Using data-driven models to estimate the energy use of buildings based on a building stock model

Evelien Verellen, Karen Allacker
KU Leuven, Faculty of Engineering Science, Department of Architecture, Kasteelpark Arenberg 1 – box 2431, 3001 Leuven, Belgium

evelien.verellen@kuleuven.be

Abstract. In order to increase the renovation rate in Belgium, an approach is needed to refurbish clusters of buildings rather than individual buildings. To allow for a meaningful clustering of buildings, the energy performance of the existing buildings should be known. Nowadays all energy related data at building level in Belgium are confidential and cannot be shared with municipalities, private institutions or researchers. Crucial information regarding the energy use of the existing buildings is hence lacking to allow for such clustering. Using different machine learning techniques, i.e. decision trees, random forest models and k-NN models, these missing energy data for the buildings of the city of Leuven are predicted.

1. Introduction
In 2011, the European Commission communicated the Energy Roadmap 2050\(^1\), in which the European Commission committed itself to work towards a low-carbon economy in 2050 by reducing the greenhouse gas (GHG) emissions with 80% to 95% compared to 1990. The construction sector plays a major role in achieving this goal, since the construction sector emits 36% of all GHG and uses 40% of the global generated energy.\(^2\) During the past years, the global energy use however increased, which can partly be explained by the increasing world population. This increase in population is mainly manifesting in cities. It is estimated that by 2050 two third of the world population will live in cities compared to only 39% in 1980 and 54% in 2015.\(^3\) In order to achieve these European targets, but also to achieve the UN sustainable development goal 11 - Sustainable cities and communities, cities will need to take the lead in decreasing their GHG emissions and become more sustainable. Different cities already actively contribute by formulating their own roadmaps.\(^4\) In 2019, Roadmap 2025|2035|2050 of Leuven was published, including 80 clusters of projects with intermediate steps by 2025, 2035 and 2050 contributing to a climate neutral city by 2050.\(^5\) One major challenge in the Roadmap of Leuven is to increase the energy renovation rate of buildings which is currently less than 1% per year.\(^6\) The current approach, mainly consisting of renovating one building at a time, does not allow to reach the set target of Leuven, i.e. an increased annual renovation rate of 2.5%. A potential alternative approach is to renovate multiple buildings at once. To organise a large-scale renovation using a top-down approach, clusters of buildings in need of renovation should be identified. The first step towards such clustering, is setting up a building stock model of the city including all buildings characteristics, spatial information and energy use. These data need to be linked to the physical location of the buildings in order to allow to include the location of the buildings in the clustering approach.
Various data about buildings are typically available at different sources and differ in level of detail, which makes it challenging to combine these. Important information is moreover only available for part of the stock, or is not available at all. Missing data can often be (partly) filled by transforming the available data using statistics, geographic information systems (GIS) or data-driven models. Data-driven models were used in previous studies to identify representative buildings or to group buildings by building classification, building clustering (e.g. K-means clustering, partitioning around medoids (PAM), hierarchical agglomerative and hierarchical divisive) and predictive modelling (e.g. random forest, artificial neural networks). Using data-driven models instead of statistical approaches, makes it possible to analyse multiple parameters in parallel. Instead of using registered energy data, also physical models can be used to estimate energy use based on building characteristics. Although physical models are more suitable to accurately predict energy use and the influence of energy saving measures, necessary data are often lacking. In this case, prediction models as artificial neural networks can also be used to predict the building energy consumption of entire cities. Such prediction models are also used in this paper to estimate the energy use of buildings for which data are lacking (see subsequent section).

