Predicting CO2 Emissions for Buildings Using Regression and Classification

Alexia Avramidou, Christos Tjortjis

Predicting CO_2 Emissions for Buildings Using Regression and Classification

Avramidou Alexia[1] and Christos Tjortjis[1][0000-0001-8263-9024]

[1]The Data Mining and Analytics Research Group, School of Science & Technology, International Hellenic University, Thessaloniki, Greece
c.tjortjis@ihu.edu.gr

Abstract. This paper presents the development of regression and classification algorithms to predict greenhouse gas emissions caused by the building sector, and identify key building characteristics, which lead to excessive emissions. More specifically, two problems are addressed: the prediction of metric tons of CO_2 emitted annually by a building, and building compliance to environmental laws according to its physical characteristics, such as energy, fuel, and water consumption. The experimental results show that energy use intensity and natural gas use are significant factors for decarbonizing the building sector.

Keywords: Greenhouse gas (GHG) emissions prediction, Machine Learning (ML), Data Mining, buildings, energy consumption, regression, classification

1 Introduction

It is widely known that climate change is a global threat; immediate actions need to be taken to limit its most important side effects. The operation of buildings accounts for approximately 40% of primary energy consumption globally, drawing the attention of governments to act instantly by adopting energy policies and carbon emission measures **Error! Reference source not found.**]. Given this reality, countries and cities have already set strict long-term energy efficiency and carbon reduction goals for existing and new buildings. To support global and city-scale decarbonization goals, energy disclosure directives are a significant policy tool to accelerate the transition towards climate neutrality [2].

This work evaluates several regression and classification algorithms, for predicting the annual greenhouse gas (GHG) emissions using properties reported at energy disclosure records. It proposes a methodology for emissions prediction utilizing feature engineering on energy and fuel consumption data, collected from large residential and public buildings in New York. The feature selection and engineering phase includes grouping buildings into 9 main categories according to their type and applying a logarithmic transformation to the Total GHG emissions to eliminate outliers. Also, high correlated features are removed from the analysis. In addition, this work analyzes various classification algorithms for predicting compliance to environmental laws. For this problem, the same data source is used, combined with emissions limits provided by the Local Law 97 (LL97), for two compliance periods.

This work differs from standard data-driven predictive models for the building sector in several ways. Although there are numerous studies discussing energy waste and performance predictive models for buildings, there is limited research focusing on forecasting emissions. Also, emissions prediction at a building-scale level has not been used so far, as most of the relevant studies mainly present their results at a city-scale or country-scale level [21, 22]. So, this work tries to fill this gap in the literature by predicting GHG emissions caused by the building sector at a building-scale level, analyzing their spatial characteristics and behavior. Additionally, our contribution could help governments and building owners understand the environmental footprint of buildings and take actions for energy efficiency and decarbonization.

The paper is structured as follows: Section 2 presents background information and reviews the literature, Section 3 provides the problem definition along with a brief description of the datasets, the pre-processing steps and the methods used. Section 4 presents exploratory data analysis results along with predictions. In Section 5 results are discussed and evaluated and Section 6 concludes the paper with future directions.

2 Literature Review

Building performance and energy consumption has been the subject of abundant academic research, driven by the need for a "greener" building sector. Unsupervised data analytics and clustering techniques are considered more practical and promising in discovering knowledge given limited prior information, concerning building operational and consumption data [3].

K-means clustering has been used to identify buildings with similar temporal energy performance patterns [4,5]. Also, clustering tenants' behavior has proven that there is a strong relationship between the number of bedrooms and energy consumption, as well as home working [6]. In addition, K-means has been applied to group educational buildings according to their energy performance for space heating and evaluate energy savings in the building sector [7]. Another study used K-means to cluster school buildings to create a priority list for retrofit measures [8].

