

# FARADAY: SYNTHETIC SMART METER GENERATOR FOR THE SMART GRID

**Sheng Chai & Gus Chadney**

Centre for Net Zero

{sheng.chai, gus.chadney}@centrefornetzero.org

## ABSTRACT

Access to smart meter data is essential to rapid and successful transitions to electrified grids, underpinned by flexibility delivered by low carbon technologies, such as electric vehicles (EV) and heat pumps, and powered by renewable energy. Yet little of this data is available for research and modelling purposes due to consumer privacy protections. Whilst many are calling for raw datasets to be unlocked through regulatory changes, we believe this approach will take too long. Synthetic data addresses these challenges directly by overcoming privacy issues. In this paper, we present Faraday, a Variational Auto-encoder (VAE)-based model trained over 300 million smart meter data readings from an energy supplier in the UK, with information such as property type and low carbon technologies (LCTs) ownership. The model produces household-level synthetic load profiles conditioned on these labels, and we compare its outputs against actual substation readings to show how the model can be used for real-world applications by grid modellers interested in modelling energy grids of the future.

## 1 INTRODUCTION

A huge part of the global transition to net zero involves increasing the share of electricity from renewable sources and concomitantly electrifying heating and transport. Given the variability of renewables and growing adoption of low carbon technologies (LCTs), there are new challenges to our electricity grid such as creating new demand peaks, grid constraints and mismatches between demand and supply. Households with LCTs however can provide flexibility through automation to solve these challenges (National Grid, 2021; Ofgem, 2021; Xilas, 2023; Cheesecake Energy Ltd., 2022).

With access to granular household-level electricity consumption data, especially of households with LCTs, we can build better bottom-up grid models of future energy systems to model scenarios, such as varying frequencies of adverse weather conditions or different heat pump adoption rates to understand grid resiliency, identify grid constraints, and plan for grid reinforcement projects (Damianakis et al., 2023; Chen et al., 2023). Many countries have rolled out smart meters to collect electricity consumption at 15 or 30-minute intervals. Access to this data is limited due to privacy concerns. An example of a smart meter data repository in the United Kingdom is the Smart Energy Research Lab (Elam et al., 2023) which has a strict approval criteria. Calls to liberalise smart meter data for public good (Energy Systems Catapult, 2023) are growing, but they too involve lengthy bureaucracy and strict governance processes. Generating synthetic smart meter data circumnavigates these issues.

There is a growing number of literature (Zhang et al., 2019; Liang & Wang, 2022; Gu et al., 2019) that have applied generative artificial intelligence technologies to create synthetic smart meter data. Yet many of them are only theoretical and lack the ability to condition outputs on the type of LCTs that households own. Understanding how different households with LCTs consume electricity is important in modelling grid systems of the future.

In this paper, we present Faraday - a model trained using a combination of Variational Auto-encoder and Gaussian Mixture Model. We demonstrate how it has been evaluated and tested in the real world. The model is trained on a proprietary dataset of UK households with metadata on the type of LCTs they own and as such is capable of conditioning outputs by this information. The model is

currently deployed live as a web-app and an API in closed-alpha phase, available to about 50 alpha testers from academia and industry.

## 2 METHODOLOGY

### 2.1 ABOUT THE DATASET

We have access to proprietary data belonging to an energy supplier in the United Kingdom. The dataset consists of metered half-hourly electricity consumption data of 20 thousand households in 2021 and 2022, alongside attributes such as the type of LCTs they own, the property type and the property’s energy rating. The total number of smart meter readings in this dataset is in the excess of 300 million.

The access to this rich dataset means Faraday is able to generate household-level synthetic smart meter profiles that are conditioned upon inputs that the user provides, such as whether the household owns an electric vehicle (EV), their capacity to shift around consumption in response to grid conditions (i.e. whether they have a smart tariff) and the type of property (e.g. bungalows, terraced or flats). The model’s outputs are valuable to grid modellers who want to model future energy systems that rely on low carbon technologies and renewable energy sources.

