Synthetic profile generation for energy system planning using Artificial Intelligence

Master's Thesis

for the degree of

Master of Science (M.Sc.)

Data Science

at the Faculty of Natural Sciences
Friedrich-Alexander-Universität Erlangen-Nürnberg

submitted on April 21, 2026

by **Aniruddha Maiti**

Matriculation: 23195586

Supervisors: Prof. Dr. Enrique Zuazua, Dr. Zhengping Ji

With funding from the:


Federal Ministry
of Research, Technology
and Space

Acknowledgments

I would like to thank everyone who supported me to complete this Masters thesis.

I am grateful to M.Eng. Kerstin Oetringer. Thank you for the time you dedicated and the effort you invested; your advice and suggestions were valuable in bringing this project to a conclusion.

Many thanks to Dr. Zhengping Ji and Prof. Dr. DhC. Enrique Zuazua for helping and supporting me throughout the process of writing this thesis manuscript.

Special thanks to Fraunhofer IEG and Federal Ministry of Research, Technology and Space for helping me set up the foundation of my career, providing it a direction and supporting me financially.

This Masters thesis and degree program could not have been completed without your support. Ganz Herzlichen Dank!

Contents

Abstract

Introduction

Rapid urbanization and population growth have led to a significant increase in energy consumption, especially in densely populated metropolitan areas. Simultaneously, the European Union and Germany have set ambitious goals to achieve carbon neutrality by the year 2045, demanding a large-scale transition to renewable energy systems. In this context, accurate modeling of demand profiles of consumers in a fine-grained level becomes an essential and fundamental step for efficient urban planning, grid stability and resource allocation. Increasing variability in both demand and supply further emphasizes the need for detailed and reliable consumption profiles that reflect real-world energy consumption patterns.

Objectives

This study aims to address the lack of high-resolution energy demand modeling by developing detailed consumption profiles for residential buildings. The primary objective is to generate realistic synthetic demand data using Generative Artificial Intelligence, specifically Variational Autoencoders, to imitate consumer behavior under varying conditions. Additionally, the study evaluates the similarity between real and synthetic data using statistical and distribution-based metrics, including Mean, Standard Deviation, Entropy, and Maximum Mean Discrepancy. A further objective is to analyze the Simultaneity Factor, which quantifies the relationship between peak demand and aggregated consumption, and to assess whether synthetic data preserves this characteristic. As a secondary task classical time-series forecasting methods are explored to predict short-term demand and evaluate their suitability for such applications.

Methodology

The study utilizes building-level datasets containing electricity and heat consumption measured at fixed temporal resolutions. Data pre-processing steps include cleaning, normalization, and alignment with external variables such as temperature. To capture the stochastic nature of energy demand, Variational Autoencoders are trained on conditioned inputs, enabling the generation of synthetic consumption profiles that reflect underlying temporal and statistical patterns.

Model evaluation is conducted using a combination of statistical metrics and visual analysis. Kernel Density Estimation plots are employed to compare probability distributions, while temporal segments on weekly basis are analyzed to assess structural similarity and support visual inspection. The Maximum Mean Discrepancy is used to quantify differences between real and generated data distributions, and entropy measures are used to capture variability, randomness and uncertainty in the datasets. In addition, the simultaneity factor is computed for both real and synthetic datasets to evaluate the preservation of aggregated demand characteristics.

For the secondary objective, Autoregressive Integrated Moving Average models are implemented to forecast short-term demand. Model selection is guided by autocorrelation analysis and stationarity transformations, and performance is assessed through comparison with actual observed values.

Results

The Variational Autoencoder-based approach successfully generates synthetic demand profiles that closely match the statistical properties of real consumption data. Distributional similarity is confirmed through Kernel density distribution analysis, where synthetic data aligns well with real data across different consumption ranges. Temporal evaluation further demonstrates that the generated profiles capture recurring patterns and variability observed in daily and weekly demand cycles.

Quantitative metrics support these findings, with comparable mean and standard deviation values between real and synthetic datasets, and low Maximum mean discrepancy scores. Entropy-based analysis indicates that the generated data retains a similar level of randomness and uncertainty as the real dataset. Importantly, the simultaneity factor computed for synthetic data closely approximates that of real data, suggesting that aggregated demand behavior is effectively preserved.

In contrast, the ARIMA-based forecasting approach exhibits limited performance. While the model captures short-term correlations, it tends to produce smoothed predictions with reduced variability and systematic underestimation of peak demand.

Conclusion

This study demonstrates that generative models, specifically Variational Autoencoders, provide a robust framework for generating realistic synthetic energy demand profiles at high temporal resolution. The ability to preserve both statistical characteristics and aggregated demand behavior highlights their potential for applications in energy system planning and simulation.

At the same time, the limitations of classical time-series models such as ARIMA emphasize the challenges associated with modeling complex, high-frequency energy data using linear approaches. These findings underline the importance of adopting advanced, data-driven methodologies for accurate demand representation. Overall, the integration of generative modeling techniques offers a promising direction for addressing data scarcity and improving decision-making in modern energy systems.

Abbreviations

VAE	Variational Autoencoder
ARIMA	AutoRegressive Integrated Moving Average
SF	Simultaneity Factor
MMD	Maximum Mean Discrepancy
DS	Data Science
ML	Machine Learning
GAN	Generative Adverserial Networks
AI	Artificial Intelligence
ODH	Open District Hub
KDE	Kernel Density Estimation
CVAE	Conditional Variational Autoencoder
TSGM	Time Series Generative Modeling

Chapter 1

Introduction

The rapid expansion of cities combined with continuous population growth has placed increasing pressure on existing energy systems. At the same time, European countries, particularly Germany, are pursuing a long-term transition toward climate neutrality, aiming to significantly reduce greenhouse gas emissions by 2050 (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen). This shift requires not only cleaner energy sources but also a much more precise understanding of how energy is consumed within urban environments.

Traditional aggregated consumption models are often insufficient to capture the variability and complexity of modern energy usage. Daily fluctuations, seasonal effects, and behavioral differences across households and buildings create highly dynamic demand patterns that must be represented at a minute resolution. Developing accurate and realistic demand profiles is therefore a key requirement for improving grid reliability, supporting renewable integration, and enabling data-driven urban energy planning.

This Master's thesis contributes to the field of neighborhood planning. Neighborhood planning and Artificial Intelligence together form an interdisciplinary field (Yue et al., 2025) known as Geospatial Artificial Intelligence (GeoAI) or Intelligent Urban Informatics that seeks to shape and improve the existing neighborhood plan of a locality or region. This field merges data science, machine learning, and geographic information systems to analyze urban patterns, optimize land use, and create predictive models for sustainable, human-centered city development. It consists of a wide range of activities that are important for sustainability, growth, and quality of life (Yue et al., 2025). It involves efficient land use and zoning, defining safe public transportation networks, sewage and wastewater treatment systems, energy-efficient building standards and utilities, social infrastructure, and recreational spaces (Ouchra et al., 2023).

Keys aspects of this field (Cina et al., 2025) include:

- **Pattern Recognition & Prediction:** AI helps identify hidden patterns in urban data, enabling planners to forecast growth, land value, and infrastructure needs.
- **Generative Planning:** Tools such as GANs are used to simulate and optimize

complex urban configurations.

- **Smart City Optimization:** AI technologies manage urban services, enhance sustainability, and improve quality of life.
- **Human-Centered Design:** The focus is on using AI to improve, rather than replace, human planning, ensuring equitable and sustainable neighborhoods.

This convergence enables the creation of *digital twins* (Bibri et al., 2024) of cities, allowing simulation of urban changes before they are implemented.

All activities mentioned above play a vital role when it comes to defining a good neighborhood plan. Each and every city has a unique plan according to its demographic factors, geographical location, environmental factors, land availability, infrastructure utilities, social and cultural lifestyle of citizens, and government priorities.

Citizens of any country, city, town, or village consume several kinds of energy such as electricity for their electronic gadgets and appliances, heating energy for water, space, and cooking. Momentarily, European Union has introduced plans to build sustainable energy-efficient cities, which would help authorities to optimize energy generation and supply and build climate-friendly neighborhoods (European Commission). Energy-efficient buildings can protect the climate, reduce carbon emissions, predict and forecast energy demand optimally; both consumers and producers can financially benefit from this plan. (Baumeister, 2025)

Germany has committed to achieving climate neutrality by 2045 via transforming their present Kommunale Wärmeplanung. (Municipal heat planning) (Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen) A considerable portion of heat demand in Germany is directly dependent on the use of fossil fuels such as oil and gas. The federal government has introduced laws and regulations for municipalities to develop comprehensive Kommulane wärmeplanung. The plans serve as fundamental tools to develop sustainable and optimal heat supply networks, assess the local energy demands, and predict them as accurately as possible (Fraunhofer Institute for Applied Information Technology FIT, 2023).

Precise modeling of heat and electricity of all kinds of buildings in the neighborhood plays an important role in this context. Municipalities require detailed, accurate, and reliable data of consumption patterns of residential, industrial, and commercial buildings in order to integrate decentralized technologies such as heat pumps, district heating, and building renovations.

Considering the growth of population, boom in immigration, and globalization have increased and changed energy requirements in various cities and regions across Germany and Europe (Gans, 2018). Present Kommunal Wärmeplanungs are momentarily accomodating the requirements of the consumers, but they would become inefficient in the long run. Emerging technologies in Artificial Intelligence, particularly Generative Artificial Intelligence, could obtain realistic and precise demand predictions of consumers in a locality. The ability to accurately reconstruct demand prediction from a granular level

to a regional level can assist municipalities in taking informed decisions that are both economically feasible and environmentally sustainable (Deutsche Energie-Agentur (dena), 2024).

1.1 Keys problems in neighborhood planning

- **Strong interaction between components and complexity**: Modern neighborhoods are no longer single-energy systems. Electricity, heating, and cooling are increasingly interconnected, creating nonlinear feedback loops across sectors. For instance, electric vehicles increase electricity demand (Kachirayil et al., 2022)

  These interactions significantly increase the complexity of planning and operation, as they require the consideration of external environmental factors such as weather, snow, wind-speed, humidity and nonlinear system dynamics .

- **Multi-scale networks**:

  A fundamental challenge in neighborhood energy planning is the coupling between multiple spatial scales. Decisions made at the building level directly influence district-level and even city-level energy outcomes as shown in figure 3.6. Energy systems set up in a single building will affect the buildings sharing the same heat and electricity; other systems, such as sewage networks, can affect an entire block of buildings, which may consist of other commercial buildings as well.

- **Uncertainty in demand, behavior, and climate**: Energy system planning is inherently affected by multiple sources of uncertainty. These include variability in occupant behavior, which significantly influences consumption patterns, fluctuations in weather conditions affecting heating and cooling demand, and uncertainty in future technology costs and policy developments. (Leprince et al., 2023)

  Among these, occupant behavior has been identified as one of the most influential sources of uncertainty, often exceeding the impact of technical parameters in determining optimal system design. This highlights the need for stochastic and data-driven approaches that can explicitly account for uncertainty in both demand and system operation.

The increasing interconnection of energy sectors such as electricity, heating, and mobility represents a central challenge in transitioning into renewable sources of energy. The Open District Hub at Jülich addresses this challenge by developing integrated, data-driven planning and simulation tools for urban districts. Its main objective is to enable sector-coupled energy systems that support the transition towards carbon-neutral neighbourhoods by jointly modelling and optimizing cross-sector interactions. Unlike traditional planning approaches that treat energy domains separately, ODH at Jülich focuses on holistic district-level representations, where renewable energy sources can be efficiently distributed across multiple end-use sectors.

A key contribution of the project is the development of a digital planning environment that integrates technical, economic, and operational aspects of urban energy systems. This allows planners and decision-makers to evaluate long-term investment strategies and operational scenarios under realistic, multi-energy system constraints. In this context, ODH can be interpreted as a foundation for future digital twin approaches in urban energy systems, where complex inter-dependencies between infrastructure components are explicitly simulated.

However, the strong connection between sectors introduces significant modeling complexity, as interaction between electricity demand and heating systems patterns must be accurately represented. This highlights the need for advanced computational methods and data-driven approaches capable of capturing such nonlinear dependencies in consumption behavior. In this regard, AI models such as generative models and machine learning-based forecasting methods can complement part of the objective of ODH by improving pattern recognition, demand prediction, and scenario generation. Consequently, the integration of AI-based models with platform-based planning tools such as ODH represents a promising direction for the development of intelligent, scalable, and holistic neighborhood energy systems. (ODH@Jülich, 2025)

This Master's thesis is focused on synthetic profile generation and contributes to the new Kommunal Wärmeplanung by developing synthetic data for accurately modeling demand profiles on several temporal resolutions: hour, minute, and second. Several municipalities do not have access to high-resolution data on a granular regional level. This lack of data hinders the creation of realistic demand profiles (Nilashi et al.), captures seasonal patterns, detects anomalies, assesses maximum and minimum load behavior, and integrates flexible renewable energy technologies.

The synthetic data generated with the assistance of Generative AI provides the municipalities with realistic demand profiles and a privacy-preserving alternative to real measured data. The AI models have been optimized and trained to mimic consumption behavior (Madanchian, 2024) with consideration of the locality's climatic conditions and seasonal patterns . Furthermore, these AI models can reveal the frequency of usage of appliances or electronic gadgets, thereby enabling a further detailed description of actual electricity consumption in a residential building.

1.2 Problem statement and research gap

Existing approaches pose a fundamental limitation in neighborhood energy planning as they often rely on physical simulation models which are computation expensive, difficult to scale, they can ignore spatial proximity leading to a trade-off between model fidelity and computational efficiency. Furthermore, they may ignore occupant behavior, external environment factors. (Kleiminger and Beckel, 2016)

To address these challenges, this thesis positions itself at the intersection of energy system modeling and data-driven machine learning approaches. In particular, generative models such as Variational Autoencoders (Pan, 2019) offer a promising direction for learning

realistic consumption patterns and producing synthetic load profiles that preserve statistical and temporal characteristics of real-world energy data. By integrating such models into neighborhood energy planning frameworks, it becomes possible to improve scenario generation, enhance uncertainty representation, and support more robust optimization of multi-energy systems (Bolluk et al., 2025).

1.3 Objectives

Accurate modeling of energy demand at the neighborhood level requires high-resolution and complete consumption data. However, in practical applications, such data is often unavailable due to privacy regulations, incomplete measurement infrastructure, or technical limitations in data collection systems. In particular, regulations such as the General Data Protection Regulation (GDPR) restrict access to detailed consumption data, while sensor failures and communication issues frequently lead to missing values in time-series datasets. These challenges limit the applicability of conventional data-driven methods for energy system planning, forecasting, and optimization.

To address these limitations, this thesis attempts to achieve the following objectives:

1. Synthetic profile generation
2. Reconstruction of missing values

Synthetic profile generation focuses on creating artificial yet realistic energy consumption data that preserves the statistical and temporal properties of real-world measurements. (Giudice et al., 2025) Instead of relying solely on limited or sensitive datasets, generative models such as Variational Autoencoders are employed to learn the underlying distribution of observed consumption patterns. These models enable the generation of diverse consumption scenarios (Fu et al., 2024). To enhance realism, the generated profiles are conditioned on relevant contextual features, for example weather features of the city, allowing the model to capture seasonal effects and behavioral patterns (Wang et al., 2022).

The resulting synthetic datasets provide a scalable and privacy-preserving alternative to real measurements (Kotal and Joshi, 2022). They can be used to simulate energy demand across different types of buildings, support demand estimation for buildings and districts, and enable scenario-based analysis for renewable energy integration. In the context of neighborhood energy planning, such synthetic profiles are particularly useful, as they allow planners to evaluate system performance under varying demand conditions without requiring extensive real-world data collection.

In parallel **handling missing values** in time-series data is secondary objective to ensure data integrity and model reliability. Energy datasets are especially sensitive to missing entries, as even small gaps can distort temporal dependencies and impact negatively. Models such as ARIMA and SARIMA models are trained to reconstruct missing values. (Nassir et al., 2018)

By combining synthetic data generation with missing value reconstruction, this work establishes a reliable data foundation for subsequent energy system modeling tasks. This integrated approach directly addresses key challenges in neighborhood energy planning, namely data scarcity, uncertainty, and incomplete observations, and supports the development of more realistic and scalable demand models.

Chapter 2

Methodology

This chapter presents the fundamental concepts and methodological framework underlying synthetic profile generation and the application of Generative Artificial Intelligence in this Thesis.

2.1 Generative Artificial Intelligence

Generative Artificial Intelligence is a subfield of Artificial Intelligence that focuses on the development of models capable of generating new data samples that resemble a given dataset. These models learn the underlying statistical distributions, patterns, and structure of the input data and utilize this learned representation to produce realistic synthetic outputs, such as text, images, audio, and time-series data (Banh and Strobel, 2023)

In contrast, discriminative models are designed to distinguish between different classes or predict target variables based on input features. While discriminative approaches focus on learning decision boundaries, generative models aim to approximate the joint probability distribution of the data, enabling them to recreate data instances that are statistically similar to the original dataset Nguyen et al. (2023). This fundamental difference makes generative models particularly suitable for tasks involving data synthesis and simulation.

Generative models have been widely adopted across multiple domains, including finance, healthcare, marketing, and computer vision, where the ability to generate realistic synthetic data is of significant practical value (Harries et al., 2025). In recent years, their application has expanded into the field of energy systems, where data availability and privacy concerns often limit the use of high-resolution consumption data (Böcking et al., 2024). Some common generative modeling approaches include, Generative Adversarial Networks, Autoencoder, Diffusion models (Gargary and De Cristofaro, 2024).

In the context of energy demand modeling, Generative Artificial Intelligence enables the creation of synthetic electricity and heat load profiles that preserve the statistical and temporal characteristics of real consumption data. This capability is particularly important for urban energy planning, where access to fine-grained building-level data is often

restricted (Wang and Hong, 2020). By generating realistic demand scenarios, generative models support simulation, forecasting, and optimization tasks while maintaining data privacy and addressing data scarcity challenges.

2.2 Generative AI approaches

2.2.1 Variational Autoencoders

Recent advances in Generative Artificial Intelligence have enabled data-driven synthesis of realistic electricity load profiles. Among these methods, Variational Autoencoders and conditions in the forms of vectors have gained attention due to their probabilistic formulation and ability to learn latent representations of complex consumption patterns.

A Variational Autoencoder is an extension of Autoencoder, which is a type of unsupervised neural network that learns to efficiently compress (encode) data into a lower-dimensional representation (latent space) and then reconstruct (decode) it back to its original form, aiming to minimize the reconstruction error (Kingma and Welling, 2019). The advantage of a Variational Autoencoder over the vanilla Autoencoder is that a Variational Autoencoder learns a distribution (mean and variance) in the latent space, enabling smooth interpolation and generation of new data points.

