POLITECNICO DI TORINO
Repository ISTITUZIONALE

Generation of synthetic load profiles for different typologies of residential users through metadata-driven generative AI models

Original
Generation of synthetic load profiles for different typologies of residential users through metadata-driven generative AI models / Giudice, R., Amico, S., Piscitelli, M.S., Capozzoli, A.. - In: BUILDING SIMULATION CONFERENCE PROCEEDINGS. - ISSN 2522-2708. - 19:(2025). (19th IBPSA Conference on Building Simulation, BS 2025 aus 2025) [10.26868/25222708.2025.1673].

Availability:
This version is available at: 11583/3010070 since: 2026-04-19T15:46:21Z

Publisher:
International Building Performance Simulation Association

Published
DOI:10.26868/25222708.2025.1673

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

02 August 2026

Generation of synthetic load profiles for different typologies of residential users through metadata-driven generative AI models

Rocco Giudice[1], Simone Amico[1], Marco Savino Piscitelli[1,*], Alfonso Capozzoli[1]

[1]Department of Energy “Galileo Ferraris”, TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, 10129, Italy

**Corresponding Author*: marco.piscitelli@polito.it

Abstract

Generating synthetic yet realistic electrical load profiles for residential users is essential for energy scenario analysis, planning, and optimal design. While generative deep learning models have shown promise in producing synthetic load profiles, they often overlook contextual information, limiting their ability to capture the diversity of user typologies and behaviours. To address this, a Conditional Variational Autoencoder (CVAE) is proposed, integrating user metadata, historical consumption patterns from 35 real residential units (e.g., apartments, single-family houses), and exogenous factors to generate realistic daily electricity profiles for both consumers and prosumers.

By capturing the probabilistic variability of residential electrical loads, the model accurately represents energy usage and production patterns while preserving macroscopic and long-term relationships. Its effectiveness was further demonstrated in a data-scarce context by simulating the operation of a small Renewable Energy Community (REC) in Italy, where it successfully captured load variability driven by occupancy patterns, appliance presence, and seasonality. The results outperformed those obtained with a conventional REC estimation approach, highlighting the model potential for improving energy planning and management.

Key Innovations

- A novel approach for generating electrical load profiles of residential users using deep learning techniques was proposed.
- The model exploits simple building metadata and contextual information for enhancing the accuracy of the generation process.
- An interpretable approach was designed to explain key factors that affect load profile generation.

Practical Implications

This work contributes to addressing planning and operation aspects of communities of residential users by leveraging a deep learning-based probabilistic approach to generate synthetic load profiles. The trained model can be integrated into planning and design tools and decision support systems for such communities, including Renewable Energy Communities (RECs), enabling more accurate estimations of energy flows and consumption patterns. This supports strategies to enhance self-consumption and self-sufficiency at the community level by promoting behavioral changes among members and optimizing energy-sharing dynamics.

Introduction

A realistic estimation and synthetic generation of residential electricity load profiles is a fundamental task for energy planning, grid management, and the development of sustainable energy systems. In the context of communities of residential users, such as Renewable Energy Communities (RECs), accurate load profiles enable scenario analysis to optimize energy-sharing strategies (Schito et al., 2024; Gallo et al., 2024) and support the definition of new incentive mechanisms (Giudice et al., 2025) to enhance self-consumption and self-sufficiency. Moreover, realistic load profiles are essential for distribution grid management, ensuring efficient demand response planning, optimal tariff design, and effective demand-side management strategies. The generation of synthetic load profiles for residential buildings has been extensively explored in the literature, primarily through two distinct approaches: ·

- Simulation-Based Methods that use detailed physical modeling to simulate energy consumption, considering factors such as building characteristics, occupant behavior, and appliance usage.
- Data-Driven Methods that employ statistical and machine learning techniques to generate synthetic load profiles based on historical consumption data. ·

Both approaches aim to generate realistic load profiles that capture the variability of residential electricity consumption. Mathematical models have traditionally been used to create synthetic profiles without real consumption data, typically through bottom-up simulations that aggregate appliance-level or household consumption (Gallo et al., 2023). A well-known example is the Markov model, which simulates energy usage through probabilistic state transitions, incorporating factors such as appliance usage (Diao et al., 2017), building characteristics (Ge et al., 2018), human activity patterns (Arbulu et al., 2024), and weather conditions (Binderbauer et al., 2022). While offering high granularity, these models require extensive domain knowledge and often produce discrepancies between simulated and actual consumption.

With the rise of smart meter data, data-driven approaches have gained popularity for generating realistic load

https://doi.org/10.26868/25222708.2025.1673

profiles. Traditional methods include regression models trained on real consumption data (Giannuzzo et al., 2024) and clustering techniques that group users by load patterns (Charwand et al., 2020). In Italy, a reference method developed by ARERA synthesizes monthly load profiles for residential users with the same contracted power, categorizing them by day type (weekdays, Saturdays, Sundays) using an averaging method (ARERA, 2023). However, these approaches struggle to capture daily variability, often producing overly smoothed representations.

