GIS-based Multi-scale Residential Building Energy Performance Prediction using a Data-driven Approach

Usman Ali[1], Mohammad Haris Shamsi[1], Mark Bohacek[3], Karl Purcell[3], Cathal Hoare[1], Eleni Mangina[2],James O'Donnell[1]
[1]School of Mechanical and Materials Eng and UCD Energy Institute, UCD, Dublin, Ireland
[2]School of Computer Science and UCD Energy Institute, UCD, Dublin, Ireland
[3]Sustainable Energy Authority Of Ireland, Dublin, Ireland

Abstract

Urban planning and development strategies are undergoing a transformation from conventional design to more innovative approaches in order to combat climate change. As such, city planners often develop strategic sustainable energy plans to minimize overall energy consumption and CO_2 emissions. Planning at such scales could be informed by spatial analysis of the building stock using Geographic Information Systems (GIS) based mapping. A data-driven methodology could aid identification of building energy performance using existing available building data. However, existing studies in literature focus on either a single building or a limited number of buildings for energy performance prediction, thus, ignoring multiple scales. This paper develops a methodology for GIS-based residential building energy performance prediction at multi-scale using a data-driven approach. The machine-learning algorithm predicts building energy ratings from local to national scale using a bottom-up approach. The multi-scale mapping process integrates the predictive modeling results with GIS. This study demonstrates the methodology for the Irish residential building stock to evaluate the energy rating at multiple scales. Modeling results indicate priority geographical areas that have the greatest potential for energy savings.

Key Innovations

- Generalized methodology to predict building energy performance at multiple scales
- Data-driven approaches for GIS-based building energy modeling.
- Formulated GIS maps identify areas with energy savings potential.

Practical Implications

The proposed research helps the urban planners, local authorities, and energy policymakers to predict the multi-scale residential building energy performance. Furthermore, identify the priority geographical areas that have the greatest potential for building energy savings by targeting community-based campaigns to increase retrofitting activity.

Introduction

The global energy consumption from the building sector accounts for more than 40% of the total energy consumption in developed countries (EU-Energy, 2018; EESI, 2018). Climate change is one of the main factors contributing to the increase in energy usage, and has a direct influence on the demand for heating and cooling in buildings (Zheng and Weng, 2019). This growth of annual energy consumption, especially in urban areas, will eventually lead to a substantial increase in carbon emissions (Güneralp et al., 2017). Therefore, the building sector is receiving increased attention with the aim of reducing overall energy consumption and emissions. Stakeholders (urban planners and policymakers) are looking at innovative sustainability strategies to transform existing buildings into more sustainable forms. As such, city planners often develop strategic sustainable energy plans to minimize overall energy consumption and CO_2 emissions. Planning at such scales can be informed by spatial analysis of the building stock using Geographic Information Systems (GIS) based mapping (Ali et al., 2020).

GIS modeling techniques help to visually analyze, manipulate, and manage a large amount of data embedded in a geographical context (Zheng and Weng, 2019). GIS modeling often requires a large amount of building stock data for analysis, for instance buildings' characteristics and their actual energy performance. A significant challenge for stakeholders is to collect the building data with limited resources. The European Union (EU) has mandated the Energy Performance of Buildings Directive (EPBD) to ensure that the member states develop a building database comprising of building energy performance in the form of Energy Performance Certificates (EPCs) (EU, 2018). However, building stock databases normally represent 30 - 50% of the entire building stock (Pasichnyi et al., 2019). Furthermore, available data are insufficient for urban stakeholders to formulate sustainable energy conservation measures.

Various data-driven approaches have been implemented over the past few years in the domain of building energy demand prediction (Hong et al., 2020; Zhao et al., 2020). These approaches use historical data to formulate data-driven models based on statistical and machine learning (ML) algorithms (Wang et al., 2020; Sun et al., 2020). Machine learning algorithms are broadly divisible into supervised and unsupervised learning techniques (Amasyali and El-Gohary, 2018; Abbasabadi and Ashayeri, 2019). Supervised learning algorithms can be further divided into regression and classification algorithms. Regression algorithms are optimal when the output variable is a real value, for instance, energy consumption (Deb et al., 2017). Classification algorithms suit applications where the output variable is a label, for instance, energy rating and building type (Benavente-Peces and Ibadah, 2020). Commonly used supervised learning algorithms include the nearest neighbor, naive Bayes, rule induction, deep learning, Support Vector Machines (SVM) and neural networks (Abbasabadi and Ashayeri, 2019). Unsupervised learning techniques are suitable in the absence of any corresponding output variable for the inputs (Ali et al., 2018). Commonly implemented unsupervised learning algorithms include k-means clustering and association rules (Sun et al., 2020).