2. Problem definition

This study focuses on the building stock of Leuven and aims at estimating the heating energy of each building based on various building parameters. Estimating the heating energy through a physical model using all building characteristics in a detailed way was not possible as not all necessary data were available. This means that the energy use at building level has to be estimated using rough building data such as the building area, building typology, construction year and roof type. These recorded building parameters differ from city to city, so the level of detail and possible applications of the model can’t be generalised. Furthermore, these data can only be used to make rough estimations of the energy use, but it is assumed that the level of accuracy can be increased to a sufficient level for the purpose of this research, i.e. estimating the renovation needs of the buildings. The goal of this paper is hence to improve the estimation of the energy use by enriching the known data at building level with building information which is not known at individual building level, more specifically, with the EPC (Energy Performance Certificate) data. In Flanders EPC data are known in an anonymised way, i.e. not linked to the geographical location of the building. By combining these EPC datasets with the location-precise data of Leuven, predictions about the energy use of the buildings in the city of Leuven can be made using data-driven models. In a next step, the predicted values should be validated with real data (e.g. through sample testing - respecting privacy), especially when to be used by policy makers for action taking regarding renovations.

3. Methodology

3.1. General

The goal of this paper is hence to predict the heating demand of residential buildings in Leuven partly based on a limited number of building specific GIS information and partly based on general anonymised EPC datasets using a data-driven model. The EPC database of energy demands includes more extensive building parameters than known at the building specific level (GIS), but the geographical location of these buildings is not known. The challenge hence is to use the detailed data of the EPC dataset to predict the heating demand of all residential buildings in Leuven where only a limited amount of building parameters are known. Previous studies mostly focussed on predicting the energy demand very precisely using very detailed building parameters, in which neural networks seemed the most appropriate solution. However in this paper, appropriate data-driven models are searched that allow to identify similarities between building characteristics from the anonymised EPC data and the building characteristics in the GIS data of Leuven. The general scheme to develop a data-driven model is visualised in Figure 1 and each of the steps are discussed in more detail in section 3.2-3.5.
3.2. Data Collection
Two types of data are collected in order to train the predictive model. The first type is the building specific data: the municipality of Leuven provided GIS data of the building stock including geometric data (e.g. roof type, number of floors), functional data (e.g. function of the building, number of housing units) and administrative data (e.g. construction year). Besides the data of the municipality of Leuven, Flanders also provides publicly available GIS data (GRBgis) considering the geometry of the building (e.g. building footprint, building height) and the street addresses of the buildings. Both these datasets are used to predict the energy demand needed for heating using the predictive model.\(^\text{14}\)

The second data type is the general EPC data without known location provided by the Flemish Energy Agency (VEA). In these datasets all input data needed to create the energy performance certificate are mentioned excluding the building address, but including the energy demand needed for heating, administrative data, postal code, geometry data, building element data and installation data. These datasets are used to train, validate and test the predictive model.

3.3. Data Pre-processing
Before the data can be used to train the predictive model, the data need to be pre-processed by cleaning, reducing, integrating or transforming the data.\(^\text{11}\) In a first step the parameters which will be used to train the model are selected in the EPC datasets of almost two million housing units. The following twelve parameters were selected matching with the known building parameters in the city of Leuven: construction year, building typology, roof type, building area, protected volume, ownership, building U-value, U-value of the roof, area of the roof, postal code, construction year of the heating installation and the energy demand for heating. These parameters will be included in the predictive models, and the importance of each of the features will be determined. Next, the buildings of both the Leuven and EPC datasets where one of the selected parameters are missing are deleted. This results in a dataset of almost 300,000 housing units. Finally, data with discrete values were transformed into discrete numeric values using the LabelEncoder in Python, in order to ensure that the predictive models can interpret all values.\(^\text{15}\)
The energy demand for heating was transformed into ‘energy classes’ in order to allow to use both regression models and classifying models to predict the energy demand for heating. Different classification systems were tested: no classification (real values), 26 classes (A-Z), 13 classes (A-M) and finally 7 classes (A-G). In Figure 2 it becomes clear that the classification system with seven classes gives the highest accuracy.

Furthermore the data which are used to make the predictions are pre-processed. In a first step, data gaps were filled using the methodology described in Verellen et al. (2019).16 In a second step, the building area of the buildings in Leuven are roughly estimated as this important data was lacking. Only the ground perimeter and number of floors are known in the GIS data. The following is assumed:

\[ A = N \cdot A_r + 0.7 \cdot N_r \]

With:
- \( A \): building floor area (m²)
- \( N \): number of full floors (-)
- \( A_r \): Projected area of the roof (m²)
- \( N_r \): number of floors under the roof (-)

Finally the EPC dataset is prepared for the modelling, therefore the dataset is split into a training set (80%), a cross validation set (16%) and a test set (4%). For this subdivision, different ratios were tested using a decision tree model (see 3.4.1), and this 80% subdivision resulted in the highest accuracy.10 The training set is used to train the model, the cross validation set is used to optimise the model properties and the test set is used to finally test the model.

