



HAL
open science

What insights can we draw from our residential energy models?: guidelines for future modelling exercises

Matthias Heinrich, Marie Ruellan, Jean-Pierre Lévy, Allou Samé, Latifa Oukhellou

► To cite this version:

Matthias Heinrich, Marie Ruellan, Jean-Pierre Lévy, Allou Samé, Latifa Oukhellou. What insights can we draw from our residential energy models?: guidelines for future modelling exercises. The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Nov 2022, Boston, United States. 10.1145/3563357.3567833 . hal-03892639

HAL Id: hal-03892639

<https://hal.science/hal-03892639v1>

Submitted on 23 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

What insights can we draw from our residential energy models? Guidelines for future modelling exercises

Matthias Heinrich[†]
LATTS, COSYS-GRETTIA
Ecole des Ponts, Univ
Gustave Eiffel, CNRS
Marne-la-Vallée, France
matthias.heinrich@enpc.fr

Marie Ruellan
SATIE
CY Cergy Paris Université
Cergy, France
marie.ruellan@cyu.fr

Jean-Pierre Lévy
LATTS
Ecole des Ponts, Univ
Gustave Eiffel, CNRS
Marne-la-Vallée, France
jean-pierre.levy@enpc.fr

Allou Samé
COSYS-GRETTIA
Université Gustave Eiffel
Marne-la-Vallée, France
allou.same@univ-eiffel.fr

Latifa Oukhellou
COSYS-GRETTIA
Université Gustave Eiffel
Marne-la-Vallée, France
latifa.oukhellou@univ-eiffel.fr

ABSTRACT

Understanding energy consumption in the residential sector remains challenging. In current regression models, prediction performance on a test dataset is frequently used to justify the ability of the associated model to correctly describe the energy consumption. However, prediction and explanation are often conflicting goals in modelling and most models convey heuristics rather than a precise representation of the processes involved. From this point of view, comparing and analyzing several modelling strategies is a way to improve our comprehension. In this work, we studied both the prediction performance and the heuristics provided by nine models of housing energy consumption built from survey data of 2000 French households. Concerning prediction, the eXtreme Gradient Boosting algorithm based on features extracted from surveys provided the best predictor with an average MAPE of 38%. Residual analysis showed that the models generally make the same type of errors, with absolute errors being much larger for houses than for apartments. This suggests that the improvement in predictive performance relies more on the introduction of supplementary predictive variables than on increasing the model's complexity. In terms of interpretation, all models identify building type, floor area, household composition and heating energy as key determinants of final energy consumption. We illustrate, through t-SNE mapping, that these variables are in fact very much interrelated. The study therefore underlines the high interest of studying the interaction between the variables. Explanatory and predictive models could therefore be based on an adapted segmentation of dwellings and households. This type of modelling could also join the work in the sociology of energy which largely describes the interweaving of these variables within lifestyles.

KEYWORDS

Buildings, Energy model, Heuristics, XGBoost, Prediction

ACM Reference format:

Matthias Heinrich, Marie Ruellan, Allou Samé, Jean Pierre Lévy, Latifa Oukhellou. November 2022. What insights can we draw from our residential energy models? Guidelines for prediction and interpretation. In *Proceedings of ACM International Conference on Systems for Energy-Efficient Built Environments (2022)*. ACM, Boston, USA, 4 pages.

1 Introduction

1.1 Problem & context

The building sector accounts for nearly 40% of global energy consumption and 20% of greenhouse gas emissions and, as stipulated by the Paris Agreement, this sector must implement drastic reductions in consumption [1]. Numerical modelling is a major tool in this sector as it allows one to better understand and quantify the effect of different parameters such as insulation, inhabitants' behaviours, climate, and equipment [2]–[4]. However, models have difficulty in estimating energy consumption accurately - a phenomenon known as the *energy performance gap* [5], which reflects a limited understanding of the processes at work, limited data, and imperfect models. As a result, there is a large body of literature in various disciplines including data-science, engineering, anthropology and sociology at all spatial and temporal scales [5], [6].