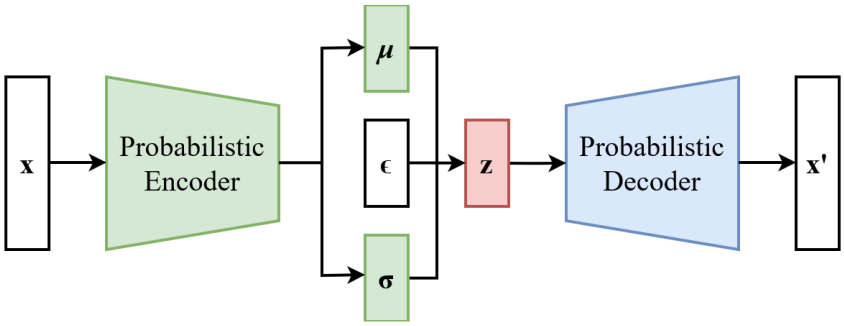

Figure 2.1: Variational Autoencoder. Source: (Kingma and Welling, 2019).

In the context of load profile generation, VAEs learn the underlying distribution of historical electricity consumption time series and can generate new synthetic profiles by sampling from the learned latent dimension. This allows for the creation of diverse yet statistically consistent demand profiles without duplication of real user data, which is particularly valuable in privacy-sensitive energy applications.

An important extension of this neural network is the Conditional Variational Autoencoder (CVAE), which incorporates contextual variables in the form of vectors. Conditioning enables the generative model to produce load profiles that are consistent with specific external characteristics (Harvey et al., 2021), such as seasonality and building features.

2.3 Simultaneity Factor

In energy system planning, accurately estimating demand is a fundamental requirement for developing reliable and efficient design of electrical infrastructure. Meanwhile, aggregate energy consumption of individual buildings can be determined with relative ease, the temporal misalignment of individual demand introduces complexity. This temporal overlap of consumption is commonly referred to as *Simultaneity*, it plays a decisive role in determining peak loads and system stress. To capture this effect, the concept of **Simultaneity factor** is widely used in electrical engineering and energy demand modeling (Rehau AG, 2012)

The *Simultaneity factor* describes the degree to which multiple consumers draw power at the same time. It is defined as the ratio between the maximum coincident demand of a group of consumers and the sum of their individual maximum demands (Winter et al., 2001).

$$f_{\text{sim}} = \frac{P_{\text{max, group}}}{\sum_{i=1}^{N} P_{\text{max},i}} \tag{2.1}$$

A *Simultaneity factor* close to 1 indicates that many consumers reach their peak demand simultaneously, while lower values reflect more dispersed or staggered consumption patterns. The concept is particularly important when scaling demand from individual units to a neighborhood, district, or entire urban areas (Dickert and Schegner).

In practice, the simultaneity factor enables planners and researchers to avoid systematic overestimation of peak demand that would arise from assuming full coincidence of individual maxima. Especially in residential energy systems, where consumption behavior is strongly influenced by human activity patterns, weather conditions, and appliance usage, simultaneity is inherently limited. Consequently, incorporating simultaneity factors leads to more realistic demand profiles and more cost-effective infrastructure sizing, such as transformer capacities, distribution lines, and storage systems (Gust et al., 2024).

In this thesis, the *simultaneity factor* aims to capture how simultaneity evolves with aggregation level, time resolution and user behavior. This perspective allows for a more flexible and data-driven representation of essential demand coincidence for designing and optimization of future energy systems.

2.4 Architectures

The following subsection explains the architecture of VAE, ARIMA models thoroughly.

2.4.1 Variational Autoencoders

A variational autoencoder is an artificial neural network architecture introduced in 2013 . It is an extension of the vanilla Autoencoder that learns a smooth, probabilistic latent space by compressing and reconstructing data to produce entirely new data. VAEs capture the underlying structure of a dataset to produce outputs that closely resemble the original data (Doersch, 2016).

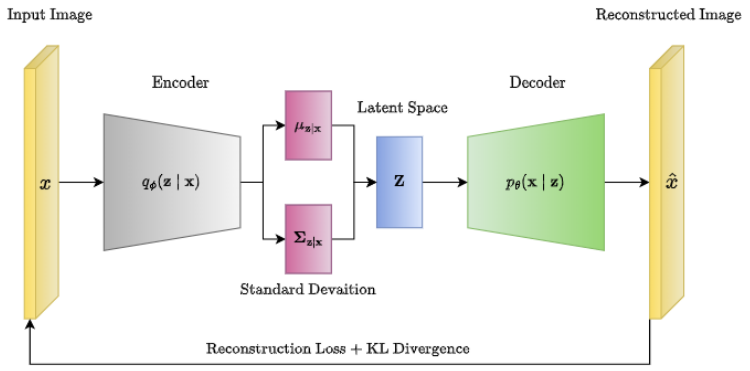

Figure 2.2: Architecture of a Variational Autoencoder. Source: (Kingma and Welling, 2019)

A VAE consists of two neural networks:

- **Encoder:** Maps input data $\mathbf{x}$ to the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ of the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$.
- **Decoder:** Maps latent samples $\mathbf{z}$ back to the data space via $p_\theta(\mathbf{x} \mid \mathbf{z})$.

Generative Modeling Framework

Let $\mathbf{x} \in \mathbb{R}^d$ denote an observed data sample and $\mathbf{z} \in \mathbb{R}^k$ a latent variable drawn from a prior distribution $p(\mathbf{z})$, typically chosen as a standard multivariate Gaussian:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.2}$$

The generative process is defined as:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) \tag{2.3}$$

where $p_\theta(\mathbf{x} \mid \mathbf{z})$ represents the likelihood modeled by a neural network parameterized by θ.

VAE Loss Functions

The training of a Variational Autoencoder is based on maximizing the Evidence Lower Bound (ELBO) on the marginal log-likelihood of the observed data. Equivalently, this can

be interpreted as minimizing a loss function composed of two terms: a reconstruction loss and a regularization loss.

(i) **Reconstruction Loss**

The reconstruction loss measures how well the decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ can reconstruct the original data $\mathbf{x}$ from the latent code $\mathbf{Z}$ sampled from the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$:

$$\mathscr{L}_{\text{rec}} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] \tag{2.4}$$

This term encourages the VAE to produce output that are close to the input data.

(ii) **Regularization Loss (KL Divergence)**

The regularization term enforces that the approximate posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ remains close to the prior distribution $p(\mathbf{z})$, typically a standard multivariate Gaussian:

$$\mathscr{L}_{\text{KL}} = \text{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z})}\right] \tag{2.5}$$

This term regularizes the latent space, promoting smoothness and enabling meaningful sampling.

Total VAE Loss

The total loss function for training the VAE combines the reconstruction and regularization components:

$$\mathscr{L}_{\text{VAE}} = \mathscr{L}_{\text{rec}} + \mathscr{L}_{\text{KL}} = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] + \text{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big) \tag{2.6}$$

Minimizing $\mathscr{L}_{\text{VAE}}$ corresponds to maximizing the ELBO, thereby simultaneously encouraging accurate reconstruction of data (Pinheiro Cinelli et al., 2021).

2.4.2 Statistical Time Series Models: ARIMA and SARIMA

ARIMA model and Seasonal ARIMA are used for forecasting time series data and analyzing sequential data. In the context of energy demand modeling, they are widely adopted and provide a baseline for capturing linear dependencies and seasonality patterns.

Let y_t denote a univariate time series. After differencing d times, the ARIMA model can be expressed as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{2.7}$$

Autoregressive Integrated Moving Average (ARIMA)

The components of ARIMA allow the model to capture patterns such as trends and behavior over a period of time helping to predict future values based on historical data (Hyndman and Athanasopoulos, 2018). It combines three key components to model data:

- p is the order of the autogressive (AR) component,
- d is the degree of differencing applied to achieve stationarity,
- q is the order of the moving average (MA) component.

(i) **Autoregression (AR):**

The Autogressive part (AR) of an ARIMA model is represented by the parameter p. It signifies the dependence of the current observation on its previous values. Mathematically, an $AR(p)$ model can be represented as follows:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_p Y_{t-p} + \varepsilon_t \tag{2.8}$$

Here:

- Y_t is the current observation,
- c is a constant,
- ϕ_1 and ϕ_2 are the autoregressive parameters.
- ε_t represents the error term at time t.

(ii) **Differencing (I):**

The differencing part of ARIMA is represented by the parameter d. It involves transforming a non-stationary time series into a stationary one by differencing consecutive observations . We can apply the differencing operation multiple times until stationarity is achieved. The formula for differencing is as follows:

$$Y_t' = Y_t - Y_{t-1} \tag{2.9}$$

Here:

- Y_t' is the differenced series at time t,
- Y_t is the original series at time t,
- Y_{t-1} is the value of the series at the previous time step.

The differencing process is typically applied multiple times until stationarity is achieved. The notation $I(d)$ indicates the order of differencing required for stationarity.

(iii) **Moving Average (MA):**

The moving average part (MA) of an ARIMA model is represented by the parameter q. It indicates the dependence of the current observation on the previous forecast errors. Mathematically, an $MA(q)$ model can be represented as follows:

$$Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.10}$$

Here:

- Y_t is the current observation,
- c is a constant,
- ε_t is the error at time t,
- θ_1 to θ_q are the moving average parameters.

Seasonal ARIMA

While ARIMA models are effective in capturing short term temporal dependencies, they are limited in representing repetitive seasonal patterns commonly observed in energy demand data. To address this limitation, the Seasonal ARIMA (SARIMA) model extends ARIMA by explicitly modeling seasonal autoregressive and moving average components.

A SARIMA model is denoted as:

$$\text{SARIMA}(p, d, q)(P, D, Q)_s$$

Here:

- (p, d, q) are the non-seasonal ARIMA parameters,
- (P, D, Q) are the seasonal autoregressive, differencing, and moving average orders,
- s is the length of the seasonal cycle,
- p is the Autoregressive order,
- d is the number of non-seasonal differences,
- q is the moving average order,
- P is the seasonal autoregressive order,
- D is the seasonal differencing order,
- Q is the seasonal moving average order.

Before applying SARIMA, seasonal differencing is often required to make the data stationary. The process involves subtracting the current observation from one that corresponds to the same season in the previous cycle. Seasonal differencing helps remove the seasonal patterns from the data, enabling more accurate forecasting.

The SARIMA model is expressed mathematically as follows:

$$\Phi_P(B^s)\,\phi_p(B)\,(1-B)^d(1-B^s)^D Y_t = \Theta_Q(B^s)\,\theta_q(B)\,\varepsilon_t \tag{2.11}$$

Here:

- y_t: the observed time series at time t,
- B: the backshift (lag) operator,
- ϕ_1: the non-seasonal autoregressive coefficient,
- Φ_1: the seasonal autoregressive coefficient,
- θ_1: the non-seasonal moving average coefficient,
- Θ_1: the seasonal moving average coefficient,
- s: the seasonal period,
- ε_t: the white noise error term.

2.5 Time Series Generative Modeling

Time Series Generative Modeling is an open-source framework (Nikitin et al., 2024) for synthetic time series dataset generation and evaluation. It builds on open-source libraries and implements various methods, such as GANs, VAEs, for synthetic time series simulation. Moreover, TSGM provides many approaches for evaluating synthetic time series data.

It is designed to facilitate the development, training, and evaluation of generative models for sequential data. Unlike conventional machine learning libraries that primarily target static tabular data, TSGM explicitly addresses the challenges associated with time series generation, such as temporal dependencies, seasonality, and stochastic variability.

The library provides a unified interface for implementing a range of generative modeling approaches, including both probabilistic latent-variable models and adversarial learning frameworks. By abstracting common preprocessing steps, training routines, and evaluation procedures, TSGM enables consistent comparison across different generative architectures while reducing implementation complexity.

In the context of energy systems, where electricity demand exhibits strong temporal structure across multiple time scales, TSGM offers a structured environment for synthetic load profile generation.

2.5.1 Supported Generative Models

TSGM supports multiple classes of generative models for specifically modeling time series data. These include VAEs and CVAEs, GANs, that combine adversarial training. Several models implemented in the library are specifically adapted for sequential data, incorporating recurrent or 1-dimensional convolutional components to capture both short and long term dependencies (Nikitin et al., 2023).

The following list contains the generative and disrminative models provided by the TSGM library:

- **Generative models:**
  - `vae_conv5`: Convolutional Variational Autoencoder with five convolutional layers
  - `cvae_conv5`: Conditional Variational Autoencoder with five convolutional layers
  - `cgan_base_c4_l1`: Conditional GAN with a four-layer convolutional architecture
  - `t-cgan_c4`: Temporal Conditional GAN with convolutional layers
  - `cgan_lstm_n`: Conditional GAN incorporating LSTM-based temporal modeling
  - `cgan_lstm_3`: Hybrid Conditional GAN combining LSTM and convolutional layers
  - `wavegan`: WaveGAN architecture adapted for time series generation
  - `ddpm_denoiser`: Denoising diffusion probabilistic model (DDPM) with convolutional denoiser
- **Downstream and evaluation models:**
  - `recurrent`: Basic recurrent neural network architecture
  - `clf_transformer`: Transformer-based classification architecture

The modular design of TSGM allows different model architectures to be trained with specific conditions desired by the user. This is particularly important for evaluating trade-offs between model stability, sample diversity, and temporal structure when generating synthetic time series. Furthermore, the framework includes built-in utilities for scaling, windowing, and reshaping time series data, which are essential preprocessing steps for generative learning.

2.5.2 Use case for synthetic load profile generation

Synthetic electricity load profile generation presents several challenges, including non-stationarity, strong daily and seasonal patterns, and the presence of rare but critical peak

demand events. Generative models must therefore balance statistical fidelity with temporal structure, consumer behavior and robustness.

TSGM is well-suited for this task as it enables the application of advanced generative architectures to time series data while maintaining a consistent modeling pipeline. In this thesis, TSGM is used as the primary framework for implementing and evaluating generative models for synthetic electricity and heat demand profiles. The library facilitates systematic experimentation with different architectures and supports the generation of realistic load profiles that preserve key statistical and temporal characteristics of the original dataset.

By leveraging TSGM, this work focuses on analyzing model behavior and output quality rather than low-level implementation details, thereby improving reproducibility and methodological clarity.

2.6 Evaluation Metrics

To assess the quality of the generated synthetic energy demand profiles, several statistical and distribution-based metrics were employed. These metrics quantify the similarity between real and synthetic data in terms of central tendency, randomness and distributional characteristics.

Mean

The mean represents the average value of the dataset and provides a measure of central tendency. Hurley and Tenny (2023) It is defined as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.12}$$

where x_i denotes individual observations and N is the total number of samples. Comparing the mean values of real and synthetic datasets helps evaluate whether the overall level of energy consumption is preserved. Wang and Hong (2020)

Standard Deviation

The standard deviation measures the dispersion of data around the mean and reflects the variability of the time series. (El Omda and Sergent, 2024)

Standard deviation is formulated as follows:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \tag{2.13}$$

A close match in standard deviation between real and synthetic datasets would indicate that the generated profiles capture inherent fluctuations in energy demand. (**?**)

Maximum Mean Discrepancy (MMD)

Maximum mean discrepancy is a non-parametric metric used to compare the distributions of two datasets (Johnson, 2020). It measures the distance between the mean embeddings of the real and synthetic samples in a reproducing Kernel Hilbert Space (RKHS). The MMD is defined as follows:

$$\mathrm{MMD}^2(X,Y) = \mathbb{E}x, x' \sim X[k(x,x')] + \mathbb{E}y, y' \sim Y[k(y,y')] - 2\mathbb{E}_{x\sim X, y\sim Y}[k(x,y)] \tag{2.14}$$

where X and Y represent the real and synthetic datasets, respectively. Lower MMD values indicate a higher similarity between the two distributions.

Kernel Selection and Gamma parameter for MMD

In the computation of MMD, the choice of kernel function plays a crucial role. In this thesis, the Gaussian (Radial Basis function) kernel is used, which is defined as follows:

$$k(x,y) = \exp\left(-\gamma|x-y|^2\right) \tag{2.15}$$

where, γ is the kernel bandwidth parameter that controls the sensitivity of the kernel to differences between samples.

A commonly used approach for selecting an appropriate value of γ is based on the median heuristic. In this method, γ is defined as follows:

$$\gamma = \frac{1}{2 \cdot \mathrm{median}\left(|x_i x_j|^2\right)} \tag{2.16}$$

where $|x_i x_j|^2$ represents the pairwise squared distances between samples in the dataset. This heuristic adapts the kernel width to the scale of the data and provides a robust choice without requiring manual tuning (Gretton et al., 2012).

Alternatively, γ can be expressed in terms of the standard deviation σ of the data:

$$\gamma = \frac{1}{2\sigma^2} \tag{2.17}$$

In practice, selecting an appropriate γ ensures that the kernel captures meaningful similarities between real and synthetic samples, thereby increasing the reliability of the MMD-based.

Differential Entropy

Differential entropy is used to quantify the uncertainty or randomness of continuous data distributions. Unlike discrete entropy, it is defined for continuous variables and captures the spread and unpredictability of the data (Liu et al., 2021). It is expressed as:

$$H(X) = -\int p(x)\log p(x)dx \tag{2.18}$$

where $p(x)$ denotes the probability density function of the variable X. In this study, differential entropy, differential entropy is used to compare the variability and complexity of real and synthetic energy demand profiles. Similar entropy scores suggest that the generated data preserves the level of uncertainty present in the original dataset.

Overall, these metrics provide complementary insights into the statistical fidelity of the generated data, enabling a comprehensive evaluation of both distributional similarity and temporal variability.

Chapter 3

Implementation

The development of generative models for synthetic profile generation follows a structured workflow, encompassing data exploration, pre-processing, model development, and evaluation. This chapter describes the practical implementation steps undertaken in this thesis, including data preparation, model configuration, and training procedures. Each step plays an important role to generate reliable synthetic demand profiles.

The following models were implemented for the generation and analysis of synthetic electricity and heat demand data:

1. Conditional Variational Autoencoder (CVAE)
2. Autoregressive Integrated Moving Average (ARIMA)

In order to build the Generative models, an open source framework called 'Time series generative modeling' (TSGM) (Nikitin et al., 2024) has been utilized to generate synthetic time series data. As TSGM library provides a comprehensive collection of pre-built generative models, enabling efficient development and experimentation without the need to implement models from scratch.

The framework allows users to import, configure, and customize model architectures based on specific requirements. In this thesis, TSGM was employed to implement CVAE for synthetic profile generation, offering flexibility in hyperparameter tuning and model setup.