The rise of Generative Artificial Intelligence (GenAI) has introduced new opportunities for generating synthetic building load profiles. These models excel at producing synthetic data that accurately reflect the statistical and structural characteristics of real consumption patterns. Among the most studied architectures, Generative Adversarial Networks (GANs) have shown promise in modeling residential energy consumption across different time scales. A GAN consists of two neural networks—a generator, which creates synthetic data, and a discriminator, which assesses its authenticity—trained adversarially to enhance the realism of generated outputs. These models effectively capture both the cyclical and stochastic components of residential energy consumption, ensuring a realistic data distribution and greater diversity among generated profiles (Gu et al., 2019). Variants like Wasserstein GANs (WGANs) improve training stability and performance, making them well-suited for energy applications (Li et al., 2021). However, GANs suffer from training instability and mode collapse, limiting their reliability in scenarios requiring diverse and consistent load profile generation.

An alternative to GANs is Variational Autoencoders (VAEs), which take a probabilistic approach to generating consumption profiles. Differently from traditional autoencoders, VAEs map input data into a continuous latent distribution, allowing new synthetic instances to be sampled. This property helps preserve temporal correlations in the original data, ensuring that generated load profiles maintain consistent temporal patterns (Xia et al., 2024).

In residential load profile estimation, factors such as seasonality, occupant behavior, climate, geography, and appliance types (Thorve et al., 2023) are crucial for accurately characterizing energy consumption patterns. However, these factors also add complexity to the modeling process, making precise and adaptable estimation methods challenging to develop. Many existing studies overlook these variables when training generative models, limiting their ability to capture real-world consumption dynamics. For instance, identifying a reversible heat pump in a building unit helps explain higher electricity use in extreme temperatures, while occupancy schedules justify demand peaks at specific times of the day. Integrating metadata into generative models improves their ability to capture real-world electrical load patterns, enhancing the representation of seasonal variations and daily consumption trends. Without this information, generated profiles tend to smooth out consumption patterns, making it difficult to reflect behaviors such as seasonal heat pump usage (Claeys et al., 2024). Incorporating building characteristics, system specifications, and climatic data acts as a constraint, leading to more realistic load profile representations (Fu et al., 2024). Conditioned models, such as Conditional Variational Autoencoders (CVAEs) and Conditional Generative Adversarial Networks (CGANs), have proven effective in integrating metadata into synthetic load profile generation. Differently from their non-conditional counterparts, these models introduce conditional inputs, allowing them to generate profiles that are contextually guided by user and building characteristics. This improves their ability to capture real-world variability and accurately predict rare events, such as load peaks, outperforming traditional generative methods (Wang et al., 2022; Baasch et al., 2021).

To enhance accuracy, some studies used cluster labels derived from historical load profiles to guide generation (Yang et al., 2022; Tian et al., 2022; Wang and Hong, 2020). While clustering improves synthetic profile accuracy by grouping similar consumption patterns, it assumes detailed consumption data are already available that often is not the case in real-world applications.

In data-scarce environments, where synthetic profiles are most valuable, identifying meaningful cluster labels could be impractical. The only available data often consists of high-level descriptors such as the presence of thermal-sensitive appliances, aggregated energy consumption from utility bills, or general household characteristics. Despite being easy to collect, these attributes provide essential contextual information, reinforcing their importance for realistic load profile generation.

This study investigates the impact of simple, widely accessible metadata on synthetic residential energy profile generation using a Conditional Variational Autoencoder (CVAE). Leveraging a heterogeneous dataset of Italian households, the model integrates exogenous factors such as weather conditions, PV and battery capacity, presence of electric heating/cooling, usage patterns, and monthly consumption statistics. These features, obtainable via surveys or utility bills, enable realistic profile generation in data-scarce settings. By conditioning on metadata and environmental variables, the model generates profiles that better reflect real consumption dynamics, improving aggregated REC energy-sharing estimates. Performance is benchmarked against a reference approach combining ARERA standard load profiles and PVGIS-based generation estimates.

Case study

The dataset comprises daily load profiles from 35 residential units across Italy, including apartments and single-family houses. Each unit is characterized by numerical and categorical metadata, historical grid exchange data, and environmental conditions. Metadata includes user type (13 prosumers and 22 consumers), presence of electrical heating/cooling (17 units equipped with electrical heating systems, while 27 have electrical

https://doi.org/10.26868/25222708.2025.1673

cooling systems), contractual power (31 out of 35 have contracted power of 3 kW, while the remaining units are contracted for 4.5 kW or 6 kW), photovoltaic nominal power (only for prosumer) and battery storage capacity (only for 5 users).

To capture external influences on energy use, historical time series of outdoor air temperature, global horizontal irradiance (GHI), and direct normal irradiance (DNI) were collected for each location. The dataset spans non-contiguous periods from 2020 to 2024, with a total of 9,276 daily load profiles. Grid power exchange, the model's target variable, was recorded at a 15-minute resolution, with positive values for withdrawals and negative for surplus injections.

