Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Energy Performance Certificate Estimation at Large Scale Based on Open Data

To cite this article: Frédéric Montet *et al* 2023 *J. Phys.: Conf. Ser.* **2600** 032009

View the article online for updates and enhancements.

You may also like

- Design and performance evaluation of a substitution solution for spiral casing of pico-hydroelectric plants
Ludovic Favre, Maxime Chiarelli, Nicolas El Hayek et al.
- Open Data and Data Analysis Preservation Services for LHC Experiments
J Cowton, S Dallmeier-Tiessen, P Fokianos et al.
- Promoting Open Data services to decision-makers: Providing interactive data through Web Maps and Web Applications for Oradea city and Bihor county
I M Pârvu, I C Cuibac Picu, P D Dragomir et al.

This content was downloaded from IP address 194.230.158.25 on 18/12/2023 at 09:15

Energy Performance Certificate Estimation at Large Scale Based on Open Data

Frédéric Montet[1], Alessandro Pongelli[2], Stefanie Schwab[3], Mylène Devaux[4], Thomas Jusselme[2], Jean Hennebert[1]

iCoSys Institute[1], *HEIA-FR*, Fribourg, Switzerland
ENERGY Institute[2], *HEIA-FR*, Fribourg, Switzerland
TRANSFORM Institute[3], *HEIA-FR*, Fribourg, Switzerland
iTEC Institute[4], *HEIA-FR*, Fribourg, Switzerland

E-mail: frederic.montet@hefr.ch, alessandro.pongelli@hefr.ch, stefanie.schwab@hefr.ch, mylene.devaux@hefr.ch, thomas.jusselme@hefr.ch, jean.hennebert@hefr.ch

Abstract. This paper presents an innovative methodology for enhancing energy efficiency assessment procedures in the built environment, with a focus on the Switzerland's Energy Strategy 2050. The current methodology necessitates intensive expert surveys, leading to substantial time and cost implications. Also, such a process can't be scaled to a large number of buildings.

Using machine learning techniques, the estimation process is augmented and exploit open data resources. Utilizing a robust dataset exceeding 70'000 energy performance certificates (CECB), the method devises a two-stage ML approach to forecast energy performance. The first phase involves data reconstruction from online repositories, while the second employs a regression algorithm to estimate the energy efficiency.

The proposed approach addresses the limitations of existing machine learning methods by offering finer prediction granularity and incorporating readily available data. The results show a commendable degree of prediction accuracy, particularly for single-family residences. Despite this, the study reveals a demand for further granular data, and underlines privacy concerns associated with such data collection. In summary, this investigation provides a significant contribution to the enhancement of energy efficiency assessment methodologies and policy-making.

1. Introduction

Globally, each country is setting energy limits in order to be able to reduce its energy consumption and to increase the self-consumption of its own production. In Switzerland, these targets are set by the 'Energy Strategy 2050'[1]. One of the goals of this strategy is to reduce building-related consumption by financing building renovation. This financial help may increase the renovation rate of old buildings in the next years .

In Switzerland, energy performance certificates (EPC) are called Cantonal Energy Certificate for Buildings (CECB)[2] and allow for the assessment of the current state of a building. Such EPCs provide an evaluation in the form of a global energy performance label and an envelope efficiency label.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

The global efficiency label takes into account the consumption and type of equipment used for heat production and domestic hot water, as well as the consumption and own production of electricity. The second label describes the quality of the thermal envelope by taking into account the insulation of the roof, walls, windows, doors and floor, as well as all thermal bridges present and the shape of the building.

Despite being one of the key tool to provide insights on a building, the delivery of those EPCs is not optimal. First, it requires an expert to travel on-site in order to survey all building characteristics for a complete analysis, increasing the time and cost required for the analysis of large building stock. Second, the evaluation at large scale isn't possible given the underlying manual work. Finally, online data and manually gathered data in databases such as the CECB are not leveraged.

In the literature, some work do exist to make EPCs calculation automatic with the help of machine learning (ML) approaches [3, 4, 5]. However, the reported predictions do not yet reach a level allowing to use them by policy-makers.

From these problems, the following research questions were laid out :

RQ 1 How to improve the previous estimation methods so that it allows for more prediction details?