The majority of data-driven studies in literature focuses on either a single building or a limited number of buildings for prediction of energy consumption Ali et al. (2020). A small number of studies implement machine learning models for building energy prediction at a large scale where the focus is on GIS-based energy modeling (Kontokosta and Tull, 2017; Abbasabadi and Azari, 2019). For instance, Ma et al. devised an approach to estimate the energy use intensity of 3640 multi-family residential buildings in New York City by integrating GIS and big-data technology (Ma and Cheng, 2016). A methodology to formulate a data-driven predictive model to map city-scale energy use in buildings was proposed by (Kontokosta and Tull, 2017). However, existing urban scale research uses synthetic data to generate and train the data-driven models (Nutkiewicz et al., 2018; Abbasabadi and Azari, 2019). Nutkiewicz et al. (2018) devised a framework to integrate engineering (physics-based) simulation and machine learning methods in a multi-scale urban energy modeling workflow. This use of synthetic data is due to the lack of large quantities of high-quality data required to train prediction models. Therefore, building stock modeling requires a robust GIS-based modeling approach that predicts the energy performance of the entire building stock data using limited resources for complex decision analysis. Significant opportunities exist to build upon existing research and thereby, develop a generalized methodology for GIS-based multi-scale modeling at an urban scale.

The novelty of this study proposes a GIS-based data-driven generalized methodology to predict and map the residential building energy performance at a multi-scale with limited available resources. This work introduces a methodology to facilitate GIS-based building energy performance prediction using supervised machine learning algorithms. The main aim of this paper is to formulate an intelligent machine learning model that can be used to predict building energy performance with spatial attributes by using a bottom-up approach. Therefore, the proposed methodology integrates data driven, GIS and bottom-up approaches. As a multitude of machine learning algorithms exist, this research also compares these different algorithms in terms of prediction accuracy when applied to predict building energy ratings using existing building stock data. Furthermore, the research also maps the prediction results at multi-scale to identify the priority geographical areas that have the greatest potential for energy savings.

The paper comprise following sections: Section 2 provides detailed description of the methodology for GIS-based residential building energy performance prediction. Section 3 discusses the Irish residential building stock case study to generate GIS-based maps at multi-scale using data driven prediction models. Section 4 presents the conclusions and future work.

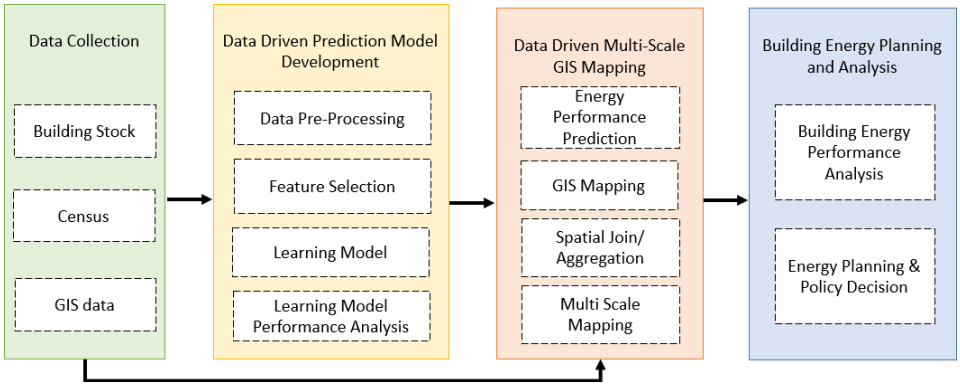

Figure 1: Methodology for GIS-based multi-scale residential building energy performance prediction using a data driven approach.

Methodology

The devised methodology facilitates GIS-based building energy performance prediction using supervised machine learning algorithms. As majority of the data-driven studies in literature focus on either a sin- gle building or a limited number of buildings, the proposed technique accounts for building performance prediction at multiple scales.

GIS-based prediction of building energy performance follows four crucial processes, namely, data collection, data driven prediction model development, multi-scale GIS mapping and building energy planning and analysis (Figure 1). The data collection process identifies the required sources of data for data driven modeling. The model development process involves data pre-processing, feature selection, learning model formulation and model performance analysis to formulate and identify suitable data-driven prediction models. The GIS mapping process links the energy performance predictions to specific geographical regions using spatial join and aggregation techniques. This process also involves the generation of multi-scale maps based on certain scenarios. Building energy planning and analysis uses these multi-scale maps to trace the energy footprint of different areas, which could further aid the energy planning and policy making process.

Data Collection

GIS-based modeling at multiple scales requires a combination of several data inputs, namely, building stock, census, and GIS data. The modeling requires these databases to geographically associate and map building energy performance at multiple scales. Building stock data include currently available building characteristics information. Generally, Energy Performance Certificate (EPC) database contains the crucial information related to the building stock. Similarly, census data provide quantitative information about buildings at a national scale. GIS data constitute the geographical details such as buildings' footprint and boundaries of regions (neighborhood, districts, counties and cities) in the form of a shapefiles, which can spatially describe vector features: points, lines, and polygons.