Methods

This study employs CVAE model to generate synthetic daily load profiles for residential users, leveraging building metadata and exogenous variables. By capturing dependencies between these factors and net electricity exchange, the CVAE ensures realistic energy consumption and production patterns. Extending the Variational Autoencoders , this model learns probabilistic data representations for improved profile generation. The VAE consists of an encoder-decoder architecture, where the encoder compresses input data x into a structured latent representation z, parameterized as a probability distribution q, typically assumed to be a standard normal distribution $q(x|z)$ with mean μ and variance σ^2. A latent variable is sampled from this distribution and passed through the decoder, which reconstructs the original input as a function of $p(x|z)$. Usually, the optimization problem is composed by two components: the reconstruction error, using a metric such as the Mean Square Error (MSE), and the Kullback-Leibler (KL) divergence, which ensures that the predicted output distribution is close to the original one. Differently from VAEs, which rely solely on input data for encoding, CVAEs integrate conditional variables c that guide both the encoding and decoding processes, ensuring that generated samples align with predefined conditions rather than being purely random. In this case, the CVAE introduces the conditional distribution $q(z|x,c)$ in the encoder side, and $p(x|z,c)$ in the decoder side. A reference of this explanation is presented in Figure 1.

Methodology

This study demonstrates the importance of using building metadata, usage information and exogenous variables (e.g., outdoor air temperature, solar radiation) to generate more contextual daily load profiles of residential users. To this purpose, a CVAE model was employed, which conditions the generation process using different families of input features. In particular, the following ones were employed in this study:

- **Building metadata**: Includes user type (consumer, prosumer, or prosumer with storage), presence of electrical heating/cooling, building electrification level (low, medium, high) – intended as a qualitative metric that reflects the extent to which the building relies on electricity for its energy needs – photovoltaic system rated power, electrical storage capacity (if applicable), and contracted power – the maximum allowable grid draw agreed with the electricity provider.
- **Building usage features**: these features describe macroscopic energy usage patterns at monthly level and the most probable peak consumption hours at daily scale. Monthly energy usage patterns were derived from energy bills. In Italy, energy bills typically report consumption based on three standard Time-of-Use (ToU) tariffs: F1, F2, and F3. Following this classification, the extracted features include F1, representing the total monthly energy consumed during daylight hours on weekdays (8 AM – 6 PM); F2, which accounts for the total monthly energy consumed during residential ramp-up hours on weekdays (7 AM and 7–10 PM) and on Saturdays from 7 AM to 10 PM; and F3, capturing the total monthly energy used on Sundays and during night-time hours on weekdays. Additional features include the ratios F1/F2, F1/F3, and F2/F3. When reconstructing daily load profiles within the same month, the CVAE model applies a consistent set of these extracted values.

  On the other hand, information about peak consumption hours was retrived from the electrical load time series by identifying the hours of the day when occupants are likely at home and consuming the highest amount of daily energy. This was determined based on periods when net electricity demand exceeds a statistical threshold, referred to as *On-peak hours*. For each residential user, this threshold is defined for each day as the median net demand plus half the difference between the peak thereshold (95th percentile) and the minimum thereshold (5th percentile). The resulting feature is a 24-hour time series of boolean values (0,1), marking the hours when the user exhibits the highest energy usage. Although derived from historical data, this information can be inferred from interviews and electricity bills in data-scarce contexts, eliminating the need for detailed consumption monitoring.
- **Temporal features**: include features such as the hour of the day, the day of the week and the month for the daily net electricity exchange profile to generate. Those features are transformed into cyclical features by extracting the $sin(x)$ and $cos(x)$ of each x feature listed above.
- **Exogenous features**: include the outdoor air temperature, the global horizontal irradiance and the direct normal irradiance for the whole day with hourly detail.

A schema of the CVAE employed and the different families of features used is shown in Figure 1.

The model architecture follows a fully connected structure with linear layers in both the encoder and decoder. Due to the limited sample size and diversity in the case study, both encoder and decoder include only two hidden layers. The second hidden layer in the encoder and

https://doi.org/10.26868/25222708.2025.1673

the first in the decoder have half the dimension of their adjacent layers. All hidden layers use the ReLU activation function. For training, a custom optimization function was implemented instead of the standard reconstruction loss with MSE and KL divergence typically used in CVAEs.

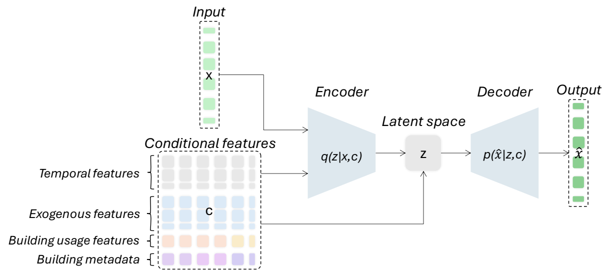

Figure 1: CVAE model architecture employed in this study.