RQ 2 Is an estimator using easily accessible open-data already provides good enough information for some usages ?

If those questions would be answered, estimation of energy efficiency would be faster, more generic and allow for numerous applications in subfields where buildings are at the core of the discussion. To mention a few examples, refurbishment planning could be evaluated at a country, district or city level and real estate companies could automatically add efficiency indices to their advertisements and so on.

In this work, a two-stage approach is carried out based on the EPC of the CECB. The first phase consists of using the data available online to reconstruct the missing parts of the data needed to create an EPC. The second phase consists of applying an algorithm to estimate the energy performance of the building.

Section 2 of the paper presents the methodology used to make online available data usable, as well as the estimation method to get the energy efficiency of a building. Section 3 presents the results according to the metrics proposed to evaluate their performance. Finally, section 4 and section 5 discuss and conclude the results and future considerations.

2. Method

Overall, the objective of the method is to compute global and envelope EPC labels from the data available online. Figure 1 summarizes the different steps from data collection to batch estimation that are covered in the current methodology.

First, online data is gathered and formatted to create a set of incomplete EPC samples. Those values provide the necessary information to reconstruct samples as accurately as possible. Second, an estimation of the CECB global and envelope efficiency is computed, which gets plotted on a map. For the estimation step, a regression-based approach on the value in $kWh/(m^2 * year)$ is used to compute the label.

In the next sections, each block of the aforementioned figure is individually explained. As the order of development of this pipeline is different than its usage, the explanation in the next sections is provided as it was developed.

2.1. Dataset Characteristics

The dataset consists of more than 70'000 certificates and in the previous work an exploration of the data was carried out to understand their nature and completeness[5].

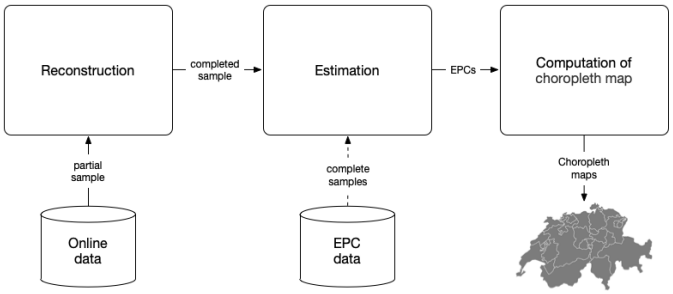


Figure 1: Overview of the EPC estimation pipeline—This diagram shows a way to leverage online data (usually few variables) together with EPC data (usually many variables) so that a large scale estimation of EPC becomes possible.

There are three types of certificates present in the dataset. The first is the *CECB Standard (G)* which was standard between 2009 and 2016, then the *CECB Plus (GP)* certificate was introduced in 2012 which is more detailed and has more parameters for the characterisation of the label. In addition, the *CECB New building (GN)* certificate was introduced, which is as detailed as the *CECB Plus* and provides the label for new construction. The number of samples per subdivision can be seen in the Table 1.

Type of certificate	**Number of samples**
CECB Standard (G)	14'430
CECB Plus (GP)	55'018
CECB New building (GN)	4'215

Table 1: Representation of the various CECB types in the CECB dataset

The subdivision by age of construction of the building was also checked to give an insight and a subdivision of the dataset according to the periods in the Table 2.

Period of construction	**Number of samples**
<1919	12'296
1919-1945	6'295
1946-1960	9'308
1961-1970	9'424
1971-1980	10'912
1981-1990	9'860
1991-2000	5'500
2001-2010	4'399
>2010	5'659

Table 2: Representation of the various period of construction in the CECB dataset

In the dataset, for each certificate, there are values such as consumption values divided by energy source, surface area values, U-values of all elements, thermal bridge values, year of construction, number of floors, number of appartments, location of the building, weather station used as a reference, the two certificate labels, some building improvement proposals for some

certificates and other building data. This data is collected or estimated (as in the case of some U-values) by experts.

2.2. Estimation of building efficiency
To estimate the building energy consumption, the dataset from section 2.1 is used. In the later, 32 variables are used to predict the target variables (envelope and global efficiency) with a regression model. For training, the CatBoost library is used, hyper-parameters (Number of iterations, learning rate, depth and L2 regularization) are tuned with a randomized grid search and a cross-validation with a k-fold where $k = 3$ is performed.