Furthermore, Artificial Neural Networks (ANNs) are commonly used in such problems because of their high predictive power [9]. However, their implementation is challenging because several hyperparameters need to be adjusted for accurate results [10]. ANNs are often compared with ensemble methods like Random Forest. In [11] it is mentioned that ANNs performed marginally better than Random Forest in predicting hourly HVAC energy consumption, but ensemble methods tend to deal with multidimensional data better. Also, fuzzy systems and ANNs using occupancy data are used to describe how energy is consumed within a building [12].

Another study compares ANNs with Support Vector Machines (SVM) for predicting building energy consumption in four office buildings [13]. The results have shown that SVM performs better than ANNs and the reason could be the small data pool used in this study, thus abnormal data were not so frequent. Also, when applying SVM for prediction someone needs less hyperparameters to optimize compared to ANNs. Additionally, Support Vector Regression (SVR) has been used to develop

sensor-based forecasting models for residential buildings [14] and to improve energy efficiency of HVAC systems analyzing historical data for buildings [15].

A common practice in predicting electricity consumption is to transform a regression model to a binary classification problem with 'high' and 'low' target labels [16]. It is stated that turning the regression problem to a binary one, achieves better results when the point of separation is the mean of all instances [17,18].

Another work focuses on generalizing self-reported energy data from a small sample of buildings to a city-scale level [19]. Three different Machine Learning (ML) algorithms are used, namely Linear Regression, Random Forest, and SVR, along with feature selection techniques to make predictions from the Local Law 84 (LL84) self-reported energy disclosure data for large buildings. The results showed that Linear Regression performs better when predicting total building energy consumption at the zip code-level for the entire city, while SVR performs better in terms of accuracy when estimating energy use within the sample of LL84 buildings. Also, building size, use and morphology seem to be significant attributes for energy use prediction at the building and zip code levels. Larger buildings are found to have smaller Energy Use Intensity (EUI), while taller ones are more intensive.

Energy benchmarking is often used to evaluate the energy performance of buildings and is a crucial step towards reducing emissions. Comparability is a vital element to the success of a benchmarking system and has been the subject of many studies. In order to improve the comparability of benchmarking the energy performance of English schools was examined, assessing the impact of various features, such as built form or occupancy [20]. By analyzing the dataset using ANNs, the floor area and the number of pupils seemed to be very important determinants of school energy use.

Another work presented a method for energy classification and rating of school based on fuzzy clustering techniques compared with frequency rating techniques [21]. The fuzzy clustering method forms more robust classes avoiding imbalanced classes and classifies the buildings more precisely according to their common characteristics and similarities. The results indicated that school buildings should improve considerably their energy consumption and environmental quality.

In another study a new methodology for buildings energy benchmarking is discussed [22]. It comprises feature selection, clustering algorithm adaptation, result validation and interpretation. In comparison with the energy star approach, the proposed methodology was able to provide a more comprehensive benchmarking approach. This is because the clustering approach incorporates various building characteristics which affect energy usage, while the Energy Star approach classifies the buildings according to their use type.

Several studies observed factors affecting CO_2 emissions in the building sector and proposed methods for predicting building environmental footprint. More specifically, a Back Propagation (BP) ANN has been utilized for predicting CO_2 emissions caused by the Chinese commercial sector [23]. The most affecting indicators for emissions associated with the building sector were energy intensity, coal consumption, second industry GDP, education level, total population, business sector GDP and imports.

Other works focused on estimating indirect building carbon emissions within the boundaries of various types of Local Climate Zones (LCZs) [24]. The aim was to

discover interesting patterns and help improving energy management in specific regions. The authors conclude that it is necessary to include not only morphological parameters, which are used in this study, but also information about occupancy, HVAC systems, building use, materials and more.

3 Approach

Several studies have been conducted to predict energy consumption patterns and evaluate the factors that affect energy waste, both in existing buildings and new constructions. Despite the significance of the afore mentioned studies, there is limited research focusing on forecasting carbon emissions caused by the building sector and which factors contribute most to the environmental footprint of a building.

3.1 Problem Definition

This work analyzes an energy disclosure dataset with the primary purpose of predicting the total GHG emissions of a building and focused on discovering any useful information about factors causing excessive emissions. Also, this work can give insights to building owners and decision makers on whether a building complies or not to the specific requirements of decarbonization legislations.