Many physical and data-based numerical models have been developed. The purpose can be to predict residential consumption at a given spatial and temporal scale [4], [6], or to propose an explanation by inferring an understanding of the phenomenon [7]–[9]. Several reviews have been carried out in recent years [4], [6], [10], [11] to assess the state-of-the-art in the construction of predictive models. These studies divide the models into 3 categories: engineering methods such as dynamic thermal simulation, statistical methods such as multiple linear regression (MLR), conditional demand analysis, and artificial intelligence methods such as artificial neural networks [12], support vector machine [13], random forest [14], or extreme gradient boosting (XGB) [15]. The review papers point out that a minority of the work reviewed has focused on residential buildings notably because of data availability. A large part of the literature focuses on short term prediction based on sensor data [16]. Other papers mobilize building, household and behaviour data to model long term energy forecasts [2], [17]. These can either focus on consumption related to end uses, such as heating, cooling, ventilation, or lighting, or on a type of energy (electricity, wood, gas, fuel oil, etc.). The investigations discuss architecture selection and dimension reduction through feature extraction strategies to improve

prediction: they apply linear techniques such as factor analysis, or non-linear techniques such as auto-encoding using neural networks. As stated above, these predictive models suffer from considerable imprecision, which is greater for predictive models with a short time horizon (hour, day), or reduced spatial resolution compared to annual models or models at a higher spatial scale. In general, machine learning modelling provides better predictors than engineering modelling.

These models are often used also to provide explanations about energy consumption in dwellings. Regression models are widely used to identify and quantify the causal effects of explanatory variables such as household, housing, and behavioural characteristics on energy consumption [2], [18]–[21] because they generally provide a good predictive performance, thus making validation possible, and because they carry with them a strong theoretical framework based on the additivity of individual effects "all other things being equal" [22]. An example is the work of Belaïd [18] who used a structural equation modelling approach to estimate the energy consumption of 2356 French households. To do so, he used data from the 2013 PHEBUS survey conducted by the French Ministry of Ecology. The data describe households and dwellings, behaviours, and final energy consumption. After validating the model, the article quantified the direct and indirect effects of household and housing characteristics on their energy consumption to support proposals for public action, in particular the thermal renovation of housing, which is shown to have a greater impact than household characteristics. Machine learning models such as decision trees and ensemble models can also provide useful information [23]. Decision trees provide a set of rules based on variables and thresholds. The variables used in ensemble models such as random forest and XGB can be compared in terms of their contribution to the accuracy of the algorithm.

However, it is important to mention that there is no consensus in the scientific community about the interest of validating or not the predictive performances of the models before drawing any conclusions from them. Shmueli [22] reminds us that predicting and explaining are two different goals that should not be confused. He shows that it is sometimes possible to build a "false" model capable of providing better predictions than the "true" model. Furthermore, in a seminal article published in *Science* in 1994, Oreskes [24] recalled that: "Models can only be evaluated in relative terms, and their predictive value is always open to question. The primary value of models is heuristic." In this perspective, the predictive performance of the models occupies a more nuanced place: rather than simply justifying an interpretation or validating a set of processes, it informs on the model's capacity to reproduce the correlations through a set of heuristics. At this point, it is possible to formulate two different questions for these models: what are the most suitable modelling structures and features to build a good predictor? What insights are provided by models of diverse complexity?

We propose in this work first to compare both the predictive and explanation performance from models classically used in the literature. This includes the comparison of 3 models of different complexity (decision tree, multiple linear regression, XGB model) and 3 feature extraction strategies (raw variables, principal components, and variables from autoencoder training). Then, by analysing these results, confronting them with a t-SNE observation of the data and by mobilising the literature in the sociology of energy, we try to justify the interest of a segment-based approach to build more accurate and realistic models.