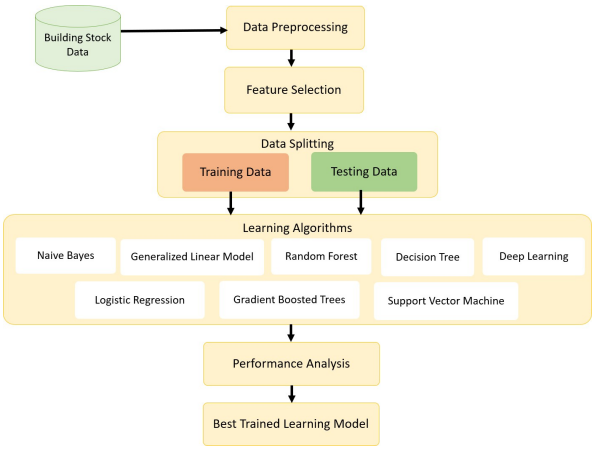

Figure 2: Methodology for data driven building energy performance prediction to identify the optimal learning model.

Data-Driven Prediction Model Development

The development of data-driven prediction model to evaluate the building energy performance involves data pre-processing, feature selection, data splitting, learning algorithm implementation and model performance analysis (Figure 2). These processes help in the identification of an intelligent machine learning algorithm capable of predicting building energy performance.

Building data are often obtained through extensive surveys and therefore, data may contain incomplete, missing or inconsistent information. This necessitates the processing of raw data to remove noise, errors or outliers and hence enhance the suitability of the data. Data pre-processing eliminates the data inconsistencies before it can act as an input to formulate the learning model. Some of the important pre-processing techniques are data cleaning, data transformation and outlier detection (Ali et al., 2016). Data cleaning processes eliminates data that are incomplete, incorrect, inaccurate, or improperly formatted. Data transformation involves the conversion of nominal data types to numerical data types. Outlier detection involves the identification and treatment of data points with exceptionally different distributions and statistically significant deviations. This process usually employs outlier detection algorithms such as distance-based, density-based and Local Outlier Factor (LOF). This work implements the LOF algorithm for outlier treatment of the building stock dataset as the algorithm is optimal for large datasets (Ali et al., 2019). The LOF algorithm measures the density of objects between each other using the nearest neighbors distance formula.

The feature selection process involves filtering of the entire variable list to determine and define the optimal set of features. This process removes irrelevant or redundant variables and retains only those variables that influence model performance. The feature selection process optimizes and reduces the dimensionality of model inputs, thereby, significantly reducing complexity and computational load. Feature selection generally involves the use of statistical or engineering methods. Statistical feature selection methods determine key features by performing various statistical tests on the dataset. The engineering feature based selection methods identifies features based on published existing literature and expert/survey reports. However, at a large scale, feature selection mainly depends on the availability of data. The values of the entire feature set are often not available for each and every building in the dataset. Hence, this process em-

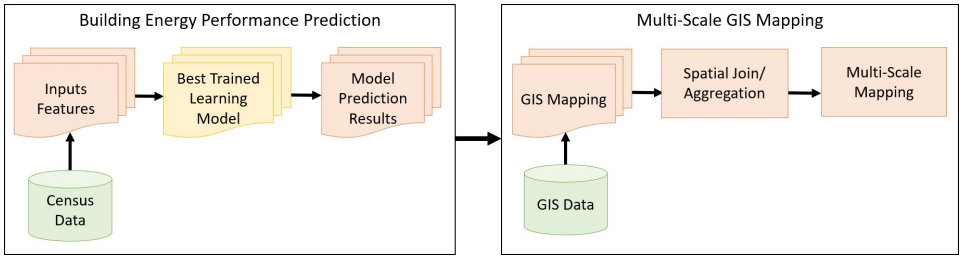

Figure 3: Process workflow to map the building energy performance at multi-scale using a data-driven approach.

ploys only those features that cover the entire span of the considered building stock.

The data splitting process divides a dataset into two, a training set to train the model and a test set to test the trained model (Ali et al., 2019). This process usually employs a random splitting technique or a cross-validation technique. The random splitting technique splits the data into train and test subsets using a 80/20 ratio split. The cross-validation technique divides the data into k number of subsets, which follows the implementation of data splitting to each of the subsets. Cross validation provides a balance between minimal bias and variance of the training model. Each iteration uses the k^{th} subset for testing while (k-1) subsets are used to train the models.