3.2 Data Description

Two data sources were used for this study. LL84, or the NYC Benchmarking Law requires annual benchmarking and disclosure of energy and water usage information. LL84 covers properties with a single building with a gross floor area greater than 50000 square feet and lots having more than one building with a gross floor area greater than 100000 square feet. This dataset includes information about energy use by fuel type, physical descriptors, as well as information concerning occupancy, water use and GHG emissions. We chose data for 2017, which is the latest version publicly available.

The second data source is LL97. LL97 sets detailed requirements for two initial compliance periods: 2024-29 and 2030-34. Buildings over 25000 square feet are required to meet annual carbon intensity limits during each compliance period based on building type. To comply, building owners must submit an emissions intensity report every year or pay substantial fines. In this work, we aim to predict whether a building complies or not for a compliance period, using the LL84 dataset, combined with the carbon emissions intensity limits provided by LL97. The emissions intensity limits are listed in **Table 1**.

3.3 Data Processing

LL84 data are self-reported, therefore many data fields suffer from missing values and outliers. Several cleaning and filtering steps were conducted prior to analysis. First,

entries with duplicate or missing Borough, Block and Lot (BBL) number were removed, because BBL is a unique spatial identifier for properties in NYC. Then observations with zero or missing values either in their reported weather normalized source EUI or in their total GHG emissions were dropped, which consisted almost 19% of the total number of observations. Also, some features were removed because they either suffered from a high percentage of missing values, reaching almost 95% for some features or were not relevant to our predictions, like street name and number or the date of submission of the report.

Table 1. Carbon emissions intensity limits by property type and period.

Space Use	Carbon Limit 2024-29 (kgCO2e/sf)	Carbon Limit 2030-34 (kgCO2e/sf)
Medical Office	23.81	11.93
Retail	11.81	4.3
Assembly	10.74	4.2
Hotel	9.87	5.26
Office	8.46	4.53
School	7.58	3.44
Multifamily Housing	6.75	4.07
Factory	5.74	4.67
Storage/Warehouse	4.26	1.1

Then, for the remaining features, missing entries have been replaced with the mean value of the respective column. Additionally, some features were excluded from the analysis, as they were highly correlated with other features, like energy consumption fields with different units. Finally, one hot encoding has been performed for the Primary Property Type feature to fit our data to several algorithms. **Fig. 1** shows the feature selection process.

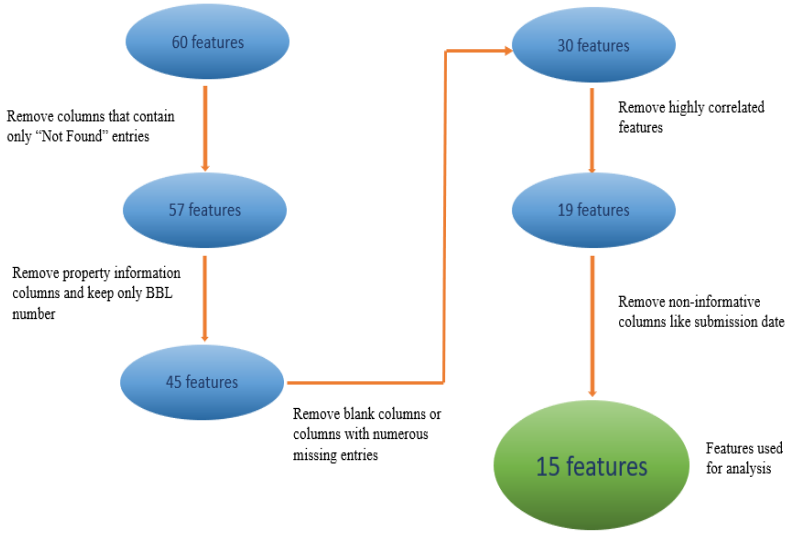

Fig. 1. Feature selection process.

