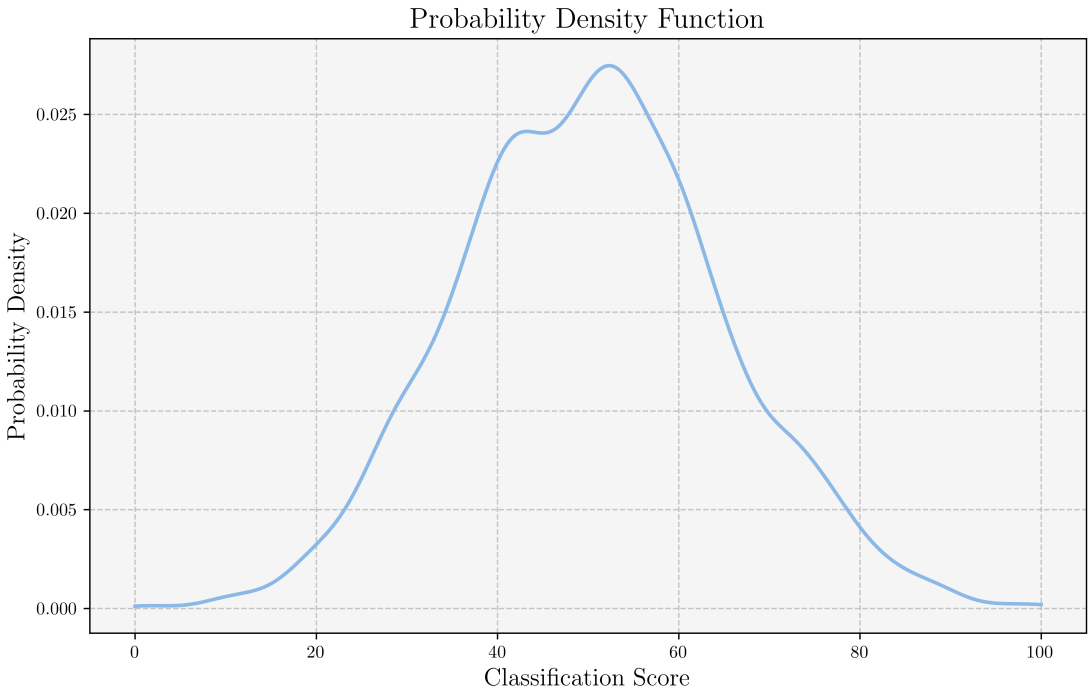

Figure 3.4: Sample probability density function (PDF) computed using scores from subsection 3.2.1.

- $P(\mathcal{I})$ is the **prior probability**: the initial probability of any profile being a true input, based on the proportion of true inputs in the dataset.
- $P(\text{Score}(p_i))$ is the **marginal probability**: the overall probability of observing the identifiability score $\text{Score}(p_i)$ across the entire dataset.

The following outlines our process for integrating our computed quantities of inference with Bayes' theorem for an ultimate *probability of identification.*

Computing the Probability of Identification

1. **Prior Probability ($P(\mathcal{I})$)**:

   The prior probability represents our initial belief about the likelihood of any profile being an input to the model, assuming no additional information. It is calculated as the ratio

of the size K of the input profile set $\mathcal{I}$ to the size of the pool $|S|$:

$$P(\mathcal{I}) = \frac{K}{|S|} \tag{3.8}$$

2. **Likelihood ($P(\textbf{Score}(p_i) \mid \mathcal{I})$)**:

   The likelihood represents the probability of observing a specific identifiability score, $\text{Score}(p_i)$, for a profile p_i, given that it is a true input. This is derived from the PDF (Figure 3.4), which shows the distribution of identifiability scores:

$$P(\text{Score}(p_i) \mid \mathcal{I}) = \text{PDF}(\text{Score}(p_i)) \tag{3.9}$$

3. **Marginal Probability ($P(\textbf{Score}(p_i))$)**:

   To ensure the posterior probabilities sum to 1, we normalize them by the total probability of the scores, or the marginal probability:

$$P(\text{Score}(p_i)) = \sum_j P(\text{Score}(p_j) \mid \mathcal{I}) \cdot P(\mathcal{I}) \tag{3.10}$$

4. **Posterior Probability ($P(\mathcal{I} \mid \textbf{Score}(p_i))$)**:

   The posterior probability updates the prior probability with the likelihood to reflect the new evidence provided by the scores:

$$P(\mathcal{I} \mid \text{Score}(p_i)) = \frac{P(\text{Score}(p_i) \mid \mathcal{I}) \cdot P(\mathcal{I})}{P(\text{Score}(p_i))} \tag{3.11}$$

$$\implies P(\mathcal{I} \mid \text{Score}(p_i)) = \frac{P(\text{Score}(p_i) \mid \mathcal{I}) \cdot P(\mathcal{I})}{\sum_j P(\text{Score}(p_j) \mid \mathcal{I}) \cdot P(\mathcal{I})} \tag{3.12}$$

We interpret $P(\mathcal{I} \mid \text{Score}(p_i))$ as the probability that profile p_i contributed to some data representation, which we refer to as the *probability of identification*.

With both input and non-input profiles assigned some *probability of identification*, we can isolate the input set and their probabilities for an aggregate probability of identifying an

input, specifically. With this, we have a general metric with which competitive models can be compared for their privacy.

3.3 Quantifying the Efficacy of Aggregation

In this section we use our *probability of identification* metric to better understand the efficacy of the incumbent aggregation model. This section describes our methodological approach to assess the impact of the aggregation size and the length of profiles being aggregated; we also inject artificial outliers in the input set to quantitatively assess aggregation's ability to anonymise outlier, 'werewolf' profiles.

3.3.1 Average Identifiabiltiy

The average performance of aggregation across the entire EDOL dataset is assessed by repeating our *probability of identification* methodology across multiple trials, each with a unique input subset. For each trial, after assigning each profile's respective probabilities, we compute two evaluative metrics: precision at K, and mean rank.

Precision at K ($P@K$) measures the proportion of true input profiles which fall within the top K ranked profiles across all trials, where K is the size of the input set. Let $\mathcal{I}$ be the set of true input profiles, and let $\mathcal{R}_K$ be the set of the top K profiles ranked by the probabilities obtained from Bayesian inference. We define precision at K as:

$$P@K = \frac{|\mathcal{I} \cap \mathcal{R}_K|}{K} \tag{3.13}$$

From here, the model is holistically assessed based on the average $P@K$ across all trials and the mean input rank, or the average of the input profiles ranks across all trials.

3.3.2 Aggregation Size Investigation

As discussed in Section 2.2.2, all enforced minimum aggregation sizes are not based on quantitative evidence; here, we lay quantitative foundation to these policies by testing the impact on identifiabiltiy of varying aggregation size. Using aggregation sizes $K = \{2, 3, \ldots, 19\}$, we define $\binom{|S|}{K}$ unique trials, where $|S|$ represents the total number of 1-day load profiles parsed from the EDOL dataset and K is the aggregation size, each trial consists of a unique aggregation of K inputs, and the remaining non-input profiles are included in the consideration pool. Every value of K yields $\binom{|S|}{K}$ $P@K$ values, which are averaged and represent the model's relative privacy performance at that aggregation level.

3.3.3 Profile Duration Investigation

With this test, we reprocess the EDOL dataset such that, for $l\_p = \{1$ day, 2 days, 7 days, 30 days, 90 days, 365 days$\}$, each year-long EDOL profile is parsed into 365 days$//l_p$ individual sub-profiles of length l_p. From here, the same methodology described in subsection 3.3.2 is repeated for all values of l_p and aggregation sizes $K = \{2, ..., 10\}$ so as to investigate the impact of profile size on identifiability with a variable aggregation size.

3.3.4 Outlier Profile Identifiability Investigation

With this test we assess the aggregation models ability to anonymise profiles with distinct features; we repeat the methodology described in subsection 3.3.2, however for every trial, one input profile is randomly selected to have artificial outlier features added to it. The mean rank of the outlier profile is then compared to the mean rank of all other input profiles as a means of assessing whether the outlier profile is more or less identifiable, on average.

While it is infeasible to replicate all possible variability in real-world smart meter readings, our objective here is to assess whether outlier features result in increased identifiability, and what the impact of the number of outlier features present in a single profile is on that profiles

identifiability. We synthesise the artificial outlier features by identifying the maximum half-hourly reading across the EDOL dataset, r_{max}, randomly selecting an hour-long period within the selected input profile, and setting this periods reading to be some random value, rounded to the nearest MWh, within the range $(r_{max} \times 2, r_{max} \times 3)$. This process is repeated with a variable number of outlier peaks, ranging from 1 to 8.

3.4 Synthetic Load Profile Generation

A primary output of this dissertation is a comparison between the aggregation model and an implementation of a synthetic load profile generative artificial intelligence. We train a Python-based generative AI on the EDOL profiles and their primary metadata features. The model design is based on [60] which develops *Faraday*, a generative AI capable of generating synthetic load profile data based on conditional metadata.

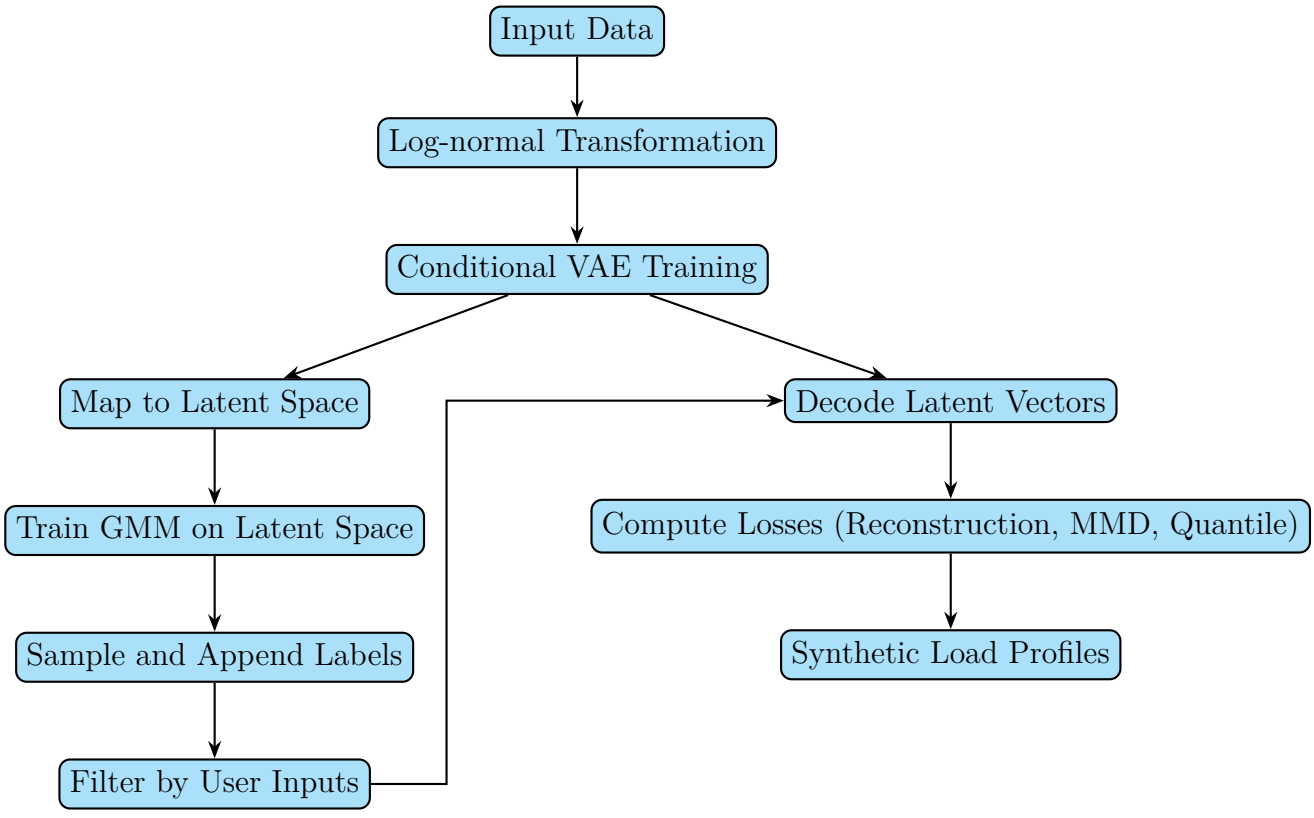

Figure 3.5: Flowchart depicting the information flow for the synthetic load profile generation process.