3.4. Model Training

For the prediction of the heating demand of buildings, different models were tested in order to select the most suitable one. As for the training set of the EPC data both input (building parameters) and output (energy demand for heating) are known for each of the buildings, only supervised learning models were considered. In this study three different types of models are investigated, both as a classifier and as a regressor: decision tree, random forest and K-Nearest Neighbors (k-NN).

3.4.1. Decision tree algorithm. The first data-driven prediction model which was considered is the Decision Tree. This model learns simple decision rules based on the given building parameters. This first model can be visualised and is relatively easy to interpret, which makes it a good model to validate the contribution of the different building parameters. The different building parameters can be easily tested since this model can also handle discrete non-numerical values.15 This algorithm was applied as a regression model where the predicted value is the numeric value for the heating demand and as a classifier model where the predicted value is a class for the heating demand. The model itself was optimised using a 7-fold cross validation. The cross validation dataset (see section 3.3) was used to optimise the model parameters of the tree (min_samples_leaf and max_leaf_nodes).

In order to find the most relevant building parameters, various combinations of parameters were tested both for the regressor and classifier model, and the resulting accuracies were compared. The accuracy of the model proved not to be high (0.002 – 0.31), which can be explained by the fact that decision trees are sensitive to overfitting the training data. Furthermore, the visualisations of the decision trees showed that some of the building parameters are dominating the trees. It could hence be concluded that the decision tree is not suited for these datasets and goal of study.

3.4.2. Random Forest algorithm. The second data-driven model applied is the random forest algorithm. This model is an ensemble method, combining the predictions of multiple prediction models, in this case decision trees. The model subdivides the training dataset in various sub-datasets and fits a number of decision trees on these sub-sets. Subsequently, the average of the results is made to cancel out errors. By combining the results of the different decision trees the model is less prone to overfitting than when a single decision tree is used, resulting in a better predictive accuracy and a smaller variance. The model
itself was optimised using a 7-fold cross validation. The cross validation dataset (see section 3.3) was used to optimise the model parameters of the tree (max_depth, random_state and n_estimators).7,15

Using the random forest algorithm, the importance of each of the building parameters can be defined. Table 1 summarises the contributions for each of the applied building parameter combinations for both regressor and (3) classifier approaches. It can be seen that the building area and the average U-value are having a much higher impact on the end result than the construction year or the roof type. Furthermore, it becomes clear that the classifier approach (A-G) results in the highest accuracy, the regressor approach results in the lowest accuracy.

Table 1 Contribution of the various building parameters to the random forest prediction and accuracy of the different combinations of building parameters

| Building parameter       | Regressor  | Regressor (A-G) | Classifier (A-G) | Classifier (A-G) | Classifier (A-G) | Classifier (A-G) | Classifier (A-G) | Classifier (A-Z) | Classifier (A-Z) | Classifier (A-M) | Classifier (A-M) |
|--------------------------|------------|-----------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Construction year        | 0.0006     | 0.116           | 0.002            | 0.017            | 0.013            | 0.006            | 0.004            | 0.166            | 0.015            |                   |                   |
| Tyology                  | 0.0003     | 0.035           | 0.040            | 0.163            | 0.116            | 0.069            | 0.086            | 0.090            | 0.121            |                   |                   |
| Roof type                | 0          | 0.011           | 0.001            | 0.007            | 0.006            | 0.002            | 0.003            | 0.036            | 0.006            |                   |                   |
| Building area            | 0.002      | 0.811           | 0.342            | 0.791            | 0.866            | 0.552            | 0.515            | 0.709            | 0.857            |                   |                   |
| Heated volume            | 0.006      | -               | 0.240            | -                | 0.372            | -                |                   |                   |                   |                   |                   |
| Rented/owner             | 0.0001     | 0.027           | 0.007            | 0.022            | -                | -                |                   |                   |                   |                   |                   |
| Average U-value          | 0.989      | -               | 0.365            | -                | -                | 0.392            | -                |                   |                   |                   |                   |
| U-value roof             | 0.0006     | -               | 0.001            | -                | -                | -                | -                |                   |                   |                   |                   |
| Construction year        | 0.002      | -               | 0.002            | -                | -                | -                | -                |                   |                   |                   |                   |
| Installations            |            |                 |                  |                  |                  |                  |                  |                   |                   |                   |                   |
| Accuracy                 | 0.167      | 0.0033          | 0.3044           | 0.2612           | 0.2642           | 0.2622           | 0.299            | 0.0621           | 0.1573           |                   |                   |