Learning algorithms predict the output classifiers of a given set of data points. The output classifiers are often referred to as labels or categories, for instance, energy rating. This research implements eight different algorithms for energy rating prediction. Previous literature suggests that these algorithms deliver excellent performance when used for energy forecasting (Wei et al., 2018; Amasyali and El-Gohary, 2018). The implemented algorithms include Naive Bayes (NB), Generalized Linear Model (GLM), Random Forest(RF), Decision Tree (DT), Deep Learning (DL), Logistic Regression (LR), Gradient Boosted Trees (GBT) and Support Vector Machine (SVM).

This study establishes the effectiveness of learning prediction models using adopted performance indices such as model development time, Classification Error (CE) and ACCuracy (ACC) (Wei et al., 2018). Accuracy is the ratio of number of correct predictions to the total number of input data that evaluates the performance of prediction algorithm. The index represents a percentage of the correct number of predictions in the entire result. Classification error describes the percentage of incorrect predictions in the complete result. Furthermore, a confusion matrix is appropriate to visualize algorithm performance and to summarize prediction results. These performance indices decide the respective efficacy of individual learning prediction models for GIS mapping.

Data Driven Multi-Scale GIS Mapping

GIS analysis maps predicted building energy performance values onto the appropriate geographical areas, thereby providing context for the results. Data driven multi-scale GIS mapping involves two steps. The first step involves building energy performance prediction at multi-scale. At such large scales, the building quantification data collection determines the number of buildings in the area. Quantification data can be extracted from national census databases. However, the desired information associated with these data is often not available for large scale areas. Therefore, a building's data contains the input features required to predict a building's energy performance. These features act as inputs to the best-trained learning models to get final prediction results. The input features could be a single set or multiple sets selected in a feature selection process. Each set of features comprise individually trained learning models. The mapping process combines these individual models to map the predicted performance values.

In the second step, the multiple scales GIS mapping process maps the predicted building energy performance results using the GIS data for the desired area. The GIS data is available in a shapefiles format, which contains the geographical information of the target area. This study implements a bottom-up approach for GIS mapping to depict neighborhoods and small areas (groups of buildings), which constitute the lowest map scale. The spatial join or aggregation technique merges the lower scale to the higher scale for multi-scale mapping (Figure 3).

Building Energy Planning and Analysis

Building energy planning and analysis involves the use of mapping results to identify the priority areas for improvement in terms of energy consumption. The stakeholders can analyze and determine adequate policy decisions based on different areas. The mapping would further help urban planners to integrate the socio-economic and demographic data with energy consumption patterns.

Case Study

The main objective of this paper is to develop a GIS-based building energy performance calculation methodology for an entire building stock. The methodology integrates the data driven approach with bottom-up modeling to predict the building energy performance at multiple scales. The national stock of Ireland decomposes into counties, cities, districts and small areas in order to analyze building energy performance across different regions.

Year of Construction									
Classfication	Performance Measure	NB	GLM	LR	DL	DT	RF	GBT	SVM
A1,A2...E,F,G	ACC	27%	13%	27%	23%	23%	23%	23%	23%
	Time	0:06	1:10	1:41	26:49	0:05	2:30	35:39	1:58
A,B,C,D,E,F,G	ACC	47%	36%	48%	43%	44%	44%	44%	46%
	Time	0:04	0:32	0:46	26:10	0:04	1:18	17:35	1:40
A,B,CD,EFG	ACC	72%	61%	73%	70%	70%	70%	70%	70%
	Time	0:05	0:26	0:29	26:36	0:03	1:00	10:33	1:03
AB,CD,EFG	ACC	72%	71%	73%	71%	71%	71%	71%	71%
	Time	0:03	0:21	0:22	26:22	0:04	0:57	8:27	1:02

Building Type									
Classfication	Performance Measure	NB	GLM	LR	DL	DT	RF	GBT	SVM
A1,A2...E,F,G	ACC	24%	13%	23%	13%	13%	13%	14%	13%
	Time	0:07	5:26	1:46	24:17	0:04	2:00	34:35	2:27
A,B,C,D,E,F,G	ACC	43%	36%	43%	25%	36%	36%	36%	35%
	Time	0:03	0:36	0:50	24:03	0:03	0:54	16:24	1:57
A,B,CD,EFG	ACC	67%	61%	66%	61%	61%	61%	62%	61%
	Time	0:03	0:27	0:30	26:49	0:03	0:38	10:21	1:21
AB,CD,EFG	ACC	67%	61%	67%	61%	61%	61%	61%	61%
	Time	0:04	0:24	0:23	24:31	0:03	0:33	7:57	1:24