At a high level, our recreated version of Faraday begins by training a conditional Variational Autoencoder (VAE) to encode the input dataset into a latent space and training a Gaussian Mixture Model (GMM) on this latent space to capture its distribution. During inference, latent vectors are randomly sampled from the GMM and decoded using the trained VAE decoder. To enable conditional sampling, labels are included with the latent codes during GMM training (see Figure 3.5). While [60] has not released code nor detailed its implementation, this dissertation's recreation was shared with the author of [60] who confirmed the approach and parameter assumptions are correct with small implementation adjustments recommended. The following details our technical implementation with training, encoding, and decoding being conducted using the popular *PyTorch* library:

i. Conditional Variational Autoencoder (VAE) Training

The Conditional Variational Autoencoder (VAE) consists of an encoder, a latent space, and a decoder:

Encoder and Latent Space

The encoder maps the input data x and conditional metadata y to the latent space z. The encoder is parameterized by ϕ, which outputs the parameters of the latent distribution, like the mean $\mu(x, y)$ and variance $\sigma(x, y)$ of a Gaussian distribution:

$$z \sim q_\phi(z|x, y) \tag{3.14}$$

Where $q_\phi(z|x, y)$ represents the distribution of z given x and y.

To enable backpropagation through the stochastic sampling process, the reparameterization trick is employed. The reparameterization trick expresses the sampling operation z as a deterministic function of $\mu(x, y)$, $\sigma(x, y)$, and an auxiliary noise variable ϵ, which is drawn

from a standard normal distribution:

$$z = \mu(x, y) + \sigma(x, y) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{3.15}$$

This trick allows gradients to be passed through the sampling process, facilitating the optimization of the VAE.

Decoder

The decoder is parameterized by θ and reconstructs the input data x from the latent representation z and conditional metadata y:

$$\hat{x} = p_\theta(x|z, y) \tag{3.16}$$

Where $\hat{x}$ is the reconstructed data.

ii. Gaussian Mixture Model (GMM) Training

In traditional VAEs, the latent space is modeled using a unimodal Gaussian distribution. For the Faraday model, a Gaussian Mixture Model (GMM) is used to better capture the complex distribution of the latent space. A GMM represents a data distribution as a combination of multiple Gaussian distributions, each with its own mean and variance, which allows the model to capture complex, multimodal distributions that enable the model to better approximate and separate diverse patterns within time series data. The GMM is trained on latent representations of data, where the distribution of the latent space z is modeled as:

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z|\mu_k, \Sigma_k) \tag{3.17}$$

Where:

- K is the number of mixture components,

- π_k are the mixture weights,
- μ_k are the means,
- Σ_k are the covariance matrices of the Gaussian components.

iii. Inference and Synthetic Data Generation

During inference, new latent vectors are sampled from the GMM, $\tilde{z} \sim p(z)$, such that the generated samples match the distribution of the training data. The decoder network then transforms the latent vectors back into the data space:

$$\tilde{x} = \text{Decoder}_\theta(\tilde{z}, y) \tag{3.18}$$

Where $\tilde{x}$ represents the generated synthetic data.

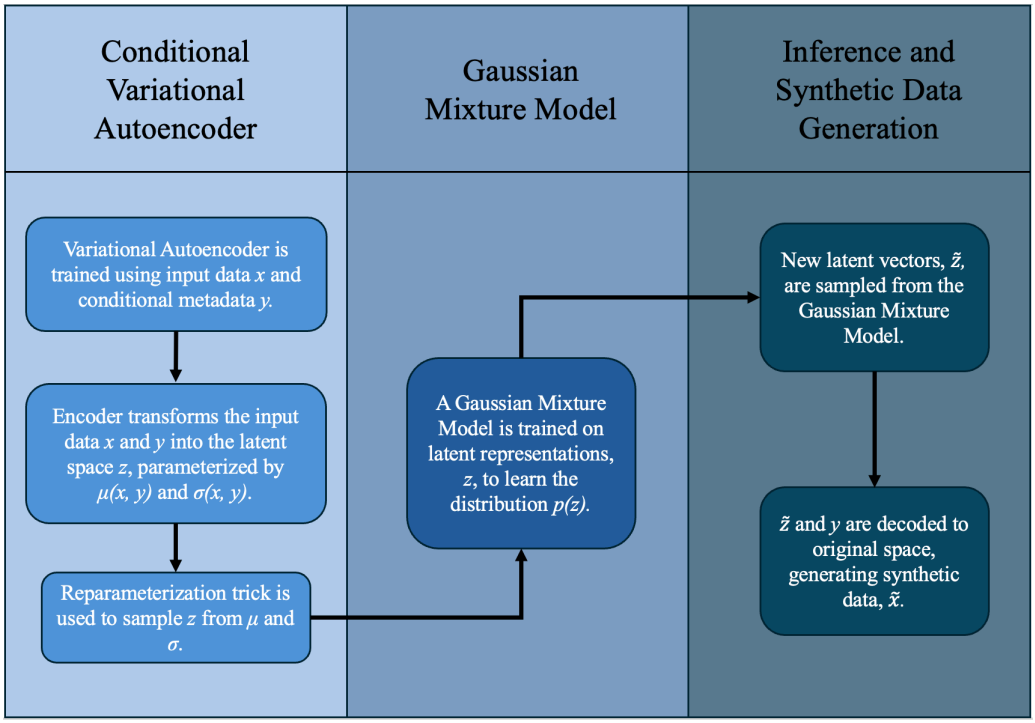

Figure 3.6: Graphic of our synthetic generation methodology, with the mathematical process summarized. The output of this system are unique, synthetic profiles which are representative of a profile whose metadata matches the specified model input.

3.5 Comparing Aggregate and Synthetic Data

In our comparison of aggregation and synthetic data, we must benchmark model output from both a quality and privacy standpoint, as a model with enhanced privacy over aggregation is only valuable if quality is maintained or improved, and vice-versa.

3.5.1 Quality Comparison

From [60], we know the quality of output for a data release model depends on fidelity and utility. An investigation of the real-world utility of generated data is beyond the scope of this dissertation as the primary shortcoming of aggregation which we aim to address is the loss of critical features [62], therefore here we assess quality strictly from a fidelity standpoint. Our evaluation of the fidelity of synthetic data relative to aggregated data measures differences in quantile loss and the statistical similarity between input and output.

Quantile Loss

Quantile loss measures how well the synthetic data captures the distribution of the real data at various quantiles. We calculate the quantile values for both real and synthetic datasets and compute the absolute differences between them using the 5th, 50th, and 95th percentiles.

Statistical Similarity

The peak-to-mean ratio (PMR) and variance (σ^2) for the datasets are used to capture statistical similarity, where variance is computed as follows:

$$\sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_X)^2 \tag{3.19}$$

Comparison Implementation

Fidelity metrics are computed for the models' output and each of their inputs; the difference between the average of the input metrics and the output metrics is used to benchmark data quality, and these differences were averaged across all trials which consisted of every unique input combination as in subsection 3.3. The trial sets and synthetic generation method are carried over from 3.5.2 where input profiles have identical metadata and the synthesized data was generated using this metadata as the prompt for each trial.

3.5.2 Anonymity Comparison

$P@K$ and outlier mean rank are used to compare privacy preservation across the two models, as demonstrated in subsections 3.3.2 and 3.3.4, respectively. 10-profile aggregations are used in this comparison as this is a common minimum for published databases in the UK. The dataset is reorganised by survey answers such that for each trial, 10 random profiles whose participants share the same responses to the three primary questions (household income, home size, and number of residents) are added to the pool as the input set in addition to all others who have different answers, with the remaining profiles whose metadata is the same as the input set being discarded.

The synthetic model is designed such that when profiles are generated based on specific conditions, the model primarily relies on training data with identical metadata to construct the representation. However, it also incorporates features from training data with partial metadata matches to introduce diversity. In our testing, we consider only those training profiles with metadata that exactly matches the AI prompt as true inputs, meaning they are the only profiles at risk of being identified in the model output.

Precision at K Comparison

Using the newly organised trial sets, we perform a $P@K$ test as in 3.3.2 whereby all of the trials' $P@K$ outputs are averaged to a single metric. This test is conducted once using the

generative AI as their anonymising model and once using aggregation.

Outlier Identifiability Comparison

Outlier identifiability is assessed using the same trial sets and the same synthetic generation procedure as in the previous test; the outlier injections and identifiabiltiy assessments are conducted following the same method as described in 3.3.4 for both models.

3.6 Conclusion

This chapter defines two novel methodologies used in this dissertation's analyses: a method for quantifying the average anonymising efficacy of a smart meter data release model given some data, and a method for using this quantification framework to compare two competitive release models. We use the framework to quantify the relative anonymising performance of aggregation with a variable aggregation size, variable input profile length, and an injected outlier, all to lay scientific foundation to the current regulation. We develop a functional generative AI for producing synthetic load profiles, and compare the fidelity and anonymity of its output with aggregation to test whether the incumbent privacy model is truly competitive.

4

Results

In this section we present two sets of results: the results of applying our privacy assessment framework to aggregation while varying model parameters, and the results of our comparison of our generative AI with a 10-house aggregation model in terms of fidelity and privacy. This chapter's results not only demonstrate the superiority of the generative AI developed for this dissertation, but formalise previously assumed behaviors of aggregation, such as the specific impact of increasing aggregation size.

