Energy Performance Certificate Estimation at Large Scale Based on Open Data

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

[1]iCoSys Institute, [2]ENERGY Institute, [3]TRANSFORM Institute, [4]iTEC Institute

Frédéric Montet

Our method improves energy efficiency assessments in Switzerland's built environment, aligned with the nation's Energy Strategy 2050. Traditional methods are time-consuming and costly with expert survey.

Using two machine learning models and more than 70,000 energy performance certificates, our model paves the way to a country-wide built park performance estimation, which would drive renovation strategy at large scale.

Short Introduction

This research tackles inefficiencies in Switzerland's current Energy Performance Certificates (EPCs), which are crucial for the country's Energy Strategy 2050 goals. We explore how to make these assessments faster, cheaper, and scalable using machine learning.

Two key questions guide the study :

RQ1 How to enhance existing methods for better prediction ?

RQ2 How does open data car serve as a reliable estimator ?

In this work, a two-stage approach fills in data gaps using online resources and then uses an algorithm for energy performance estimation as a surrogate for official calculations.

Success in this research has the potential to improve building energy assessments by offering a scalable solution that could be applied to national policy and various industry applications.

A Method to Use Open Data

Overall, the objective of the method is to compute global and envelope EPC values 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 partial EPC samples. Those values provide the basis to reconstruct full samples as accurately as possible.

Second, a numerical estimation of the EPC in kWh/m²/year is computed with a regression. This value can be used for downstream tasks, like computing a label like a letter or a choropleth map.

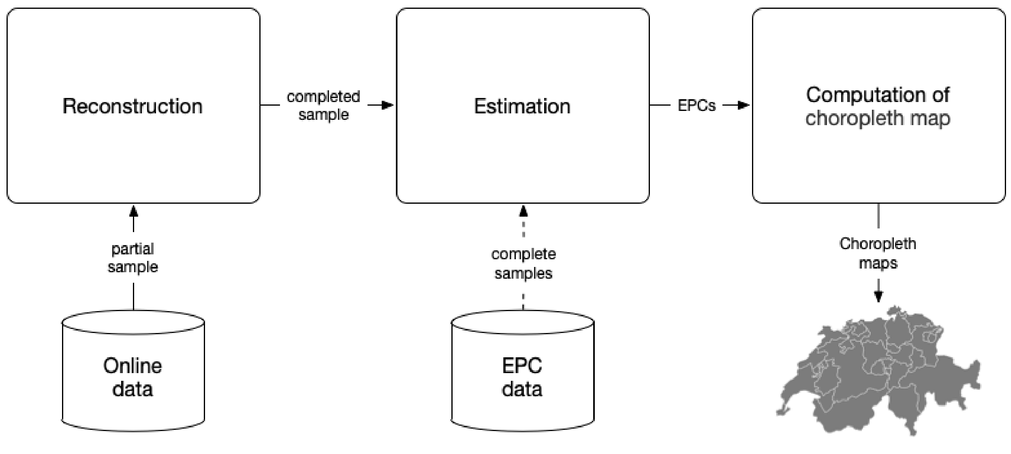

Figure 1. *Overview of the EPC estimation pipeline*

Used Dataset

The dataset consists of ~70'000 certificates. In a previous work, an exploration of the data was carried out to understand its nature and completeness [1]. This data is collected or estimated (as in the case of some U-values) by CECB experts.

Three types of certificates are present in the dataset. The first is the *CECB Standard* (14'430 samples) used between 2009 and 2016. Then, the *CECB Plus* (55'018 samples) certificate was introduced in 2012 until today. This EPC is more detailed and has more parameters for the characterisation of the label. In addition, the *CECB New building* (4'215 samples) certificate was introduced, which is as detailed as the *CECB Plus* and provides a label for new constructions.

Each certificate in this dataset contains values such as consumption split by energy source, surface area values, U-values of all elements, envelope efficiency and global efficiency, as well as other building data.

Sample Reconstruction

Based on the main dataset, a set of function is used to generate partial samples. Those samples are generated to train a reconstruction model and check the reconstruction performance given the publicly available variables.

For this reconstruction model, the training routine is similar to the one used for estimation to the exception that the input and output are different.

Performance Estimation

To estimate the yearly surface area consumptions, the main dataset is used. In the later, 56 variables are used to predict the target variables (envelope and global efficiency) with a regression model.

Once the consumption is obtained given the results from previous section, a letter is attributed given equations from the CECB and SIA norm for the envelope score and global score, which already represents a downstream task.

Results of the Study

Large scale built park performance estimation is at our doorstep. Results show that with the publicly available data, EPCs can be averaged with a median error below ten percent. Further validations are in progress.

As presented on table 1, when put together, the reconstruction and estimation model do perform well with all variables initially available. When using only the variables available publicly, without surprise the scores get degraded quite strongly. With such results, is it possible to use the models for some applications ?