Heating Fuel Type									
Classfication	Performance Measure	NB	GLM	LR	DL	DT	RF	GBT	SVM
A1,A2...E,F,G	ACC	25%	13%	25%	17%	13%	17%	18%	17%
	Time	0:07	1:03	1:46	26:30	0:05	2:20	35:38	2:14
A,B,C,D,E,F,G	ACC	46%	36%	46%	39%	40%	39%	40%	38%
	Time	0:03	0:34	0:51	24:53	0:04	1:13	16:29	1:49
A,B,CD,EFG	ACC	69%	64%	70%	64%	64%	64%	65%	64%
	Time	0:04	0:26	0:29	25:21	0:03	0:53	10:40	1:15
AB,CD,EFG	ACC	70%	64%	71%	64%	64%	63%	65%	64%
	Time	0:03	0:20	0:22	24:48	0:03	0:46	8:32	1:14

High Low

Note: Navie Bayes (NB), Generalized Linear Model (GLM), Random Forest(RF), Decision Tree (DT), Deep Learning (DL), Logistic Regression (LR), Gradient Boosted Trees (GBT) and Support Vector Machine (SVM)

Figure 4: Learning algorithms' accuracy comparison of the different building EPC rating classification to predict building energy performance

Year of Construction					
	true A	true B	true CD	true EFG	class precision
pred. A	4496	448	22	3	90.48%
pred. B	332	9026	5459	247	59.92%
pred. CD	289	8118	89791	18637	76.85%
pred. EFG	28	1052	18210	30641	61.37%
class recall	87.39%	48.41%	79.12%	61.87%	
Building Type					
	true A	true B	true CD	true EFG	class precision
pred. A	2760	577	962	330	59.62%
pred. B	417	6976	3789	601	59.20%
pred. CD	1563	9746	93624	27099	70.91%
pred. EFG	334	1348	15221	21452	55.93%
class recall	54.39%	37.41%	82.42%	43.35%	
Fuel Type					
	true A	true B	true CD	true EFG	class precision
pred. A	3061	489	887	317	64.39%
pred. B	525	7450	4060	475	59.55%
pred. CD	1168	9534	94165	23699	73.24%
pred. EFG	399	957	14602	25011	61.05%
class recall	59.40%	40.42%	82.81%	50.53%	

Figure 5: Confusion matrix of the Naive Bayes predictive model for building energy rating prediction.

This research proposes a GIS-based framework for multi-scale mapping of residential building energy performance that would act as a visualization aid for energy policymakers. The proposed methodology uses the Irish residential building stock to map the energy performance of buildings. The Irish Energy Performance Certificate (EPC) dataset contains the building related information for residential buildings in Ireland. The EPC rating relates to the overall energy building performance measured in terms of energy consumption and carbon dioxide emissions. The rating scale comprises energy labels (A1 to G) which represent the energy usage intensities in increasing order of magnitude. The energy rating calculation uses the official Dwelling Energy Assessment Procedure (DEAP) software. Publicly available EPC dataset contains more than 695,000 Irish residential buildings' data. Since there are more than 1,983,715 residential buildings in Ireland this means there is EPC data is available for only 39% of residential building stock (SEAI, 2018). This study employs machine learning algorithms to predict the energy rating of the remaining 61% of the stock by using limited variables.

This study uses the small areas concept for GIS mapping. Each small area represents groups of buildings and a collection of small areas constitute one district. The mapping process associates the predicted building energy rating to the small areas. For the demonstrated case study, small area mapping involves the Dublin city local authority scale and uses the Irish Small Area Population Statistics (SAPS) datasets (CSO, 2016) published in 2016. Based on information available from Ireland's Central Statistics Office, Ireland comprises 34 counties or cities, 139 districts, and 18,641 small areas with more than two million residential buildings. The small area concept allows the mapping of building energy rating at a granular level to support large scale analysis. GIS data extracts the required data from the Irish Central Statistics Office (CSO) dataset administered by Ordnance Survey Ireland (OSI). The dataset constitutes the Irish geographical details such as buildings' footprint and regional boundaries (neighborhood, districts, counties and cities) in the form of a shapefile (CSO, 2016).

After data collection, the next step is building energy performance prediction model development. This processes involves data pre-processing, feature selection and learning model implementation. Often gathered using surveys and questionnaires, the EPC data need to be pre-processed for formulation of learning models. The data pre-processing steps filter and remove the inconsistent, irrelevant and incomplete variable values in the EPC dataset. This procedure either removes or substitutes the missing/zero values with their averages. This process further employs the LOF outlier detection (using Euclidean distance) algorithm to remove outliers from the data. The energy rating is the label or output variable used as an output classifier. Following the feature selection process, a data transformation technique deduces several combinations of rating classifiers (for instance, AB, CD and EFG) from the existing rating labels (A1,A2,..., E, F, G). The classifiers generate clusters of nearby energy ratings. For instance, the classifier labeled 'EFG' comprises the individual rating labels E, F and G. Reducing the number of classifiers helps in achieving a better prediction accuracy.