The standard approach, which minimizes overall reconstruction error, tends to generate overly smooth predictions, failing to capture the peaks in real residential load profiles and to preserve the total daily energy used. To overcome this issue, a custom loss function was developed with three main components:

- **KL divergence**: this term remains unchanged from the classical CVAE formulation.
- **Peak preservation loss**: this term introduces a penalty to align both peak magnitudes and time intervals of the real and generated load profiles, applied on the reconstructed error using a logarithmic scaling to control its impact, ensuring that the model learn the high variability of the residential loads.
- **Energy conservation loss**: this term introduces a penalty when the total daily energy consumption derived from the generated synthetic hourly load profile is significantly different from the original one. Even this term is log-scaled to better control its impact. This component ensures that the synthetic profiles maintain the correct energy magnitude, preventing systematic under- or overestimation of the energy patterns.

The training process included an early stopping strategy to prevent overfitting and improve computational efficiency. Instead of a fixed number of epochs, training stopped dynamically based on validation loss. Within a range of 15 to 50 epochs, if validation loss increased by more than 0.01, training was stopped early. The batch size was set at 128 samples per batch.

The dataset, consisting of daily load profiles of the considered 35 residential users, was split into groups. Training used 80% of the profiles, 20% were set aside for testing, and 10% of the training data was used for validation. Each residential user was represented, by means of its own daily load profiles, in all subsets, ensuring as much as possible a proportional distribution. Input and output features were scaled between 0 and 1 to ensure consistency in the training process.

Input features were scaled using global maximum and minimum values in the training set for numerical variables, while categorical variables were one-hot encoded. For the output variable – the net energy exchanged with the grid – a diversified approach was followed considering user typology. For consumers, the contracted power was set as the maximum value of the output variable, while 0 was used as the minimum, reflecting the absence of energy injection into the grid. For prosumers, the contracted power was also set as the maximum value, but the minimum was defined as the negative PV rated power. This choice accounts for the theoretical maximum energy injection into the grid, which occurs when the PV system operates at full capacity without local consumption or storage.

A hyperparameter optimization approach determined the optimal hidden layer size, latent space size, and learning rate. Hidden layers were selected from {1024, 512, 256, 128}, latent space from {32, 64, 128}, and learning rate from {1e-4, 1e-3, 1e-2}. Model performance was assessed using Jensen–Shannon divergence, a smoother and more stable alternative to KL divergence.

After training, the model was applied to reconstruct hourly energy exchange for out-of-sample users in a real REC in Italy during January 2025. The synthetic profiles were used to estimate *virtually shared energy*, defined as the hourly minimum between total prosumer grid injection and total consumer grid withdrawal, a key metric for determining REC financial incentives (Ministro dell'Ambiente e della Sicurezza Energetica, 2024).

The test case involved five residential users, including two prosumers with photovoltaic systems. They differ in terms of electrical appliances and configurations resulting in varied consumption behaviors. Model performance was evaluated against actual community energy data and a commonly used baseline method in Italy, referred to as "PVGIS+ARERA." This approach combines standard ARERA load profiles, which are selected by contracted power, day type, and location, with photovoltaic generation simulated using PVGIS based on system parameters and geographic data. Net electricity exchange was obtained by subtracting simulated PV production from estimated consumption. Shared energy values were then computed and compared across actual data, the CVAE model, and the PVGIS+ARERA method.

Results

This section presents the training results and demonstrates the CVAE model's ability to capture key electricity patterns and user typologies. Model performance is further evaluated within a real energy community, assessing its accuracy in estimating shared energy among users.

The hyperparameter optimization process identified the optimal CVAE architecture with a hidden layer size of 512 neurons, a latent space of 64 neurons, and a learning rate of 1e-3, achieving a final JS value of 9e-4. Figure 2 shows the loss function trend over training steps, depicting mean loss values for the training and validation sets at each epoch, marked by dashed vertical grey lines. With this configuration, training concluded at epoch 17 when the mean training loss reached the same order of

https://doi.org/10.26868/25222708.2025.1673

magnitude as the validation loss. Estimation results on instances in the test set led to a MAE of 204 Wh and a RMSE of 365 Wh.

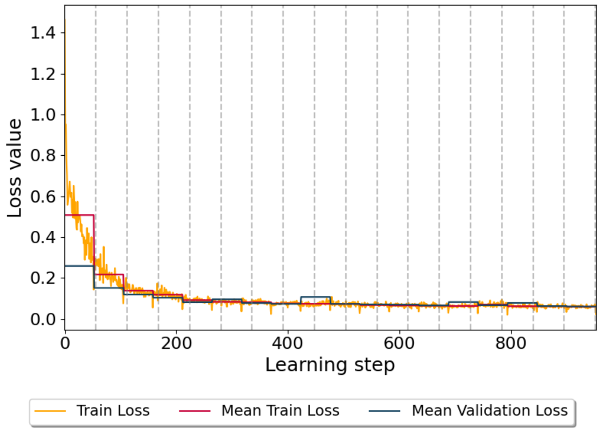

Figure 2: Trend of the loss function over the learning steps. The vertical dashed lines represent the epochs.