The final dataset kept for analysis consists of 15 features listed below:

Weather Normalized Site EUI (kBtu/ft^2)	Weather Normalized Site Electricity Intensity (kWh/ft^2)
Self-Reported Gross Floor Area (ft^2)	Weather Normalized Site Natural Gas Intensity (kWh/ft^2)
Primary Property Type Self-Selected	Water Use Intensity (All Water Sources) (gal/ ft$^{2)}$
Year Built	Total GHG Emissions (Metric Tons CO_2e)
Number of Buildings	Weather Normalized Source EUI (kBtu/ft^2)
ENERGY STAR Score	Weather Normalizer Site Natural Gas Use (therms)
Occupancy	Electricity Use- Grid Purchase (kWh)
Borough	

The next step was to group building type values to be compatible with the building types listed in LL97. More specifically, the building types were grouped into 9 main categories: Office, Educational, Hotel, Residential, Warehouse, Public Building, Retail, Hospital, and Other. Most of the buildings were residential, only 30% being non-residential properties. To filter our data from misreported or anomalous entries, a logarithmic transformation to the Total GHG values was applied for each building type. The aim was to approximate the normal distribution given that a log-normal distribution was observed in the raw data, as shown in **Fig. 2**.Observations were excluded from the analysis if they outside the threshold of two standard deviations from the logged mean. The percentage of outliers eliminated from the analysis was 30.5%.

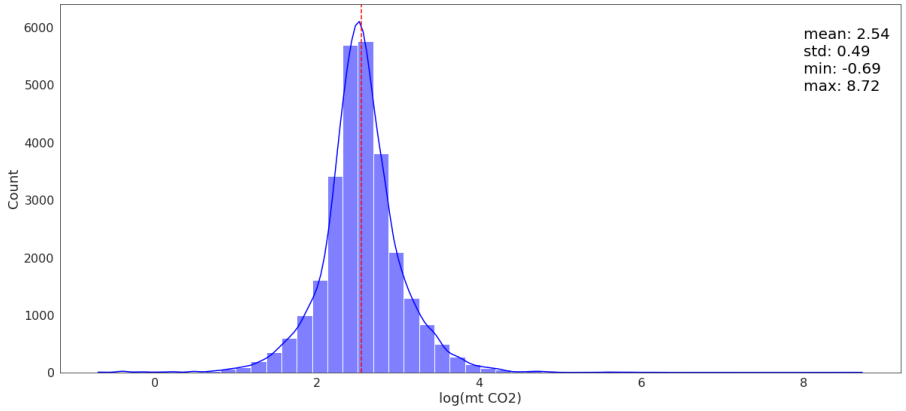

Fig. 2. Histogram of log transformed GHG emissions. The red line shows the log sample mean.

4 Experimental Results

This section presents prediction results for the two problems; predicting the annual CO_2 emissions (in metric tons emitted) using regression and the second determines which buildings comply to emissions limits via classification.

4.1 Predictions for total GHG emissions

Using regression, the following results were achieved (**Table 2**). Before applying any regression algorithm, a train test split was performed, keeping 25% of the data for testing. The algorithms examined were Linear Regression, SVR, Random Forest, XGBoost, CatBoost and ANNs. Also, a hyperparameter tuning was conducted to improve model performance. The hyperparameter grid was selected by exploring the documentation for each algorithm and also by taking into consideration the values suggested in a relevant study [19]. Three evaluation metrics were used, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2. The ANN model used in the study is a feed forward MLP with 3 hidden layers and a Rectified Linear Unit (ReLu) activation function. The scores achieved are shown in Table 2. The best performing algorithm was ANNs achieving the lowest RMSE and highest R^2. Also, CatBoost performed very well, resulting in the lowest MAE among all algorithms.