4.1 Quantifying Aggregation Privacy

Using our average *probability of identification* test, we measure the impact on identifiability of tuning aggregation parameters so as to lay quantitative foundation to current legislation and to enable adaptations to current protocols without sacrificing privacy standards.

4.1.1 Impact of Aggregation Size

First, we investigate aggregation size, or the number of profiles being aggregated to produce an output, as it is assumed to have the greatest impact on the quality of an aggregate output [25].

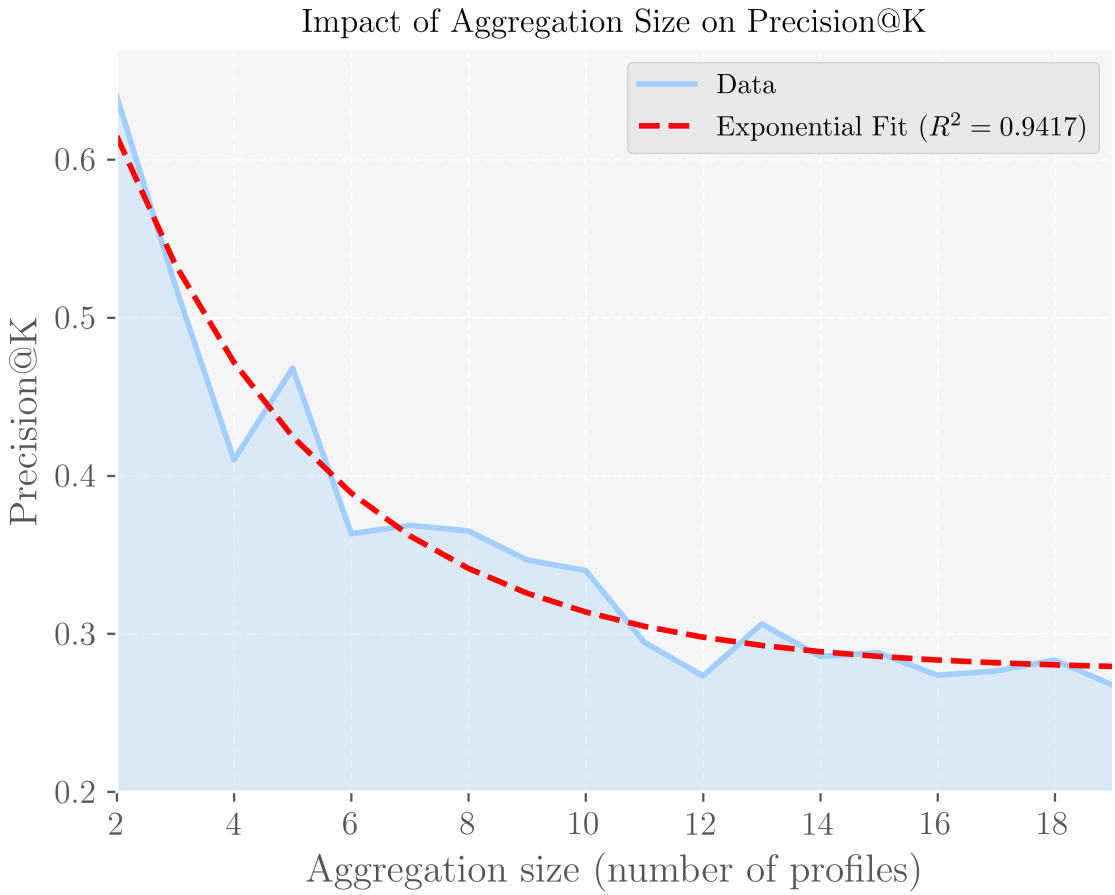

Figure 4.1: The average $P@K$ metric for all trials at each aggregation size is plotted, with a fit negative exponential curve superimposed. $P@K$ represents the number of true input profiles whose probabilities of identification are ranked in the top K across all profiles which were evaluated, on average across all trials where K is the aggregation size. Here, a low $P@K$ indicates a low risk of identification.

From Figure 4.1 we can qualitatively see an exponential relationship between aggregation size and the identifiability of aggregation input; modelled as an exponential, we get the following fit negative relationship between $P@K$ and aggregation size

$$y = 0.5865 \cdot \exp(-0.2747 \cdot x) + 0.2761 \tag{4.1}$$

The fit, which produces a strong aggreement of $R^2 = 0.9417$, indicates privacy and aggregation size are positively correlated as a low $P@K$ indicates low identification risk. The negative exponential shape indicates a negative correlation with a diminishing return, meaning small changes in the size of aggregation at the low end have a more substantial impact on identifiability than those at the high end.

4.1.2 Impact of Input Profile Length

Here, we vary the length of the profiles used as input to the aggregation and test the impact of this parameter across multiple aggregation sizes.

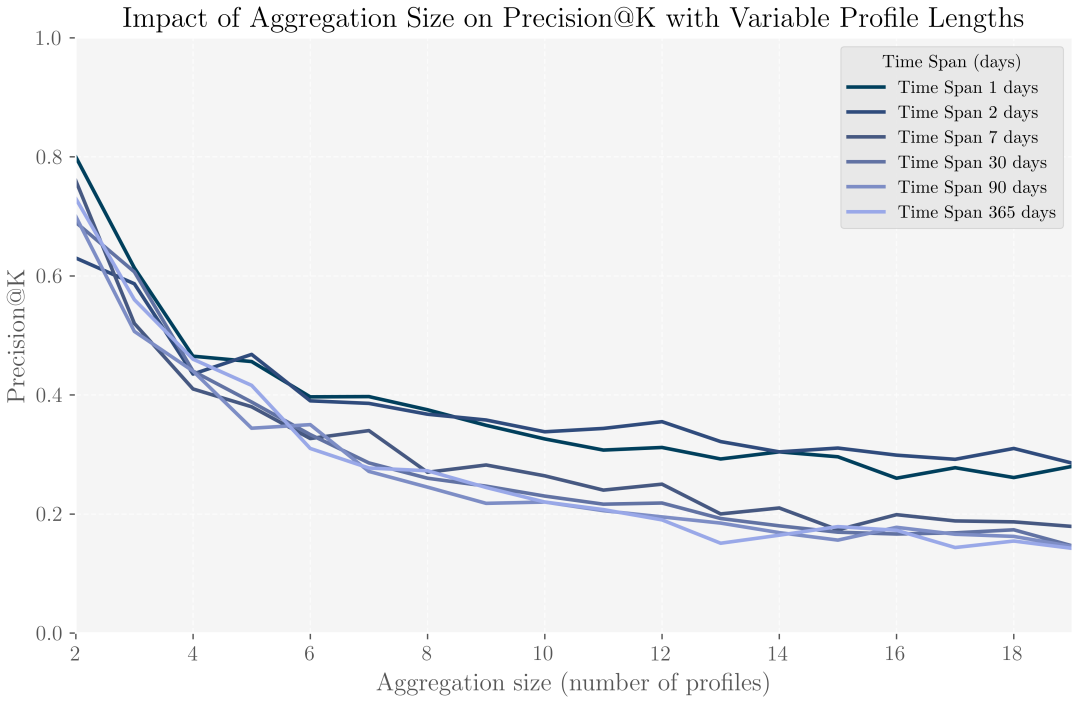

Figure 4.2: The average $P@K$ metric for all trials at each aggregation size is plotted, with each curve corresponding to trial sets consisting of profiles of some specified length. Here, a low $P@K$ indicates a low risk of identification.

From Figure 4.2 we can qualitatively see that as profile length increases, identifiability decreases; this trend is observed most prominently at high aggregation sizes. More precisely, across all aggregation sizes we get an average Pearson correlation coefficient between $P@K$ and

profile length of -0.68, indicating a strong negative correlation.

4.1.3 Identifiability of Outlier Profiles Within Aggregations

A common criticism of aggregation is its inconsistency at preserving anonymity of data which is an outlier within the dataset being aggregated [63, 28]. Here, we test aggregation's ability to anonymise profiles with strong outlier features and whether the number of these outlier features impacts their identifiability.

For each trial, we compute the difference between the identifiabiltiy rank of the profile selected for outlier features to be added and the average non-outlier input profile, and we average this metric across all trials for every number of added outlier features.

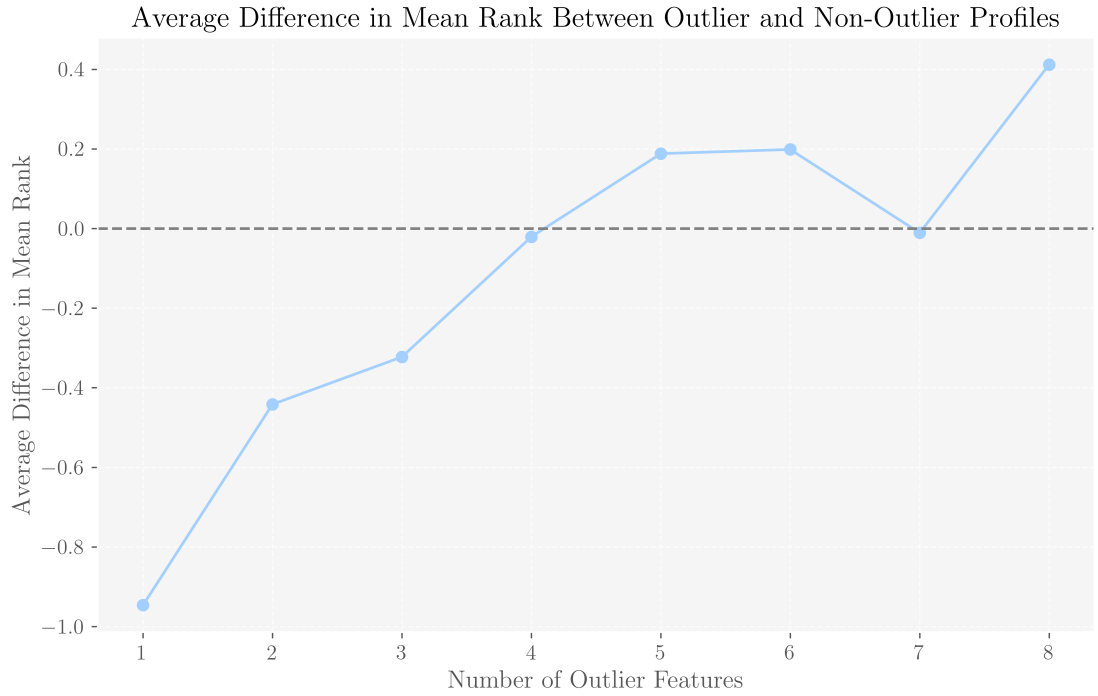

Figure 4.3: The difference between the mean rank of the outlier profile and the mean rank of all other non-outlier profiles is plotted against the number of outlier features artificially added to the selected input profile. The dashed horizontal line at 0 represents the threshold above which the outlier profile is less identifiable than the average input profile.