The EPC data comprises a number of input vari-

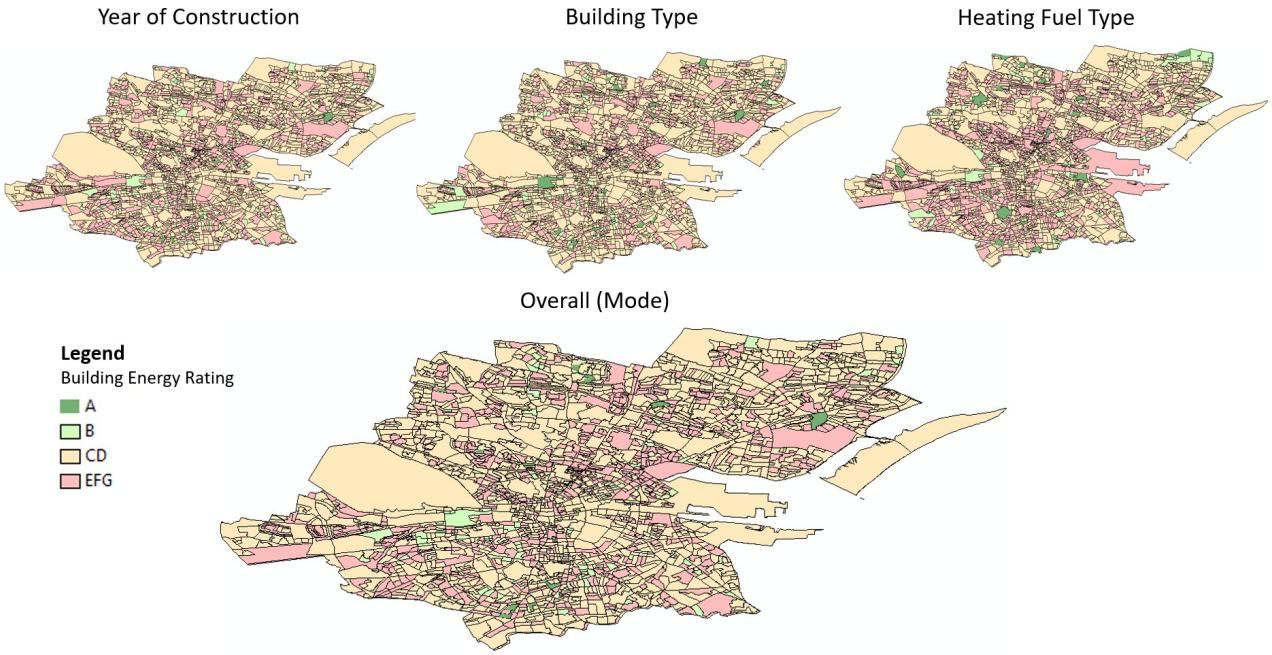

Figure 6: GIS mapping of Dublin City small area shows the most repeated number of building energy rating prediction results.

ables to calculate the energy rating of a building. The devised model in this study predicts building energy performance at a small area/neighborhood scale mainly due to privacy issues. The feature selection process depends upon the availability of data for the entire building stock. Therefore, the trained learning model uses the input features that are easily available at the small area/neighborhood scale for the entire stock, for instance, the prediction features for this particular case study include the year of construction, dwelling type, and heating fuel. After the feature selection procedure, the data are split into two parts to create training and testing data using the cross-validation algorithm. The data splitting process splits the EPC data into ten subsets of equal size.

The training process uses eight different algorithms to train the EPC data. In this paper, the deep learning model is based on a multi-layer feed-forward artificial neural network. Learning algorithm performance indices determine the optimal learning model for GIS mapping. Furthermore, the training process also considers different classification of energy ratings. The results show that the Naive Bayes algorithm performs significantly better in term of accuracy and time. Due to limited input features, the learning algorithm does not perform well with the detailed classification of energy ratings such as A1,A2,...G. Therefore, we further aggregate the lower energy rating band to test the improvement in accuracy. The results indicate that classification A, B, CD, EFG represents a significant improvement in accuracy (Figure 4). The classification is acceptable for stakeholders because the goal is often to identify the building energy rating with significantly poor performance. The highest accuracy values achieved with different features, namely, year of construction, building type and fuel type are 72%, 67% and 69% respectively (Figure 4). Similarly, the lowest classification error values achieved with different features, namely, year of construction, building type and fuel type are 28%, 33% and 31% respectively. Furthermore, a confusion matrix summarises the total number of correct and incorrect A, B, CD and EFG rating predictions using the highest accuracy Naive Bayes model (Figure 5).