2.3. Attribution of a EPC-like letter
Once the estimation of the energy consumption is obtained given the results from previous section, a letter is attributed given equations from CECB Norm and SIA norm for the envelope score and global score.

2.4. Reconstruction of an EPC sample
Based on the dataset from 2.1, a set of utility function is used to generate partial samples. Datasets are generated to check the reconstruction performance given two scenarios : (1) with random variable deletion, (2) with publicly available variables only. For each dataset, 1000 samples are randomly selected for each number of missing values going from 1 to N, the total number of variables in a sample. N can change given the scenario.

Once the dataset is generated, the training routine is similar to the one used for estimation to the exception that the input and output variables are different. The CatBoost library is used and the training is performed with an hyper-parameter search on the number of iterations, learning rate, depth and l2 leaf reg using a randomized grid search to look for the best score. A cross-validation with a k-fold where $k = 3$ is done.

3. Results
In this section, results of the method are presented. First, with an evaluation of the regression model. Second, the implementation of the EPC letter attribution is introduced. Finally, with the performance of the reconstruction model.

3.1. Estimation of building efficiency
The same training as been performed on two variations of the datasets : unbalanced data and balanced data given the SMOGN rebalancing strategy[6]. Given the aforementioned dataset variations, two different performances are obtained. The Table 3 presents the performance given usual regression metrics.

	Global		**Envelope**	
	Unbalanced	Balanced	Unbalanced	Balanced
R2	0.75	0.84	0.8	0.85
MAPE	0.18	0.14	0.17	0.15
MSE	3057	3498	793	985
RMSE	55.29	59.14	28.16	31.38
MAE	33.16	38.76	18.74	22.45

Table 3: Numerical results of the regression performance on global and envelope efficiency prediction. The data displayed shows scores unbalanced and balanced with the SMOGN balancing strategy.

3.2. Attribution of a EPC like letter

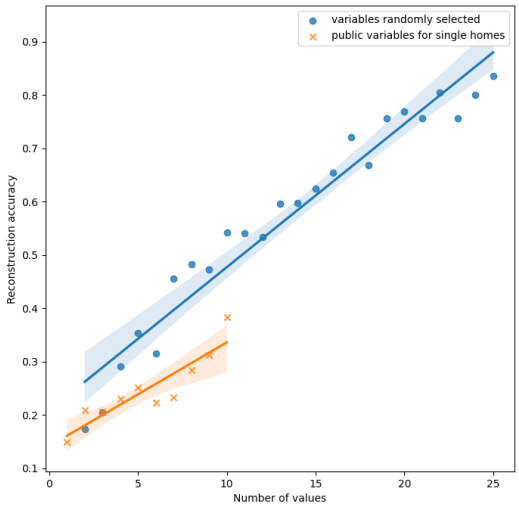
The result of this part of the method is represented by a set of Python modules allowing to take the output from the regression model and output the EPC letter. Two main functions are developed: one for the global efficiency and a second for the envelope efficiency.

To make a correct CECB calculation, this code implies specific domain knowledge acquired in collaboration with the Energy institute from the HEIA-FR. The developed code takes a subset from the SIA 380-10 Norm and CECB Normalisation document in v5.1.1 and allows for the attribution of CECB labels on single classes. The implementation of the CECB formulas is available on request.

3.3. Reconstruction of an EPC sample

As shown in Figure 2a, the reconstruction of an EPC sample follows a linear evolution as more variables are added to facilitate the reconstruction. The variables selected randomly are showing a higher score that the publicly available variables focused on single homes only, thus highlighting the difference of importance for some variables.

When showing the Table 2b, the low R^2 score indicates a large variability of the prediction below 10 available variables for the random variable deletion. The reconstruction of single homes CECB samples based only on public data reaches 0.38, which shows the absence of strongly explanatory available variables.



(a) Plot of the reconstruction evolution from 1-2 variables available to all of them.

	random	**single homes**
2	0.17	0.21
5	0.35	0.25
10	0.54	0.38
20	0.77	—
25	0.84	—

(b) Data table showing the R^2 score given different number of available variables.