1.2 Methodology, data, and models

The methodology followed is summarised in the figure below.

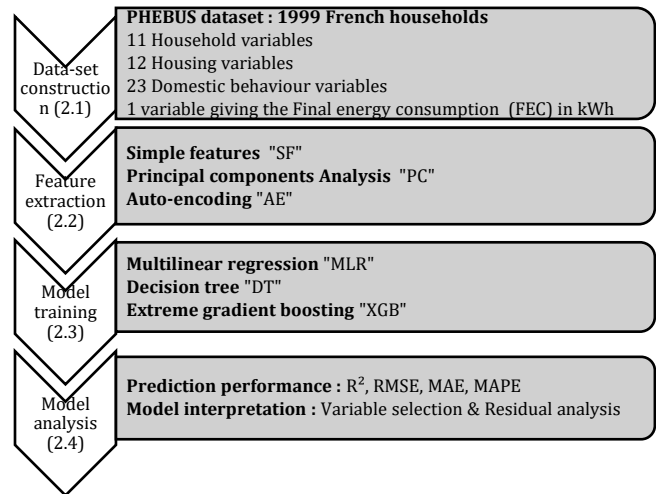


Figure 1: Methodology of this work

The study uses data from the 2013 PHEBUS survey conducted by the statistical department of the current French Ministry of Ecological Transition [25]. The sample consists of 8000 dwellings representative of French regions, climatic zones, housing types and years of construction. The data collected describe the occupied dwellings, including heating and renovations that have been carried out, the socio-demographic characteristics of households, energy consumption and energy behaviour. We used 47 variables from this survey, 11 of which describe households (income, savings, composition, age of the reference person, number of person and consumption units, socio-professional group, ...), 1 final energy consumption (noted FEC, as Final Energy Consumption), 12 dwellings (type, surface, number of rooms and bedrooms, occupancy status, date of construction, urban area, insulation work, Unified Degree Days, type of heater and age of heater, ...) and 23 residential behaviour (including average attendance, number of equipment's, declared heating temperature, number of showers per day, heating regulation level, heating deprivation,...). The dataset comprises 1999 households for which the data are complete. The data used are not freely available but can be requested with the agreement of the producer and the consent of a statistical confidentiality committee. In this modelling work the dependent variable is the logarithm of final energy consumption. The numerical variables are standardised and centred. In this study we

built a total of 9 final energy consumption models. For that purpose, we combined 3 feature extraction strategies (Simple Features “SF”, Principal Components “PC”, and Autoencoding “AE”) and 3 model frameworks (decision tree “DT” [23], multiple linear regression model “MLR” [23], eXtreme Gradient Boosting “XGB” [26]). A few details on the features extraction strategies are given here. The SF are coded as dummy variables for the multiple linear regression (MLR) and by one hot coding (OHE) for the XGB training. PC feature extraction uses Factor Analysis on Mixed Data (FAMD)

[27]. By performing three FAMDs on the groups of household, housing, and energy behaviour variables, we generated a total of 52 synthetic variables to capture 80% of the variance each time. The calculations were performed on R using the *FactoMineR* library. The AE variables were generated by neural networks called auto-encoders. As the data are sparse, we used a cosine loss function [28]. The size of the hidden layer was chosen to be 15 to achieve a threshold of -0.85 of the loss function. The encoding calculations were performed in Python using the Keras library [29].

Performance metric	R ² (%)			RMSE (MWh)			MAE (MWh)			MAPE (%)		
	SF	PC	AE	SF	PC	AE	SF	PC	AE	SF	PC	AE
Feature extraction												
Decision tree	51.7	50.2	35.7	7.8	7.9	9.1	5.0	5.1	6.0	43.4	44.1	53.4
Multilinear regression	60.7	58.5	42.8	7.0	7.2	8.5	4.5	4.7	5.6	38.4	39.8	47.3
Extreme Gradient Boosting	63.1	57.8	46.4	6.8	7.3	8.3	4.4	4.7	5.4	37.7	39.6	46.0

Table 1: Average prediction performance for all models (DT, MLR, XGB) and feature extraction strategies (SF, PC, AE). The average indicators are calculated from 50 runs with random selection of test datasets. Top scores per column are shown in bold. (Data: Secondary processing from PHEBUS data).