3.1 Overview of workflow

The implementation process follows a structured workflow consisting of data preparation, model development, and evaluation. Initially, the datasets are cleaned, transformed, and aligned with relevant external features such as temperature. Subsequently, generative models, particularly CVAEs, are configured and trained to learn the underlying patterns of energy demand. In parallel, ARIMA models are implemented to predict missing energy demand values in a dataset. Finally, the generated synthetic data is evaluated against real

data using statistical metrics and visual analysis to assess the similarity between real and the newly generated synthetic energy demand profiles.

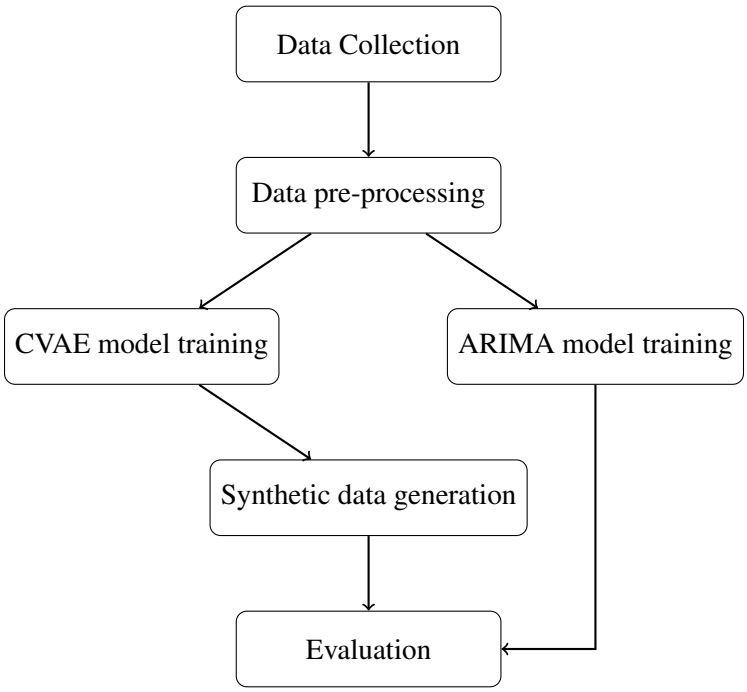

Figure 3.1: Workflow of the implementation process

3.2 Dataset description

This section presents the datasets used in this study for modeling and evaluation. The work is based on two main sources of energy consumption data: electricity demand and heat demand. Both datasets contain time-series measurements that reflect real-world consumption patterns over time. In addition, relevant external features, such as weather information, are incorporated where applicable to improve model performance and enable conditional learning.

Data source	Energy type	Sampling rate	Region
MFRED	Electricity	15 minutes	Manhattan, New York, USA
Residential Energy – Denmark (RED-DK)	Heat	1 Hour	Aalborg, Denmark
TABULA	Heat	1 hour	Germany

Additionally, weather data have been obtained to make better predictions for the RED-DK dataset.

- **Geographical coordinates:** 57.04°N, 9.93°E

The weather dataset for all the datasets mentioned above have been obtained from the following API:

- **Data source:** Open-Meteo Historical Weather API (`https://open-meteo.com/en/docs/historical-weather-api`)

The following subsections provide a detailed overview of each dataset, including their structure, temporal resolution, and key characteristics relevant to the proposed modeling approach.

3.2.1 MFRED dataset

The Multifamily Residential Electricity Dataset (MFRED) Meinrenken et al. (2020) is a high-resolution real-world energy dataset designed to support research in residential electricity consumption, demand modeling. It contains detailed measurements of real and reactive power collected from 390 apartments located in multifamily residential buildings in Manhattan, New York. The dataset spans a full year of observations and provides measurements at an hourly temporal resolution, making it suitable for the objective of the thesis: Generating synthetic profile generation.

MFRED spans across a large number of residential units, being divided into 26 clusters of buildings, with 15 apartments within each cluster summing up to 390 apartments. In addition to active power consumption, it also includes reactive power measurements, enabling deeper analysis of electrical load characteristics beyond simple energy usage. To preserve privacy, individual apartment data are aggregated into groups of 15 units following established utility data disclosure standards.

The real demand profile is shown together with the corresponding temperature data to provide additional context regarding its dependence on weather conditions.

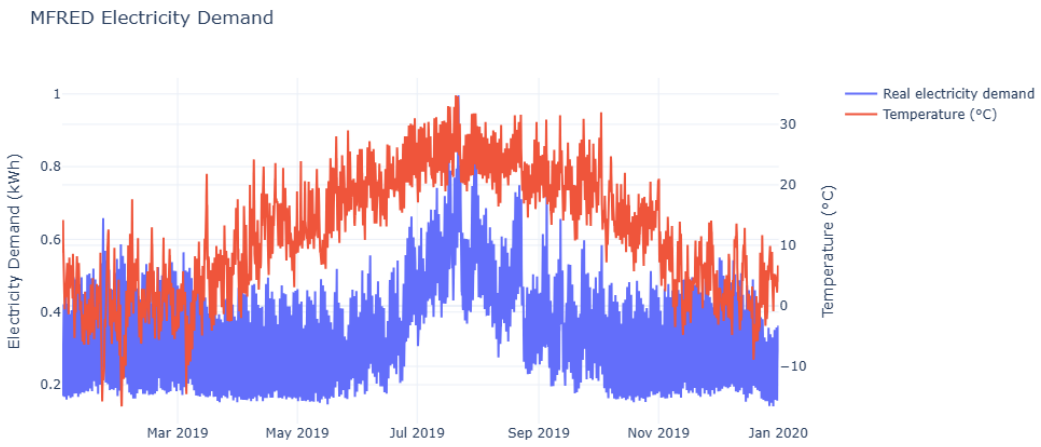

Figure 3.2: Real electricity demand profile of MFRED dataset

Simultaneity Factor Score of the MFRED Dataset

The SF score was computed for both the real and synthetic MFRED datasets. The results indicate a high level of coincidence in peak demand, with the synthetic data slightly overestimating this effect.

Dataset	SF Score
Real dataset	0.85
Synthetic dataset	0.89

Table 3.1: Comparison of simultaneity factor scores for real and synthetic MFRED data

Limitations in Calculating Cluster-Level Simultaneity Factors

While the simultaneity factor was successfully computed for the aggregated dataset, it is not possible to calculate the simultaneity factor for all 26 clusters within the MFRED dataset. This limitation arises due to the privacy-preserving aggregation applied to the dataset.

The MFRED dataset follows a *15/15 privacy rule*, which ensures that electricity consumption data cannot be traced back to individual households. Under this rule, the demand profiles of fifteen apartments from each cluster are aggregated into a single combined load profile before being released. As a result, the dataset only provides aggregated demand data for each building group rather than the individual demand of each apartment.

To calculate the simultaneity factor of a particular cluster, the peak demand of each individual apartment within that cluster must be known. According to the definition of the simultaneity factor presented in Equation 3.4, the denominator requires the sum of the individual non-coincident peak demands of all apartments in the group.

For a cluster consisting of fifteen apartments, this quantity would be expressed as:

$$P_{\text{sum\_peaks}} = \sum_{i=1}^{15} P_{\text{max},i} \tag{3.1}$$

where:

- $P_{\text{max},i}$ represents the maximum demand of the i-th apartment,
- i denotes the apartment index,
- 15 represents the total number of apartments within the cluster.

This expression represents the theoretical sum of the individual peak demands that would occur if each apartment reached its maximum demand independently. The simultaneity factor then compares this quantity with the coincident peak demand observed for the aggregated load profile.

However, due to the privacy aggregation applied in the MFRED dataset, the individual peak demands $P_{\max,i}$ are not available. Instead, the dataset only provides the combined load profile of the fifteen apartments. Consequently, the individual peaks of the apartments cannot be reconstructed from the aggregated signal.

This means that only the coincident peak demand of the aggregated group is observable, while the true denominator of the simultaneity factor equation remains unknown. Without access to the individual smart meter data, it is therefore impossible to compute the internal simultaneity factor for the apartments within a cluster.

As a result, the simultaneity factor analysis presented in this study is limited to the aggregated demand profiles available in the dataset. While this still provides valuable insight into the overall synchronization behavior of electricity demand across building groups, it does not allow for the evaluation of simultaneity effects at the individual household level.

3.2.2 TABULA dataset

The TABULA project (Typology Approach for Building Stock Energy Assessment) provides a standardized European building stock dataset developed to support energy efficiency analysis and policy development. It defines representative residential building typologies across multiple European countries, including Germany, based on construction period, building size, and renovation status. Each typology includes detailed information on building envelope characteristics such as insulation levels, window types, and heating systems (Loga et al., 2016).

Unlike measured consumption datasets, TABULA has simulated datasets that provides typical energy demand values for space heating, domestic hot water, and overall primary energy use. These standardized profiles enable comparative analysis of energy performance across different building categories and support the generation of synthetic demand profiles.

Simulated dataset

Simulated data refers to artificially generated information created through computer simulations, mathematical models, or algorithms, rather than collected from real-world events. It is designed to mimic the statistical patterns, relationships, and characteristics of real-world systems, making it highly useful for testing scenarios, training AI models, and protecting privacy. (ScienceDirect)

In the context of the TABULA dataset, building energy demand profiles are derived from standardized building typologies, physical properties (e.g., insulation levels, floor area), and usage patterns. These inputs are processed through engineering-based or statistical models to produce synthetic energy consumption profiles that approximate realistic behavior.

The simulated TABULA dataset was divided into two representative clusters based on building characteristics such as age, type, refurbishment status, and energy label.

For clarity and consistency throughout the remainder of this thesis, Cluster 4203 is referred to as Cluster A, while Cluster 4195 is referred to as Cluster B.

Cluster	Construction year	Energy label	Refurbishment status	Type
Cluster 4203	2015	D	Ambitious	Terraced
Cluster 4195	2010-2015	C	Standard	Terraced

Table 3.2: Summary of TABULA clusters used in the analysis

The real demand profile is shown together with the corresponding temperature data to provide additional context regarding its dependence on weather conditions.

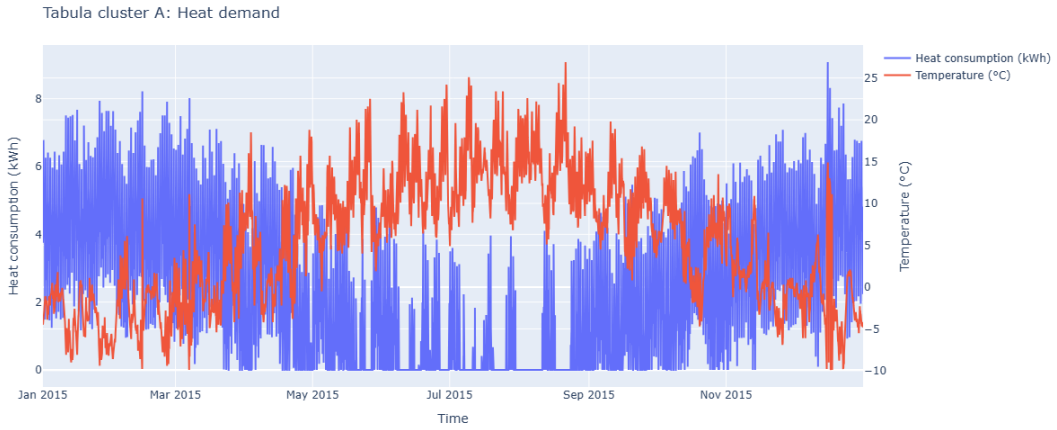

Figure 3.3: Cluster A heat demand

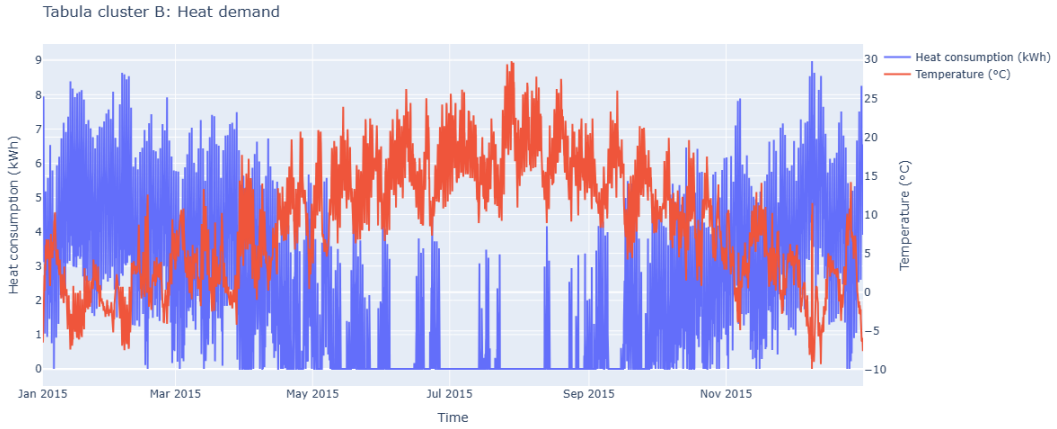

Figure 3.4: Cluster B heat demand

3.2.3 RED-DK dataset

This dataset consists of multiple building types with varying construction years and characteristics. However, due to the presence of frequent missing values across the time series, the dataset was primarily utilized for the purpose of missing value prediction rather than full-scale modeling.

For this task, a subset of the data was selected, covering the period from August 1 to August 12, with hourly resolution. Based on this observed data, missing value prediction was performed for a subsequent period of three days. This setup allows for the evaluation of forecasting performance by comparing predicted values with known observations.

The characteristics of the selected building used for this analysis are summarized below:

Table 3.3: Building characteristics used for missing value prediction

Attribute	Value
Building ID	300
Construction year	2003
Energy label	D
Building type	Apartment

This selection provides a representative case for evaluating the effectiveness of forecasting models under conditions of incomplete data.

3.3 Data pre-processing

This section describes the preprocessing steps applied to the datasets used in this thesis. Since the data sources differ in structure and origin, separate pre-processing pipelines were developed for the MFRED dataset and the TABULA dataset, they have been discussed separately.

3.3.1 Data cleaning

MFRED dataset

The MFRED dataset obtained from the open-source repository was cleaned and reformatted to ensure consistency in time-series structure and compatibility with the modelling pipeline. The final processed format is shown in Table 3.4.

DateTime (UTC)	Time Resolution	Energy (kWh)
2021-01-01 00:00:00	1 hour	1.20

Table 3.4: Structure of the MFRED dataset after preprocessing.

The energy column represented a cumulative energy consumption in kilowatt-hours (kWh). To obtain hourly demand values, a differencing operation was applied to convert cumulative readings into non-cumulative consumption values per hour. The first value in the resulting series was back-filled using the second observation to avoid introducing artificial zero values at the start of the sequence.

The dataset contained a small number of missing values, primarily occurring between 09 July 2019, 14:30 and 21:30 UTC, due to temporary meter outages. These missing segments

were removed as they accounted for only a negligible proportion of the overall dataset and did not significantly affect temporal patterns.

TABULA dataset

The TABULA dataset consists of a large collection of residential building typologies across Europe, covering a wide range of construction periods, building types, and renovation states (International Organization for Standardization, 2017). The dataset includes buildings ranging from newly constructed units (post-2015) to historical structures built as early as 1860. Building types include apartments, multi-family houses, terraced houses, and other residential categories with varying refurbishment statuses (Loga et al., 2016).

Due to the large number of buildings and associated data points, it was not computationally feasible to process the entire dataset simultaneously on a standard 8 GB memory system. Therefore, the dataset was processed on one building at a time, each being treated as an independent time series for model training and evaluation.

The dataset follows the structure shown in Table 3.5.

DateTime (UTC)	**Time Resolution**	**Energy (Wh)**
2021-01-01 00:00:00	1 hour	52000

Table 3.5: Structure of the TABULA dataset after preprocessing.

Energy values in the TABULA dataset were originally provided in watt-hours (Wh). These were converted into kilowatt-hours (kWh) by dividing by 1000 to ensure consistency with the MFRED dataset. The resulting values were then used for model training and evaluation.

3.3.2 Cluster dendogram

A cluster dendrogram was used to analyze and visualize similarities among the demand profiles within the MFRED dataset. Specifically, it was applied to identify hierarchical relationships between the 26 consumer groups based on their electricity consumption patterns. This approach enables the grouping of similar load profiles without requiring prior labeling, making it suitable for exploratory analysis of residential energy behavior.

The dendrogram is constructed using hierarchical clustering, where individual consumer groups are iteratively merged based on their similarity in demand characteristics. The resulting tree structure provides a clear visual representation of how closely related different clusters are, allowing the identification of distinct consumption patterns within the dataset.

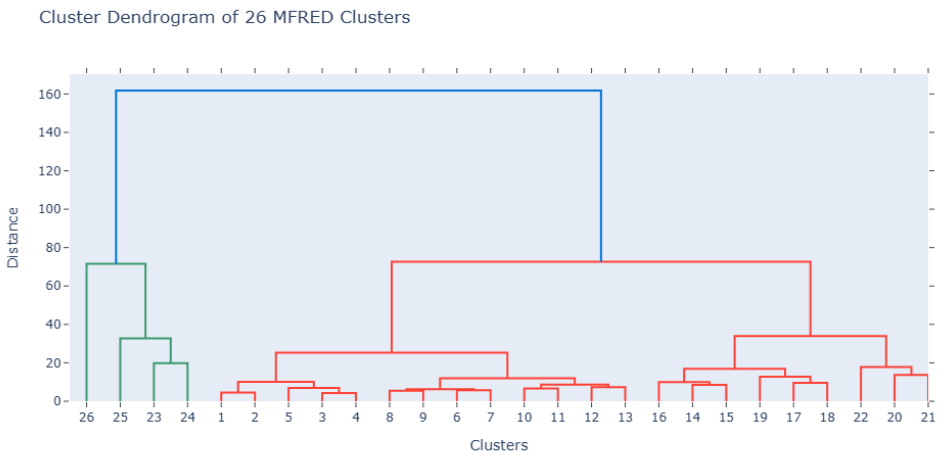

Figure 3.5: Hierarchical clustering dendrogram of the MFRED dataset showing similarity relationships among 26 consumer groups based on their demand profiles.

3.4 CVAE implementation

The Conditional Variational Autoencoder (CVAE) was implemented using the `tsgm` library, which provides pre-built architectures for time-series generative modeling. Rather than implementing the model from scratch, the focus of this thesis was on adapting the input pipeline, defining appropriate training configurations, and designing an effective conditioning strategy for energy demand generation.