### 2.2 FARADAY ARCHITECTURE

Faraday is a model based on Conditional Variational Auto-encoder (VAE) (Sohn et al., 2015) and Gaussian Mixture Model (GMM) algorithms. It works by first training a conditional VAE. The training data is then mapped to the latent space using the trained auto-encoder on which a GMM is trained to learn the distribution of the latent space. During inference, a random sample of latent vectors is drawn from the GMM and decoded using the trained decoder. To support conditional sampling, labels are appended to the latent codes when training the GMM. During sampling, random samples are drawn and labels that do not match the user’s inputs are discarded.

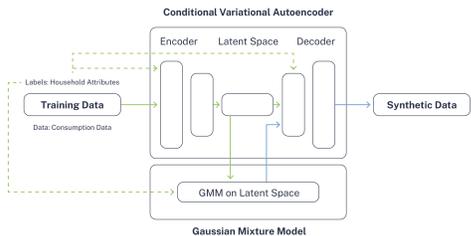


Figure 1: Faraday architecture

#### 2.2.1 MODIFICATIONS TO VAE

A traditional VAE’s reparametrization layer uses a normal distribution via the Kullback-Leibler (KL) Divergence loss to approximate the latent space. As smart meter data is non-normal and heavily positively skewed, KL-Divergence loss is not ineffective for learning this distribution. Applying log-normal transformation to the input data helps but comes at the cost of the fidelity at the higher quantiles (the peak loads) which is an area of interest when it comes to grid modelling. Instead of KL-Divergence loss, the Maximum Mean Divergence (MMD) loss is used in similar fashion to InfoVAE (Zhao et al., 2018). In addition, quantile losses at 5<sup>th</sup>, 50<sup>th</sup> and 95<sup>th</sup> are added to the loss function to help improve fidelity of synthetic outputs at various quantiles. The loss function used is therefore the sum of: a. the reconstruction loss (mean-squared error), b. the MMD (instead of KL-Divergence) loss and c. the three quantile losses.

#### 2.2.2 GAUSSIAN SAMPLING OF LATENT SPACE

The latent space of a traditional VAE is modelled using a unimodal Gaussian distribution. However, this is insufficient to capture the distribution of the latent space. To guarantee that the distribution of the generated samples matches that of the real data, a large random sample of the training data was drawn and encoded to the latent space with the trained encoder. A GMM is then trained to learn the distribution of the latent space where the training data occupies. During inference, random samples are drawn using the GMM and samples are decoded using the trained decoder. This is different to

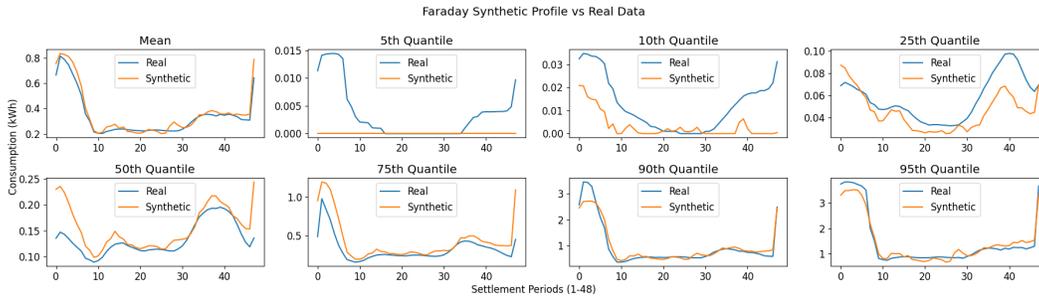


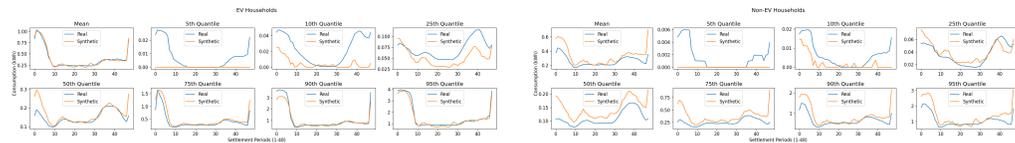
Figure 2: Faraday outputs at various quantiles. Y-axis is the kWh consumption. X-axis is the half-hourly periods where 1 is 00:00hrs and 48 is 23:30 hrs.

the original GM-VAE (Aguilera et al., 2023) where the Gaussian Mixture Model replaces the prior distribution of a unimodal Gaussian to learn the distribution of the latent space implicitly.