2 Model performance

2.1 Prediction performance

Each model was trained on 80% of the database, i.e. 1600 households. They were then evaluated on their predictive performance, i.e. their ability to predict the annual final energy consumption of the 20% of households not included in the training set. To reduce the dependence of this evaluation on data selection, the models were trained and evaluated 50 times on randomly drawn samples. 4 indicators were calculated from the predictions of the models (R², Root Mean Square Error “RMSE” in kWh, Mean Absolute Error “MAE” in kWh, Mean Absolute Percentage Error “MAPE” in percent). The average of these 4 indicators over the 50 trainings are given in Table 1. We used the *caret* library implemented in R. As a result, the choice of the metric has little or no impact on the ranking of the models. We see that the XGB-SF model offers the best predictive performance, slightly ahead of the MLR-SF multiple linear model. Secondly, we observe that feature extraction and the resulting loss of information degrades the predictive performance of the models.

2.2 Model interpretation

The analysis of the variables selected by the models generally makes it possible to draw insights about the phenomenon which is modelled. After computing several models, we noticed that with two notable exceptions (the variables describing the amount of sparings and the declared heating temperature) all the variables selected by DT were also selected by the XGB model. And all the variables selected by XGB were also selected by the MLR model. Whatever the feature extraction strategy and the model, the variables describing the dwelling, the heating energy, and the size of the dwelling were selected first. This confirms the important role of building characteristics on the FEC. However, the variables characterising the household and domestic behaviours occupy higher or lower ranks, depending on the training sample, but also

on the feature extraction strategy. We compared the residuals of the MLR-SF, DT-SF and XGB-SF models to observe whether the modelling structure allows for a change in the nature of the errors or whether, on the contrary, all the errors are the same, which may then highlight a problem with the data or the modelling assumptions made. We observed that the residuals of the three models were similar and that the models tend to underestimate (resp. overestimate) low (resp. high) consumption levels.

3 Conclusion and key messages : How to design more efficient energy models?

Modellers would like to be able to provide a generic answer to explain and predict housing energy consumption through the abstractions and variables of the physical and social phenomena that models carry out. Obviously, the performance of current predictive models is modest as underlined by the indicators of the best predictor XGB-SF: R² ~ 63.1%, RMSE ~ 6.8 MWh, MAE ~ 4.4 MWh, MAPE ~ 37.7% and the modelling interpretation says not much more than that the characteristics of the dwelling are determinant. Depending on the target, we would consider two key messages for future energy modelling exercises.

For prediction: search for more (and better) data. We found in this work that making the features and structure of the regression model more complex brought relatively little prediction performance. Furthermore, the errors in the models were of the same type, which suggests that the “raw” variables are good but that some are missing or that the data were not good enough.

For explanation: Consider the significant correlations between the variables. The calculation of independence tests between the characteristic variables of the household (composition of the household, age of the reference person, occupation status), the dwelling (surface, type, type of heating) and the domestic behaviours give very significant p-values (<5%). To observe these correlations, we project the households and dwellings in a two-dimensional space. The space is constructed by t-SNE (see Maaten