3.4.1 Training procedure

The training pipeline required restructuring the raw time-series data into the format expected by the `tsgm` framework (Nikitin et al., 2024). Specifically, the data was transformed into fixed-length sequences suitable for processing. Prior to training, all input features were normalized using the pre-built functions of the library to ensure numerical stability and to improve convergence during optimization.

Different learning rates were tested within a predefined range to identify stable training behavior and minimize reconstruction loss.

3.4.2 Optimization algorithms and hyper parameters

Multiple optimization algorithms were evaluated during training, including Stochastic Gradient Descent (SGD), and AdamW. These were selected due to their widespread use in deep generative models and their ability to handle non-stationary gradient behavior in time-series data.

The following table lists all the hyper parameters that have been used to train the cVAE models:

Parameter	Range / Values
Model	cBetaVAE
Optimizer	SGD, AdamW
Latent Dimension	8 – 16
Learning Rate	0.001 – 0.005
Beta (β)	0.1 – 1.0
Batch Size	32
Epochs	5

Table 3.6: Hyperparameter space used for cBetaVAE training using TSGM

3.4.3 Conditioning strategy

A key component of the CVAE design was the conditioning mechanism used to incorporate external contextual information. In this thesis, outdoor temperature was used as the conditioning variable, as it has a strong influence on energy demand patterns, particularly in heating-related consumption.

The temperature signal was merged with the energy consumption values of the dataset before training, allowing the model to learn conditional dependencies between weather conditions and energy usage. This enabled the generation of synthetic demand profiles that are not only statistically similar to the original data but also responsive to external environmental conditions.

3.5 ARIMA implementation

The ARIMA modelling approach was used as a classical statistical baseline for short-term energy demand forecasting. The implementation followed three main stages: stationarity analysis, model configuration, and training/validation.

3.5.1 Stationarity check

The available dataset was first analyzed to assess its suitability for ARIMA modeling. Since 12 days of data were available, a forecasting horizon of 3 days was selected. To determine whether the time series satisfied the stationarity assumption required by ARIMA, the Augmented Dickey-Fuller (ADF) test was applied. The results indicated that the raw series was non-stationary. Consequently, differencing was applied to remove trend components and stabilize the mean. Additionally, autocorrelation (ACF) and partial autocorrelation (PACF) plots were used to analyze temporal dependencies and guide parameter selection.

Weekly and daily aggregation plots were also examined to identify repeating seasonal patterns in the data.

3.5.2 Model configuration

Based on the stationarity analysis and correlation plots, ARIMA and SARIMA models were configured. The parameters (p, d, q) were selected using insights from the ACF and PACF behavior. In addition, different configurations were tested, including $(1, 1, 1)$ and $(1, 1, 72)$, where the latter was introduced to capture longer temporal dependencies. However, the selection of large lag values was heuristic and motivated by observed periodic patterns rather than strict statistical optimization.

3.5.3 Training and evaluation

The models were trained on subsets of the available data and evaluated using both direct forecasting and back-testing strategies. In the first experiment, the model was trained on approximately 15 days of data to predict the subsequent 3 days. In a second validation setup, 9 days of historical data were used to forecast the next 3 days, which were already available for comparison.

Across both configurations, the ARIMA and SARIMA models produced overly smoothed predictions that failed to capture peak demand behaviour, often converging towards near-constant or flattened output curves. Even with adjusted parameter settings such as $(1, 1, 72)$, the models consistently underperformed in representing variability in the data and tended to underestimate peak values. This indicated limitations of classical linear models in capturing the complex, non-linear structure of the energy demand series.

3.6 Evaluation

The evaluation of the proposed models was conducted separately for the MFRED and TABULA datasets in order to assess the quality of the generated synthetic time-series data. The assessment focuses on both statistical similarity and distributional alignment between real and synthetic samples.

A combination of statistical metrics and visual analytics was performed to ensure a comprehensive evaluation. The statistical measures include the mean and standard deviation, differential entropy was used to quantify the uncertainty and complexity of the generated distributions. The MMD metric was further applied to measure the distance between real and synthetic data distributions in a kernel-based feature space. In order to compute the MMD score in a consistent and fair manner, the kernel bandwidth parameter γ was carefully selected. Since MMD is highly sensitive to the choice of kernel width, an inappropriate value can lead to either over-smoothing or under-separation of the distributions.

In this work, the optimal γ value was determined using a data-driven approach based on the median heuristic (Gretton et al., 2012). Specifically, the pairwise squared Euclidean distances between samples were computed, and γ was defined as:

$$\gamma = \frac{1}{2 \cdot \text{median}\left(\|x_i - x_j\|^2\right)} \tag{3.2}$$

where x_i and x_j represent samples drawn from the combined real and synthetic datasets. This choice ensures that the kernel is appropriately scaled to the intrinsic structure of the data.

The Gaussian kernel used in the MMD computation is then defined as:

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{3.3}$$

The resulting optimal γ values for the considered datasets are summarized below. The variation in γ values across datasets reflects differences in data scale and distribution.

Dataset	Optimal γ
MFRED	43.0
Cluster A	0.14
Cluster B	0.12

Table 3.7: Optimal γ values for different datasets

This adaptive selection of γ improves the robustness of the MMD metric and ensures a more reliable comparison between real and generated distributions.

Beyond numerical metrics, KDE plots were used to visually compare the distributional overlap between real and synthetic datasets. This provided an intuitive understanding of how well the generative models replicated the underlying data distribution.

Finally, a peak-load SF metric was introduced to evaluate the ability of the models to capture extreme demand events and peak consumption behavior, which is particularly important for energy system planning and reliability analysis.

3.7 Simultaneity factor

What is Simultaneity Factor

A simultaneity factor (or coincidence factor) is a multiplier, typically between 0.1 and 1.0, used in electrical and utility engineering to estimate the actual peak load of multiple devices by accounting for the fact that not all devices operate at full capacity simultaneously. It ensures that infrastructure is not oversize or overloaded. Mathematically, it is expressed as follows:

$$k_s = \frac{P_{\text{coincident}}}{\sum_{i=1}^{n} P_{i,\max}} \tag{3.4}$$

where:

- $P_{\text{coincident}}$ represents the coincident peak demand of the entire system,
- $P_{i,\max}$ represents the maximum demand of individual consumer i,
- n represents the total number of consumers.

The diversity factor is defined as the inverse of the simultaneity factor:

$$\text{Diversity Factor} = \frac{1}{k_s} \tag{3.5}$$

A lower simultaneity factor indicates greater diversity among individual peak loads, while a higher value suggests that peak demands occur more synchronously.

In energy system planning, SF is used to estimate the effective peak demand of an aggregated set of consumers. It plays a crucial role in the design and sizing of electrical infrastructure such as transformers, distribution lines, and district-level energy systems. By accounting for the fact that individual peak loads do not occur simultaneously, the SF prevents overestimation of system components while ensuring adequate supply reliability. In general, the SF ranges between 0 and 1, where values close to 1 indicate highly synchronized consumption patterns, typically observed in small or homogeneous systems, while lower values reflect increased diversity in load behavior, common in larger or more heterogeneous populations. However, the SF is influenced by temporal demand patterns, user behavior, and climatic conditions, and therefore represents an aggregated approximation rather than a precise physical constant.

3.7.1 Limitations of the Simultaneity Factor

While the SF provides a useful approximation of aggregated demand behavior, its application in real-world energy systems is associated with some practical limitations. These limitations become evident in use cases such as electric vehicle (EV) charging infrastructure (Silber et al.).

Key Limitations of the Simultaneity Factor

- **Risk of underestimation (Grid Stress):** An underestimated simultaneity factor may lead to insufficient system sizing, resulting in inadequate capacity to meet peak power demand and potentially causing grid instability.

- **User Restrictions:** Designing systems based on low simultaneity factors may require operational constraints, such as limiting electric vehicle charging power or delaying usage, in order to prevent system overloads.
- **Inaccuracy in Large-Scale Systems:** In large collective systems, such as district heating networks or residential clusters, the duration and temporal distribution of peak demand vary significantly. Representing such behavior using a single scaling factor may lead to inaccuracies, particularly in thermal or electrical system design applications (Winter et al., 2001).

3.7.2 Bottom-Up and Top-Down Modeling Capabilities

It is important to recognize that not all methods for generating energy demand profiles are equally suitable across different levels of aggregation. Depending on the intended application, a modeling approach may require either bottom-up or top-down capabilities, or a combination of both. These capabilities determine whether a method can accurately represent demand at individual building level or at aggregated system level (ODH@Jülich, 2025)

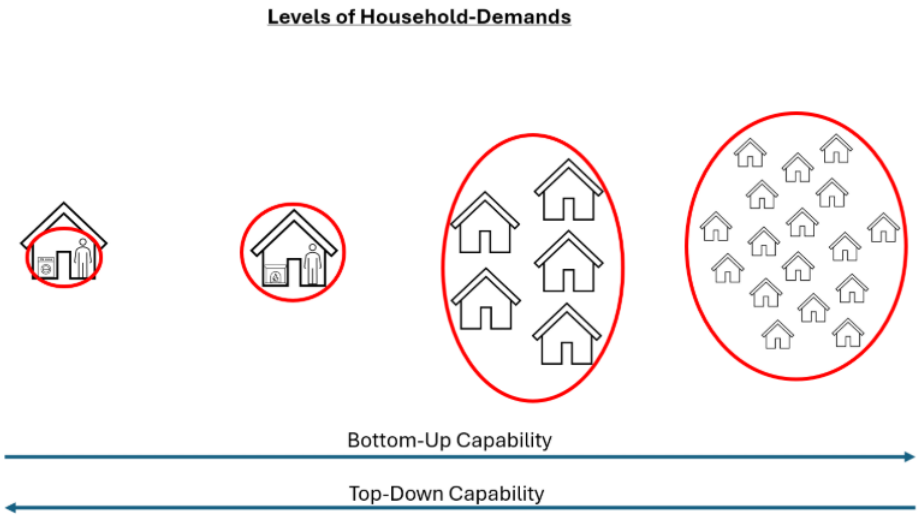

Figure 3.6: Bottom-up and top-down modeling capabilities across different spatial resolutions.

Bottom-up approaches focus on modeling individual entities, such as single households or buildings, and subsequently aggregating them to obtain system-level demand. These methods are particularly useful for capturing detailed consumption behavior, including variations due to occupancy, appliance usage, and operational patterns. However, a key limitation of bottom-up approaches is that the aggregation of individually generated profiles does not always guarantee realistic system-level characteristics, such as the expected simultaneity of peak demand.

In contrast, top-down approaches generate demand profiles based on aggregated data, often derived from large populations of consumers. One example is the Standard Load Profile (SLP) method, which is constructed using averaged consumption patterns from a large number of buildings. While such methods provide reliable estimates at the district or regional level, they lack the resolution required for modeling individual buildings or small groups of consumers.

In practical energy systems, demand does not occur simultaneously across all users, even during peak periods. This is due to variations in occupancy patterns. Therefore, an effective modeling approach should ideally balance both bottom-up and top-down characteristics to ensure realistic representation across different scales of aggregation.

Chapter 4

Results

This section presents the results and observations obtained from the Simultaneity factor and synthetic profiles generated from the CVAEs. The aim is to compare the synthetic profiles with the real profiles, evaluate, assess the performance of the implemented models and draw insights from the Simultaneity factors.

Furthermore, the results derived from the ARIMA-based forecasting models are also presented and analyzed. Visualizations and statistical measures are used to assess the similarity between real and generated data.

4.1 Analysis of MFRED synthetic demand profiles

In order to further evaluate the quality of the generated demand data, several graphical comparisons were performed between the real electricity demand profiles from the MFRED dataset and the synthetic profiles generated using the CVAE model.

Comparison of Real and Synthetic Load Profiles

To evaluate whether the synthetic profiles preserve the characteristic temporal patterns of electricity consumption, representative clusters were selected based on the hierarchical clustering structure. For each selected cluster, the real and synthetic demand profiles were plotted side by side for direct comparison.

The following figure presents a comparison between the real and synthetic electricity demand profiles for the entire neighborhood, consisting of all 26 clusters in the dataset.

For better interpretability, the comparison of both demand profiles are visualized as subplots, enabling a side-by-side comparison of real and synthetic datasets.

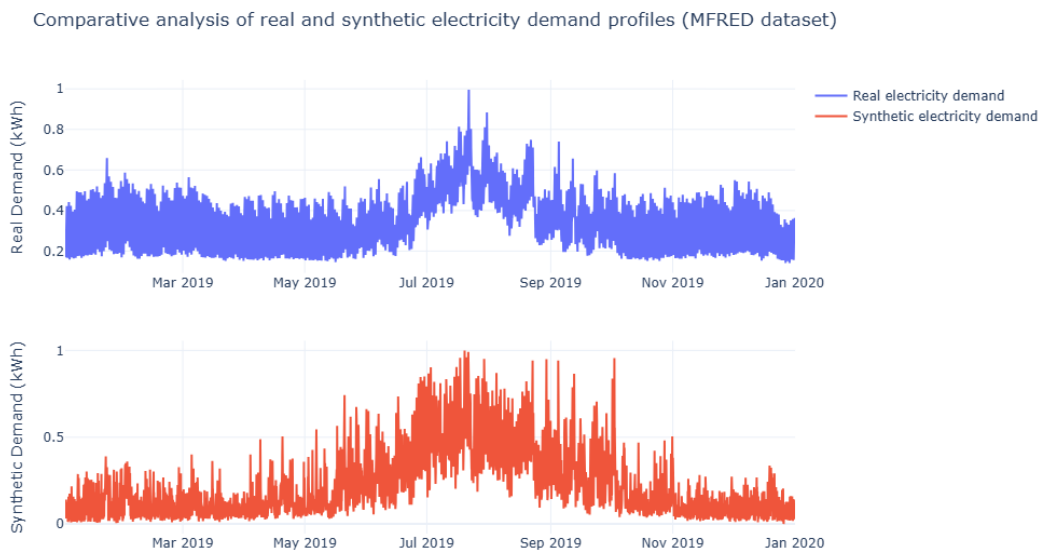

Figure 4.1: Real vs Synthetic electricity demand profile of MFRED dataset

4.1.1 Evaluation metrics for MFRED dataset

To quantitatively assess the similarity between real and synthetic electricity demand profiles, statistical and distribution based metrics have been computed. These metrics support the visual analysis of time-series plots by providing a mathematical comparison of key properties of the data.

The selected evaluation metrics include the mean, standard deviation, which capture the central tendency and variability of the demand profiles, respectively. In addition, entropy is used to quantify the information and randomness of the distributions. Finally, the MMD metric is applied as a kernel-based metric to measure the distance between the real and synthetic data distributions.

The following tables summarizes the computed evaluation metrics for both real and synthetic datasets.

Metric	Real Dataset	Synthetic Dataset
Mean	0.34	0.22
Standard Deviation	0.12	0.19
Entropy	-0.82	-0.62

Table 4.1: Comparison of evaluation metrics for real and synthetic MFRED datasets.

Dataset	MMD
MFRED Dataset	0.37

Table 4.2: MMD score

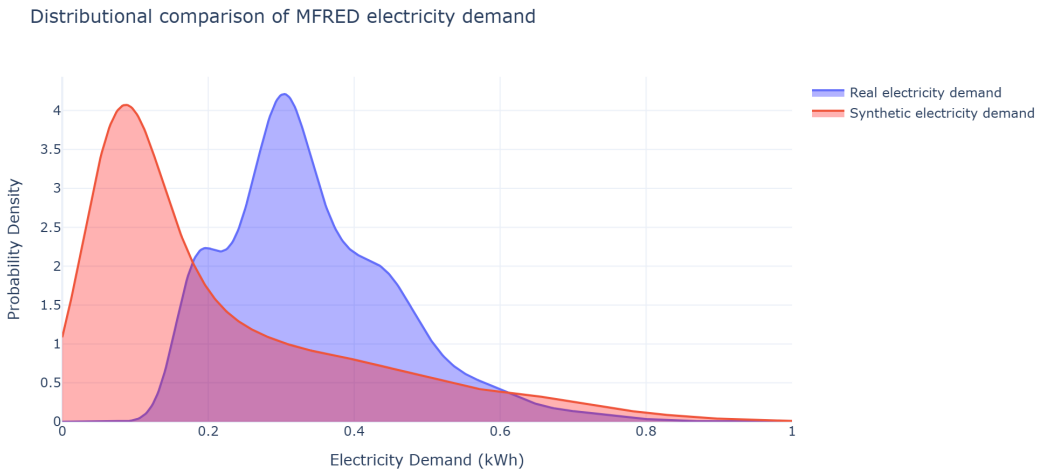

Figure 4.2: Distribution comparison of real and synthetic profiles.

4.1.2 Distributional analysis

The comparison of the mean and standard deviation values indicates that the synthetic dataset underestimates the average energy consumption relative to the real dataset, this suggests a slight leftward shift and lack of peak values in certain periods of time, which is also confirmed by the KDE plot above.

Despite the shift, the KDE analysis reveals that the synthetic data successfully captures the overall shape of the distribution, including the presence of peak consumption regions. Although the exact peak consumption magnitudes are not perfectly reproduced, the model retrains the general structure of high-consumption events, indicating that the model has understood the temporal dynamics and preserved the important distributional characteristics.

The standard deviation of the synthetic dataset is higher than that of the real dataset, indicating a broader spread of consumption values. This suggests that the VAE introduces additional variability into the generated data. This observation is consistent with the KDE plots, where the synthetic distribution spans across a wider range of values compared to the real dataset.

The differential entropy of the synthetic dataset is higher (less negative), indicating increased randomness and more dispersion compared to real dataset. This aligns with observations from the time-series behavior of the dataset. The real dataset exhibits smoother

and more structured transitions in electricity consumption, reflecting predictable usage patterns over time. In contrast, the synthetic data exhibits lesser fluctuations, resulting in a less stable signal.

This suggests that the VAE captures the global distribution, it does not fully reproduce the temporal consistency present in real consumption patterns.

4.1.3 Compute simultaneity factor

This analysis of the SF has been derived from the MFRED Manhattan electricity dataset and the synthetic demand profiles generated using CVAEs have been presented here. The objective of this analysis is to evaluate whether the synthetic data preserves the collective demand behavior observed in the real dataset, particularly with respect to peak demand coincidence across multiple buildings.

Simultaneity Factor of the Real Dataset

For the real MFRED electricity dataset, the following values were obtained:

- Sum of Individual Peaks: 30.57 kWh
- Coincident System Peak: 25.91 kWh

Using Equation 3.4, the simultaneity factor is calculated as:

$$k_s = \frac{25.91}{30.57} \approx 0.85 \tag{4.1}$$

The corresponding diversity factor is:

$$\frac{1}{k_s} = 1.18 \tag{4.2}$$

A simultaneity factor of 0.85 indicates that approximately 85% of the total individual peak demand occurs simultaneously. This suggests that although some diversity exists among individual demand profiles, a substantial portion of the peak loads overlap in time.