	Global		Envelope	
	All variables	Public variables	All variables	Public variables
$R2$	0.86	0.48	0.95	0.31
$MAPE$	0.13	0.39	0.09	0.47
MAE	25.47	49.63	7.53	34.21

Table 1. *Data showing the model scores when all variables are available and only with the one available publicly*

Let's visualize

On figures 2 and 4, efficiency values by municipality in the canton of Fribourg (Switzerland) are presented on four choropleth maps.

Those values are computed for each building and then, averaged per area. When aggregated, the error of each individual buildings gets averaged.

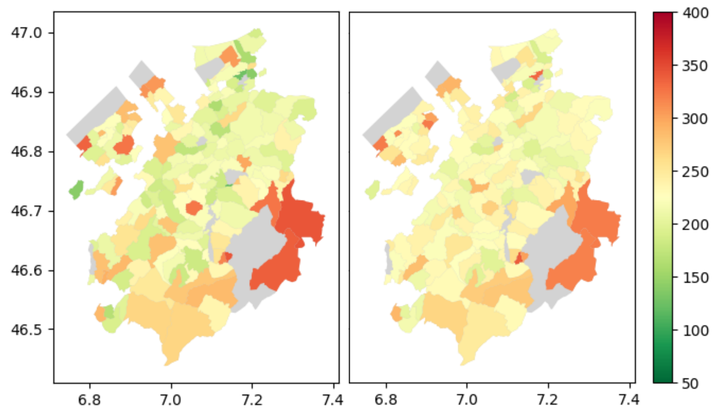

Figure 2. *Estimation of the global efficiency by municipality (in kWh/m²/year). On the left the true values, on the right the one estimated with public data.*

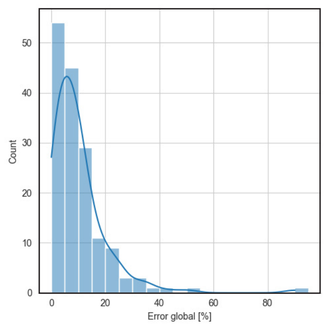

Figure 3. *Distribution of the error (in %) of average global efficiency estimation values. The median error is 7.6 %*

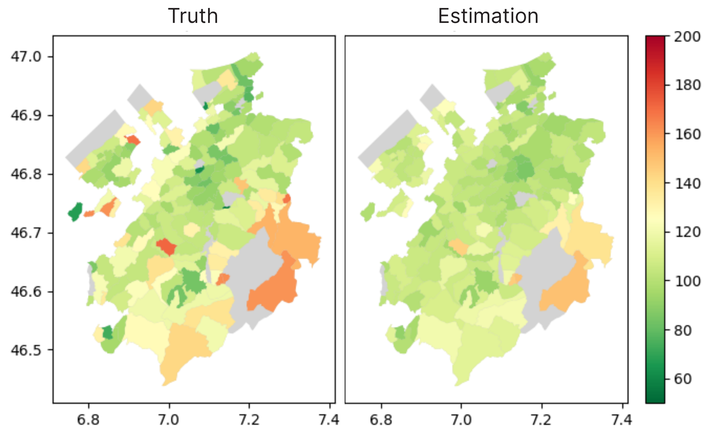

Figure 4. *Estimation of the envelope efficiency by municipality (in kWh/m²/year). On the left the true values, on the right the ones estimated with public data.*

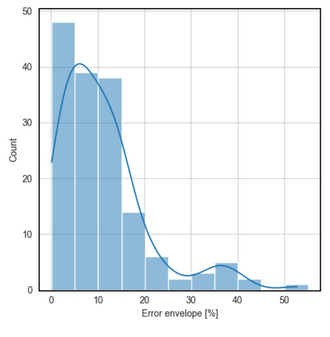

Figure 5. *Distribution of the error (in %) of averaged envelope efficiency estimation values. The median error is 8.55 %*

Compared to the true value (on the left), the estimated average per municipality reaches a median error below 10%, as in the captions from figures 3 and 5.

Today, those results seem promising and could be the basis for further work. Nevertheless, they need to be tested more in-depth to be fully validated for decision making processes.

Future Directions

Future work should focus on the validation of the results. Also, predicting and recreating key parameters, such as U-values and building size to improve the accuracy of energy efficiency estimates could help. Finally, consolidation efforts of cantonal building registries into a single federal register (RegBL) could allow for Swiss scale estimation.

Contact Us

Frédéric Montet
frederic.montet@hefr.ch

Alessandro Pongelli
alessandro.pongelli@hefr.ch

Michal Bryxí
michal.bryxi@vast.ch

References

[1] 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*

[2] 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 Lu´ıs Torgo P B and Moniz N (PMLR) pp 36–50 URL *https://proceedings.mlr.press/v74/branco17a.html*

[3] 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