### 3 RESULTS

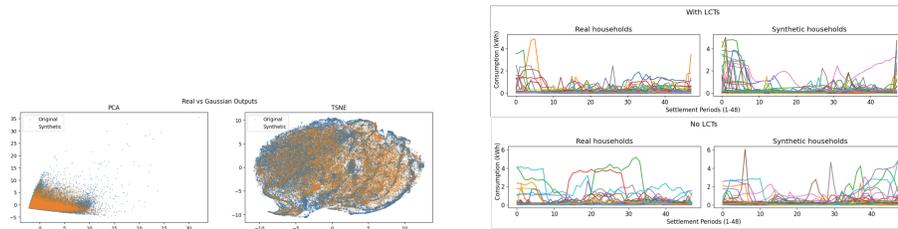
In this section we analyse the outputs of Faraday through these three qualities Jordon et al. (2022): 1) Fidelity 2) Utility and 3) Privacy. We also compare the results to other GAN-based models for utility and demonstrate that Faraday outputs were of higher utility than GAN-based outputs.

#### 3.1 FIDELITY



(a) Load profile at various quantiles for Households with EV. (b) Load profile at various quantiles for Households without EV.

Figure 3: Faraday outputs at various quantiles conditioned by whether households own an electric vehicle (EV).



(a) Distribution of Faraday outputs visualised with PCA and T-SNE plots. (b) Comparison between real vs Faraday outputs at individual household level.

Figure 4: Fidelity of Faraday outputs.

Fidelity refers to the statistical similarity between synthetic and real data. This can be done quantitatively by comparing the statistical metrics (such as mean and quantile values), or qualitatively by comparing the distributions visually via t-stochastic neighbour embedding (TSNE) or principal component analysis (PCA) plots (Yoon et al., 2019). Figures 2 and 4a show high fidelity of Faraday outputs except at the 5<sup>th</sup> quantile where synthetic data is clipped to a minimum of zero. More tuning could be done to improve performance at the 5<sup>th</sup> quantile but comes at the expense of performance at the 95<sup>th</sup> quantile. Due to grid modellers’ interest in studying peak loads, performance at

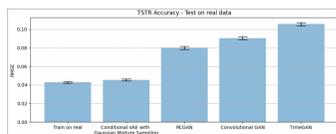
higher quantiles is prioritised over performance at lower quantiles. Figures 3a and 4b shows the load profiles at various quantiles of households with or without electric vehicles whilst figure 4b shows individual samples of load profiles.

### 3.2 UTILITY

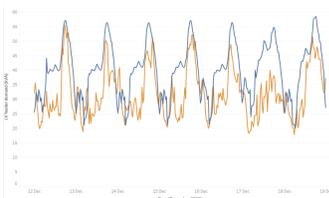
Utility looks at how useful generated samples are in real life applications. In the RCGAN paper, Esteban et al. (2017) proposes the “Train on Synthetic, Test on Real” (TSTR) framework where two competing models for an evaluation task are trained: one on synthetic data and one on real data. The intuition is that if the synthetic data is useful, then a model trained on synthetic data should perform as well as the model trained with real data.

For this utility task, we used real 2021 consumption data to generate synthetic consumption profiles for 2021. Two forecasting models are then trained, one on real 2021 data and one on synthetic 2021 data, to predict 2022 consumption. Figure 5a shows the performance of Faraday outputs is similar to the forecasting model trained on real data. We also compared the utility of Faraday outputs against other popular GAN-based models and show that Faraday outputs have higher utility in this forecasting task.