Simultaneity Factor of the Synthetic Dataset

The same analysis was performed on the synthetic demand profiles generated by the CVAE model. The results are summarized as follows:

- Sum of Individual Peaks: 17.76 kWh

- Coincident System Peak: 15.76 kWh

The simultaneity factor for the synthetic dataset is therefore:

$$k_s = \frac{15.76}{17.76} \approx 0.89 \tag{4.3}$$

The corresponding diversity factor is:

$$\frac{1}{k_s} = 1.13 \tag{4.4}$$

The calculated value indicates that approximately 89% of the individual peak demand occurs simultaneously in the synthetic dataset.

Comparison Between Real and Synthetic Data

A comparison of the simultaneity factors shows that the synthetic dataset reproduces the overall level of demand synchronization observed in the real data.

Dataset	Simultaneity Factor (k_s)	Diversity Factor
Real dataset	0.85	1.18
Synthetic dataset	0.89	1.13

The similarity between the two values suggests that the generative model has closely captured the collective demand behavior of the system. In particular, the synthetic data preserves the relative timing and occurrence of peak demand events.

However, it can also be observed that both the sum of individual peaks and the coincident peak in the synthetic dataset are approximately half of those observed in the real dataset.

Consequently:

- Individual maximum loads become smaller compared to the real dataset.
- The overall coincident system peak is also reduced.

Despite this reduction in absolute peak values, the ratio between the coincident peak and the sum of individual peaks remains similar. Since the simultaneity factor is defined as this ratio, the resulting SF values remain close to those observed in the real dataset.

Error Analysis

To further evaluate the difference between the real and synthetic simultaneity factors, the relative error was calculated.

- Real SF: 0.85
- Synthetic SF: 0.89
- Difference (ΔSF): +0.04

The relative error can be expressed as:

$$\text{Relative Error} = \frac{SF_{synthetic} - SF_{real}}{SF_{real}} \times 100 \tag{4.5}$$

$$\text{Relative Error} \approx 4.71\% \tag{4.6}$$

The positive difference indicates that the synthetic dataset slightly overestimates the simultaneity factor.

Interpretation

The observed overestimation suggests a mild increase in synchronization among the generated demand profiles. In other words, the model tends to align peak demand events slightly more than what is observed in the real dataset. This behavior may arise from the smoothing effect commonly associated with generative models such as CVAEs, where the latent representation introduces variability and fluctuations.

Nevertheless, the relatively small difference between the two simultaneity factors indicates that the synthetic data preserves the general simultaneity characteristics of the real electricity demand profiles. Therefore, the generated profiles can still be considered representative for analyzing aggregated demand behavior and system-level peak dynamics.

4.1.4 Visual analysis on a weekly basis

This subsection covers the visualization of the synthetic dataset on a weekly basis for better comprehension across all clusters. It is based on weekly frequency.

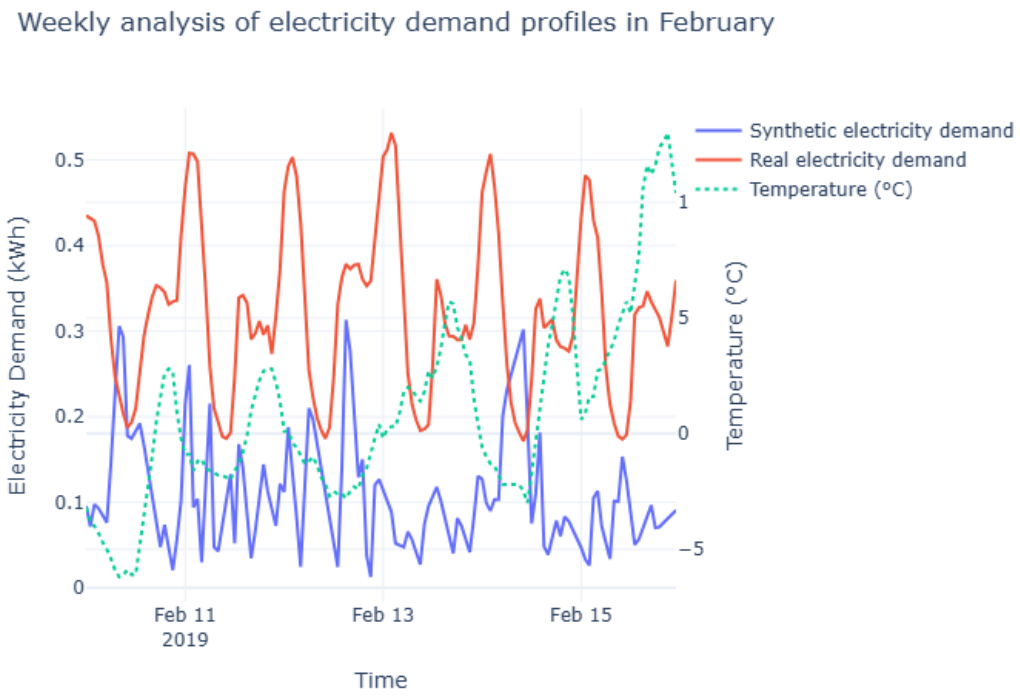

Figure 4.3: MFRED dataset electricity demand profile comparison across February.

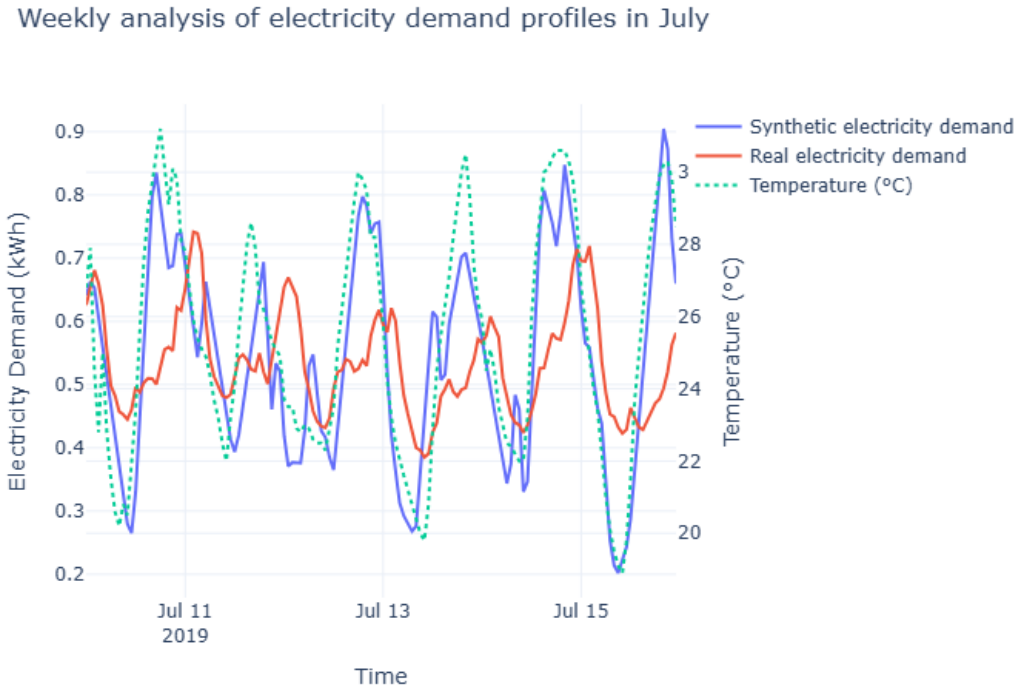

Figure 4.4: MFRED dataset electricity demand profile comparison across July.

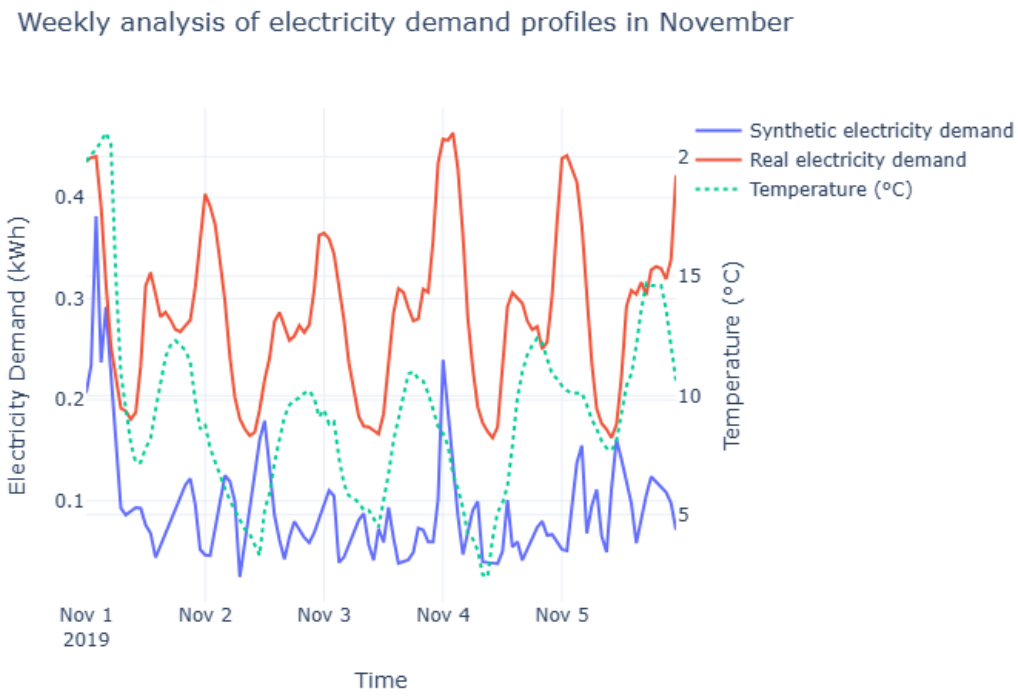

Figure 4.5: MFRED dataset electricity demand profile comparison across November.

Discussion and analysis

The plots indicate that the synthetic demand closely follows the overall temperature trend, suggesting that the model has successfully captured the general dependency between heating demand and temperature. However, it tends to underestimate, or sometimes overestimate peak loads, indicating a loss of extreme variability in the generated data. Overall, while the synthetic series preserves the global seasonal pattern, it smooths out local fluctuations compared to the real demand.

The following figures represent the demand profiles of three representative clusters. Each subplot displays the real profile on the top and the corresponding synthetic profile below them. This layout allows for a clear visual assessment of the generative model's ability to reproduce both the shape and relative timing of daily demand patterns.

4.1.5 Cluster level analysis

Three clusters were selected based on the dendrogram analysis shown in Figure 3.5, as they exhibit distinctly different distribution patterns compared to the remaining groups. The selected clusters are Cluster 9, Cluster 10, and Cluster 13, which represent the most structurally diverse segments within the dataset.

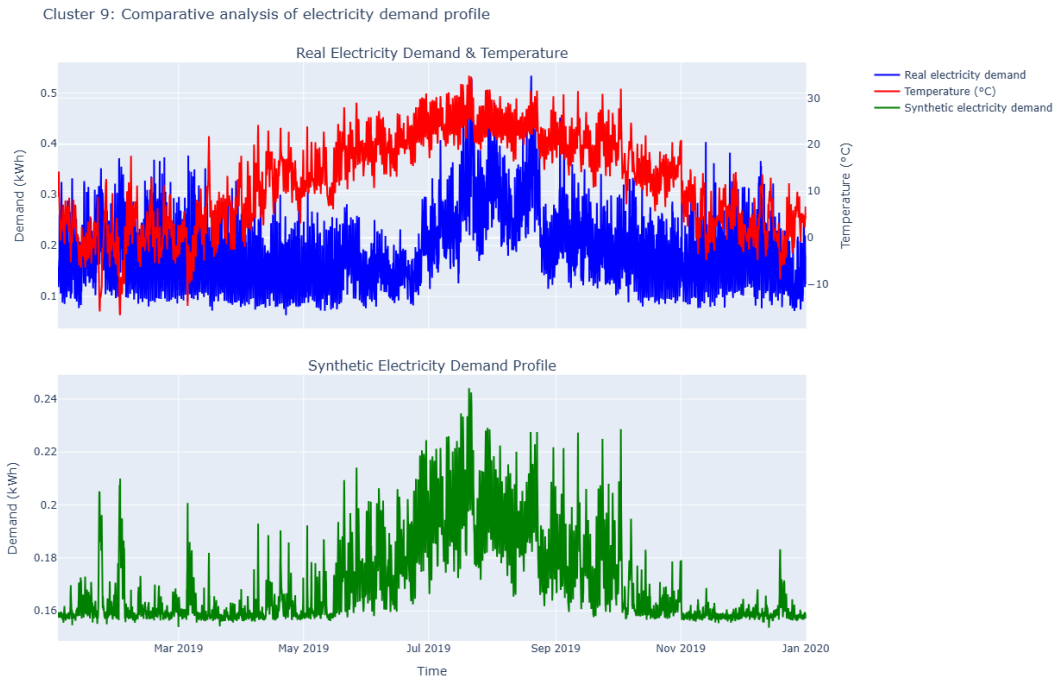

Figure 4.6: Cluster 9: Real vs Synthetic electricity demand profile.

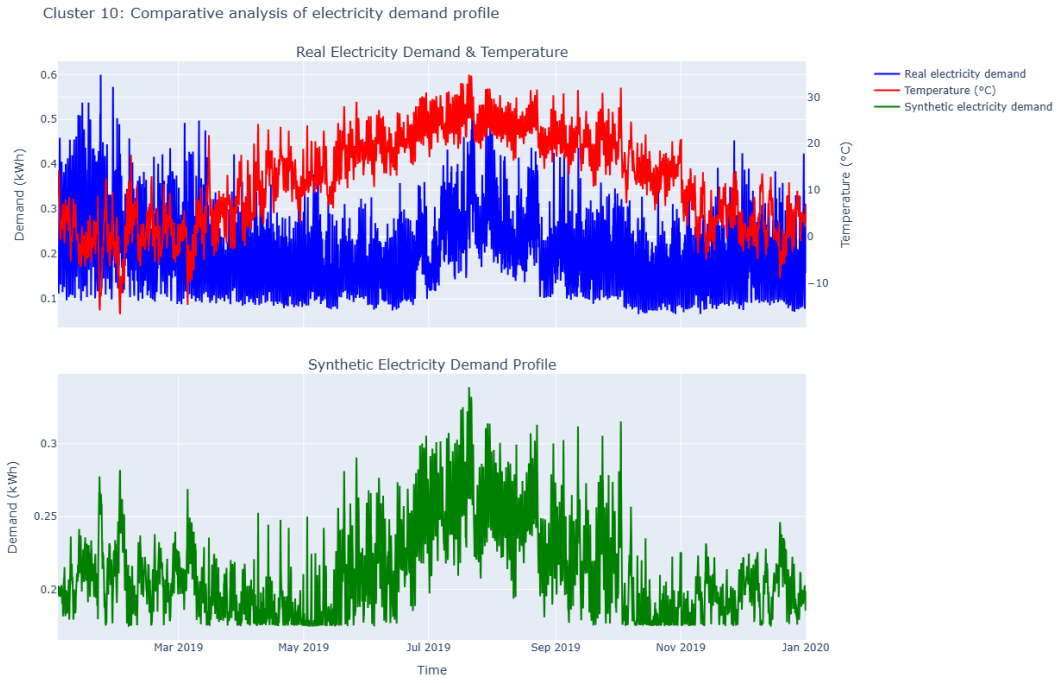

Figure 4.7: Cluster 10: Real vs Synthetic electricity demand profile.

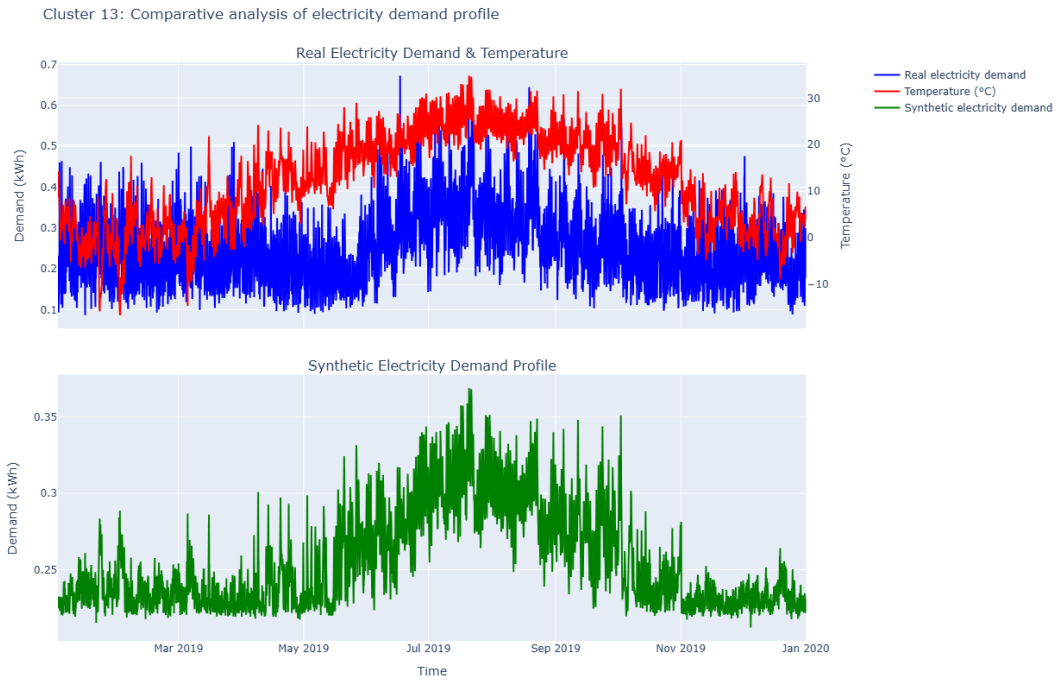

Figure 4.8: Cluster 13: Real vs Synthetic electricity demand profile.

They generally follow the overall structural pattern of the real demand profiles; however, they occasionally fail to capture the local patterns. This suggests that while the global temporal dynamics are well learned, some of the minute patterns are not fully reproduced.

To further investigate these differences, a visual analysis is conducted based on representative weekly patterns across different periods of the year. This allows for a more detailed comparison of seasonal dynamics and short-term variability between real and synthetic data.