The next step implements data-driven multi-scale GIS mapping of energy rating prediction results. This process maps the predictions (calculated using different input features) onto small areas inside the Dublin city local authority. Extracted from the census database of Dublin city local authority, these features act as inputs to the Naive Bayes trained learning model. The final mapping results further improve with spatial aggregation which maps the most repeated energy rating in a small area between different modeling results (Figure 6). Furthermore, the spatial join approach aggregates the small area to county or city-scale for multi-scale mapping.

The case study further analyses the opportunities for building energy planning through the identification of areas offering significant energy savings. For instance, the multi-scale GIS map could be used to evaluate the percentage of EFG rated buildings at different scales. The map represents that Dublin City contains the highest proportion of EFG rated buildings. The results can be further extended to identify specific districts with a higher proportion of EFG rated buildings. This will eventually help the stakeholders to identify the priority areas requiring energy retrofits.

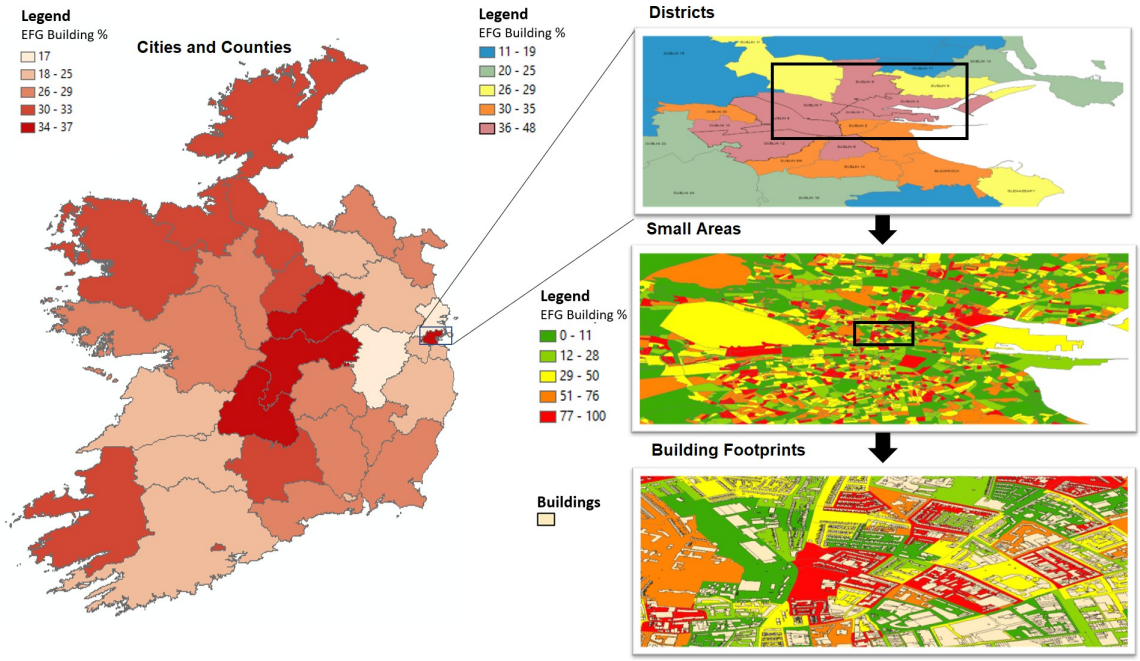

Figure 7: GIS mapping of percentage of EFG rating buildings at multiple scale.

The multi-scale mapping results also help in planning further integration of socio-economic, demographic, and correlated data by using multi-criteria decision analysis (Figure 7).

Conclusion

This study proposes a GIS-based data-driven generalized methodology to predict and map the residential building energy performance at multi-scale. GIS-based energy performance maps can provide a firm foundation to stakeholders when formulating policy measures aimed at reducing energy consumption and CO_2 emissions. Often, large-scale energy performance calculations require numerous resources both in terms of data input and computational load. The devised small area approach provides a granular level for detailed analysis, which can be further aggregated to produce the rating profiles at city or national scale.

The proposed methodology implements machine learning algorithms that provide a two-fold benefit; firstly, the technique reduces the required amount of inputs and secondly, renders an enhanced efficiency to the entire process in terms of computational load. Furthermore, building stock databases normally represent ≳0 - 50% of the entire building stock. Therefore, this approach will allow the stakeholders (local authorities, energy policymakers, and urban planners) to predict the energy performance for the rest of the building stock, thereby, reducing the uncertainty in the overall decision making process. Alongside, the planning and deployment of urban scale retrofit measures will be better informed and effective in reducing the overall consumption.

This paper uses a limited number of input features due to the unavailability of detailed GIS and census survey data. Hence, the obtained results could further be improved by using more detailed building type quantification data. The future work might also include more features for detailed analysis.