Faraday outputs were also compared against real-world data measured at substations. Day & Wilson (2023) used Faraday outputs in a digital twin project in Birmingham, UK, and compared Faraday outputs to substation measured data and found strong alignment between the two. This was done by collecting property information of houses served by the substation, inputting this information into Faraday to produce household-level profiles and aggregating the profiles to the substation level. Results show that Faraday outputs have similar peaks in terms of magnitude and time, but have higher ‘base load’. This could be due to a skew in population between this dataset and the average UK population (e.g. in terms of income and demographics etc).



(a) TSTR performance.



(b) Faraday compared to substation data. Blue is Faraday outputs. Orange are substation measured data.

Figure 5: Utility of Faraday outputs.

### 3.3 PRIVACY

Privacy refers to the risk of leaking private data because of the model overfitting on the training data. By design, Faraday samples from a distribution to generate synthetic samples during inference and its outputs are therefore ‘synthetic’. However, there could still be a risk of the distribution being overfitted and thus leaking private data, especially in the case of outliers (van Breugel et al., 2023). Some implicit measures have been implemented to mitigate privacy risks, such as:

1. Applying a k-anonymity of 3 such that at the finest granularity of dimensions there are still at least three households.
2. Exposing Faraday only in a partial black-box setting; users submit inputs via an API to generate outputs. No outputs are generated during sampling if the households matching users’ inputs represent a small proportion of the training data.
3. Only outputting daily profiles (as opposed to weekly or monthly profiles) which limits re-identification risk.

There is a trade-off between privacy and utility. Because of the inherent privacy risks such as the potential re-identification of individuals, Faraday only outputs daily profiles - hence limiting the

utility of its data to ‘intraday’ analysis at household level. However, outputs are still useful for inter-day purposes on an aggregated or population level as seen from the comparison to substation outputs in figure 5b.

## 4 FUTURE WORK AND CONCLUSION

Privacy is a central concern when it comes to sharing smart meter data as it contains highly sensitive information. Existing literature on generating synthetic smart meter data places emphasis on fidelity and utility metrics, but there is limited commentary on the privacy of synthetic smart meter data. Whilst Faraday has implicit measures implemented to guard against privacy risks, more work should be undertaken to explicitly quantify the privacy risks of generating synthetic smart meter data. This could include implementing differential privacy (Abadi et al., 2016), or explicit evaluation tasks such as membership inference or reconstruction attacks. Doing so could give us the confidence to output time series of longer horizons, such as weekly or monthly, which would have even higher utility for grid research.

### MODEL AVAILABILITY

As Faraday is trained on proprietary data, the model is only available as a partial black-box access via a web-app and an API. The model is still in closed alpha phase, but we welcome requests for research purposes. For access to Faraday, please contact [faraday@centrefornetzero.org](mailto:faraday@centrefornetzero.org) with your use case.