Figure 3 displays an example of load profiles generated on the test set for two different user types: a prosumer (Figure 3a) and a consumer (Figure 3b).

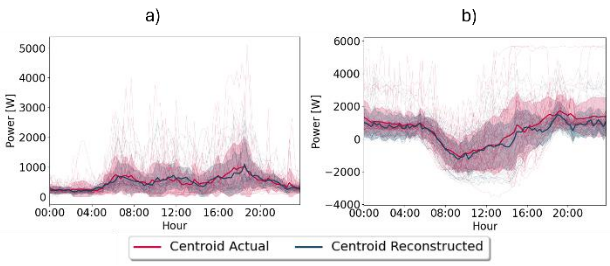

Figure 3: Comparison between real load profiles and reconstructed load profiles for a prosumer (Figure 3a) and a consumer (Figure 3b). The bold line represents the centroid, while the area corresponds to a one standard deviation.

The plot highlights the real centroid profile (in red) with the reconstructed centroid profile generated by the CVAE model (in blue), including the standard deviation at each hour. As shown, the overall distribution of the real and reconstructed load profiles is similar, demonstrating the model's ability to capture both daily and seasonal variations in energy consumption. This reflects its effective use and generalization of covariate information, such as irradiance dependence for the prosumer and peak consumption hours for the consumer.

To improve result interpretability and demonstrate the ability of the model to capture complex relationships, this section presents qualitative analyses on four key aspects. (i) Long-term dependencies between electrical energy consumption and external air temperature for users with thermally sensitive electric devices, such as electric heating and cooling systems. (ii) Long-term dependencies between surplus energy injected into the grid and external conditions, including solar irradiance and PV system size, for users equipped with PV systems. (iii) Preservation of total daily energy exchanged between buildings and the grid. (iv) Temporal alignment of peak demand events with actual usage patterns.

The relationship between electrical consumption and outdoor air temperature was analyzed to assess the model ability to capture thermal dependencies. Figure 4 presents this analysis on a daily scale, showing the correlation between mean daily external air temperature and daily energy consumption for four representative user typologies with different heating and cooling system configurations. Daily energy consumption was obtained by summing hourly values of energy withdrawn from the grid, as estimated by the CVAE model. The considered results pertain to the test set and to consumers without PV systems. Figure 4a shows that the model accurately captured the dependency between average daily outdoor air temperature and electricity consumption for users with an electrical cooling system, with the estimated trend closely matching the actual one. A similar correlation appears for users with electrical heating (Figure 4b) and those with both electrical heating and cooling (Figure 4c). In contrast, no significant correlation was found for buildings without electric heating or cooling (Figure 4d), confirming the model ability to effectively exploit metadata to condition the estimation process.

Secondly, the relationship between total daily irradiance and surplus energy exchanged with the grid was analyzed for users with a PV system. Since the CVAE model estimated net energy exchange, daily injected surplus was computed as the absolute sum of negative energy exchange values, representing the portion of PV-generated energy that exceeded user demand and was fed into the grid. Figure 5 presents the results for two prosumers using test set data. The first had a 6 kW PV system (Figure 5a), while the second a 3.2 kW system (Figure 5b). The model accurately estimated the overall magnitude of surplus energy injected into the grid,

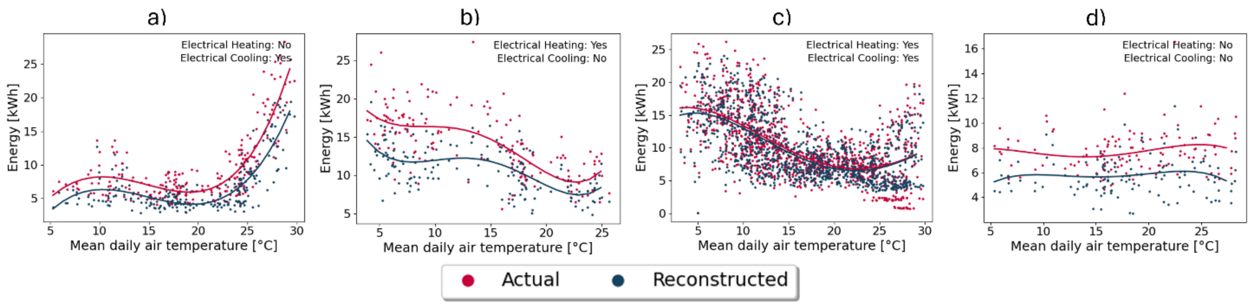

Figure 4: Thermal dependency analysis for four residential user types, differentiated by the presence of electrical heating or cooling systems. Actual data are shown as red dots, while estimated data are shown as blue dots.

https://doi.org/10.26868/25222708.2025.1673

capturing the trend for buildings with larger PV systems with minimal bias.