Table 2. Regression results after the selection of the optimal hyperparameters

	MAE	RMSE	R^2
Linear Regression	33.59	63.67	0.8563
SVR	13.6	32.62	0.9622
Random Forest	9.37	19.68	0.9862
XGBoost	11.09	19.3	0.9868
CatBoost	8.81	17.26	0.9894
ANN	9.15	16.75	0.9900

4.2 Predicting compliance for 2024-29

The goal of the experiments in this section was to predict compliance for properties contained in the LL84 dataset, using the LL97 carbon limits for each building type. This problem is a binary classification one, determining whether a certain building complies or not to the LL97 regulation.

For this binary classification problem, the feature “self-reported gross floor area” was excluded, as it was used to calculate the limits for compliance and thus it would affect predictions. We kept the same features for prediction as these used for the regression problem. A train/test split was conducted keeping 25% of the data to evaluate the predictions. The algorithms used were Random Forest, XGBoost, CatBoost and ANNs. For this problem, a feed forward MLP was developed, with 1 hidden layer. The metrics used for evaluation were accuracy and F-score. A grid search was

conducted to determine if there are any hyperparameters which could enhance model accuracy. **Table 3** illustrates the models' prediction performance tested on unknown data. Random Forest is the most powerful predictor with 98,7% accuracy and 0.9918 F-score. All algorithms performed well with insignificant differences between them.

Table 3. Classification results for the period 2024-29

	Accuracy	F-score
Random Forest	0.987	0.9918
XGBoost	0.983	0.9896
CatBoost	0.985	0.9911
ANN	0.984	0.9902

Table 4 shows the confusion matrix using the best performing algorithm for this period which is Random Forest. Rows represent the actual target values and columns the predicted labels.

Table 4. Confusion matrix for the period 2024-29

		Predicted	Predicted
		0	**1**
Actual	**0**	947	33
Actual	**1**	30	3845

4.3 Predicting compliance for 2030-34

The acceptable CO_2 limits emitted from buildings for the period 2030-34 are much lower than the limits of the previous period examined, but the procedure is almost identical. The results are shown in **Table 5**. In this case, CatBoost appears to be the best predictor, but again the scores are very similar for all algorithms.

Table 5. Classification results for the period 2030-34

	Accuracy	F-score
Random Forest	0.981	0.9538
XGBoost	0.981	0.9548
CatBoost	0.982	0.9561
ANN	0.978	0.9472

Table 6 shows the confusion matrix using the best performing algorithm for this period, which is CatBoost. Now most of the buildings do not comply with the regulations and thus belong to class 0.

Table 6. Confusion matrix for the period 2030-34

		Predicted	Predicted
		0	**1**
Actual	**0**	3808	60
Actual	**1**	28	987

5 Discussion

Results on the regression problem indicate that tree-based algorithms perform very well, with CatBoost appearing to be the best among them for all metrics. However, the differences are insignificant observing the R^2 score, which is almost 0.99 for all tree-based algorithms. Thus, ensemble methods seem to be more promising for this kind of problem, compared with traditional regression algorithms like Linear Regression or SVR which resulted in less accurate predictions. Indeed, for Linear Regression MAE and RMSE scores were 3 times bigger compared with ANNs and the ensemble models. However, SVR overall performance was good reaching an R^2 score of 0.96. It is worth mentioning, that increasing the number of trees for Random Forest, XGB and CatBoost from 100 to 1000 during hypertuning led to lower errors in all metrics. In addition, ANNs gave the best results in terms of RMSE and R^2, but it was more difficult to find the best hyperparameters and their execution time is longer compared with tree-based algorithms.

Concerning the pre-processing phase, the outlier detection process, which was proposed in previous studies [5,17], significantly improved model performance. Observations that fell out of the threshold of two standard deviations of their logged mean were excluded from the analysis. Also, missing values have been replaced with the mean value of the respective column. This procedure limited the number of data entries which were drastically misreported and narrowed the range of the target value between 123 and 805 metric tons of CO_2 emitted.

Another interesting finding is that the most important predictors are the gross floor area, source and site energy use intensity, electricity, and natural gas use. That means that these characteristics could be the key for decarbonizing buildings. On the contrary, building type does not seem to play a significant role for carbon emissions. This could be explained by highlighting that almost 70% of the buildings reported in the dataset were residential, so it was difficult to draw conclusions for other building types, which appeared only about 5% of times. Therefore, a more balanced dataset regarding the property type may have been more informative about the environmental footprint of different types.