and Hinton [30]) on the variables after one-hot encoding with a perplexity parameter of 50, which provided stable and relevant results. Note that the criterion for constructing the space is the relative position of individuals, which means that the axes cannot be interpreted with the initial variables. In figure 2, we have plotted the households map by intersecting it with variables such as surface area, type of dwelling, household composition, logarithm of final energy consumption, number of persons. The variable FEC was not used to create this synthetic space. Several remarks can be made. Firstly, the map allows several distinct and homogeneous zones to be distinguished for each variable. For example, the map describing the housing type highlights the fact that flats are mainly located on the left while single-family houses are mainly on the right. Additionally, regarding household composition, we can see that families with children are present in two distinct blocks on the right and left of the graph (in fact, it is the most affluent households on the right and the most deprived households on the left). Single people are positioned in the center while couples without children are located on the lower part. When plotting the maps by highlighting the level of energy consumption, we can see a strong separation between the less energy consuming flats and the more energy consuming detached houses. The visualization is interesting because it allows for going beyond a simple correlation and independence analysis: in particular, it suggests exploring a

database segmentation exercise. This approach would make it possible to carry out separate models for flats and single-family houses, since the dynamics in terms of consumption seem to be very different. Also, modelling by segment would make it possible to account for the recurrent associations between the modalities of the different variables observed here in the data (e.g. couples without children living in a small - most likely old - gas-heated detached house). These associations are consistent with results from the sociology of energy, which describes the association between types of household, types of housing, types of energy [31] and behaviour [32]. We believe that it is not certain that segment-based modelling will significantly improve the predictive performance of housing energy consumption models. However, segment-based models make it possible, if not to quantify precisely, to evaluate in relative terms the effect of any changes in this segment (insulation, regulation behaviour, energy efficiency of equipment, etc.).

ACKNOWLEDGMENTS

This work was partially funded by the French Ministry of Ecological Transition. We would like to thank the reviewers and the editor for their comments and suggestions which allowed to significantly improve the article.