The following figures show the consumption behavior patterns on a weekly basis for clusters 10 and 13:

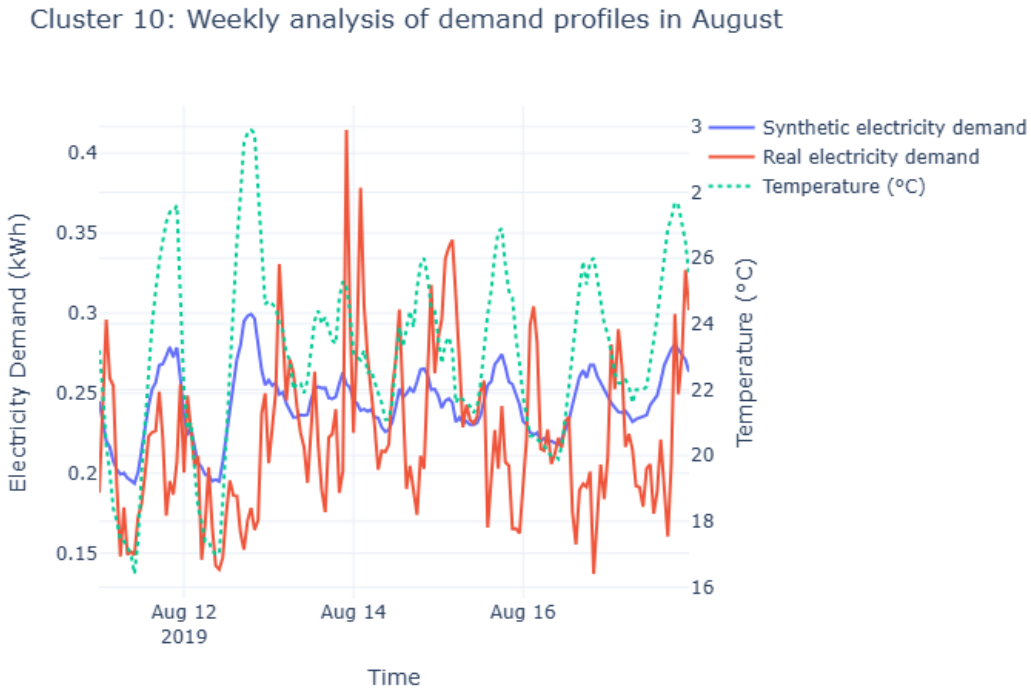

Figure 4.9: MFRED dataset electricity demand profile comparison across JULY.

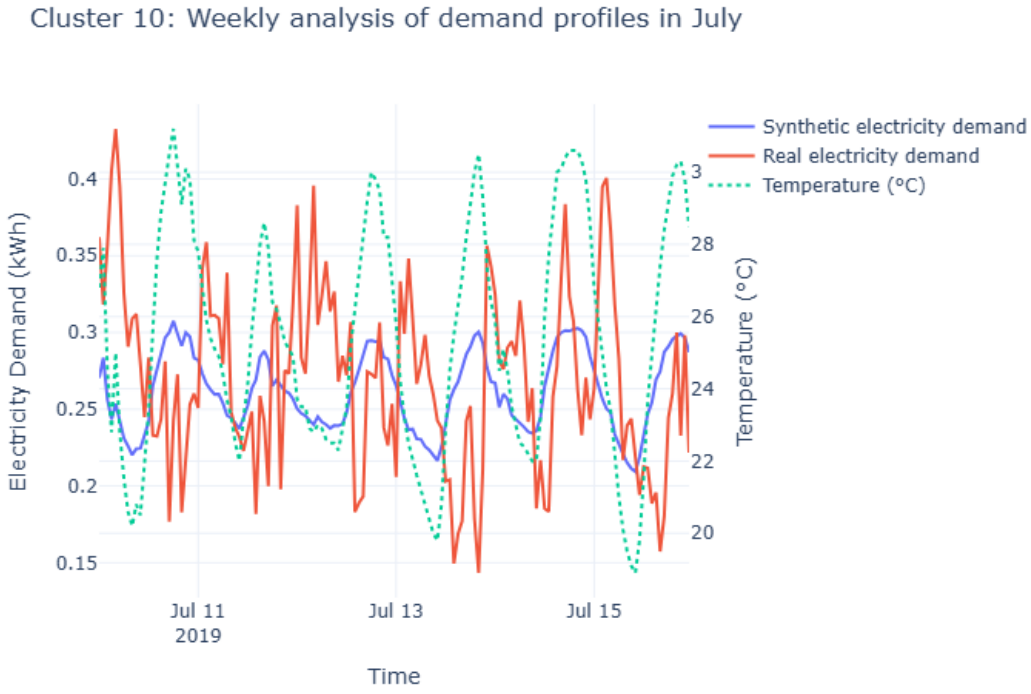

Figure 4.10: MFRED dataset electricity demand profile comparison across JULY.

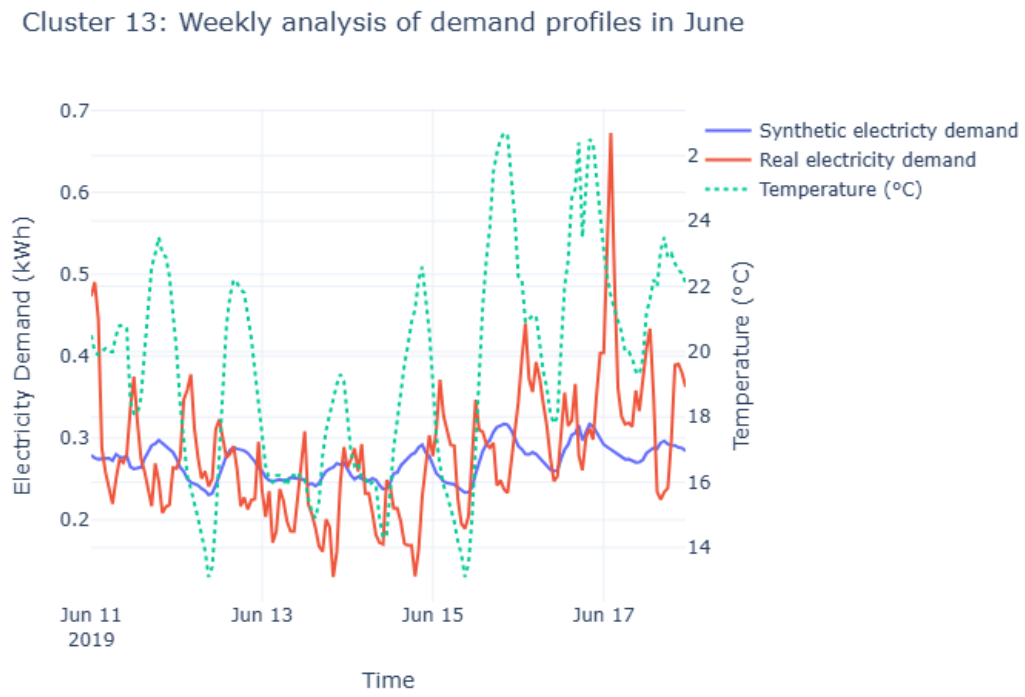

Figure 4.11: MFRED dataset electricity demand profile comparison across JULY.

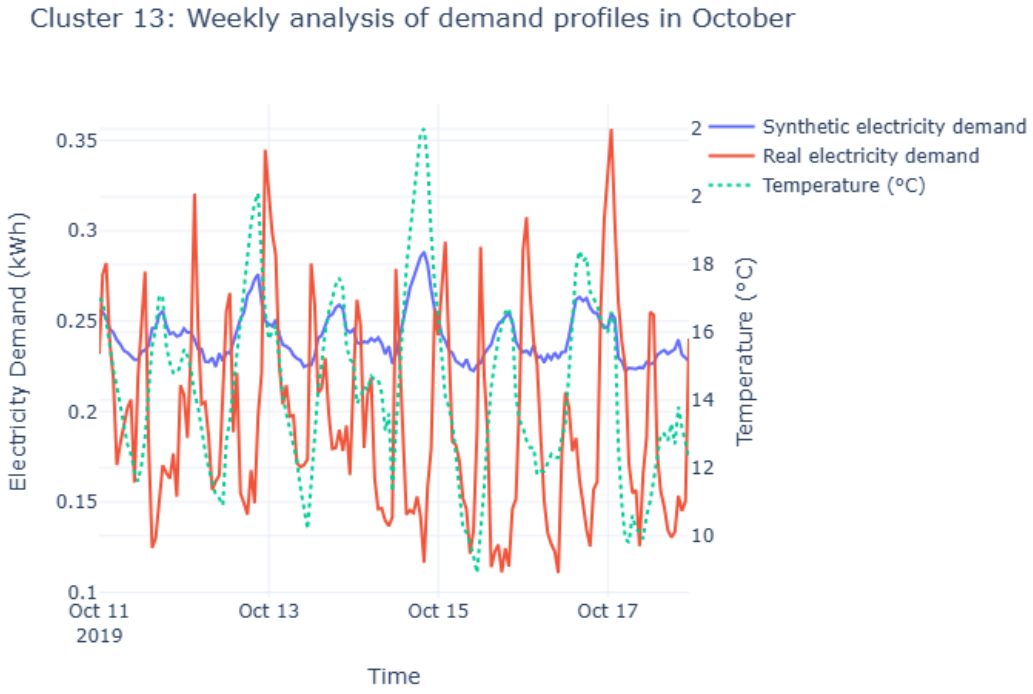

Figure 4.12: MFRED dataset electricity demand profile comparison across JULY.

Discussion and analysis

The weekly figures above follow a similar pattern to those shown in Figures 4.3–4.5 from the global MFRED dataset. The synthetic profiles are largely temperature dominated and tend to miss local peak loads; however, the model still successfully reproduces the overall global seasonal structure and long-term trend of the demand profiles.

4.2 Analysis of TABULA synthetic demand profiles

In order to further evaluate the quality of the generated demand data, graphical comparisons were performed between the real electricity demand profiles for Clusters A and B, the synthetic profiles.

The following figures represent the comparison between real and synthetic demand profiles for the selected buildings.

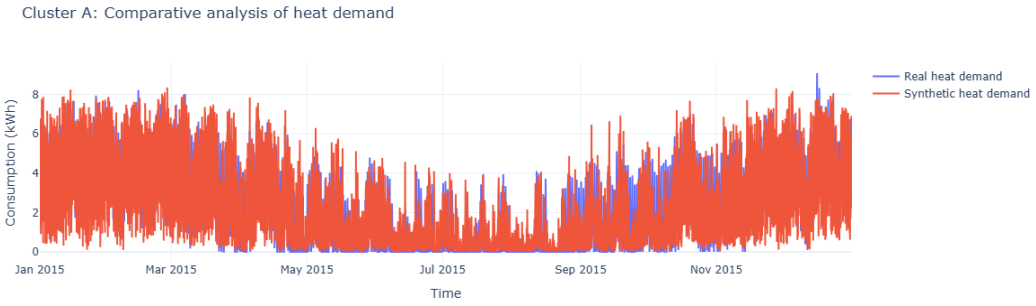

Figure 4.13: Cluster A: Real vs Synthetic heat demand profile.

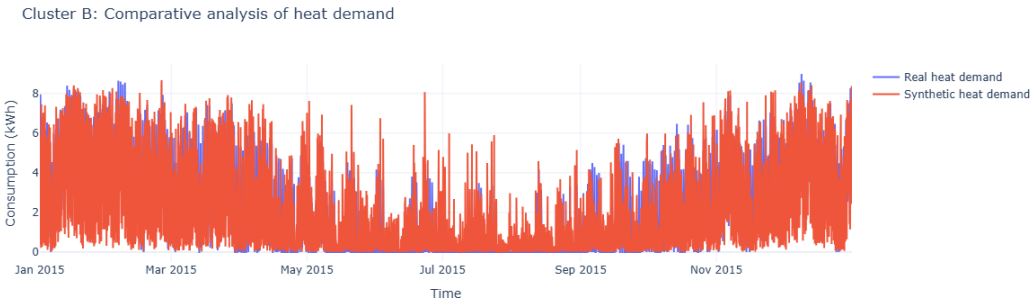

Figure 4.14: Cluster B: Real vs Synthetic heat demand profile.

4.2.1 Evaluation metrics of TABULA dataset

To evaluate the capability of CVAEs in generating realistic electricity demand profiles for each cluster, the real and corresponding synthetic demand profiles are compared to assess the model's ability to reproduce temporal dynamics and consumption characteristics.

Figures 4.13 and 4.14 present the comparison between real and synthetic demand profiles for the selected buildings.

As observed, the synthetic profiles successfully capture the general temporal structure of the real demand data, including the timing of daily peaks and the overall load shape. The model is able to reproduce consistent daily patterns, indicating that the underlying temporal dependencies have been effectively learned and understood by the Variational Autoencoder.

However, it can also be observed that the synthetic profiles exhibit reduced peak magnitudes in certain time periods. The behavior suggests that the CVAE tends to overestimate or sometimes underestimate the peak demand during those periods, leading to slight change in the overall variability. Despite this, the preservation of peak timing and overall demand structure indicates that the model captures the essential characteristics of electricity consumption.

Overall, the results demonstrate that the CVAE is capable of understanding the underlying consumption patterns, while exhibiting slight reduction in peak consumption intervals.

Evaluation metrics for cluster A

Metric	Real Dataset	Synthetic Dataset
Mean	2.16	1.84
Standard Deviation	2.23	2.04
Differential entropy	3.07	3.51

Table 4.3: Comparison of evaluation metrics for Cluster A.

Dataset	MMD
Cluster A	0.012

Table 4.4: MMD score

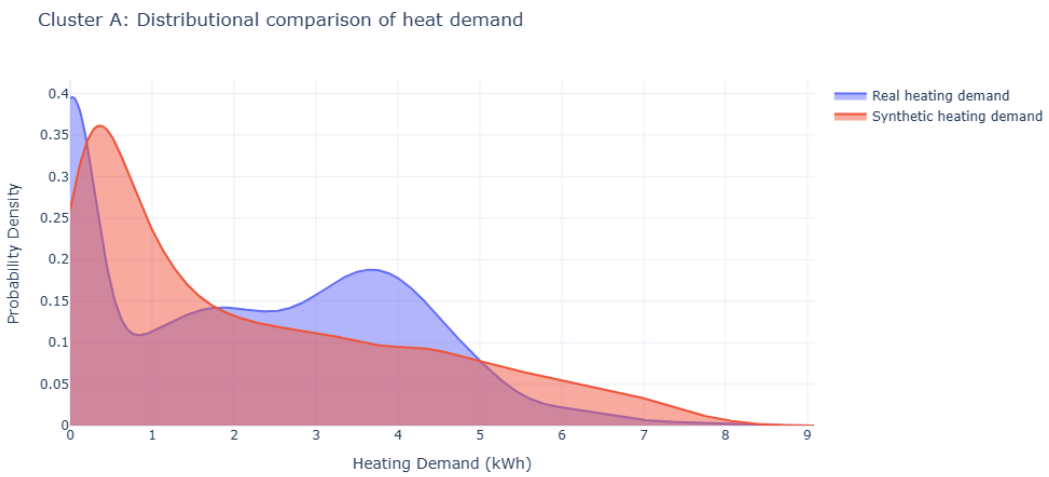

Figure 4.15: Cluster A: KDE PLOT.

Explanation:

The results show that the synthetic dataset closely follows the statistical structure of the real dataset, despite some noticeable deviations in central tendency and variability.

The mean value of synthetic dataset consumption is lower than that of the real consumption, indicating a slight underestimation of heating demand. Similarly, the standard deviation is also lower, indicating that the synthetic demand values exhibit reduced variability.

The entropy demonstrates that the synthetic demand values also contain more randomness and comparatively more dispersed, which can be confirmed with a visual analysis of the weekly graphs.

Despite these differences, the MMD value of 0.012 indicates an extremely close alignment between the real and synthetic distributions, KDE plot figure 4.15 demonstrates the same behavior. This suggests that the generative model is able to capture the overall distributional structure very effectively, even if minor deviations, dispersion and more randomness exist in specific statistical properties.

Summary

Although some small biases are present in mean and variance, the extremely low MMD value confirms that the model understands the underlying structure of the heating demand distribution for this cluster.

Evaluation metrics for Cluster B

The following table summarizes the computed evaluation metrics for cluster B.

Metric	Real Dataset	Synthetic Dataset
Mean	2.03	1.91
Standard Deviation	1.99	2.04
Differential entropy	2.88	3.27

Table 4.5: Comparison of evaluation metrics for cluster B.

Dataset	MMD
Cluster B	0.016

Table 4.6: MMD score

Here are the probability density plots for the TABULA Cluster B datasets

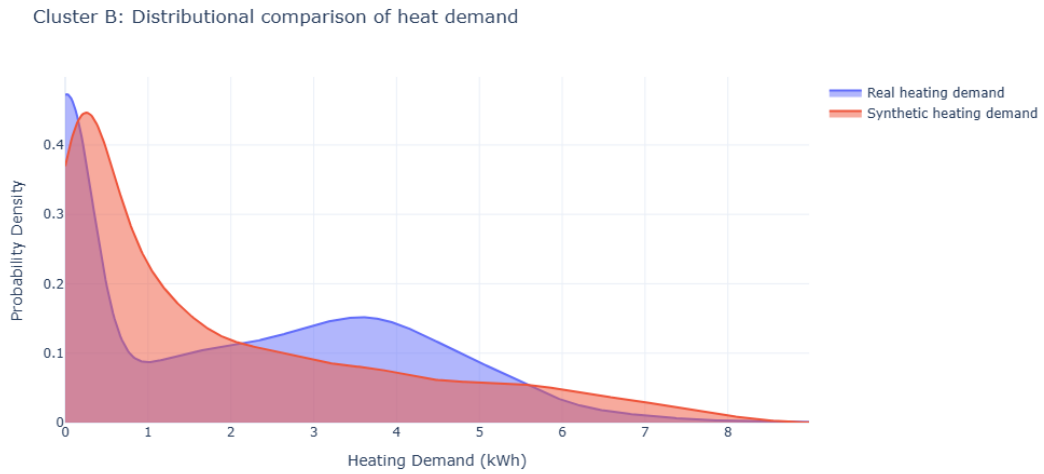

Figure 4.16: Cluster B: KDE PLOT.

Explanation: Cluster B follows similar trend as Cluster A, it has more randomness and dispersion, contains slight deviation in central tendency but carries a good MMD score to indicate understanding of the overall dataset.

4.2.2 Distributional analysis

The comparison of statistical metrics shows that the synthetic dataset closely follows the real dataset. The mean values are nearly identical, suggesting that the variability in the synthetic data is well-preserved. The KDE plot in figure 4.16 above exhibits the same behavior as well.