Acknowledgment

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under the SFI Strategic Partnership Programme Grant number SFI/15/SPP/E3125. We acknowledge the SEAI for access to anonymised datasets. The opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the SFI and SEAI. The authors would also like to acknowledge IBPSA Project 1

References

Abbasabadi, N. and M. Ashayeri (2019). Urban energy use modeling methods and tools; a review and an outlook for future tools. *Building and Environment*, 106270.

Abbasabadi, N. and R. Azari (2019). A framework for urban building energy use modelling. *ARCC Journal of Architectural Research*.

Ali, U., C. Buccella, and C. Cecati (2016). Households electricity consumption analysis with data mining techniques. In *Industrial Electronics Society, IECON 2016-42nd Annual Conference of the IEEE*, pp. 3966–3971. IEEE.

Ali, U., M. H. Shamsi, F. Alshehri, E. Mangina, and J. O'Donnell (2018). Comparative analysis of machine learning algorithms for building archetypes development inurban building energy modeling. In

Building Performance Modeling Conference and SimBuild.

Ali, U., M. H. Shamsi, M. Bohacek, K. Purcell, C. Hoare, E. Mangina, and J. O'Donnell (2020). A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making. *Applied Energy 279*, 115834.

Ali, U., M. H. Shamsi, C. Hoare, E. Mangina, and J. O'Donnell (2019). Application of intelligent algorithms for residential building energy performance rating prediction. In *16th IBPSA International Conference Building Simulation, Italy*. International Building Performance Simulation Association.

Amasyali, K. and N. M. El-Gohary (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews 81*, 1192–1205.

Benavente-Peces, C. and N. Ibadah (2020). Buildings energy efficiency analysis and classification using various machine learning technique classifiers. *Energies 13* (13), 3497.

CSO (2016). Census of Population 2016 - Profile 1 Housing in Ireland by Central Statistics Office. https://www.cso.ie/en/releasesandpublications/ep/p-cp1hii/cp1hii/hs/. [Online; accessed 15-Jan-2021].

Deb, C., F. Zhang, J. Yang, S. E. Lee, and K. W. Shah (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews 74*, 902–924.

EESI (2018). Building and climate change by environmental and energy study institute. http://www.eesi.org/. [Online; accessed 15-Jan-2021].

EU (2018). Directive (eu) 2018/844 of the european parliament and of the council of 30 may 2018 amending directive 2010/31/eu on the energy performance of buildings and directive 2012/27/eu on energy efficiency. *Official Journal of the European Union 61*.

EU-Energy (2018). Energy for europe by european commission. https://ec.europa.eu/energy/en. [Online; accessed 15-Jan-2021].

Güneralp, B., Y. Zhou, D. Ürge-Vorsatz, M. Gupta, S. Yu, P. L. Patel, M. Fragkias, X. Li, and K. C. Seto (2017). Global scenarios of urban density and its impacts on building energy use through 2050. *Proceedings of the National Academy of Sciences 114* (34), 8945–8950.

Hong, T., Y. Chen, X. Luo, N. Luo, and S. H. Lee (2020). Ten questions on urban building energy modeling. *Building and Environment 168*, 106508.

Kontokosta, C. E. and C. Tull (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied energy 197*, 303–317.

Ma, J. and J. C. Cheng (2016). Estimation of the building energy use intensity in the urban scale by integrating gis and big data technology. *Applied Energy 183*, 182–192.

Nutkiewicz, A., Z. Yang, and R. K. Jain (2018). Data-driven urban energy simulation (due-s): A framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. *Applied energy 225*, 1176–1189.

Pasichnyi, O., J. Wallin, F. Levihn, H. Shahrokni, and O. Kordas (2019). Energy performance certificates—new opportunities for data-enabled urban energy policy instruments? *Energy Policy 127*, 486–499.

SEAI (2018). Energy in the residential sector report. https://www.seai.ie. [Online; accessed 15-Jan-2021].

Sun, Y., F. Haghighat, and B. C. Fung (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings*, 110022.

Wang, Y., T. Wu, H. Li, M. Skitmore, and B. Su (2020). A statistics-based method to quantify residential energy consumption and stock at the city level in china: The case of the guangdong-hong kong-macao greater bay area cities. *Journal of Cleaner Production 251*, 119637.

Wei, Y., X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews 82*, 1027–1047.

Zhao, Y., C. Zhang, Y. Zhang, Z. Wang, and J. Li (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment 1* (2), 149–164.

Zheng, Y. and Q. Weng (2019). Modeling the effect of climate change on building energy demand in los angeles county by using a gis-based high spatial- and temporal-resolution approach. *Energy 176*, 641–655.