3.4.3. K-Nearest Neighbors (k-NN) algorithm. The K-NN is an example of instance-based learning, where no general model is developed, but instead the instances of the training data are stored. In a next step, for each point the nearest neighbours with the most representative building parameters are defined. The ‘k’ represents the number of neighbours which are considered. These neighbours are used to calculate the prediction based on uniform weights between the different neighbours. The optimal value for this ‘k’ is depending on the dataset, a larger ‘k’ decreases noise, but makes classification boundaries less clear. The optimal value for ‘k’ was defined using a 7-fold cross validation.10,15

3.4.4. Input Data. In order to validate the quality of the models, initially the models are trained using multiple building parameters of the whole EPC database of Flanders. Due to the limitation of the available building parameters at the building level in the city of Leuven, only a part of the building parameters (building area, roof type, building typology and construction year) were used to train the data, so that predictions at building level can be made. Consequently, this limited amount of building parameters results in a lower accuracy. Afterwards, the models and predictions were redefined using only the buildings from Leuven in the EPC database, but this resulted in a negligible difference.

3.5. Model Testing

3.5.1. Feature selection. The twelve parameters defined in section 3.3 are validated by their corresponding accuracy of the models. The accuracies of the random forest model associated with different combinations of building parameters are visualised in Table 1. It shows that the combination including most building parameters results in the highest model accuracy. However, the building parameter combination with the least amount of parameters corresponds to the current available data in the city of Leuven. From this it can be concluded that the reliability of the final predictions for the building stock of Leuven can be significantly improved by collecting additional data. Especially information about the building U-value could lead to an important improvement.
3.5.2. Model evaluation. The accuracy of the different models discussed in section 3.4 are evaluated based on the coefficient of determination (R²), the mean squared error (MSE) and the mean absolute error (MAE). The MAE of the different regression models is also compared to the MAE of the average value of the training set (33891). These values for the different models, both the regressor variant and the classifier variant, are visualised in Figures 3-5.

This shows that the k-NN model performs best with an accuracy (R²) of 0.8323 when using the available building parameters for the city of Leuven and of 0.8922 when using all selected building parameters in the EPC dataset. In the second situation, the MAE performs 6.3 times better than the MAE of the average value (33891), and 5.5 times better compared to the MAE of the decision tree and the random forest model.

![Figure 3 Accuracy of the different models (Decision tree, Random Forest and kNN) for both Regressor and Classifier approaches for two parameter combinations](image)

![Figure 4 MAE of the different models (Decision tree, Random Forest and kNN) for a Regressor approaches for 2 parameter combinations](image)

![Figure 5 MSE of the different models (Decision tree, Random Forest and kNN) for a Regressor approach for 2 parameter combinations](image)
4. Results

4.1. Model Selection

In section 3 various predictive models were tested against the goal of the study. The conclusion of this analysis based on their accuracy, through assessing $R^2$, MAE and MSE, is that the k-NN classifier model is the most suitable. In a final step, the predictions of the k-NN classifier for the heating demand of the test data are compared to the actual values of the corresponding buildings in the EPC dataset. In Figure 6, a big spike around value 1 becomes clear. This means that most predictions are situated within a deviation of 5% around the real heating demand, which makes the k-NN classifier model a reliable model. This distribution of the predictions compared to the real values is similar when the full EPC dataset of the whole of Flanders is used and when only the EPC dataset of Leuven is used. The predictions using only the data of Leuven are however having less noise and a higher accuracy.