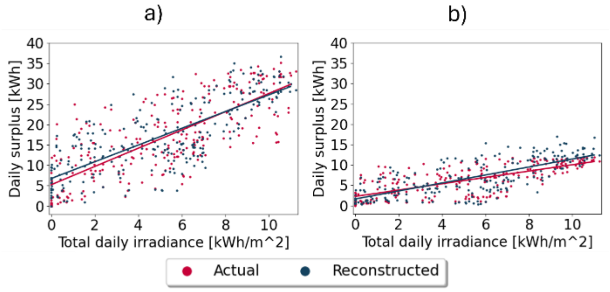

Figure 5: Relationship between total daily irradiance and surplus energy for two prosumers with PV systems of respectively 6 kW (a), and 3.2 kW (b). Actual data appears as red dots, estimated data as blue dots.

To evaluate the ability of the model to retain information on total daily energy exchanged with the grid, Figure 6 compares the distributions of total daily energy exchange across the train, validation, and test sets. The results showed that the model effectively captured the overall distribution, with the reconstructed values closely aligning with the true distributions.

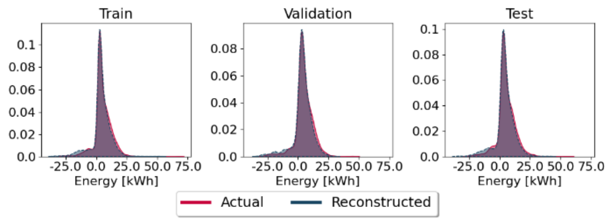

Figure 6: Distributions of the actual (red area) and estimated (blue area) total daily energy exchange with the grid across the train, validation, and test sets.

To assess the ability to estimate peak consumption hours, the alignment between On-peak hours extracted from the profile generated by the CVAE model and those identified in the actual consumption profile was evaluated. On-peak hours were determined based on the criteria described in the Methodology section. Figure 7 presents the confusion matrix for On-peak and Off-peak hour predictions on the test set. In 97.2% of cases, the timing of On-peak and Off-peak hours in the generated daily load profile aligned with the timing in the actual daily load profile, which served as the reference for model estimation.

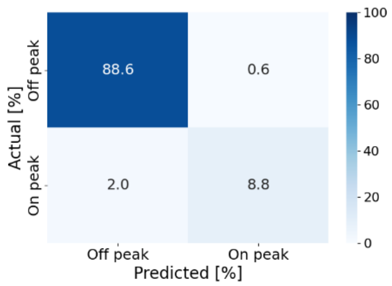

Figure 7: Confusion matrix for On-peak and Off-peak hours estimation.

To further evaluate model performance, the CVAE model was used to estimate the daily net energy exchange profiles for five energy community members (three prosumers and two consumers) and then assess the amount of virtually shared energy.

Figure 8 compares the results over the analyzed period of January 2025, showing actual community data alongside estimates from the CVAE model and the "ARERA+PVGIS" method. As previously explained, the "ARERA+PVGIS" method relies on predefined load profiles to estimate user energy demand. For users with PV systems, it uses the PVGIS tool to estimate solar energy production, which is then subtracted from the energy consumption profile to determine the net energy exchange profile.

While the "ARERA+PVGIS" method primarily captures the solar production component, it relies on averaged load profiles that do not account for the specific temporal variations in user consumption. This limitation results in a mismatch between the estimated and actual net energy exchange, as fluctuations in real demand and short-term variations are not properly reflected. In contrast, the CVAE model reconstructs the energy exchange profile by learning from actual data, allowing it to capture the dynamic nature of consumption patterns more accurately.

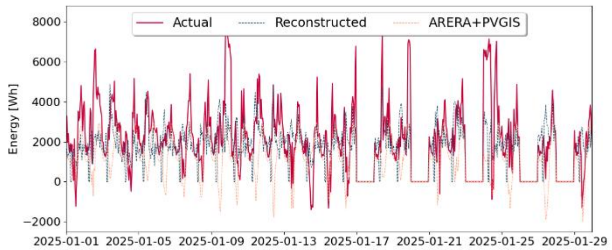

Figure 8: Comparison of net electricity exchange between the REC and the grid, considering actual values (red line), CVAE estimations (blue dashed line), and ARERA+PVGIS results (orange dashed line).

In this sense, Figure 9 presents a detailed error analysis, showing the density distribution of hourly errors (estimated – actual) for the total surplus energy exchanged between the entire community the grid (Figure 9a) and errors in estimating virtually shared energy (Figure 9b). The results confirm the strong performance of the CVAE model in both aspects.

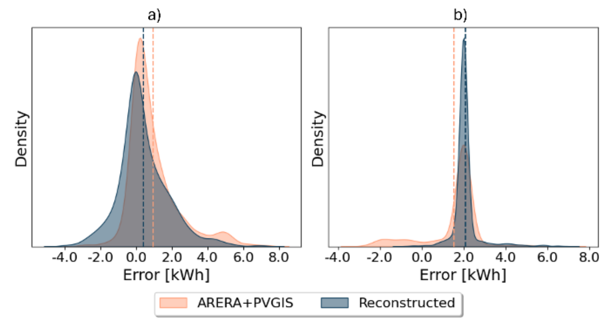

Figure 9:Distribution of errors regarding a) the total net energy exchange of the community and b) the amount of shared energy.