The differential entropy of the synthetic dataset is slightly higher than that of the real dataset, indicating a small increase in randomness. The weekly plots below provide a closer visual inspection. As more unstable consumption values have been generated by the CVAE model, the synthetic demand exhibits unstable behavior compared to the real consumption. The real dataset exhibits smoother and more structured transitions. However, this difference is small and does not significantly affect the overall distribution similarity.

This observation is strongly supported by the MMD value of 0.016, which indicates a very close alignment between the real and synthetic distributions. Such a low MMD suggests that the generative model is able to accurately reproduce the statistical structure of the original heating demand data for this dataset.

Summary

Overall, the results for Cluster B indicate a very strong match between the real and synthetic datasets. The statistical properties are matching with each other closely, and the low MMD value confirms that the probability distribution of the synthetic dataset is closely matching with the real dataset

4.2.3 Compute simultaneity factor

Initial Results ($k_s = 1.0$) Initial calculations on the Tabula dataset yielded a simultaneity factor of 1.0. This result suggests a state of perfect synchronization, which is practically unachievable in any real-world residential or commercial neighborhood.

The mathematical reasoning for this observation is associated to the dataset structure itself. Although the original source contained data 110 buildings across different foundational year, refurbishment status and type of houses. The analysis was initially performed on a single building data. In a single-load scenario with one building, the formula collapses:

$$k_s = \frac{\max(P_1(t))}{\max(P_1(t))} = 1 \tag{4.7}$$

As only the dataset of one building is available, the factor does not represent the diversity of demand in a whole neighborhood district, but rather the maximum demand value of the consumer for that complete year.

4.2.4 Visual analysis on a weekly basis

The following graphs provide a weekly visualization plot to investigate minute differences:

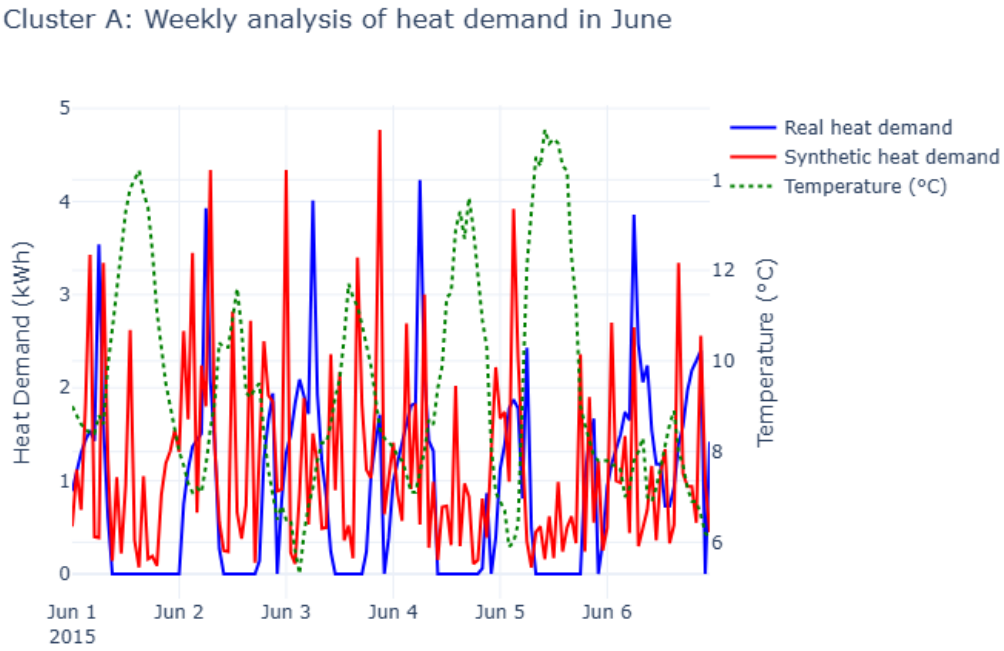

Figure 4.17: Cluster A: Weekly comparison for June profile.

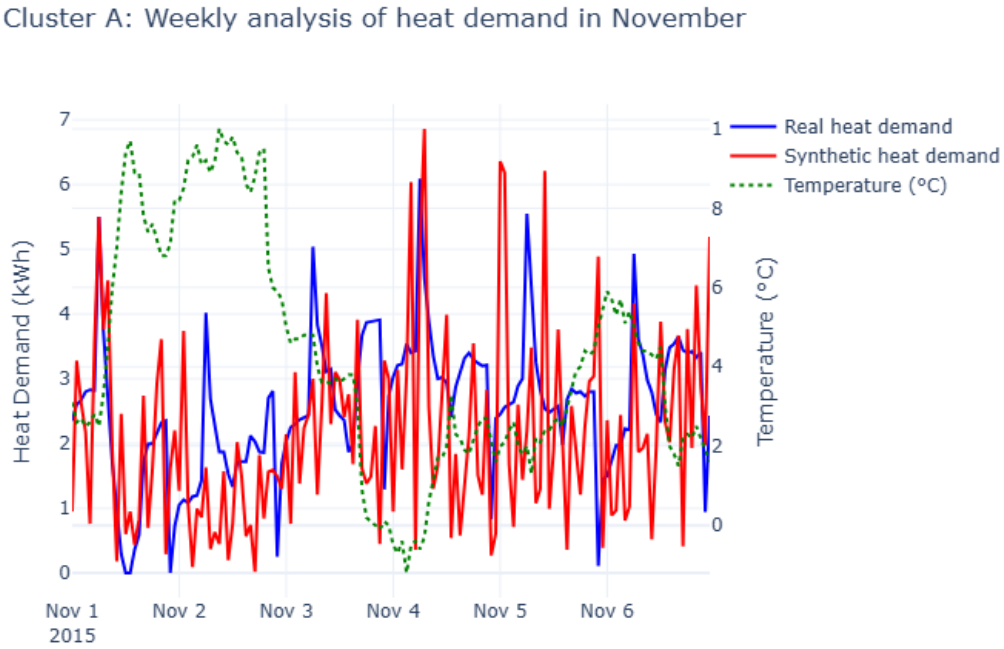

Figure 4.18: Cluster A: Weekly comparison for November profile.

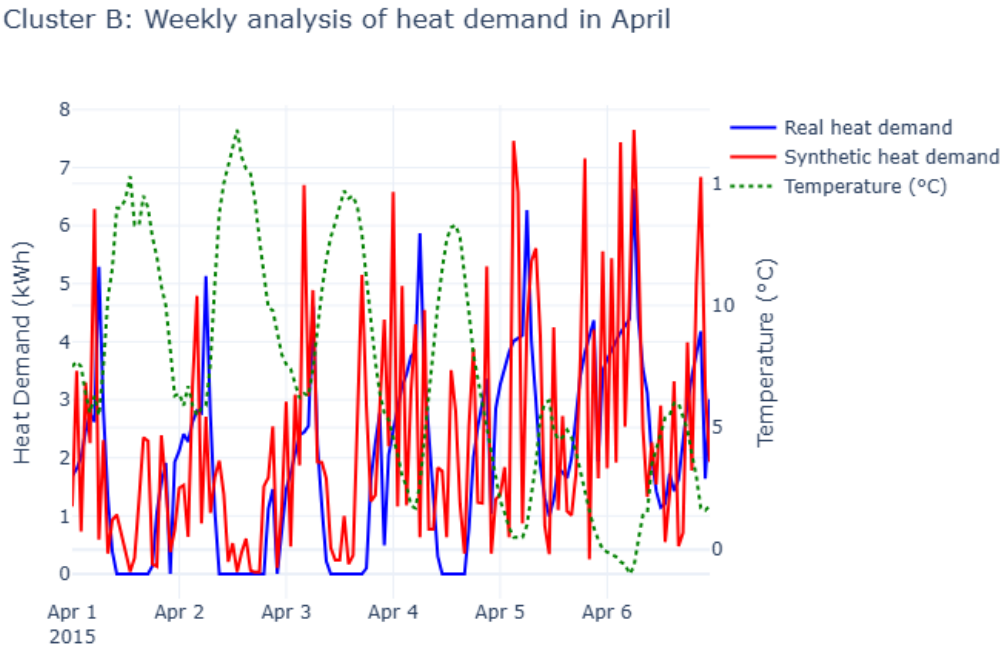

Figure 4.19: Cluster B: Weekly comparison for April profile.

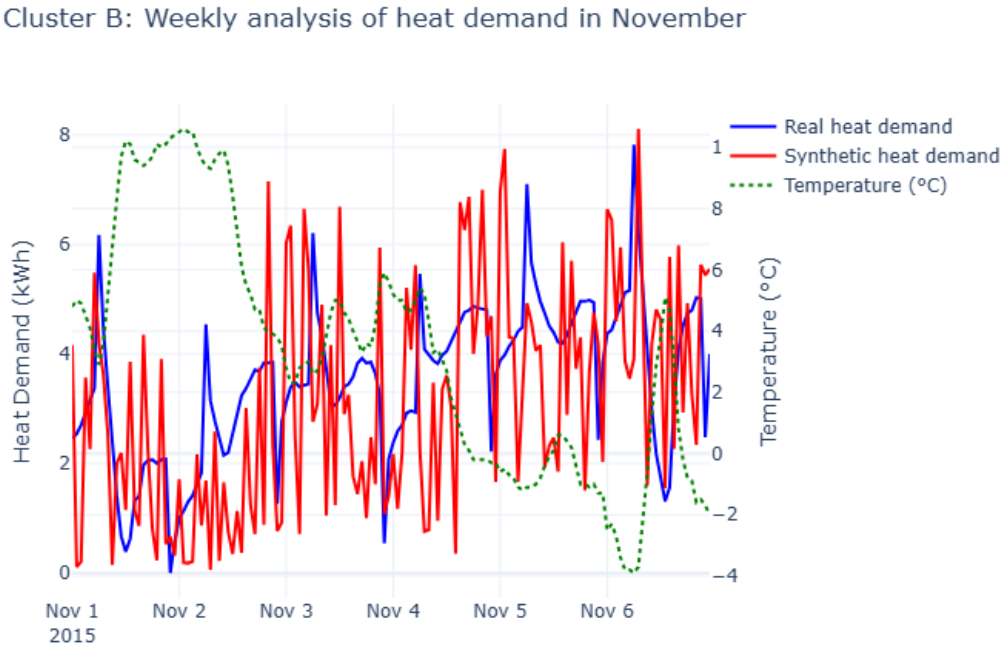

Figure 4.20: Cluster B: Weekly comparison for November profile.

Discussion and analysis

Unlike the weekly plots of MFRED dataset, the weekly TABULA datasets tend to follow and understand the local patterns. They are not significantly influenced by the temperature and tend to understand the overall structure of the heat demand values across winter and summer seasons.

4.3 Forecast models

In addition to the generative modeling of electricity demand profiles, another analysis was conducted to evaluate the capability of classical time series models in forecasting missing values. While this task does not support the primary objective of the thesis, it provides a complementary perspective on modeling of temporal electricity consumption patterns. The objective of this task was to predict the heat demand values for the next 3 days (72 hours) using historical observations.

The initial analysis revealed that that the time series was non-stationary. The Autocorrelation function (ACF) and Augmented Dickey Fuller (ADF) testing determined a gradual decay, indicating that the data is non-stationary Consequently, first-order differencing ($d = 1$) was applied to make the dataset stationary.

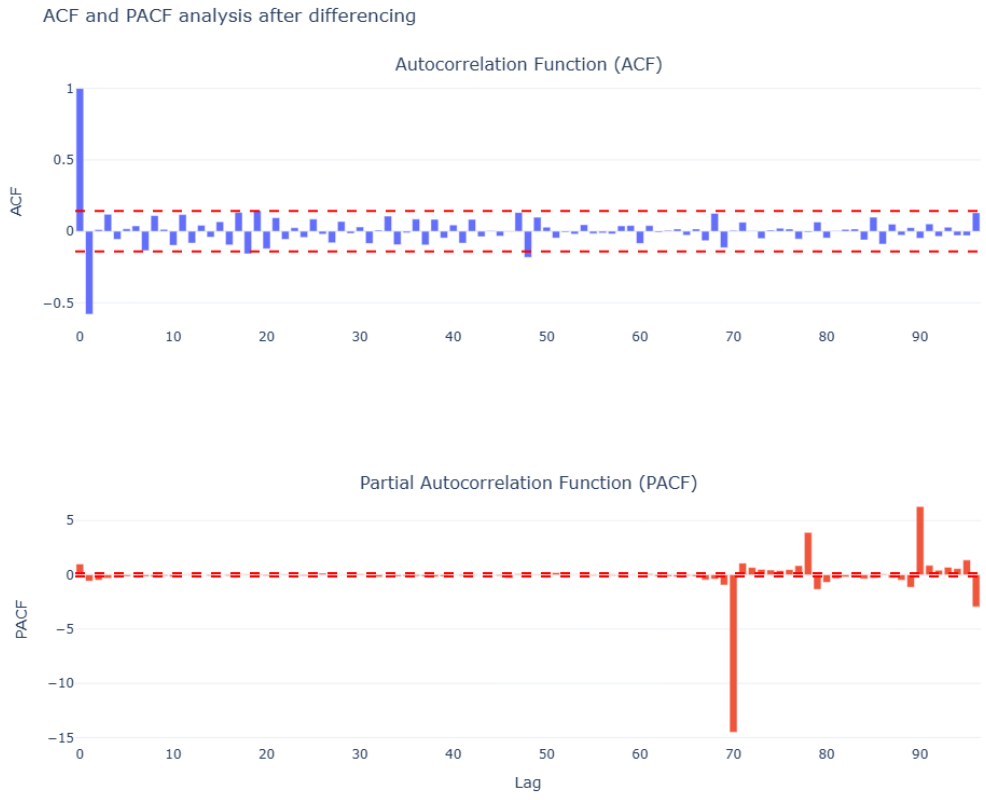

Figure 4.21: ACF and PACF results after differencing.

Following the stationarity adjustment, several model configurations were explored, a baseline ARIMA (1,1,1) was implemented; however this configuration resulted in a nearly constant forecast with negligeble variability. To improve the model performance, higher order moving average components were introduced, such as ARIMA(1,1,72) and ARIMA(6,0,72). The intention was to capture longer temporal dependencies, while thesee models introduced slight variability, they did not follow the consumption pattern of the available values on a weekly basis.

4.3.1 Comparative forecast validation

To further evaluate the reliability of the forecasts, a comparative validation approach was employed. Two forecasting scenarios were considered:

- **Standard Forecast:** Prediction of unknown future values based on historical observations.
- **Verification Forecast:** Prediction of a known future period, allowing direct comparison with ground truth data.

To better understand the structure of the dataset, the time series was analyzed at multiple aggregation levels, including hourly, daily, and weekly resolutions. This analysis revealed clear temporal patterns, particularly recurring daily cycles and longer-term trends.

The following graph shows the recurring daily cycles and trends:

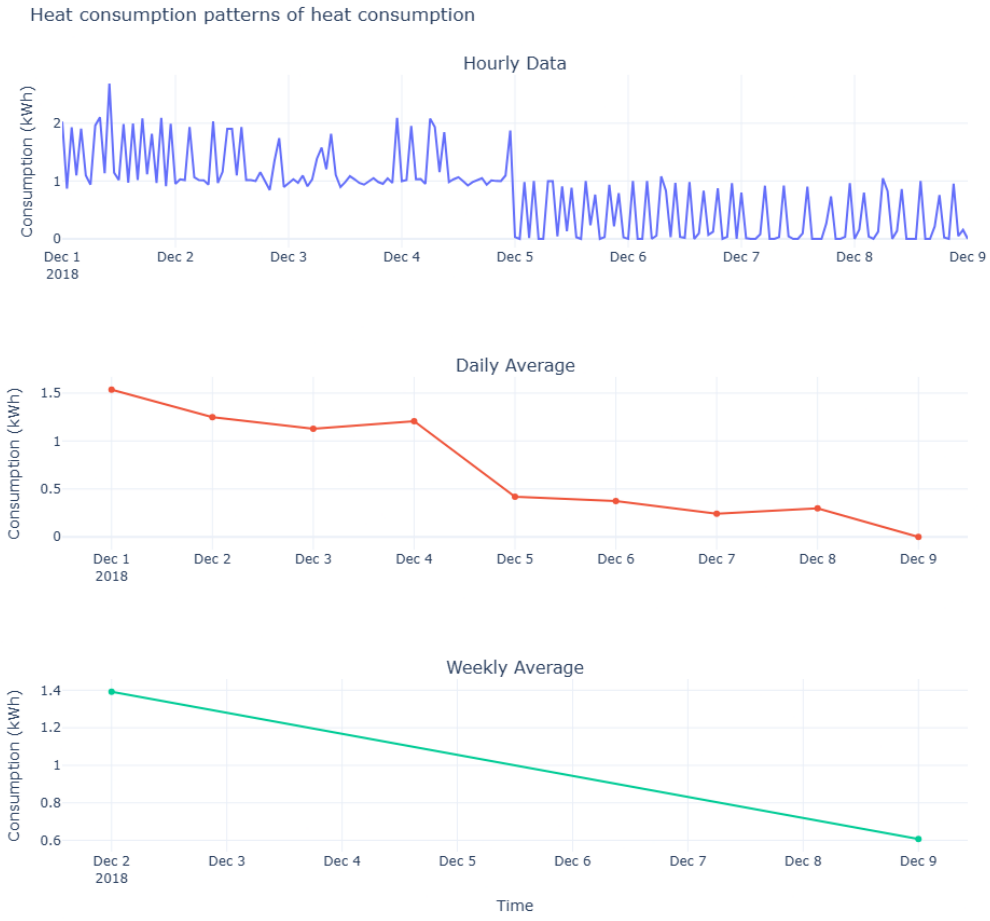

Figure 4.22: Temporal consumption patterns across different period of time.

The comparison revealed that the verification model consistently underestimates the actual power consumption values. The predicted series did not follow the temporal patterns. The consistency between both scenarios and hyper-parameters indicates that the model output is largely driven by historical observations rather than an accurate representation of underlying temporal dynamics.

The ARIMA based forecasts did not reproduce the recurring consumption behavior on a weekly cycle. Meanwhile, the real values from the dataset during that period exhibited greater peak consumption values. This discrepancy highlights the inability of the model to capture multi-scale temporal dependencies.