![Figure 6 Comparison of the predicted energy demands for heating and the actual values for the dataset of Leuven (Left) and the dataset of Flanders (right)](image)

4.2. Check with city of Leuven

Finally, these predictions of the heating demand per building for the building stock of Leuven can be visualised using GIS, as shown in Figure 7, in this figure most buildings are class A or B due to the large number of relative small building sizes. In a next step these predictions for the heating demand will be validated taking the building size into account and using top-down energy data which is available at street level from the energy distribution network operator Fluvius.

![Figure 7 Visualisation of the predicted heating demand of the buildings in the city of Leuven in seven energy classes (A-G)](image)
5. Future outlook
These predictions of the energy demand for heating of the buildings of Leuven will be used to estimate the renovation need and potential of the building stock. In a next step buildings with similar renovation needs will be clustered using a K-means model including the building parameters, the energy performance, the renovation potential and demographic characteristics.

Acknowledgement We would like to thank professor Davis from KU Leuven for his help and insights regarding the machine learning methodology; and the Flemish Energy Agency (VEA) for providing the EPC datasets.

References
[1] European Commission 2011 The roadmap for transforming the EU into a competitive, low-carbon economy by 2050 p 1–4.
[2] Van de Brug E, van Breevoort P, Pestiaux J, Cornet M, de Jager D and Vermeulen P 2014 Verkenning middellange termijn (2030) en lange termijn (2050) energie- en broeikasgasscenario’s in Vlaanderen Ecofys/Climact p 122.
[3] United Nations, Department of Economic and Social Affairs, Population Division 2015 World Urbanization Prospects: The 2014 Revision ST/ESA/SER.A/366.
[4] United Nations 2016 Sustainable cities: why they matter Available at: https://www.un.org/sustainabledevelopment/wp-content/uploads/2016/08/11.pdf. (Accessed: 18th December 2019).
[5] De Paep M, Verachtert K and Van Reeth J 2019 Roadmap 2025|2035|2050: Naar een klimaatneutraal Leuven BUUR i.o.v. vzw Leuven 2030.
[6] Vlaams Energieagentschap 2017 Renovatiepact & Woningpas VEA.
[7] Tardioli G, Kerrigan R, Oates M, O’Donnell J and Finn D P 2018 Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach Building and Environ 140 p 90-106 10.1016/J.BUILDENV.2018.05.035.
[8] Ascione F, Bianco N, De Stasio C, Mauro G M and Vanoli G P 2017 Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: A novel approach. Energy 118 p 999–1017.
[9] Arambula L R, Cappelletti F, Romagnoni P and Gasparella A 2014 Selection of Representative Buildings through Preliminary Cluster Analysis International High Performance Buildings Conference Paper 137.
[10] Wahid F and Kim D H 2016 A prediction approach for demand analysis of energy consumption using K-nearest neighbor in residential buildings Int. J. Smart Home 10 p 97–108.
[11] Amasyali K and El-Gohary N M 2018 A review of data-driven building energy consumption prediction studies Renew. Sustain. Energy Rev. 81 p 1192–1205.
[12] Chen Y, Hong T and Piette M A 2017 Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis Appl. Energy 205 p 323–335.
[13] Jovanović R, Sretenović A A and Živković B D 2015 Ensemble of various neural networks for prediction of heating energy consumption Energy Build. 94 p 189–199.
[14] Vlaamse overheid Basiskaart Vlaanderen (GRB). Available at: https://overheid.vlaanderen.be/informatie-vlaanderen/producten-diensten/basiskaart-vlaanderen-grb (Accessed: 9th August 2018).
[15] Pedregosa et al. 2011 Scikit-learn: Machine Learning in Python JMLR 12 p 2825-2830.
[16] Verellen E and Allacker K 2019 Inventory of the existing residential building stock for the purpose of environmental benchmarking IOP Conf. Ser. Earth Environ. Sci. 323.
[17] Buffat R, Froemelt A, Heeren N, Raubal M and Hellweg S 2017 Big data GIS analysis for novel approaches in building stock modelling Appl. Energy 208 p 277–290.