As shown in both figures, the mean absolute error (dashed line) of the CVAE model (Reconstructed) is closer to

https://doi.org/10.26868/25222708.2025.1673

zero, indicating low bias. In contrast, the "ARERA+PVGIS" method exhibits estimation bias, particularly for shared energy (Figure 9b), due to an overall underestimation of user energy demand.

Discussion

This paper introduces a deep learning-based approach for generating synthetic residential load profiles, leveraging key building metadata within a probabilistic framework. The results show that incorporating simple yet informative descriptors— such as the presence of electric heating and cooling systems, PV and battery storage, and usage patterns—preserves realism and variability in generated synthetic profiles. Differently from conventional methods based on average consumption patterns, the proposed model better captures residential electricity usage patterns. The model correctly identifies thermally sensitive electrical loads and accounts for on-site generation and usage patterns, leading to more accurate demand and production estimates. When applied to a real REC, the CVAE-based approach outperformed the "ARERA+PVGIS" method in estimating both total energy exchange and shared energy, that represent a reference approach for estimating the behaviour of residential users in RECs. The reduction in estimation errors confirms that integrating a metadata-driven conditioning strategy improves the quality of generated profiles, particularly valuable in data-scarce contexts where historical consumption data may be unavailable.

Several enhancements could improve the robustness and generalizability of this framework. Separating load and generation components would help distinguish energy consumption from on-site production, refining their individual contributions. Expanding the dataset with diverse residential load profiles would improve generalization across different building types and climates. Existing efforts, including those by Greater London Authority (2025), Eljand et al. (2023), and Goncalves et al. (2022), have focused on dataset development, and integrating such data would further enhance model generalizability.

Future work would explore the use of eXplainable AI (XAI) techniques to identify and refine the key features contributing to synthetic profile generation. Methods such as SHAP would be used to quantify feature importance and provide insights into the model decision-making process, thereby supporting metadata selection and improving model transparency. Additionally, exploring alternative generative models, such as diffusion models or temporal fusion transformers, which have shown promising results in other domains (Fu et al., 2024; Nazir et al., 2025). Additionally, refining loss functions, applying data augmentation to address imbalance, and incorporating advanced feature extraction methods could further improve model performance and result reliability.

Conclusions

This study demonstrated that incorporating metadata and contextual information into the modelling of residential energy usage patterns enhances the generation of realistic building electrical load profiles. A CVAE model was used for its ability to integrate structured contextual data within a probabilistic generative framework. Its simplicity and flexibility made it well-suited for modeling diverse electricity patterns across different user typologies.

The results showed that the model effectively captured net electricity exchange patterns, preserving long-term trends with temperature for thermally sensitive loads and solar radiation for PV-equipped buildings. It accurately mapped peak consumption periods based on building usage and reconstructed total daily energy exchanged with the grid. When tested on a real REC, the model outperformed a widely used method in estimating both total electricity exchange and virtually shared energy. Eventually the discussion highlighted the need for a more comprehensive approach, integrating additional datasets and refining model architecture and training techniques to improve generalization.

Acknowledgement

The work of Alfonso Capozzoli and Marco Savino Piscitelli was carried out within the project FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (Piano Nazionale di Ripresa E Resilenza (PNRR) – Missione 4 Componente 2, Investimento 1.3 – D.D. 1555 11/10/2022, PE00000013). The work of Rocco Giudice was made in the framework of the project PNRR-NGEU which has received funding from the Italian Ministry of University and Research (MUR) – DM 351/2022.

References

Arbulu, M., Perez-Bezos, S., Figueroa-Lopez, A., & Oregi, X. (2024). Opportunities and barriers of calibrating residential building performance simulation models using monitored and survey-based occupant behavioural data: A case study in Northern Spain. *Buildings, 14(7), 1911*.

ARERA. (2023). Analisi dei consumi dei clienti domestici. Retrieved March 1, 2025, from https://www.arera.it/dati-e-statistiche/dettaglio/analisi-dei-consumi-dei-clienti-domestici.

Baasch, G., Rousseau, G., & Evins, R. (2021). A Conditional Generative adversarial Network for energy use in multiple buildings using scarce data. *Energy and AI, 5, 100087*.

Binderbauer, P. J., Kienberger, T., & Staubmann, T. (2022). Synthetic load profile generation for production chains in energy intensive industrial subsectors via a bottom-up approach. *Journal of Cleaner Production, 331, 130024*.