Figure 4.3 depicts these results; we see that the presence of few outlier features tends to result in that profile being more identifiable within an aggregation, but as the number of artificial

features increases, the relative identifiability of the outlier profile decreases. Beyond 4 features, the outlier profile becomes less identifiable than the average input profile. This counterintuitive result might be due to the increased number of distinct features causing the outlier profile to blend into the aggregation more effectively, thereby diluting its overall distinctiveness and reducing its identifiability relative to the non-outlier profiles.

For each specified number of outlier features, we repeat all trials using multiple aggregation sizes. Figure 4.4 plots the overall identifiability of the model across a variable aggregation size when artificial outliers are or are not included, using $P@K$ as our benchmark, and Table 4.1 summarises these results as the difference in average $P@K$ for each experiment.

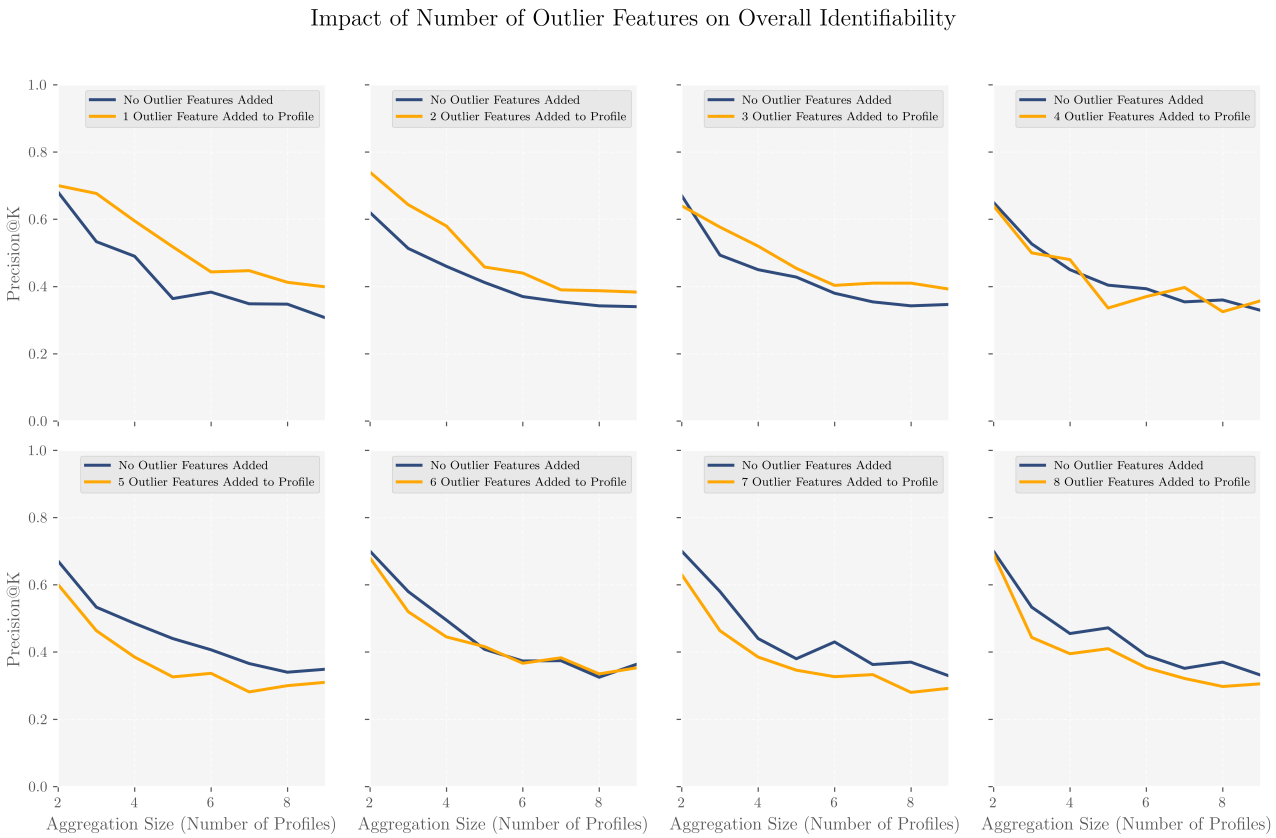

Figure 4.4: Each figure in the stack represents our results corresponding to a specified number of added outlier features. The orange curves correspond to our identifiablity assessment of trials with injected outliers, and the trials containing no injected outlier are represented by the blue curves. Here, a low $P@K$ value means a relatively low risk of identification.

	Number of Outlier Features							
	1	2	3	4	5	6	7	8
Difference in Average $P@K$	-0.09	-0.07	-0.05	0.01	0.07	0.02	0.07	0.05

Table 4.1: For each experiment represented by an individual plot in Figure 4.4, we summarise their results by computing the average $P@K$ across all trials with no injected outliers and those with the injected outlier and show their differences here. A negative difference indicates the trials with an injected outlier are more identifiable, on average.

From Figure 4.4 and Table 4.1, we see that when a profile with few outlier features is added to an input set, the model's output is, on average, more identifiable than that of the same model applied to an identical trial set without an injected outlier. At and beyond 4 artificial outlier features, the trials with the injected outlier exhibit less identifiability, on average, than those without the outlier. This reduced identifiability may result from the outlier features obscuring other input profiles' subtle characteristics, leading to a more homogenised and thus less distinguishable output.

4.2 Comparative Analysis: Aggregation and Synthetically Generated Data

4.2.1 Fidelity Retention

In this subsection we analyse the fidelity retention from input to output for both aggregate data and the synthetic profile generation model. The metrics by which fidelity is assessed are peak-to-mean ratio, standard deviation, and quantile losses.

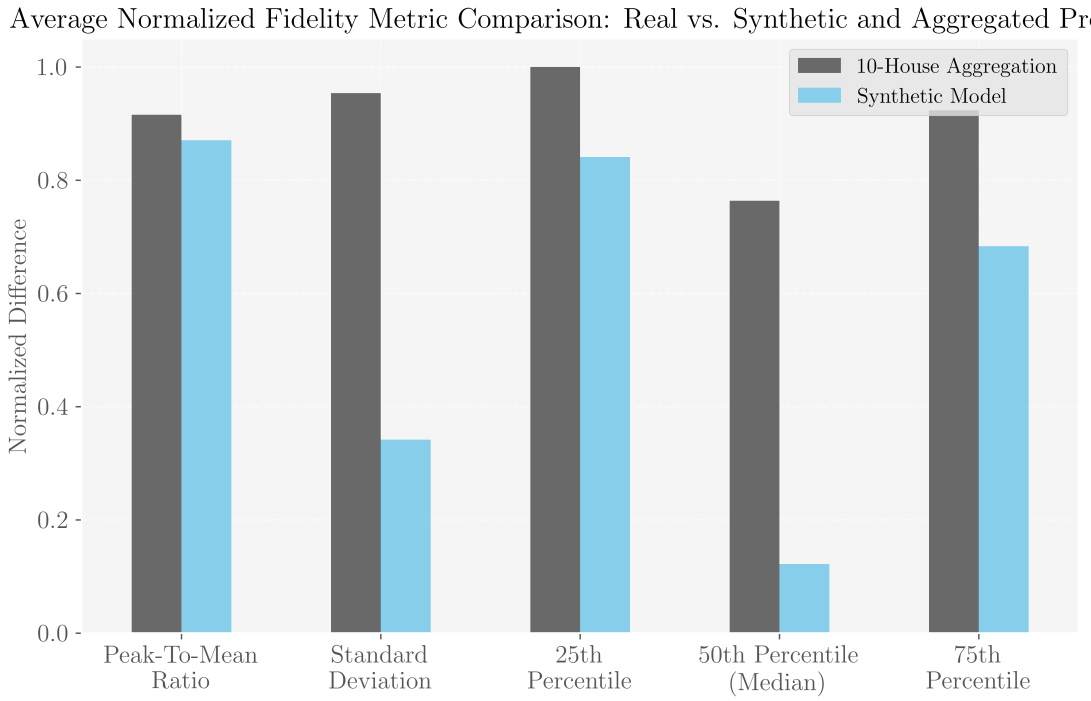

Figure 4.5: Comparison of the relative fidelity between aggregate data and our synthetic model. The *normalised difference* represents the difference between the fidelity metric computed for the model output and the average computed across the input profiles, normalised.

From Figure 4.5, we see that across all metrics the synthetic model outperforms the 10-house aggregation as the lower the normalised difference, the more fidelity retention from input to output with the comparison of their medians exhibiting the biggest discrepancy across the two models and the comparison of peak-to-mean ratio showing the least discrepancy. Figure 4.6 plots three randomly selected training profiles corresponding to a particular metadata condition, and synthetically generated profiles using the metadata condition as the prompt.