Also, site energy use intensity tends to be more influencing for building emissions than source intensity. An explanation for that could be the fact that site EUI is the amount of heat and electricity consumed by a building as reflected in utility bills. On the contrary, source EUI represents the total amount of raw fuel that is required to operate the building and it incorporates all transmission, delivery, and production losses. Consequently, it is more likely that building owners are more aware of their

site energy use by looking at their bills, but more unlikely to have calculated their source energy use properly. So, site EUI tends to be more reliable for our predictions because it has a lower possibility to be misreported in the LL84 dataset. However, the Environmental Protection Agency (EPA) suggested that source energy is the most equitable unit of evaluation and provides a complete assessment of energy efficiency in a building [25].

Comparing our findings for the classification problems, for the first period (2024-29) examined almost 80% of the buildings comply with LL97 regulations, while for the second period (2030-34) only 20% of the buildings fulfill the requirements. This indicates the need to make a transition towards greener technologies and energy efficiency refurbishments in the next few years. Additionally, there is a slight difference for model performance between the two periods, with the first period achieving higher F scores than the second, for all algorithms. Indeed, false positives for the second period experiments increased significantly compared with the first period. However, the overall performance for all algorithms was good, achieving almost $R^2 = 0.96$.

6 Conclusions

Understanding building environmental footprint is a crucial component of improving urban sustainability plans, reach carbon reduction goals, as well as achieve higher levels of energy efficiency and comfort. The analysis presented here aims to predict the total GHG emissions of buildings using the LL84 self-reported energy disclosure data from properties in New York.

Using the acceptable limits of carbon by building category, which are provided by a carbon reduction legislation applied in NYC, we tried to predict whether a building complies to the carbon law for two periods. Using six ML algorithms for the regression problem and three for the classification problems, the results suggest that the data from LL84 sample can produce reasonably accurate predictions of carbon emissions across the city at a building scale.

Overall, we found that tree-based algorithms and ANNs perform better than traditional algorithms like Logistic Regression and SVR, achieving impressively higher scores. Additionally, the preprocessing procedure seems to be very important in filtering self-reported datasets, which suffer from numerous missing and misreported values. It is also observed that building size and energy use intensity play a major role in buildings' environmental footprint.

However, some assumptions or personal estimations may affect the validity of the results. Although a preprocessing and data cleaning procedure has been followed, still it was difficult to understand if all entries are correct or detect all anomalous ones, since LL84 is self-reported. Also, most of the buildings examined were residential and the commercial buildings were very limited. This imbalance does not favor the results and makes it harder to draw conclusions about specific property types and their environmental footprint.

Future work should collect more energy disclosure data from previous years and incorporate new data, publicly available by the end of the year. Also, data from dif-

ferent regions or cities along with weather information would provide a more comprehensive view of carbon emissions caused by the urban building stock.

In addition, more ML algorithms, as well as feature selection techniques, could improve performance. Regarding ANNs implementation, a more detailed selection of hyperparameters is desirable to explore their dynamic in these types of problems. Finally, the importance of focusing on forecasting emissions is worth mentioning, as there is little research on this specific field. Combining building with transportation data could also be an idea for future research, as the transportation sector accounts for a significant amount of urban carbon emissions and could be beneficial for city-scale level sustainability plans.