Here are the predictions from standard and verification forecasts representing the performance of the ARIMA models:

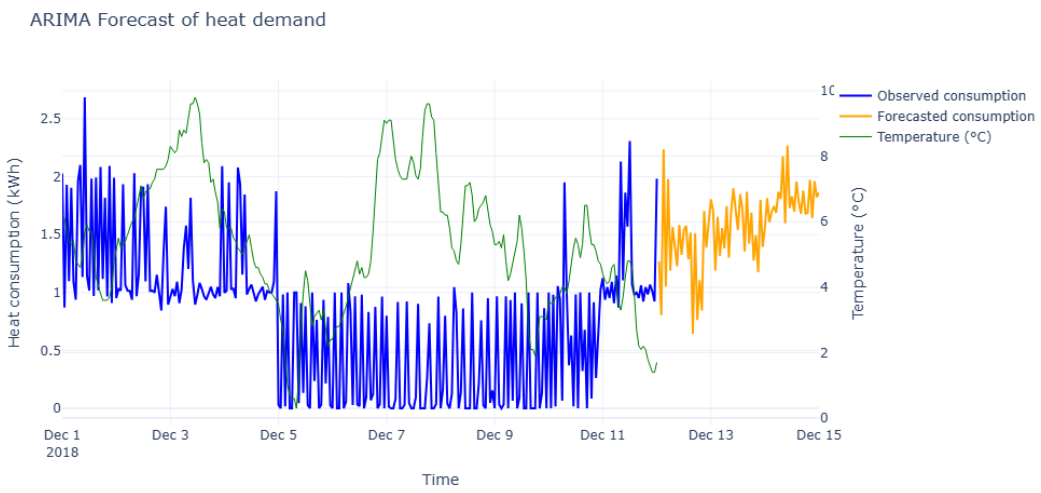

Figure 4.23: Standard model forecasts

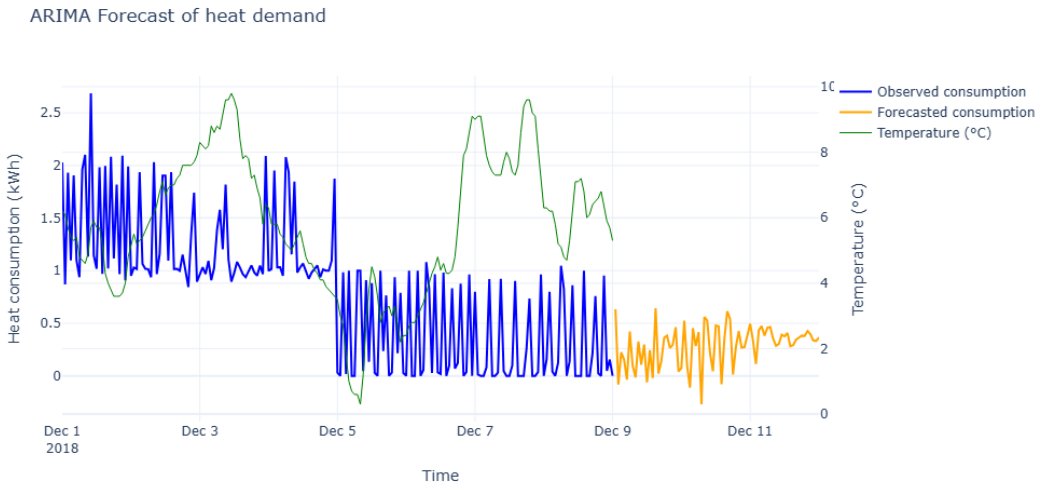

Figure 4.24: Verification model forecasts

The observed limitations of the ARIMA model can be attributed to several factors.

First, ARIMA models assume linear relationships between past and future values, whereas energy consumption data is influenced by non-linear and stochastic processes (Zhang).

Second, seasonal patterns such as daily, weekly cycles are not explicitly modeled in standard ARIMA formulations.

Third, the differencing operations may reduce meaningful structural information in the data, although it helps to achieve stationarity (Zhang).

As a result, the model exhibits a tendency to converge toward the mean behavior of the series, producing forecasts that lack variability and responsiveness to real-world

fluctuations.

4.3.2 Implications and conclusion

The findings demonstrate that classical ARIMA and SARIMA models are insufficient for accurately modeling high-resolution energy consumption data. While they provide a useful statistical baseline, they fail to capture the complexity of the temporal dynamics and multi-scale patterns inherent in the dataset.

In summary, the ARIMA based analysis highlights the challenge of applying traditional time-series models to complex real-world data. Despite appropriate pre-processing and model tuning, the forecasts remain limited in their ability to reproduce observed consumption patterns. These results emphasize the need for more flexible and expressive models in energy demand forecasting.

This limited predictive capability for the dataset motivates the exploration of more advanced approaches to capture and understand the complex consumption patterns, such as GANs or VAEs, which are comparatively more difficult to train but reap better results.

Chapter 5

Conclusion and outlook

5.1 Conclusion

This thesis focused on the development and evaluation of synthetic energy demand profiles for residential buildings using data-driven approaches. The primary objective was to generate realistic consumption patterns that preserve the statistical and temporal characteristics of real-world energy demand, while addressing challenges related to data scarcity and privacy.

A CVAE based framework was implemented to learn the underlying structure of energy consumption data and generate synthetic profiles. The generated data was evaluated using a combination of statistical metrics, including mean, standard deviation, differential entropy, and MMD, along with visual analysis through distribution and time-series plots. The results demonstrate that the synthetic data closely matches the statistical properties and variability of the original dataset. In particular, the similarity in distributional characteristics and temporal patterns indicates that the VAE successfully captures the inherent dynamics of energy demand.

Furthermore, the simultaneity factor was used as an additional evaluation measure to assess aggregated demand behavior. The comparison between real and synthetic datasets showed that the generated profiles preserve the simultaneity characteristics, suggesting that the model is capable of reproducing realistic demand aggregation patterns.

However, as per the limitations of computing the SF score on a cluster level for MFRED dataset and TABULA dataset in sections 4.2.3 and 3.2.1, it is unfortunately difficult to scale back and forth on a bottom-up and top-down approach to design a whole neighborhood plan and utilize the results presented in this thesis.

As a secondary objective, classical time-series forecasting methods based on ARIMA were explored for short-term demand prediction. While these models were able to capture basic temporal dependencies, they exhibited significant limitations in representing real-world consumption behavior. In particular, the forecasts showed reduced variability and a tendency to underestimate peak demand, indicating a bias toward mean behavior. This highlights the limitations of linear statistical models when applied to complex, high-

resolution energy data.

Overall, the findings of this study demonstrate that generative models, specifically CVAEs, provide a robust and effective approach for synthetic energy demand generation. The ability to preserve both statistical properties and aggregated demand behavior makes them a valuable tool for energy system modeling and planning applications.

5.2 Outlook

The results of this study open several directions for future research in the field of energy demand modeling and synthetic data generation. One promising extension is the application of more advanced generative models, such as Generative Adversarial Networks (GANs) and diffusion models, which have shown strong performance in capturing complex and high-dimensional data distributions. These approaches may further improve the realism and diversity of generated demand profiles.

Another important direction is the integration of energy disaggregation techniques (Kleiminger et al., 2015), which aim to decompose total household energy consumption into individual appliance-level demand. Combining generative models with disaggregation methods could enable a more detailed analysis of consumption patterns, providing insights into appliance usage and occupancy behavior while preserving user privacy.

Additionally, incorporating external influencing factors such as weather conditions, occupancy schedules, and socio-economic variables into generative models could enhance the accuracy and applicability of synthetic demand profiles. This would allow for more realistic scenario generation and improved decision-making in energy system planning.

In summary, the integration of advanced generative modeling techniques with domain-specific knowledge offers significant potential for improving the accuracy, scalability, and privacy of energy demand analysis, making it a valuable area for continued research and development.

Bibliography

L. Banh and G. Strobel. Generative artificial intelligence. *Electronic Markets*, 2023. doi: 10.1007/s12525-023-00680-1. URL https://doi.org/10.1007/s12525-023-00680-1.

A. Baumeister. Energiepositive städte: Ein neues kapitel im klimaschutz. 2025. URL https://www.techzeitgeist.de/energiepositive-staedte-ein-neues-kapitel-im-klimaschutz/.

S. E. Bibri, S. K. Jagatheesaperumal, J. Huang, and J. Krogstie. The synergistic interplay of artificial intelligence and digital twin in environmentally planning sustainable smart cities: A comprehensive systematic review. *Energy and Built Environment*, 2024. URL 10.1016/j.ese.2024.100433.

S. Bolluk, R. Aydoğan, E. Sefer, E. K. Özdemir, and S. Seyis. Synthetic data generation and energy consumption prediction in district building energy modeling. *Energy and Buildings*, 2025. doi: 10.1016/j.enbuild.2025.116716.

Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen. Kommunale wärmeplanung. URL https://www.bmwsb.bund.de.

L. Böcking, A. Michaelis, B. Schäfermeier, A. Baier, N. Kühl, M.-F. Körner, and L. Nolting. Generative artificial intelligence in the energy sector. Bayreuther arbeitspapiere zur wirtschaftsinformatik, Universität Bayreuth, 2024. URL https://epub.uni-bayreuth.de/7674/.

E. Cina, E. Elbasi, G. Elmazi, and Z. AlArnaout. The role of AI in predictive modelling for sustainable urban development: Challenges and opportunities. *Sustainability*, 06 2025. doi: 10.3390/su17115148. URL https://doi.org/10.3390/su17115148.

Deutsche Energie-Agentur (dena). Artificial intelligence brings new opportunities for district heating networks. 2024. URL https://www.dena.de/en/infocenter/translate-to-english-kuenstliche-intelligenz-bringt\protect\discretionary{\char\hyphenchar\font}{}{}-neue-chancen-fuer-fernwaermenetze/.

J. Dickert and P. Schegner. Residential load models for network planning purposes. doi: 10.1109/MEPS.2010.5721245.

C. Doersch. Tutorial on variational autoencoders, 2016. URL https://arxiv.org/abs/1606.05908.

S. El Omda and S. R. Sergent. *Standard Deviation*. StatPearls Publishing, Treasure Island (FL), 2024. URL https://www.ncbi.nlm.nih.gov/books/NBK574574/. Updated Nov 25, 2024; Accessed: April 17, 2026.

European Commission. Energy and smart cities. URL https://energy.ec.europa.eu/topics/clean-energy-transition-initiatives/energy-and-smart-cities_en.

Fraunhofer Institute for Applied Information Technology FIT. Heat planning for New AG: AI for the energy transition. 2023. URL https://www.fit.fraunhofer.de/en/publikationen/annual-report-2023/heat-planning-for-new-ag.html.

C. Fu, H. Kazmi, M. Quintana, and C. Miller. Creating synthetic energy meter data using conditional diffusion and building metadata. *Energy and Buildings*, 2024. doi: 10.1016/j.enbuild.2024.114216.

P. Gans. Urban population development in Germany (2000-2014): The contribution of migration by age and citizenship to reurbanisation. *Comparative Population Studies*, 2018. doi: 10.12765/CPoS-2017-19en. URL https://comparativepopulationstudies.de/index.php/CPoS/article/view/288.

A. V. Gargary and E. De Cristofaro. A systematic review of federated generative models. *arXiv preprint arXiv:2405.16682*, 2024. URL https://arxiv.org/abs/2405.16682.

R. Giudice, S. Amico, M. Piscitelli, and A. Capozzoli. Generation of synthetic load profiles for different typologies of residential users through metadata-driven generative ai models. *Proceedings of Building Simulation 2025*, 2025. doi: 10.26868/25222708.2025.1673.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

T. Guo, M. Bachmann, M. Kersten, and M. Kriegel. Alternative: A workflow to generate building energy demand profiles at urban scales from low-level city datasets. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4241705. URL https://ssrn.com/abstract=4241705.

G. Gust, A. Schlüter, S. Feuerriegel, I. Úbeda, J. T. Lee, and D. Neumann. Designing electricity distribution networks: The impact of demand coincidence. *European Journal of Operational Research*, 06 2024. doi: 10.1016/j.ejor.2023.11.029.

R. Harries, C. Lawson, and P. Shapira. Generative AI in science: Applications, challenges, and emerging questions. 2025. doi: 10.48550/arXiv.2507.08310.

W. Harvey, S. Naderiparizi, and F. Wood. Conditional image generation by conditioning variational auto-encoders. 2021. doi: 10.48550/arXiv.2102.12037. URL https://arxiv.org/abs/2102.12037.

M. Hurley and S. Tenny. *Mean*. StatPearls Publishing, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK546702/.

R. J. Hyndman and G. Athanasopoulos. *SAS Institure Inc.* Cary, NC, New York, 2010. URL https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm.

R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2018. URL https://otexts.com/fpp2/.

International Organization for Standardization. DIN EN ISO 52016-1: Energy performance of buildings - energy needs for heating and cooling, internal temperatures and sensible and latent heat loads, 2017.

J. E. Johnson. Maximum mean discrepancy (MMD). 2020. URL https://jejjohnson.github.io/research_journal/appendix/similarity/mmd/.

F. Kachirayil, J. M. Weinand, F. Scheller, and R. McKenna. Reviewing local and integrated energy system models: insights into flexibility and robustness challenges. 2022. URL 10.48550/arXiv.2202.13942.

D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 2019. doi: 10.1561/2200000056. URL https://arxiv.org/abs/1906.02691.

W. Kleiminger and C. Beckel. Eco data set (electricity consumption & occupancy). 2016.

W. Kleiminger, C. Beckel, and S. Santini. Household occupancy monitoring using electricity meters. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, Osaka, Japan, Sept. 2015.

A. Kotal and A. Joshi. KIPPS: Knowledge infusion in privacy preserving synthetic data generation. 2022. URL https://arxiv.org/html/2409.17315v1.

J. Leprince, A. Schledorn, D. Guericke, D. F. Dominkovic, H. Madsen, and W. Zeiler. Can occupant behaviors affect urban energy planning? distributed stochastic optimization for energy communities. 2023. URL 10.48550/arXiv.2303.03006.

H. Liu, L. Yan, Y. Song, and Q. Gao. Differential entropy. 2021. URL https://www.sciencedirect.com/topics/engineering/differential-entropy.

T. Loga, N. Stein, and B. Holliday. TABULA typology approach for building stock energy assessment. Technical report, Institut Wohnen und Umwelt (IWU), Darmstadt, Germany, 2016. URL https://www.iwu.de/forschung/.

M. Madanchian. Generative AI for consumer behavior prediction: Techniques and applications. *Sustainability*, 11 2024. doi: 10.3390/su16229963. URL https://www.mdpi.com/2071-1050/16/22/9963.

C. J. Meinrenken, N. Rauschkolb, and P. J. Culligan. Mfred, 10second interval real and reactive power for groups of 390 us apartments of varying size and vintage. *Nature Publishing Group*, 2020. URL https://doi.org/10.1038/s41597-020-00721-w.

S. Nassir, A. Badr, and W. Mousa. Estimation the missing data of meteorological variables in different iraqi cities by using arima model. 2018. URL https://www.researchgate.net/publication/326380557_Estimation_the_Missing_Data_of_Meteorological_Variables_In_Different_Iraqi_Cities_By_using_ARIMA_Model.

S. T. Nguyen, T. Tulabandhula, and M. B. Watson-Manheim. User friendly and adaptable discriminative AI: Using the lessons from the success of LLMs and image generation models. 2023. doi: 10.48550/arXiv.2312.06826. URL https://arxiv.org/abs/2312.06826.

A. Nikitin, L. Iannucci, and S. Kaski. TSGM: A flexible framework for generative modeling of synthetic time series. *arXiv preprint arXiv:2305.11567*, 2023. URL https://github.com/AlexanderVNikitin/tsgm. Version available at architectures/zoo.py.

A. Nikitin, L. Iannucci, and S. Kaski. Tsgm: A flexible framework for generative modeling of synthetic time series. 2024. URL https://arxiv.org/abs/2305.11567.

M. Nilashi, O. Boon, G. Tan, B. Lin, and R. Abumalloh. Critical data challenges in measuring the performance of sustainable development goals: Solutions and the role of big data analytics. *Harvard Data Science Review*. doi: 10.1162/99608f92.545db2cf.

ODH@Jülich. Odh@jülich. 2025. URL https://www.ieg.fraunhofer.de/en/references/odh-juelich.html.

H. Ouchra, A. Belangour, and A. Erraissi. An overview of GeoSpatial artificial intelligence technologies for city planning and development. 07 2023. doi: 10.1109/ICECCT56650.2023.10179796.

Z. Pan. Data-driven ev load profiles generation using a variational auto-encoder. *Energies*, 2019. URL 10.3390/en12050849.

L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto. *Variational Autoencoder*. Springer, Cham, 2021. URL https://doi.org/10.1007/978-3-030-70679-1_5.

Rehau AG. Gleichzeitigkeit – der unterschätzte faktor: Effiziente planung von nahwärmenetzen. 2012. URL https://www.rehau.com/downloads/109808/fachartikel-gleichzeitigkeit.pdf.

ScienceDirect. Simulated data. URL https://www.sciencedirect.com/topics/computer-science/simulated-data.

F. Silber, S. Scheubner, and A. Märtz. Analysis of the simultaneity factor of fast-charging sites using monte-carlo simulation. *International Journal of Electrical Power & Energy Systems*. doi: https://doi.org/10.1016/j.ijepes.2023.109355. URL https://www.sciencedirect.com/science/article/pii/S0142061523005975.

C. Wang, S. Tindemans, and P. Palensky. Generating contextual load profiles using a conditional variational autoencoder. 2022. doi: 10.48550/arXiv.2209.04056. URL https://arxiv.org/abs/2209.04056.

Z. Wang and T. Hong. Generating realistic building electrical load profiles through the generative adversarial network (GAN). *Energy and Buildings*, 10 2020. doi: 10.1016/j.enbuild.2020.110299.

W. Winter, T. Haslauer, and I. Obernberger. Untersuchungen der gleichzeitigkeit in kleinen und mittleren nahwarmenetzen. *Euroheat & Power*, 2001. URL https://www.verenum.ch/Dokumente/2001_Winter-Gleichzeitig.pdf. Herausgegeben vom Institut für Grundlagen der Verfahrenstechnik und Anlagentechnik, TU Graz.

Q. Wu, H. Ren, W. Gao, P. Weng, and J. Ren. Coupling optimization of urban spatial structure and neighborhood-scale distributed energy systems. *Energy*, 2017. URL 10.1016/j.energy.2017.12.076.

Y. Yue, G. Yan, and T. Lan. Shaping future sustainable cities with ai-powered urban informatics: Toward human-ai symbiosis. *npj Urban Sustainability*, 2025. URL 10.1007/s43762-025-00190-0.

G. P. Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. doi: 10.1016/S0925-2312(01)00702-0.

Declaration

I hereby certify that I have written this thesis independently and that I have not used any sources or aids other than those indicated, that all passages of the work which have been taken over verbatim or in spirit from other sources from other sources have been marked as such and that the work has not yet been submitted to any examination authority in the same or a similar form. Furthermore, LLMs were used as a paraphrasing aid to improve readability in the academic sense.

Erlangen, April 21, 2026 Signature