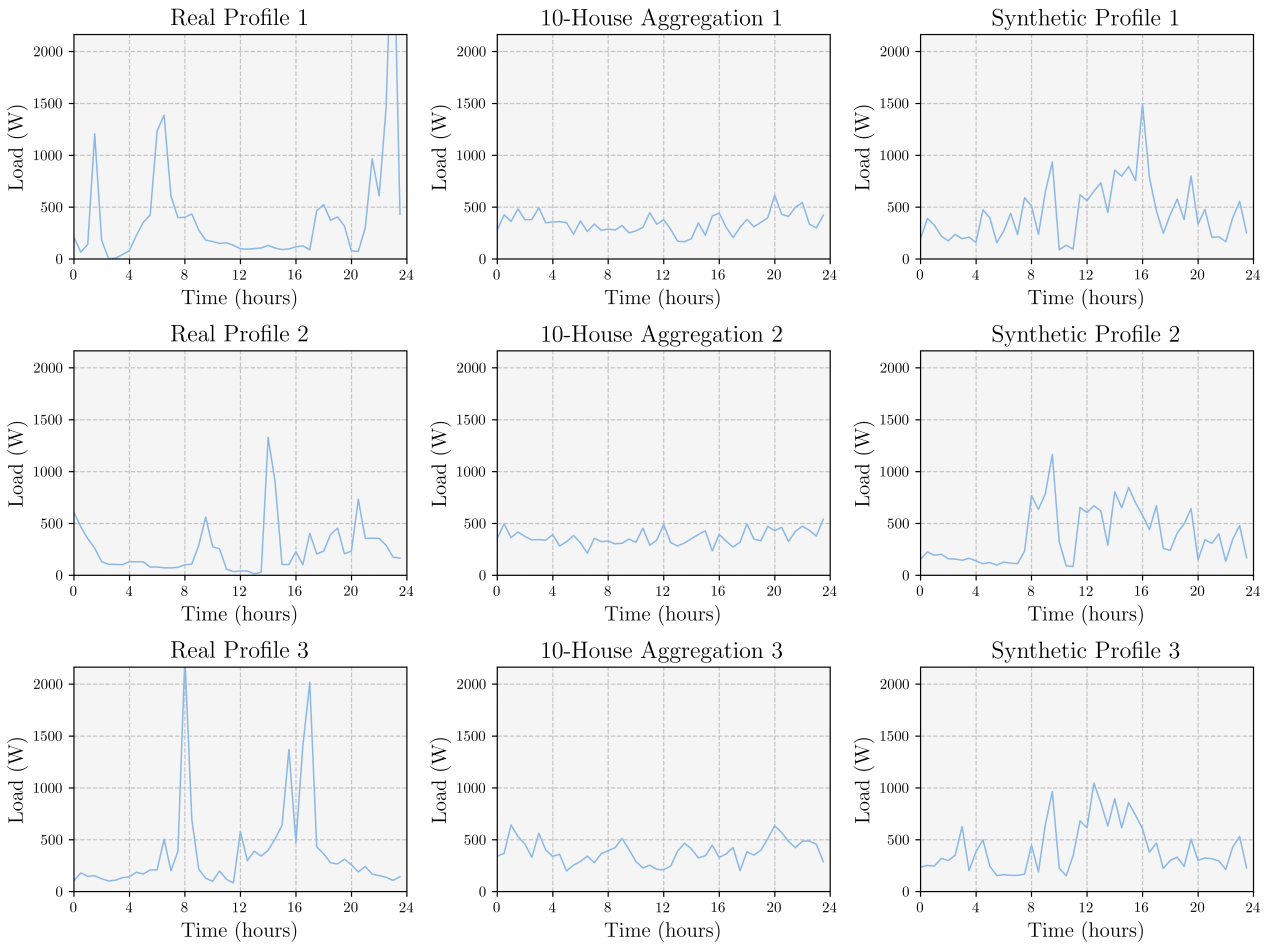

Figure 4.6: Visualisation of synthetic profile results; the leftmost column of randomly selected real training profiles, the center column depicts 3 examples of 10-house aggregations, and the rightmost column depicts 3 examples of synthetically generated profiles. All real profiles, aggregations, and synthetic profiles are derived from the set of profiles which this dissertation has permission to release.

Qualitatively, we can see clear advantages to the synthetic approach in terms of information preservation. Domestic load profiles are often characterised by distinct, large usage spikes throughout the day whose information is critical for analysts, and while we see an overall reduction in amplitude and variability in the aggregation, the synthetic data preserves high-amplitude spikes in consumption; the synthetic model therefore produces qualitatively more similar consumption patterns and profiles which are more statistically similar to their training data.

4.2.2 Privacy Comparison

Precision at K Comparison

Table 4.2 presents the results of our comparison of the average identifiability, or an average of the $P@K$ computed across all trials, between a 10-house aggregation and our synthetic model.

	10-House Aggregation	**Synthetic Model**
Average $P@K$	0.36	0.21

Table 4.2: The average $P@K$ represents the average anonymisation performance across all trials using the stated model. A low $P@K$ value indicates better anonymity in the output.

We see that on average, our synthetic model anonymises input data considerably better than the 10-house aggregation.

Outlier Anonymisation Comparison

Table 4.3 summarises the results of our comparison of the models' respective efficacy at preserving the anonymity of input profiles containing distinct outlier features, using the difference in mean rank of the injected outlier as our indication.

	Number of Outlier Features							
	1	2	3	4	5	6	7	8
Average Difference in Mean Rank	2.72	1.01	1.88	-0.38	0.67	0.81	-0.03	0.37

Table 4.3: The average difference in mean reank represents the difference between the mean rank of the injected outlier profile across all trials using the generative AI and aggregation. A positive difference indicates the synthetic model exhibited a superior perforance.

In Table 4.3 a positive difference indicates the synthetic model acheived higher anonymity

and a negative score indicates the aggregation model did; we see the synthetic model outperforms the 10-house aggregation in almost every instance, but most prominently when fewer outlier features were added. Notably, at 4 and 7 outlier features the aggregation performs better than the synthetic model by a relatively small margin.

4.3 Conclusion

This chapter quantifies the impact of specific aggregation parameters on the anonymity of aggregated data, revealing that privacy preservation comparable to or better than the 10-house standard can be achieved with smaller aggregation sizes, especially when input profile length is varied. We demonstrate that while aggregation often fails to anonymise distinct outlier profiles, beyond a certain threshold of injected variability, these outliers become less identifiable than average profiles and can even reduce the overall identifiability of the entire dataset.

In our comparative analysis, the generative AI model consistently outperforms the 10-house aggregation in both fidelity retention and average input identifiability. Additionally, the synthetic model proves superior in most outlier-injection scenarios, with only two instances where aggregation marginally outperforms it.

5

Discussion

In this chapter we interpret the results of our analyses and discuss their novel contributions to literature, their implications on policy and smart meter data privacy practices, the limitations of this study, and how our results can be complemented by future work.

5.1 Interpretation of Quantitative Results

From our review of the data privacy strategies published by UK energy providers and private data distributors [23, 8], we know implementers of aggregation make two primary assumptions, both of which are founded on intuition and qualitative assessment rather than robust scientific evidence: the anonymity afforded by aggregating data diminishes with small aggregation sizes, and strong outlier features within aggregations are most vulnerable to identification. Our results indicate that both of these assumptions are broadly true, but with new dimensions of these relationships elucidated. Our discovery of a negative exponential relationship between

risk of identification and aggregation size tells us that the addition or removal of profiles from a small aggregation is more impactful than those from a large aggregation, and from our fit function we know the marginal privacy gain with each additional profile is asymptotic with no significant impact beyond the 11-profile aggregation.

From varying the input profile length, we see the same general negative exponential relationship for all lengths, but a strong correlation which indicates longer input profiles yield an output with a lower risk of identification. This behavior is intuitive and was postulated by [46], but our results not only serve as evidence to this idea, but define the precise quantitative nature of this relationship; a likely explanation is that the added complexity of more datapoints dilutes the significance of the identification of any single feature and increases the dimensionality of the data, reducing the confidence of any membership inference.

Our testing of the relative identifiability of profiles with artificially injected outlier features reveals an interesting trend: profiles with distinct features are initially more identifiable within an aggregation, but this identifiability diminishes as additional outlier features are added. Specifically, once the number of outlier features exceeds four, the outlier profile becomes, on average, less identifiable than the non-outlier profiles within the input set. Moreover, beyond this threshold, not only does the outlier profile's identifiability decrease, but the overall identifiability of all input profiles within the aggregation decreases as well.

While the heightened identifiability of an outlier profile with distinct features is intuitive, the subsequent reduction in identifiability as more outlier features are added is less straightforward. This counterintuitive outcome can be explained by considering the dynamics of feature representation within the aggregation. When few outlier features are present, their distinctiveness allows them to dominate the aggregation, making the outlier profile highly identifiable. However, as the number of outlier features increases, these features start to overshadow and obscure the subtle characteristics of non-outlier profiles that would otherwise contribute to the aggregation. Consequently, this obscuration enhances the overall anonymity of the input set.

As for the outlier profile itself, the addition of more distinct features can paradoxically lead to a decrease in its identifiability for two main reasons. First, the increased dimensionality

spreads the unique, spiky characteristics across a broader range of variables, diluting their individual impact within the aggregation process. Second, as the outlier profile accumulates more distinctive features, it begins to diverge significantly from the common patterns present in the rest of the dataset. This divergence makes it less likely that any single feature will stand out as exceptionally unique within the aggregation, effectively 'flattening' the profile's overall distinctiveness. As a result, the outlier profile may blend into the aggregated data more effectively than profiles with fewer, but sharper, outlier features. This blending effect can lead to a scenario where the outlier profile is actually less identifiable than the average non-outlier profiles, as its numerous distinct features become a form of noise rather than a clear signal.

Our comparison of aggregation with a Python-based generative model shows that even a simple AI implementation significantly outperforms aggregation in both fidelity and anonymity. Our fidelity metrics show that more statistical and quantile information is preserved from the training data which matched the AI prompt to its output than in an aggregation of the same number of profiles. In terms of anonymity, the 10 matched training profiles are, on average, over 40% less identifiable than the input to a 10-profile aggregation. The generative model also outperforms aggregation in masking profiles with outlier features in all but 2 experiments. One possible explanation is that aggregation produces flatter, less variable profiles, a weakness that makes them less similar to the input data. By contrast, the synthetic profiles have more variability and distinct features, which the classifier recognises as closer to the input. In reality, perhaps both representations attenuate the large outlier features but the classifier deems the natural variability preserved in the synthetic data as being closer to the input which has many injected spikes. These findings highlight the importance of assessing both the model and the data together for effective anonymity, as these results would likely vary strongly with the average spikiness of the training data.

5.2 Novel Contributions

This dissertation's contributions are threefold: the novel *probability of identification* assessment methodology demonstrated, the quantitative description of the nature of the anonymity afforded by aggregation, and a direct, quantified comparison between aggregation and the alternative synthetic model, all of which is tested using a real, metadata-rich profile set.

An assessment of the identification risk of an anonymised dataset is not itself a novel concept, however it remains largely unexplored for time-series analysis. No generalised approach to assessing the identification risk of a profile anonymised by some model is found in literature, but [24] informally posits that taking a Bayesian approach with some "quantity of inference" is the correct formulation to derive such a metric; they remark that *identifiability* is not binary, but continuous and given how dependant an anonymisation model's performance is on the nature of the input data, this metric should be a function of model and data. In their article, the *quantity of inference* is left generic. Additionally, random forest classifiers have been applied to membership inference attacks and tested as a quantification of the preserved information from input to output across medical time series anonymisation models [9]. The novelty of this dissertation's approach is its combination of these distinct angles developed in literature to produce a generalisable and comparable *probability of identification*. This contribution is particularly timely given the rapid advancements in data analytics and AI, which necessitate a shift from heuristic-based privacy approaches to those grounded in empirical evidence and capable of scaling with technological advancements.

Many of our results pertaining to the quantification of the efficacy of aggregation are, in some sense, a reinforcement of the previously held intuitions expressed in policy, albeit with additional detail and relationships quantified. However, it is crucial that, as our systems become increasingly big-data driven and the large-scale distribution of sensitive data more demanded, we must protect user safety using quantified, extensible privacy models which can be methodically adapted to the quickly changing landscape. The alignment of our results, such as the relationship between anonymity and aggregation size, with the assumptions of current

policy is better described as epistemic luck for policymakers than it is a reinforcement of an existing scientific conclusion. A security strategy built on unfounded assumptions is tenuous even in cases wherein these assumptions are proven correct, as the uncertainty compounds with each corollary, leading to disastrous, unexpected effects; this was seen in 2010 when the USCB discovered significant customer data insecurity in multiple dimensions of their system which, until that point, had been built on a heuristic aggregation model whose specific limits and weaknesses were unquantified and assumed to be minor as their system evolved on this foundation.

