Statistical modelling of the energy reference area based on the Swiss building stock

Thomas Schluck\textsuperscript{1}, Kai Nino Streicher\textsuperscript{2}, Stefan Mennel\textsuperscript{1}
\textsuperscript{1}University of Applied Sciences Lucerne, Technikumstrasse 21, 6048 Horw, Switzerland; \textsuperscript{2}University of Geneva, Boulevard Carl-Vogt 66, 1205 Geneva, Switzerland

E-mail: thomas.schluck@hslu.ch

Abstract. The energy reference area \textit{ERA} is one of the most important reference values for practitioners and scientists to quantify the energetic performance of a building. Its fast and easy prediction from generally available building characteristics is a regular need. Here we present a comprehensive statistical model to predict the ERA based on a multiple regression approach and we motivate why the often chosen bivariate regression, i.e. simple regression should be disfavoured. Basis for the model development was a comprehensive data set of the Swiss buildings stock with about 30'000 buildings and 23 descriptive features. The final model only needs four building features to model the energy reference area. These features are the footprint, the living area, the number of floors and the buildings’ usages. The main key performance indicators during the model development were the coefficient of determination $R^2$ and the uncertainty given by the root-mean-square-error RMSE, which was determined on a test set (about 10% of the dataset). The here presented multiple model has an $R^2$ of 99.7% and gives low uncertainties for the predicted ERA ranging between 25% and 10% of the predicted value. Thus this work allows estimating the energy reference area and its level of uncertainty fast and easy. We are convinced that our efforts will support practitioners as well as scientists in their daily work by providing more accurate and less biased estimations of the \textit{ERA}.

1. Introduction

Buildings’ energy performances are often benchmarked on the basis of their specific energy consumption, i.e. their energy consumption normalised to their size. For instance, in the UK the floor area is used in the Standard Assessment Procedure [1], which was developed by the Building Research Establishment (BRE) to assess a building’s compliance with UK building regulations. The German passive-house initiative defines its own reference, i.e. the treated floor area [2, 3]. A comparison of these two systems can be found in Mitchell et al. [4]. In Switzerland the energy reference area \textit{ERA} (so-called \textit{Energiebezugsfläche}) is commonly used to normalise a building’s energy consumption. Due to its importance to national legislatives like the "Mustervorschriften der Kantone im Energiebereich" (MuKEn) [5] or assessment frameworks like the "Building Energy Certificate of the Cantons" (GEAK) [6] it is defined by code [7]. Especially for the GEAK the ERA is recorded by a certified expert. The drawback of this thorough approach is its high demand on resources. In principle, the energy reference area comprises all actively conditioned, i.e. heated or cooled areas of a building, but due to many specialities in its definition, its determination is complex. Thus it cannot easily be calculated from any other area and therefore practitioners as well as scientists call for more simple and
generally applicable approaches to retrieve the ERA from a few given and easily attainable buildings’ characteristics. So far the ERA has often been simply mapped to the gross floor area (GFA), which basically is retrieved from a building’s footprint (FTP) and number of floor (NOF), or the living area (LVA). The derivation of the ERA from given building data is still an open topic for discussion. We contribute to this discussion by analysing a joint dataset of the GEAK and the “Federal Register of Buildings and Dwellings” GWR [8]. The finally used dataset contains more than 30’000 residential and non-residential buildings with 23 features, including their ERA. This is the same database that was already used to analyse the thermal performance of the Swiss building stock [9, 10, 11].

2. Data and Methods
As many features of the dataset showed high collinearity the dataset was reduced to only non-collinear features. The remaining features were the footprint of the building FTP (“Bebaute Fläche”), the living area LVA (“Wohnfläche”), the number of floors NOF (“Stockwerkanzahl”), the building’s usage type BLT (“Gebäudenutzung”), the AGE of the building and its location LOC in a urban, sub-urban or rural area. The BLT differences the buildings according to the GEAK categories into Administration, Multi-Family-Home, School and Single-Family-House. Firstly, simple bivariate linear regressions were performed using ordinary-least-square and a modern, robust estimation [14] for later comparison to the multiple regression model. As ordinary-least-square and robust estimations didn’t give any significant differences in $R^2$ and RMSE – even for the multiple regressions – the ordinary-least-square was kept the standard estimation approach. To develop the multiple regression model an oversize model was assumed using the mentioned features and allowing for interaction between all features. Based on the LASSO approach in combination with a step-wise model selection [13] features and their interactions were systematically eliminated from the model (intermediary results are not shown). These approaches eliminated mainly multiple interactions of features, but also the AGE and LOC were found unnecessary. About 10% of the data was hold back for validation purposes and never used during model development, i.e. the data was split into a test and training set. Main model performance indicators were the adjusted coefficient of determination $R^2$ and the root-mean-squared-error RMSE of the predictions. The RMSE was calculated on the test set and on the training set using cross-validation [13]. Here only the on the test set calculated RMSE is presented as there were no major differences to the cross-validation based RMSE. All continuous variables were transformed, i.e. the ERA, LVA and FTP were logarithmically transformed, while the NOF was transformed using the square-root.