## REFERENCES

- Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS ’16*, pp. 308–318, New York, NY, USA, October 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL <https://doi.org/10.1145/2976749.2978318>.
- Aurora Cobo Aguilera, Pablo M. Olmos, Antonio Artés-Rodríguez, and Fernando Pérez-Cruz. Regularizing transformers with deep probabilistic layers. *Neural Networks*, 161:565–574, April 2023. ISSN 0893-6080. doi: 10.1016/j.neunet.2023.01.032. URL <https://www.sciencedirect.com/science/article/pii/S0893608023000448>.
- Cheesecake Energy Ltd. Grid constraints: The roadblock slowing electrification progress. <https://cheesecakeenergy.com/2022/04/06/grid-constraints/>, 2022. (Accessed on January 25, 2024).
- Zhiyi Chen, Ali Moradi Amani, Xinghuo Yu, and Mahdi Jalili. Control and Optimisation of Power Grids Using Smart Meter Data: A Review. *Sensors*, 23(4):2118, January 2023. ISSN 1424-8220. doi: 10.3390/s23042118. URL <https://www.mdpi.com/1424-8220/23/4/2118>.
- Nikolaos Damianakis, Gautham Ram Chandra Mouli, Pavol Bauer, and Yunhe Yu. Assessing the grid impact of Electric Vehicles, Heat Pumps & PV generation in Dutch LV distribution grids. *Applied Energy*, 352:121878, 2023. ISSN 0306-2619. doi: 10.1016/j.apenergy.2023.121878. URL <https://www.sciencedirect.com/science/article/pii/S0306261923012424>.
- Joseph Day and Grant Wilson. Evaluation of different methods to allocate buildings to energy networks p. 15-26. <https://www.birmingham.ac.uk/documents/college-eps/iidsai/teed-digitalisation-final-report-dec-23.pdf>, 2023.
- S. Elam, E. Webborn, J. Few, E. McKenna, M. Pullinger, T. Oreszczyn, B. Anderson, Communities Ministry Of Housing, European Centre For Medium-Range Weather Forecasts, and Royal Mail Group Limited. SERLSmart Energy Research Lab Observatory Data, 2019-2022: Secure Access, 2023. URL <https://beta.ukdataservice.ac.uk/datacatalogue/doi/?id=8666#6>.

- Energy Systems Catapult. Data for Good, Smart Meter Data Access. <https://es.catapult.org.uk/report/data-for-good-smart-meter-data-access/>, 2023. (Accessed on January 25, 2024).
- Cristóbal Esteban, Stephanie Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans, 2017. URL <https://arxiv.org/abs/1706.02633>.
- Yuxuan Gu, Qixin Chen, Kai Liu, Le Xie, and Chongqing Kang. GAN-based Model for Residential Load Generation Considering Typical Consumption Patterns. In *2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, February 2019. doi: 10.1109/ISGT.2019.8791575. URL <https://ieeexplore.ieee.org/document/8791575>. ISSN: 2472-8152.
- James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic Data – what, why and how?, May 2022. URL <http://arxiv.org/abs/2205.03257>. arXiv:2205.03257 [cs].
- Xinyu Liang and Hao Wang. Synthesis of realistic load data: adversarial networks for learning and generating residential load patterns. In *NeurIPS 2022 Workshop: Tackling Climate Change with Machine Learning*, pp. 1–8. Neural Information Processing Systems (NIPS), 2022. URL <https://research.monash.edu/en/publications/synthesis-of-realistic-load-data-adversarial-networks-for-learnin>.
- National Grid. Smart meter strategy (2021). <https://www.nationalgrid.co.uk/smarter-networks/smart-meter-data>, 2021. (Accessed on January 25, 2024).
- Office of Gas and Electricity Markets Ofgem. Electric vehicles: Ofgem’s priorities for a green fair future. <https://www.ofgem.gov.uk/publications/electric-vehicles-ofgems-priorities-green-fair-future>, 2021. (Accessed on January 25, 2024).
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL [https://papers.nips.cc/paper\\_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html](https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html).
- Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership Inference Attacks against Synthetic Data through Overfitting Detection, February 2023. URL <http://arxiv.org/abs/2302.12580>. arXiv:2302.12580 [cs].
- Evgenia Xilas. Britain wastes enough wind generation to power 1 million homes. <https://carbontracker.org/britain-wastes-enough-wind-generation-to-power-1-million-homes/>, 2023. (Accessed on January 25, 2024).
- Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://papers.nips.cc/paper\\_files/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html](https://papers.nips.cc/paper_files/paper/2019/hash/c9efe5f26cd17ba6216bbe2a7d26d490-Abstract.html).
- Chi Zhang, Sanmukh Kuppannagari, Rajgopal Kannan, and Viktor K. Prasanna. Generative Adversarial Network for Synthetic Time Series Data Generation in Smart Grids. Technical Report EE0008003-6, Univ. of Southern California, Los Angeles, CA (United States), October 2019. URL <https://www.osti.gov/biblio/1607585>.
- Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders, May 2018. URL <http://arxiv.org/abs/1706.02262>. arXiv:1706.02262 [cs, stat].