The novelty of our benchmarking of aggregation lies in its transformation of intuitive assumptions into scientifically parameterised principles. The discovered negative exponential relationship shows that increasing aggregation size is finite in its ability to improve the security of a system, and our outlier study demonstrates specific contexts in which aggregation is most vulnerable. Given the characteristic spiky patterns seen in domestic consumption profiles, the superiority of the synthetic model at preserving privacy with respect to these features represents a significant improvement over aggregation. Additionally, without such quantified evidence, simply doubling the mandated aggregation size to improve security or depending on aggregation to obscure especially sensitive profiles with specific, distinct features would be logical extensions of the existing assumptions, which have both been disproved. Given the ubiquity of aggregation as a privacy-preservation mechanism, it is unlikely to be replaced across all sectors overnight; as the SMIP and other data distribution systems evolve, this baseline description of their current implementations will be valuable.

Our application of the *probability of identification* framework to the direct comparison of aggregation with a synthetic model not only represents the first of its kind and a template upon which future comparisons can be made, but compelling evidence of the inadequacy of the aggregation model. The synthetic model used in this study is simplistic, unrefined, and trained on a smaller dataset compared to the Faraday model [60]. The goal of this dissertation is not to replace aggregation, but to demonstrate the benefits of investing in synthetic data and how easily aggregation can be outperformed. By directly comparing these models, we've established a baseline for the current system and shown that aggregation is no longer competitive in

today's data security environment.

5.3 Implications for Policy and Practice

5.3.1 Redefining *Privacy*

This dissertation addresses the ambiguous definition of *privacy* in both policy and practice. Legislation often equates *privacy* with *risk of identification*, but lacks a robust, objective measure of this risk, leaving data distributors without clear security standards. Absolute anonymity, as noted in the Scottish Power Energy Networks Smart Metering Data Privacy Plan, is impractical as it eliminates data utility. The challenge, then, is to maximise data utility while maintaining an acceptable level of identification risk. But what constitutes a *safe* risk? Establishing this threshold is crucial for a scientifically rigorous SMIP security strategy. Differential privacy offers a framework for such quantification, inherently setting an input-independent threshold. In contrast, models like aggregation and synthetic data require a data-driven, experimentally derived metric. Assessing the security of any system necessitates experimentation with datasets that reflect the variability of the data being protected, allowing for specific privacy guarantees based on transparent assumptions about the data.

This dissertation advocates for replacing aggregation as the standardised anonymisation model for UK smart meter data. The stagnation in current policy stems from a lack of quantitative benchmarks, but with our benchmark and the criticisms of aggregation's insecurity, we must ask: *If the security standard evolves, what should the probability of identification be for the new model?*

This can be approached in two ways: First, by setting a baseline that matches the current standard, ensuring any new model that performs as well as or better than the current practice is "safe." Preliminary tests suggest that the simple synthetic model developed here could be an acceptable replacement under this criterion. Second, by leveraging methods for quantifying a data representation's utility alongside our new privacy metric, we can optimise for the required

utility while maximising privacy. In some cases, determining the maximum permissible risk of identification might require specific attack simulations, but by quantifying both utility and privacy, we enable a more precise and efficient balance between the two.

5.3.2 Reevaluating the Role of Aggregation

Aggregation not only represents the standardised anonymisation method for smart meter data distribution in the UK, but also a common practice across many industries worldwide and a "main source" for IoT systems, Cloud environments, Artificial Intelligence and Machine Learning applications products in the Digital Single Market [64]. Its widespread use legitimises it, its ease of implementation makes it an attractive option for managers, and the notion of combining multiple sets of data to anonymise while retaining information is an intuitive concept which policymakers of any technical background can accept. Data privacy is rarely discussed in the smart meter rollout, with concerns like skepticism of government surveillance and resistance to environmental policies being more prominent and major contributors to the stagnation in new smart meter adoption. The use of aggregation has not been a major issue, and changes to the SMIP security system may expose flaws and require new customer acceptance of complex, abstract management systems; the proposal of an AI-based solution could lead to further public skepticism, making this dissertation's proposal difficult to implement under the SMIP's primary directive of full smart meter coverage in the UK.

However, mounting evidence shows the insecurity of aggregation, and this dissertation demonstrates its shortcomings in fidelity and privacy. The long-term success of the smart meter rollout relies on the utility and security of its data as well as the maintenance of public trust, making the evolution of the distribution model essential and in the best interest of all stakeholders involved. Simply replacing aggregation with a more secure alternative addresses a symptom, not the root issue. We should be working to establish probability of identification thresholds where relevant and define the security guarantees that come with them in an accessible, public-facing manner. With this bsqprivacy-first framing, the onus is

not on the model to be intuitive nor present as safe to the consumer, but on the security guarantee communicated to the public to be sufficient. This public communication of privacy-preserving methods, such as Apple's detailed public-facing explanations of their differential privacy and aggregation policies, plays a critical role in building consumer trust. Apple frames these techniques in an accessible way, emphasising how added noise and data aggregation make it impossible to trace individual data back to the user [65], which gives consumers a tangible sense of security. However, the application of AI-based models for privacy preservation presents a challenge in this regard. Unlike the straightforward concepts of noise addition or aggregation, the inner workings of AI models are often more complex and less intuitive, making it difficult to convey their security benefits to data subjects in a way that is both clear and reassuring. This gap in communication has the potential to undermine the trust that is crucial for widespread adoption of such advanced privacy-preserving techniques, and should not be trivialised.

This discussion applies broadly to the distribution of sensitive time series data, which varies widely in content and sensitivity. In some cases, aggregation may suffice, particularly when the utility of the data lies in the area under its curve, making the aggregation of multiple profiles adequate. By adopting a "privacy-first" approach, our quantified privacy framework and insights into the factors affecting aggregation's privacy allow for more targeted solutions where aggregation is one of many tools. For instance, when fidelity and granularity are less critical than privacy, a larger aggregation size may be appropriate; conversely, when granularity matters more than timescale, a longer input profile with a smaller aggregation size could be chosen. The contributions of this thesis make such adaptations feasible while maintaining comparable privacy standards. However, for smart meter data, aggregation has significant drawbacks, and based on our literature review and results, promoting synthetic data is likely a more robust solution in most applications.

5.3.3 Evaluating AI-Generated Synthetic Smart Meter Data

Among potential replacements for the current aggregation model, synthetic data generation stands out due to its enhanced utility, privacy, and adaptability. Unlike the rigid aggregation approach, synthetic models offer customisable solutions where users can query specific information while maintaining privacy through statistical similarity to real data.

This flexibility is especially beneficial when dealing with outlier features that aggregation struggles to anonymise. For example, a tailored synthetic model can preserve essential details while distorting them to meet predefined privacy thresholds. Microsoft's generative AI [66], which iteratively refines synthetic data to resemble private datasets while targeting specific, sensitive features for distortion, illustrates this concept. Although such an approach raises concerns about potential vulnerabilities like model inversion attacks, these can be managed effectively, and even in their more diluted forms, synthetic models provide greater utility than simple aggregation.

Beyond superior privacy and utility, AI-powered synthetic solutions represent a more resilient architecture to future threats. Identification attacks often rely on cross-referencing distributed data with other datasets. The transparency of aggregation can make it vulnerable as attackers have more information to narrow the scope of their cross-reference and reconstruct input data, while the black-box nature of synthetic models offers a stronger defense by limiting access to the models internal workings. This opacity makes synthetic models better equipped to handle new security challenges and reduces the risk of data identification or reconstruction [67].

From a legislative perspective, synthetic data should not be exempt from scrutiny merely because it does not contain real data. Instead, it could be recognised as an advanced form of aggregation, subject to the same rigorous privacy standards. The investment in synthetic systems, which are demonstrably more private and useful than traditional aggregation, is a key takeaway of this dissertation. Rather than advocating for a specific model, this research highlights the need for further development in this area, as shown by the superiority of even simple synthetic models over current aggregation practices.

5.3.4 Limitations and Future Directions

This study, while contributing valuable insights, has several limitations that suggest directions for future research:

Dataset Limitations: The reliance on a dataset comprising 80 residential profiles, though informative, may limit the generalisability of the findings across diverse geographic regions and socio-economic contexts.

Model Sophistication: The generative AI model, while demonstrating potential over traditional aggregation methods, is a foundational implementation; more advanced models could offer even greater improvements in privacy and utility.

Privacy Metrics: The study's focus on identification probability as the primary metric, though critical, does not fully address other important privacy concerns, such as robustness against specific attack vectors.

Data Fidelity and Utility: The evaluation of data fidelity, while indicative of the model's effectiveness, does not fully capture the synthetic data's utility in practical, context-specific applications like forecasting or policy development.

AI Implementation Challenges: A key distinction between aggregation and AI-based methods lies in their data requirements and infrastructure needs. Aggregation relies solely on the profiles it aggregates, making it more adaptable in situations with limited data. In contrast, AI models require extensive training datasets, which may not always be available, thus limiting their applicability in certain contexts. Addressing the technical challenges of scaling and implementing generative AI models in real-world applications to replace aggregation may be complex and in some cases, infeasible given current data access and infrastructure.

Building on these findings, future research should:

- Expand the dataset to include a more diverse range of profiles, enhancing the generalisability and robustness of the conclusions.
- Develop and test more sophisticated generative models to provide deeper insights into the potential of AI-driven approaches to handle complex data scenarios and maintain high utility across various applications.
- Explore additional privacy metrics, such as resistance to different attack vectors, to offer a more comprehensive assessment of the security of synthetic data models.
- Conduct longitudinal studies incorporating multiple years of data to improve understanding of seasonal and long-term trends, critical for practical applications.
- Address the technical challenges of scaling and implementing generative AI models in real-world settings, including infrastructure and expertise requirements.
- Examine potential biases in synthetic data generation to ensure fairness and applicability across different populations, leading to more equitable and effective privacy-preserving strategies.

A more thorough analysis of the uncertainty in our results is beyond the scope of this dissertation, as the random forest classification method used here is demonstrated in literature and this dissertation does not seek to quantify our analysis' efficacy at predicting the success of specific attack vectors, but to demonstrate a generic framework. The extensive database used here is a diverse set of user data, but all contributors are volunteers to an academic study, which may represent a biased sample of the British population. As this framework and new privacy models are tested on data which is increasingly representative of the population, results are likely to skew as the data changes.

6

Conclusion

This dissertation studies the dilemma between privacy and utility for smart meter data distribution in the UK, why the current privacy-preserving distribution model, aggregation, is fraught, and how to build a demonstrably better system as foundation to the national UK smart meter project. Current policies in the UK enforce that smart meter data, which is legally considered personal data, be aggregated, or averaged, with other profiles before distribution to data users to minimise the risk of customer identification. Legislation provides no threshold for a safe risk of identification; rather, its standards are defined by the practice it enforces, aggregation, which itself has no scientific basis nor does it make quantified privacy guarantees. A palpable tension has formed where critics argue aggregate data is not high enough fidelity and its enforcement is too rigid for the demands of data users, while simultaneously advocating for more security. The persistence of this tension can be attributed to the unquantified nature of current system, as we have no mechanisms or benchmarks with which the model could be safely adapted or compared to another.