References

1. Zhao, H-X., and Magoulès F. "A review on the prediction of building energy consumption." *Renewable and Sustainable Energy Reviews* 16.6 (2012): 3586-3592.
2. Kontokosta, C.E. "Energy disclosure, market behavior, and the building data ecosystem." *Annals of the New York Academy of Sciences* 1295.1 (2013): 34-43.
3. C. Fan, F. Xiao, Z. Li, J. Wang. "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review." *Energy and Buildings* 159 (2018): 296-308.
4. J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham. "k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement." *Energy and Buildings* 146 (2017): 27-37.
5. S. Papadopoulos, B. Bonczak, and C.E. Kontokosta. "Pattern recognition in building energy performance over time using energy benchmarking data." *Applied Energy* 221 (2018): 576-586.
6. K.J. Baker, and R.M. Rylatt. "Improving the prediction of UK domestic energy-demand using annual consumption-data." *Applied Energy* 85.6 (2008): 475-482.
7. N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias. "Using principal component and cluster analysis in the heating evaluation of the school building sector." *Applied Energy* 87.6 (2010): 2079-2086.
8. R.A. Lara, G. Pernigotto, F. Cappelletti, P. Romagnoni, A. Gasparella. "Energy audit of schools by means of cluster analysis." *Energy and Buildings* 95 (2015): 160-171.
9. R. Mena, F. Rodríguez, M. Castilla, M.R. Arahal. "A prediction model based on neural networks for the energy consumption of a bioclimatic building." *Energy and Buildings* 82 (2014): 142-155.
10. Seyedzadeh, S., Rahimian, F., Glesk, I. Roper M. "Machine learning for estimation of building energy consumption and performance: a review." *Visualization in Engineering* 6.1 (2018): 5.
11. A.M. Waseem, 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* 147 (2017): 77-89.
12. H. Pombeiro, R. Santos, P. Carreira, C. Silva, J.M.C. Sousa. "Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear

regression vs. fuzzy modeling vs. neural networks." *Energy and Buildings* 146 (2017): 141-151.

13. B. Dong, C. Cao, and S.E. Lee. "Applying support vector machines to predict building energy consumption in tropical region." *Energy and Buildings* 37.5 (2005): 545-553.
14. R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor. "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy." *Applied Energy* 123 (2014): 168-178.
15. D. Solomon, R. Winter, A. Boulanger, R. Anderson and L. Wu. "Forecasting energy demand in large commercial buildings using support vector machine regression." (2011).
16. K. Christantonis, C. Tjortjis, A. Manos, D.E. Filippidou and E. Christelis, 'Smart Cities Data Classification for Electricity Consumption & Traffic Prediction', *Automatics & Software Enginery*, 31(1), 2020
17. A. Mystakidis, C. Tjortjis, 'Big Data Mining for Smart Cities: Predicting Traffic Congestion using Classification', Proc.11th IEEE Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 20) 2020.
18. K. Christantonis, C. Tjortjis, A. Manos, D.E. Filippidou, E. Mougiakou and E. Christelis, 'Using Classification for Traffic Prediction in Smart Cities', 16th Int'l Conf. on Artificial Intelligence Applications and Innovations (AIAI 20) 2020.
19. C.E. Kontokosta, and C. Tull. "A data-driven predictive model of city-scale energy use in buildings." *Applied energy* 197 (2017): 303-317.
20. S.-M. Hong, G. Paterson, D. Mumovic and P. Steadman. "Improved benchmarking comparability for energy consumption in schools." *Building Research & Information* 42.1 (2014): 47-61.
21. M. Santamouris, G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, M.N. Assimakopoulos, R. Mitoula, S. Zerefos, "Using intelligent clustering techniques to classify the energy performance of school buildings." *Energy and buildings* 39.1 (2007): 45-51.
22. X. Gao, and A. Malkawi. "A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm." *Energy and Buildings* 84 (2014): 607-616.
23. L. Wen, and X. Yuan. "Forecasting CO_2 emissions in Chinas commercial department, through BP neural network based on random forest and PSO." *Science of The Total Environment* 718 (2020): 137194.
24. Y. Wu, A. Sharifi, P. Yang, H. Borjigin, Da. Murakami, Y. Yamagata. "Mapping building carbon emissions within local climate zones in Shanghai." *Energy Procedia* 152 (2018): 815-822.
25. Energystar.gov, https://www.energystar.gov/buildings/facility-owners-and managers/existing-buildings/use-portfolio-manager/understand-metrics/difference