3. Results and Discussion
Figure 1 shows the distributions of the energy reference area, the footprint and the living area for the different building usages. All variables are positively skewed, which can be best seen for the two residential categories Single-Family-Houses and Multi-Family-Homes as 99% of the buildings belong to one of these. All residual analyses indicated the transformation of the variables.

Figure 2 shows the bivariate regressions of the transformed ERA in dependence of the transformed LVA, respectively FTP and their key performance indicators. Sub-graph A shows no substantial dependency between the ERA and the LVA for the categories Administration and School in contrast to the residential uses. The high variance of School and Administration in relation to the LVA also translates into a twice as high RMSE compared to the regressions using FTP. This may be explained by the fact that the majority of both Administration (155 of 348) and School buildings (63 of 120) contain no living area. In general, the footprint of a building is an important predictor for all usages. Overall only the regression considering LVA for the category of Multi-Family-Homes perform acceptably well among all bivariate models. However,
Figure 1: Histograms of the energy reference area \( \text{ERA} \), the living area \( \text{LVA} \) and the footprint \( \text{FTP} \) in \([\text{m}^2]\) for the different building categories \textit{Administration} (total 348), \textit{Single-Family-Houses} (total 17310), \textit{Multi-Family-Homes} (total 12523) and \textit{Schools} (total 120). Binwidth is 25 \( \text{m}^2 \).

Figure 2: Sub-graphs A and B show bivariate regressions of the logarithmically transformed \( \text{ERA} \) in relation to the transformed \( \text{LVA} \) and \( \text{FTP} \). Sub-graph C shows the RMSE on the test set for the ordinary-least-square (OLS) and robust estimation (SMDM) for the different building usages. Sub-graph D gives the adjusted \( R^2 \). In comparison, the dotted line in sub-graphs C and D give the results of the final multiple regression model.
Figure 3: Residual analysis of the final model. The Tukey-Anscombe-plot (left-upper corner) and the scatter plot of the standardized residuals (left-lower corner) indicate some unresolved structures, but were deemed acceptable. The distribution of the residuals are centred and heavy-tailed (QQ-Plot right-upper corner), but approximately normally distributed. The leverage plot (right-lower corner) doesn’t indicate any large influences from extreme value.

even this best bivariate regression model under-performs compared to the final model (1), which reaches a $R^2$ of 99.7% and with 0.124 the lowest RMSE.

Figure 3 shows the most important plots of the residual analysis for the final model (1) and were found acceptable. The final model is given in equation 1.

$$Y_{ERA,i} = \beta_1 X_{LVA,i} + \beta_2 X_{FTP,i} + \beta_3 X_{NOF,i} + \left\{ \begin{array}{l} \beta_4 \text{BLT} = \text{"Admin"} \\ \beta_4 \text{BLT} = \text{"SFH"} \\ \beta_4 \text{BLT} = \text{"MFH"} \\ \beta_4 \text{BLT} = \text{"School"} \end{array} \right\} + \left\{ \begin{array}{l} \beta_5 \text{BLT} = \text{"SFH"} \cdot X_{FTP,i} \\ \beta_5 \text{BLT} = \text{"MFH"} \cdot X_{FTP,i} \\ \beta_5 \text{BLT} = \text{"School"} \cdot X_{FTP,i} \\ \beta_6 \text{BLT} = \text{"SFH"} \cdot X_{LVA,i} \\ \beta_6 \text{BLT} = \text{"MFH"} \cdot X_{LVA,i} \\ \beta_6 \text{BLT} = \text{"School"} \cdot X_{LVA,i} \end{array} \right\} + \varepsilon_i \quad (1)$$

Here $Y_{ERA,i}$ is the modelled energy reference area ERA of building $i$, $X_{LVA,i}$ the living area LVA of that building, $X_{FTP,i}$ the building’s footprint FTP and $X_{NOF,i}$ its number of floors NOF. All variables are log$_{10}$-transformed, but the number of floors, which is transformed using the square-root. The building’s usage is coded by BLT; $\varepsilon_i$ is the disturbance term. As one would expect the living area, the footprint and the number of floors appear as main effects in the model ($\beta_1$ to $\beta_3$) as well as the buildingtype, which gives a building usage dependent intercept (different $\beta_4$ for BLT categories). Depending on the building’s usage the slope for the living area $X_{LVA,i}$ and the footprint $X_{FTP,i}$ are also adapted (BLT/LVA, respectively BLT/FTP interactions).