To examine the privacy of the current system and develop a vector through which competitive systems may be compared, we develop a novel privacy assessment framework which trains a random forest classifier to quantify information preserved from input to output, and using Bayesian inference yields the *probability of identifying* the profiles used as input to some published data representation. Current publishers of smart meter data and practitioners of aggregation intuitively assume that small aggregation sizes lead to higher risk of re-identification and that profiles with large, distinct features are more vulnerable within an aggregation, but this is unsupported by concrete evidence. We use our assessment framework to quantify the anonymity afforded by aggregation, on average, given a large set of real, metadata-rich smart meter data provided by the Energy Demand Observatory and Laboratory (EDOL). We study how the number of profiles aggregated, the length of the profiles being aggregated, and the presence of input profiles with distinct outlier features impact the anonymity of the input to an aggregation.

From a survey of literature and reports from analogous international data sharing schemes, we find that aggregation, in most cases, is a dangerously vulnerable model and ill-equipped to supply the insight demanded by data users whom the UK smart meter roll out is designed to benefit, such as policymakers, researchers, and utilities. Training generative AI models to synthesise artificial consumption data based on metadata prompts emerges as the most promising replacement, so this dissertation develops such a model in Python and trains it on the aforementioned EDOL consumption data. This model's output representative load profiles are assessed for their preservation of fidelity and for their anonymity using this dissertation's novel privacy assessment framework, and they are directly compared to representations of the same data but anonymised with an aggregation.

Our tests yield two primary sets of results: quantitative metrics which demonstrate the relative efficacy of aggregation using a variety of parameters, and an objective, direct comparison of the respective merits associated with synthetic data generation and aggregation. A negative exponential relationship between risk of identification and aggregation size is discovered, as well as a strong negative correlation between input profile length and risk of identification. It is

also demonstrated that an input profile with a single outlier feature is more re-identifiable in an aggregation, but as the number of outlier features increases, its identifiability decreases to the point where beyond 4 outlier features both the injected outlier and all other input profiles are, on average, less identifiable. When we compare our aggregation tests with the output of our AI, we find that across all fidelity metrics the synthetic output more closely resembles its input data, in almost every test injected outlier profiles are more identifiable within an aggregation than the synthetic output, and overall, the input profiles to an aggregation are, on average, more than 70% more identifiable than the training input to our synthetic model.

This dissertation provides a quantitative foundation for current smart meter data distribution policies while exposing the shortcomings of the existing privacy-preservation model compared to a basic generative AI approach. The generative model this dissertation develops does not fully explore the potential of the architecture nor the peak performance of generative AI, however, even in its basic form, synthetic generation, offering greater security, flexibility, and utility, proves to be a stronger foundation for our energy system, delivering higher quality and more anonymous results. As we grow our smart grid in the UK and rely more heavily on the safe distribution of big data, this dissertation shows the significant potential of the investment in AI-powered generative systems tailored to the unique needs of individual users, and provides a novel framework with which future iterations may be compared objectively, thereby quantifying the dilemma between utility and privacy on both sides.

This study is limited by data access and the scope of the tests conducted. The utility of published smart meter data extends beyond its statistical similarity to *real* data and is context-dependent, often better assessed by its performance in tasks like forecasting or preserving specific features. The fidelity metrics used here, while indicative of our model's quality compared to aggregation, don't capture the full picture. Additionally, generative models improve with more training data and optimized parameters. Future work should focus on developing targeted generative models, testing them with context-specific utility metrics, and using large, representative datasets to ensure they anonymise data effectively while adhering to privacy standards using our *probability of identification* framework.

This dissertation underscores the critical need for a shift in the approach to smart meter data privacy in the UK. The current reliance on aggregation, without a scientific foundation or established privacy standards, has left the system vulnerable and stagnant. As our energy infrastructure increasingly depends on big data, it is imperative to adopt robust privacy metrics that balance utility and security. AI-powered generative models, as demonstrated in this research, provide a superior alternative to traditional aggregation, offering higher data quality and enhanced privacy protections. Embracing these models is not just an opportunity but a necessity for safeguarding the future of the UKs smart grid and maintaining public trust. Policymakers and practitioners must prioritize the development of stringent privacy standards and utility-maximizing models that meet these standards, ensuring that the UK's energy infrastructure remains secure and resilient in the face of evolving challenges.

Bibliography

[1] Energy Systems Catapult. *Data for Good: Smart Meter Data Access*. Tech. rep. Energy Systems Catapult, 2023. URL: https://es.catapult.org.uk/report/data-for-good-smart-meter-data-access/.

[2] House of Commons Public Accounts Committee. *Smart Meter Rollout: Progress and Costs*. Tech. rep. Accessed: 2024-08-05. UK Parliament, 2024. URL: https://publications.parliament.uk/pa/cm5803/cmselect/cmpubacc/1332/report.html.

[3] Ross Anderson et al. *Privacy and Security: The Challenges of Data Protection in Modern Networks*. Tech. rep. Accessed: 2024-08-06. University of Cambridge, 2023. URL: https://www.cl.cam.ac.uk/~rja14/Papers/JSAC-draft.pdf.

[4] Nothando Mlilo, Jason Brown, and Tony Ahfock. "Impact of intermittent renewable energy generation penetration on the power system networks A review". In: *Technology and Economics of Smart Grids and Sustainable Energy* 6 (2021), p. 25. DOI: 10.1007/s40866-021-00123-w. URL: https://doi.org/10.1007/s40866-021-00123-w.

[5] Ying Wang et al. "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges". In: *IEEE Transactions on Smart Grid* 10.3 (2019), pp. 3125–3148. DOI: 10.1109/TSG.2018.2818167. URL: https://doi.org/10.1109/TSG.2018.2818167.

[6] Energy Demand Observatory and Laboratory (EDOL). *Energy-Use Project Dataset*. 2020. URL: https://energy-use.org/Data/.

[7] National Grid Electricity Distribution. *Smart Meter Data Privacy Plan*. 2024. URL: https://www.nationalgrid.co.uk/downloads-view-reciteme/657539#:~:text=Consumption%20Data%20will%20be%20accessed%20by%20NGED%20on%20a%20periodic,connected%20to%20the%20NGED%20network.&text=1.19%20NGED%20will%20request%20a,meter%20to%20be%20provided%20automatically.

[8] Smart Energy Research Lab (SERL). *New SERL Datasets Publicly Available*. Accessed: 2024-08-06. 2024. URL: https://serl.ac.uk/new-serl-datasets-publicly-available.

[9] N. Koren et al. "Membership Inference Attacks Against Time-Series Models". In: *arXiv.Org* (2024). URL: https://doi.org/10.48550/arxiv.2407.02870.

[10] Department for Business, Energy Industrial Strategy. *Smart Metering Implementation Programme: Review of the Data Access and Privacy Framework*. Tech. rep. UK Government, 2023. URL: https://assets.publishing.service.gov.uk/media/66016b44a6c0f7f514ef9198/smart-metering-implementation-programme-review-data-access-privacy-framework.pdf.

[11] Hsiao-Dong Chen, Shun-Hsien Fan, and Shun-Yu Chang. "Fast Fault Location for Fast Restoration of Smart Electrical Distribution Grids". In: *IEEE Transactions on Smart Grid* 8.5 (2016), pp. 2226–2235. DOI: 10.1109/TSG.2016.2584611. URL: https://www.researchgate.net/publication/304489241_Fast_Fault_Location_for_Fast_Restoration_of_Smart_Electrical_Distribution_Grids.

[12] Tobias Brudermueller and Markus Kreft. *Smart Meter Data Analytics: Practical Use-Cases and Best Practices of Machine Learning Applications for Energy Data in the Residential Sector*. Tech. rep. Climate Change AI, 2023. URL: https://www.climatechange.ai/papers/iclr2023/3.

[13] Yi Wang, Qixin Chen, and Chongqing Kang. *Smart Meter Data Analytics: Electricity Consumer Behavior Modeling, Aggregation, and Forecasting*. SpringerLink, 2019. URL: https://link.springer.com/book/10.1007/978-3-030-30250-4.

[14] Katarzyna Korczak and Tadeusz Skoczkowski. "Technology Innovation System Analysis of Electricity Smart Metering in the European Union". In: *Energies* 13.4 (2020), p. 916. DOI: 10.3390/en13040916. URL: https://doi.org/10.3390/en13040916.

[15] Ofgem. *Smart Meter Reporting - Decision*. Accessed: 2024-08-14. 2023. URL: https://www.ofgem.gov.uk/sites/default/files/2023-11/Smart%20Meter%20Reporting%20-%20Decision1699465143609.pdf.

[16] Michael A Lisovich, Deirdre K Mulligan, and Stephen B Wicker. "Inferring personal information from demand-response systems". In: *IEEE Security & Privacy*. IEEE. 2010, pp. 11–20.

[17] R. Jain et al. "Privacy and Security of Smart Meter Data: A Survey". In: *Journal of Information Security* (2014).

[18] Mohamed S. Abdalzaher, Mostafa M. Fouda, and Mohamed I. Ibrahem. "Data Privacy Preservation and Security in Smart Metering Systems". In: *Energies* 15.19 (2022), p. 7419. DOI: 10.3390/en15197419. URL: https://www.mdpi.com/1996-1073/15/19/7419.

[19] Public Accounts Committee. *Delayed Smart Meter Programme Fails to Hit Targets and Secure Public Support*. Tech. rep. Accessed: 2024-08-06. UK Parliament, 2023. URL: https://committees.parliament.uk/committee/127/public-accounts-committee/news/197947/delayed-smart-meter-programme-fails-to-hit-targets-and-secure-public-support/.

[20] McKinsey. *Consumer data protection and privacy*. Tech. rep. McKinsey, 2020. URL: https://www.mckinsey.com.

[21] PrivacyEnd. *The Hidden Costs: Understanding How Data Breaches Affect Consumer Trust and Brand Reputation*. Tech. rep. PrivacyEnd, 2020. URL: https://www.privacyend.com.

[22] Center for Data Ethics and Innovation. *Maximising the Beneficial Use of Personal Information Held in the Public Sector*. Tech. rep. Accessed: 2024-08-06. UK Government, 2020. URL: https://www.gov.uk/government/publications/cdei-publishes-its-first-report-on-public-sector-data-sharing/addressing-trust-in-public-sector-data-use.

[23] SP Energy Networks. *SPEN Smart Metering Data Privacy Plan*. https://www.ofgem.gov.uk/sites/default/files/docs/2020/07/spen_smart_metering_data_privacy_plan_-_july_20_final_redacted.pdf. Accessed: 2024-08-11. July 2020.

[24] Andrew Gelman. *What I think about when I think about identifiability in Bayesian inference*. Statistical Modeling, Causal Inference, and Social Science Blog. Accessed: 2024-08-06. 2014. URL: https://statmodeling.stat.columbia.edu/2014/02/12/think-identifiability-bayesian-inference/.

[25] Zekeriya Erkin and Mustafa Erkin. "Two Is Not Enough: Privacy Assessment of Aggregation Schemes in Smart Metering". In: *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*. ACM. 2017, pp. 85–90. URL: https://www.researchgate.net/publication/320447619_Two_Is_Not_Enough_Privacy_Assessment_of_Aggregation_Schemes_in_Smart_Metering.