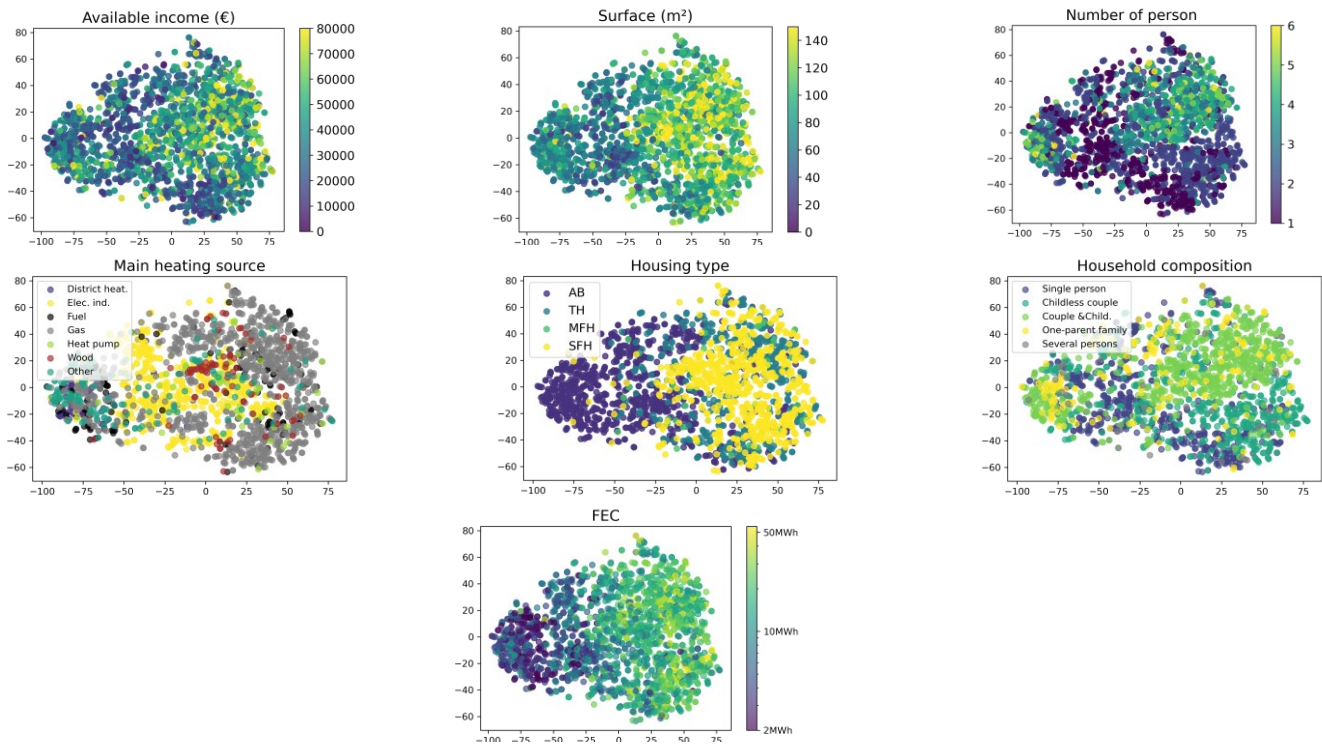


Figure 2: t-SNE visualization of household, housing, and behaviour variables. The final energy consumption variable (FEC) is not used in the calculation. (Data: Secondary processing from PHEBUS data). Abbreviations: AB: Apartment Block; TH: Terraced House; MFH: Multi-Family house; SFH: Single Family House