Charwand, M., Gitizadeh, M., Siano, P., Chicco, G., & Moshavash, Z. (2020). Clustering of electrical load patterns and time periods using uncertainty-based multi-level amplitude thresholding. International *Journal of Electrical Power & Energy Systems, 117, 105624*.

https://doi.org/10.26868/25222708.2025.1673

Claeys, R., Cleenwerck, R., Knockaert, J., & Desmet, J. (2024). Capturing multiscale temporal dynamics in synthetic residential load profiles through Generative Adversarial Networks (GANs). *Applied Energy, 360.*

Dent, I., Craig, T., Aickelin, U., & Rodden, T. (2014). Variability of behaviour in electricity load profile clustering; Who does things at the same time each day?. *In Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014.*

Diao, L., Sun, Y., Chen, Z., & Chen, J. (2017). Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings, 147, 47-66..*

Eljand, K., Laid, M., Scellier, J.-B., Dane, S., Demkin, M., & Howard, A. (2023). Enefit - Predict energy behavior of prosumers. *Kaggle.*

Fu, C., Kazmi, H., Quintana, M., & Miller, C. (2024). Creating synthetic energy meter data using conditional diffusion and building metadata. *Energy and Buildings, 312, 114216.*

Gallo, A., & Capozzoli, A. (2024). The role of advanced energy management strategies to operate flexibility sources in Renewable Energy Communities. *Energy and Buildings, 325, 115043.*

Gallo, A., Piscitelli, M. S., & Capozzoli, A. (2023). RECsim—Virtual Testbed for Control Strategies Implementation in Renewable Energy Communities. *Sustainability in Energy and Buildings 2022. SEB 2022. Smart Innovation, Systems and Technologies. Singapore.*

Ge, S., Li, J., Liu, H., Liu, X., Wang, Y., & Zhou, H. (2019). Domestic Energy Consumption Modeling per Physical Characteristics and Behavioral Factors. *Energy Procedia, 158, 2512-2517.*

Giannuzzo, L., Minuto, F. D., Schiera, D. S., & Lanzini, A. (2024). Reconstructing hourly residential electrical load profiles for Renewable Energy Communities using non-intrusive machine learning techniques. *Energy and AI, 15, 100329..*

Giudice, R., Piscitelli, M. S., Chiosa, R., & Capozzoli, A. (2025). A Data-Driven Process for Optimal Incentive Sharing in Collective Self-Consumption Groups of Residential Users. *Multiphysics and Multiscale Building Physics. IABP 2024. Lecture Notes in Civil Engineering. Singapore.*

Goncalves, C., Barreto, R., Faria, P., Gomes, L., & Vale, Z. (2022). Dataset of an energy community's generation and consumption with appliance allocation. *Data in Brief, 45*, 108590.

Greater London Authority. (2025). SmartMeter energy use data in London households. *London Datastore.* from https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

Gu, Y., Chen, Q., Liu, K., Xie, L., & Kang, C. (2019). GAN-based Model for Residential Load Generation Considering Typical Consumption Patterns. *2019 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2019.*

Li, J., Chen, Z., Cheng, L., & Liu, X. (2022). Energy data generation with Wasserstein Deep Convolutional Generative Adversarial Networks. *Energy, 257.*

Ministro dell'Ambiente e della Sicurezza Energetica. (2024). Decreto CER, https://www.mase.gov.it/comunicati/energia-mase-pubblicato-decreto-cer.

Nazir, A., Shaikh, A. K., Shah, A. S., & Khalil, A. (2023). Forecasting energy consumption demand of customers in smart grid using Temporal Fusion Transformer (TFT). *Results in Engineering, 17.*

Schito, E., Taverni, L., Conti, P., & Testi, D. (2024). Optimization of the composition of residential buildings in a renewable energy community based on monitored data. *Optimization and Engineering, 1–19.*

Thorve, S., Baek, Y. Y., Swarup, S., Mortveit, H., Marathe, A., Vullikanti, A., & Marathe, M. (2023). High resolution synthetic residential energy use profiles for the United States. *Scientific Data. Journal of Building Performance Simulation, 8*, 39–43.

Tian, C., Ye, Y., Lou, Y., Zuo, W., Zhang, G., & Li, C. (2022). Daily power demand prediction for buildings at a large scale using a hybrid of physics-based model and generative adversarial network. *Building Simulation, 15, 1685-1701*

Wang, C., Tindemans, S. H., & Palensky, P. (2022). Generating Contextual Load Profiles Using a Conditional Variational Autoencoder. *IEEE PES Innovative Smart Grid Technologies Conference Europe, 2022.*

Wang, Z., & Hong, T. (2020). Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN). *Energy and Buildings, 224, 110299.*

Xia, W., Huang, H., Duque, E. M. S., Hou, S., Palensky, P., & Vergara, P. P. (2024). Comparative assessment of generative models for transformer- and consumer-level load profiles generation. *Sustainable Energy, Grids and Networks, 38.*

Yang, W., Li, X., Chen, C., & Hong, J. (2022). Characterizing residential load patterns on multi-time scales utilizing LSTM autoencoder and electricity consumption data. *Sustainable Cities and Society*

https://doi.org/10.26868/25222708.2025.1673