Figure 2: Evaluation of the R-squared score given different variable deletion strategies for a reconstruction routine.

4. Discussion and conclusion

To answer the research questions 1 and 2, respectively : *"How to improve the previous estimation methods so that it allows for more prediction details?"* and *"Is an estimator using easily accessible open-data already provides good enough information for some usages ?"*, we tried two methods working hand in hand : (1) a reconstruction method to fill missing values in order to leverage

available open data on buildings and (2) an estimation method based on ML driven regressions to have a granular ranking.

The reconstruction is a method that is usable to some extent. The values at disposal in the public domain for single homes do allow for a R^2 score around 0.4 with 10 values, which is a partially trustable reconstruction. Given the gap between the public domain single-home reconstruction and the variables randomly selected, the addition of the most important variables could make the difference and allow for a better reconstruction as determined in [5]. Also, in comparison with the CatBoost model, a denoising auto-encoder might make better reconstruction and could be an improvement.

For the estimation step, the regression is a better approach for downstream task since numerical estimation, even if slightly wrong, won't fall in an adjacent class; for instance CECB D class instead of C. The performance are similar than in the previous study, but with an increase in usability thanks to their granularity.

In general, U-values of building components still seem to be the crux to assess the quality of a building. They have a high importance for efficiency estimation and are the base for both envelope and global estimation. Unfortunately, today there is still no dataset, apart from the CECB dataset, that contains the U-values of Swiss buildings. Being able to take a survey of all Swiss buildings by indicating all U-values in a common dataset such as RegBl would help prediction and building studies enormously.

To conclude, today's state of the art still needs improvement to estimate the energy efficiency of buildings using online data. One problem is the prerequisites for complete datasets of all Swiss buildings. The Confederation has moved in this direction and has begun the work of unifying all cantonal building registers under one location called Registre fédéral des bâtiments et des logements (RegBL)[7]. However, privacy issues poses major challenge for putting sensitive data online.

Being able to predict and recreate certain data, such as the U coefficients of various surfaces is an important point for future work. As an important parameter for predicting the energy efficiency of the building, it is therefore a possible key to improving the results obtained.

Another parameter that is now measured by hand is the size of the building. However, new online dataset like EUBUCCO provide buildings geometries across all Europe, which improve the accuracy of the model[8].

References

[1] Swiss Federal Office of Energy 2020 Stratégie énergétique 2050 `https://www.bfe.admin.ch/bfe/fr/home/politik/energiestrategie-2050.html` accessed: 2023-04-27

[2] Association GEAK-CECB-CECE Le certificat énergétique cantonal des bâtiments (cecb) `https://www.cecb.ch/` accessed: 2023-04-27

[3] García-Nieto P J, García-Gonzalo E, Paredes-Sánchez J P and Bernardo Sánchez A 2021 *Neural Computing and Applications* **33** 6627–6640 ISSN 14333058 URL `https://doi.org/10.1007/s00521-020-05427-z`

[4] Attanasio A, Savino Piscitelli M, Chiusano S, Capozzoli A and Cerquitelli T 2019 *Energies* **12** ISSN 1996-1073 URL `https://doi.org/10.3390/en12071273`

[5] Montet F, Pongelli A, Rial J, Schwab S, Hennebert J and Jusselme T 2022 *Acta Polytechnica CTU Proceedings* **38** 90–96 URL `https://doi.org/10.14311/APP.2022.38.0090`

[6] Branco P, Torgo L and Ribeiro R P 2017 SMOGN: a Pre-processing Approach for Imbalanced Regression *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications* (*Proceedings of Machine Learning Research* vol 74) ed Luís Torgo P B and Moniz N (PMLR) pp 36–50 URL `https://proceedings.mlr.press/v74/branco17a.html`

[7] Office fédéral de la statistique Registre fédéral des bâtiments et des logements (regbl) `https://www.housing-stat.ch/fr/index.html` accessed: 2023-04-27

[8] Milojevic-Dupont N, Wagner F, Nachtigall F, Hu J, Brüser G B, Zumwald M, Biljecki F, Heeren N, Kaack L H, Pichler P P *et al.* 2023 *Scientific Data* **10** 147