REFERENCES

- [1] K. Lane, 'IEA Energy Efficiency 2018 and World Energy Outlook 2018', p. 36, 2018.
- [2] A. F. Emery and C. J. Kippenhan, 'A long term study of residential home heating consumption and the effect of occupant behavior on homes in the Pacific Northwest constructed according to improved thermal standards', *Energy*, vol. 31, no. 5, pp. 677–693, Apr. 2006, doi: 10.1016/j.energy.2005.04.006.
- [3] M. Brøgger and K. B. Wittchen, 'Estimating the energy-saving potential in national building stocks – A methodology review', *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1489–1496, Feb. 2018, doi: 10.1016/j.rser.2017.05.239.
- [4] A. Fouquier, S. Robert, F. Suard, L. Stéphan, and A. Jay, 'State of the art in building modelling and energy performances prediction: A review', *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 272–288, Jul. 2013, doi: 10.1016/j.rser.2013.03.004.
- [5] P. X. W. Zou, X. Xu, J. Sanjayan, and J. Wang, 'Review of 10 years research on building energy performance gap: Life-cycle and stakeholder perspectives', *Energy and Buildings*, vol. 178, pp. 165–181, Nov. 2018, doi: 10.1016/j.enbuild.2018.08.040.
- [6] K. Amasyali and N. M. El-Gohary, 'A review of data-driven building energy consumption prediction studies', *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, Jan. 2018, doi: 10.1016/j.rser.2017.04.095.
- [7] M. Aydinlp-Koksal and V. I. Ugursal, 'Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector', *Applied Energy*, vol. 85, no. 4, pp. 271–296, Apr. 2008, doi: 10.1016/j.apenergy.2006.09.012.
- [8] F. Belaïd, S. Bakaloglou, and D. Roubaud, 'Direct rebound effect of residential gas demand: Empirical evidence from France', *Energy Policy*, vol. 115, pp. 23–31, Apr. 2018, doi: 10.1016/j.enpol.2017.12.040.
- [9] H. Estiri and E. Zagheni, 'Age matters: Ageing and household energy demand in the United States', *Energy Research & Social Science*, vol. 55, pp. 62–70, Sep. 2019, doi: 10.1016/j.erss.2019.05.006.
- [10] H. Zhao and F. Magoulès, 'A review on the prediction of building energy consumption', *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, Aug. 2012, doi: 10.1016/j.rser.2012.02.049.
- [11] Z. Liu *et al.*, 'Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction', *Energy Exploration & Exploitation*, vol. 37, no. 4, pp. 1426–1451, Jul. 2019, doi: 10.1177/0144598718822400.
- [12] M. W. Ahmad, M. Mourshed, and Y. Rezgui, 'Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption', *Energy and Buildings*, vol. 147, pp. 77–89, Jul. 2017, doi: 10.1016/j.enbuild.2017.04.038.
- [13] Z. Ma, C. Ye, H. Li, and W. Ma, 'Applying support vector machines to predict building energy consumption in China', *Energy Procedia*, vol. 152, pp. 780–786, Oct. 2018, doi: 10.1016/j.egypro.2018.09.245.
- [14] J. Ma and J. C. P. Cheng, 'Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests', *Applied Energy*, vol. 183, pp. 193–201, Dec. 2016, doi: 10.1016/j.apenergy.2016.08.096.
- [15] L. T. Le, H. Nguyen, J. Zhou, J. Dou, and H. Moayedi, 'Estimating the Heating Load of Buildings for Smart City Planning Using a Novel Artificial Intelligence Technique PSO-XGBoost', *Applied Sciences*, vol. 9, no. 13, Art. no. 13, Jan. 2019, doi: 10.3390/app9132714.
- [16] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, 'Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network', *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.
- [17] B. Allibe, 'Modélisation des consommations d'énergie du secteur résidentiel français à long terme - Amélioration du réalisme comportemental et scénarios volontaristes', p. 362, 2012.
- [18] F. Belaïd, 'Untangling the complexity of the direct and indirect determinants of the residential energy consumption in France: Quantitative analysis using a structural equation modeling approach', *Energy Policy*, vol. 110, pp. 246–256, Nov. 2017, doi: 10.1016/j.enpol.2017.08.027.
- [19] N. Fumo and M. A. Rafe Biswas, 'Regression analysis for prediction of residential energy consumption', *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, Jul. 2015, doi: 10.1016/j.rser.2015.03.035.
- [20] A. Satre-Meloy, 'Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models', *Energy*, vol. 174, pp. 148–168, May 2019, doi: 10.1016/j.energy.2019.01.157.
- [21] O. Guerra Santin, L. Itard, and H. Visscher, 'The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock', *Energy and Buildings*, vol. 41, no. 11, pp. 1223–1232, Nov. 2009, doi: 10.1016/j.enbuild.2009.07.002.
- [22] G. Shmueli, 'To Explain or to Predict?', *Statistical Science*, vol. 25, no. 3, pp. 289–310, Aug. 2010, doi: 10.1214/10-STS330.
- [23] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [24] N. Oreskes, K. Shrader-Frechette, and K. Belitz, 'Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences', *Science*, vol. 263, no. 5147, pp. 641–646, Feb. 1994, doi: 10.1126/science.263.5147.641.
- [25] 'Enquête Performance de l'Habitat, Équipements, Besoins et Usages de l'énergie (Phébus) | Données et études statistiques'. <https://www.statistiques.developpement-durable.gouv.fr/enquete-performance-de-lhabitat-equipements-besoins-et-usages-de-lenergie-phebus> (accessed Jul. 04, 2022).
- [26] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [27] J. Pagès, 'Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes', *Revue de statistique appliquée*, pp. 5–37, 2002.
- [28] Tensorflow, 'Keras Losses - CosineSimilarity', *TensorFlow*. https://www.tensorflow.org/api_docs/python/tf/keras/losses/CosineSimilarity (accessed Jul. 16, 2022).
- [29] F. Chollet, 'Keras'. 2015. [Online]. Available: <https://keras.io>
- [30] L. van der Maaten and G. Hinton, 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
- [31] J.-P. Lévy and F. Belaïd, 'The determinants of domestic energy consumption in France: Energy modes, habitat, households and life cycles', *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 2104–2114, Jan. 2018, doi: 10.1016/j.rser.2017.06.022.
- [32] M. Heinrich, M. Ruellan, L. Oukhellou, A. Samé, and J.-P. Lévy, 'From energy behaviours to lifestyles: contribution of behavioural archetypes to the description of energy consumption patterns in the residential sector', *Energy and Buildings (minor review)*, 2022.