[26] Costas Efthymiou and Georgios Kalogridis. "Smart Grid Privacy via Anonymization of Smart Metering Data". In: *2010 First IEEE International Conference on Smart Grid Communications*. IEEE, 2010, pp. 238–243. DOI: 10.1109/SMARTGRID.2010.5622050.

[27] A. Acquisti and J. Grossklags. "Privacy and rationality in individual decision making". In: *IEEE Security Privacy* 3.1 (2005), pp. 26–33. DOI: 10.1109/MSP.2005.22.

[28] Elette Lui and Rafael Pass. "Outlier Privacy". In: *Theory of Cryptography*. Ed. by Yevgeniy Dodis and Jesper Buus Nielsen. Vol. 9015. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2015, pp. 230–250. DOI: 10.1007/978-3-662-46497-7_11. URL: https://doi.org/10.1007/978-3-662-46497-7_11.

[29] National Renewable Energy Laboratory (NREL). *News Release: New Data Set Quantifies How and When Energy Is Used Across All Major U.S. Building Types and Climate Regions*. Accessed: 2024-08-27. 2021. URL: https://www.nrel.gov/news/press/2021/new-data-set-quantifies-how-when-energy-used-across-major-us-building-types-climate-regions.html.

[30] Eoghan McKenna, Ian Richardson, and Murray Thomson. "High-resolution stochastic integrated thermalelectrical domestic demand model". In: *Applied Energy* 165 (2012), pp. 445–455.

[31] Rui Yuan et al. "Unleashing the benefits of smart grids by overcoming the challenges associated with low-resolution data". In: *Cell Reports Physical Science* 5.2 (2024), p. 101830. ISSN: 2666-3864. DOI: https://doi.org/10.1016/j.xcrp.2024.101830. URL: https://www.sciencedirect.com/science/article/pii/S2666386424000559.

[32] Kingsley Ukoba et al. "Optimizing renewable energy systems through artificial intelligence: Review and future prospects". In: *Energy Environment* (2024), pp. 1–47. DOI: 10.1177/0958305X241256293. URL: https://journals.sagepub.com/doi/full/10.1177/0958305X241256293.

[33] Saad Emshagin, Wayes Koroni Halim, and Rasha Kashef. "Short-term Prediction of Household Electricity Consumption Using Customized LSTM and GRU Models". In: *arXiv preprint arXiv:2212.08757* (2022). URL: https://arxiv.org/abs/2212.08757.

[34] Aida Mehdipour Pirbazari et al. "Short-Term Load Forecasting Using Smart Meter Data: A Generalization Analysis". In: *Processes* 8.4 (2020), p. 484. DOI: 10.3390/pr8040484. URL: https://www.mdpi.com/2227-9717/8/4/484.

[35] Paraskevas Koukaras et al. "Energy Forecasting: A Comprehensive Review of Techniques and Technologies". In: *Energies* 17.7 (2024), p. 1662. DOI: 10.3390/en17071662. URL: https://www.mdpi.com/1996-1073/17/7/1662.

[36] G. Eibl and D. Engel. "Influence of Data Granularity on Non-Intrusive Appliance Load Monitoring". In: *Energy Procedia* 83 (2015), pp. 80–86. DOI: 10.1016/j.egypro.2015.12.195.

[37] Amogh Khedar et al. *Non-Intrusive Load Monitoring (NILM): A Review and Classification of Approaches*. arXiv preprint arXiv:2305.10352. Accessed: 2024-08-06. 2023. URL: https://arxiv.org/pdf/2305.10352#:~:text=Non%2DIntrusive%20Load%20Monitoring%20(NILM)%20%5B20%5D%2C,aggregated%20load%20curve%20%5B29%5D..

[38] Y. Zhao et al. "Appliance Detection Using Very Low-Frequency Smart Meter Time Series". In: *arXiv preprint arXiv:2305.10352* (2020). URL: https://arxiv.org/abs/2305.10352.

[39] Peter Francisco et al. "Bottom-up Forecasting: Applications and Limitations in Load Forecasting Using Smart-Meter Data". In: *Data-Centric Engineering* (2023). Accessed: 2024-08-06. URL: https://www.cambridge.org/core/journals/data-centric-engineering/article/bottomup-forecasting-applications-and-limitations-in-load-forecasting-using-smartmeter-data/FEE91B25F96FDFB350B3A1C17608E28C.

[40] ScienceDirect. *Identifiability*. ScienceDirect Topics. Accessed: 2024-08-06. n.d. URL: https://www.sciencedirect.com/topics/mathematics/identifiability.

[41] Vasu Jakkal. *Cybersecurity Threats are Always ChangingStaying on Top of Them is Vital, Getting Ahead of Them is Paramount*. Microsoft Security Blog. Accessed: 2024-08-07. 2022. URL: https://www.microsoft.com/en-us/security/blog/2022/02/09/cybersecurity-threats-are-always-changing-staying-on-top-of-them-is-vital-getting-ahead-of-them-is-paramount/.

[42] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Vol. 9. 3-4. Foundations and Trends in Theoretical Computer Science, 2014, pp. 211–407. DOI: 10.1561/0400000042. URL: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

[43] European Commission. *General Data Protection Regulation*. https://ec.europa.eu/info/law/law-topic/data-protection_en. Accessed: 2024-08-11. 2016.

[44] John M Abowd. "The U.S. Census Bureau adopts differential privacy". In: *Proceedings of the National Academy of Sciences* 115.13 (2018), pp. 3332–3333. DOI: 10.1073/pnas.1801846115.

[45] Fei Teng et al. *Balancing Privacy and Access to Smart Meter Data*. Tech. rep. Energy Futures Lab, Imperial College London, 2022. URL: https://www.imperial.ac.uk/energy-futures-lab/reports/briefing-papers/paper-9/.

[46] Li Gao, Zhiyong Li, and Xue Zhang. "A Study on the Effectiveness of Aggregation Techniques for Privacy Protection in Smart Meter Data". In: *Journal of Privacy and Confidentiality* 11.2 (2019), pp. 15–30.

[47] National Institute of Standards and Technology (NIST). *Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series*. https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our. Accessed: 2024-08-06. 2023.

[48] Simson Garfinkel, John M. Abowd, and Christian Martindale. "Understanding Database Reconstruction Attacks on Public Data". In: *Communications of the ACM* 62.3 (2019), pp. 46–53. URL: https://www.nist.gov/publications/differential-privacy-introduction-our-blog-series.

[49] E. McKenna, I. Richardson, and M. Thomson. "Smart meter data: Balancing consumer privacy concerns with legitimate applications". In: *Energy Policy* 41 (2012), pp. 807–814. DOI: 10.1016/j.enpol.2011.11.049.

[50] Ashwin Machanavajjhala et al. "l-Diversity: Privacy Beyond k-Anonymity". In: *22nd International Conference on Data Engineering (ICDE'06)*. 2007, pp. 24–24. DOI: 10.1109/ICDE.2006.1.

[51] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. "Workload-aware anonymization techniques for large-scale datasets". In: *ACM Transactions on Database Systems (TODS)* 33.3 (2008), pp. 1–47. DOI: 10.1145/1386118.1386123. URL: https://dl.acm.org/doi/10.1145/1386118.1386123.

[52] H. Brendan McMahan et al. "Federated Learning of Deep Networks using Model Averaging". In: *arXiv preprint arXiv:1602.05629*. 2017. URL: https://arxiv.org/abs/1602.05629.

[53] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://arxiv.org/abs/1806.04868.

[54] Shuang Chen and et al. "GAN-based Synthetic Data Generation for Accurate, Private and Scalable Smart Meter Analytics". In: *IEEE Transactions on Smart Grid* (2021). URL: https://ieeexplore.ieee.org/document/9310241.

[55] P. Ezhilarasi et al. "Smart Meter Synthetic Data Generator development in Python using FBProphet". In: *Software Impacts* 15 (2023), p. 100468. DOI: 10.1016/j.simpa.2023.100468.

[56] Research Data Scotland. *Intro to Synthetic Data*. https://www.researchdata.scot/our-work/data-explainers/intro-to-synthetic-data/. Accessed: 2024-08-07. 2024.

[57] Grand View Research. *Synthetic Data Generation Market Size, Share Trends Analysis Report By Data Type, By Modeling Type, By Offering, By Application, By End-use, By Region, And Segment Forecasts, 2023 - 2030*. Tech. rep. Grand View Research, 2023.

[58] Bappaditya Choudhury. "The Legal Challenges and Opportunities of Synthetic Data". In: *Big Data Society* 11.1 (2024), p. 20539517241231277. DOI: 10.1177/20539517241231277. URL: https://journals.sagepub.com/doi/10.1177/20539517241231277.

[59] Ana Beduschi. "Synthetic data protection: Towards a paradigm change in data regulation?" In: *Big Data Society* 11.1 (2024). DOI: 10.1177/20539517241231277. URL: https://journals.sagepub.com/doi/10.1177/20539517241231277.

[60] S. Chai and G. Chadney. "Faraday: Synthetic Smart Meter Generator for the Smart Grid". In: *Tackling Climate Change with Machine Learning, ICLR 2024 Workshop*. 2024.

[61] Jingyu Jia et al. "Total variation distance privacy: Accurately measuring inference attacks and improving utility". In: *Information Sciences* 626 (2023), pp. 537–558. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2023.01.037. URL: https://www.sciencedirect.com/science/article/pii/S0020025523000361.

[62] Ashkan Sheikhi et al. "Energy Data Analytics for Real-Time Consumer-Side Energy Management: A Review". In: *Energy Informatics* 5.1 (2022), pp. 1–24. DOI: 10.1186/s42162-022-00213-8. URL: https://energyinformatics.springeropen.com/articles/10.1186/s42162-022-00213-8.

[63] Mehdi Shateri et al. "A privacy-preserving multidimensional data aggregation scheme with secure query processing for smart grid". In: *The Journal of Supercomputing* 77.2 (2020), pp. 5174–5183. DOI: 10.1007/s11227-020-03260-6. URL: https://link.springer.com/article/10.1007/s11227-020-03260-6.

[64] Emanuela Podda. "Shedding light on the legal approach to aggregate data under the GDPR & the FFDR". In: *Expert Meeting on Statistical Data Confidentiality, Conference of European Statisticians, United Nations Economic Commission for Europe (UNECE)*. Available at: https://unece.org/sites/default/files/2021-12/SDC2021_Day1_Podda_AD.pdf. Università di Bologna. Poland, Dec. 2021.

[65] Apple Inc. *Differential Privacy Overview*. Accessed: 2024-08-27. 2017. URL: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

[66] Microsoft Research. *The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI*. https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/. Accessed: 2024-08-20. 2023.

[67] Sheng Chai et al. "Definition of Good for Synthetic Smart Meter Data". In: *IEEE Transactions on Smart Grid* (June 2024). In press.