The estimates of the final model’s coefficients and the corresponding confidence intervals are given in table 1. Finally, the finished multiple model was used to predict (the logarithmic) ERA
Table 1: Confidence intervals and estimates for the predictors of the final model.

|             | 2.5 %  | 97.5 % | estimates |
|-------------|--------|--------|-----------|
| $\beta_1$  | -0.042 | -0.019 | -0.031    |
| $\beta_2$  | 0.736  | 0.797  | 0.766     |
| $\beta_3$  | 0.226  | 0.237  | 0.231     |
| $\beta_{4,BLT="Admin"}$ | 0.515 | 0.689 | 0.602 |
| $\beta_{4,BLT="SFH"}$ | 0.503 | 0.553 | 0.528 |
| $\beta_{4,BLT="MFH"}$ | 0.402 | 0.442 | 0.422 |
| $\beta_{4,BLT="School"}$ | 0.165 | 0.496 | 0.331 |
| $\beta_{5,BLT="SFH"}$ | -0.480 | -0.416 | -0.448 |
| $\beta_{5,BLT="MFH"}$ | -0.500 | -0.436 | -0.468 |
| $\beta_{5,BLT="School"}$ | -0.001 | 0.130 | 0.065 |
| $\beta_{6,BLT="SFH"}$ | 0.357 | 0.389 | 0.373 |
| $\beta_{6,BLT="MFH"}$ | 0.480 | 0.511 | 0.496 |
| $\beta_{6,BLT="School"}$ | 0.006 | 0.054 | 0.030 |

Figure 4: Sub-graph A shows the test set values plotted against the on the test set predicted values. In red the diagonal with slope one is given, in blue a linear fit of this data. Both lines match greatly and the points scatter closely around the theoretical red diagonal. Sub-graph B shows the cumulative RMSE for the predicted and back-transformed ERA on the test data set till 5000 m².

on the test data set and calculate the cumulative RMSE distribution of the predictions. Figure 4 shows in graph A a scatter plot of the test set values against the predicted values. The points scatter closely along the theoretical diagonal with slope one (red line) as one would expect for a prediction with satisfactory accuracy. A linear fit was added for direct comparison of the red theoretical and blue observed trend. As can be seen, both lines match closely indicating a low bias of the fit. Graph B shows the cumulative RMSE of the predicted and back-transformed ERA, i.e.
it shows the empirical distribution for the RMSE, respectively the uncertainty of the predicted ERA. It shows that small areas also have small absolute, but high relative uncertainties in the range of 25%. For larger ERAs the absolute RMSE levels out towards the absolute RMSE of 568 m$^2$, such that the relative uncertainties for ERAs in the range of 5000 m$^2$ trend towards 10%.

4. Conclusion
In this work a multiple regression model (equation (1)) was presented to predict the energy reference area (ERA) of an unknown building in Switzerland. The model takes a building’s footprint, living area, number of floors and usage into account and predicts its $\log_{10}$ transformed energy reference area. It was motivated that the here presented model outperforms any bivariate regression models, which are often used by practitioners. The presented model shows a predictive quality that is low in bias and sufficiently high in accuracy. The models coefficient of determination $R^2$ is very high at 99.7%. The cumulative distribution of the RMSE shows how the uncertainty level of the predicted ERA depends on the predicted size, such that the relative uncertainty is highest with 25% for small ERAs. For ERAs in the range of 5000 m$^2$ the relative uncertainty is already in the range of 10%. These uncertainties are well acceptable for the early planning or conceptualisation phase. Thus with the here presented results anyone can estimate a building’s ERA by first calculating the $\log_{10}(\text{ERA})$ of a building using equation (1), then calculating the ERA simply by taking 10 to the power of $\log_{10}(\text{ERA})$ and finally estimating the uncertainty level from graph B of figure 4.

Acknowledgement
The authors would like to acknowledge Innosuisse through the SCCER FEEB&D, the University of Geneva and the Lucerne University of Applied Sciences and Arts for their financial support and infrastructures.

References
[1] Building Research Establishment. The government’s standard assessment procedure for energy rating of dwellings. Technical report, Building Research Establishment, 2012.
[2] W Feist, R Pflijger, and B Kaufmann. PHPP Passive house planning package – The energy balance and design tool for efficient buildings and retrofits. Darmstadt, Germany, 2015.
[3] Christina J Hopfe and Robert S McLeod. The Passivhaus Designer’s Manual: A technical guide to low and zero energy buildings. Routledge, 2015.
[4] Rachel Mitchell and Sukumar Natarajan. Normalising domestic space heating demand using post hoc models. Building Services Engineering Research and Technology, 40(3):340–359, 2019.
[5] EnDK – Konferenz Kantonaler Energiedirektoren. Mustervorschriften der Kantone im Energiebereich, 2014.
[6] Verein GEAK-CECB-CECE. GEAK - Gebäudeausweis der Kantone, 2019. [Online; accessed 04. Mai 2019].
[7] Schweizerischer Ingenieur- und Architektenverein. SIA 380 – Grundlagen für energetische Berechnungen von Gebäuden, 2015.
[8] Bundesamt für Statistik. GWR – Eidg. Gebäude- und Wohnregister, 2019. [Online; accessed 04. Mai 2019].
[9] Kai Nino Streicher, Pierreves Padey, David Parra, Meinrad C. Bürer, and Martin K. Patel. Assessment of the current thermal performance level of the swiss residential building stock: Statistical analysis of energy performance certificates. Energy and Buildings, 178:360 – 378, 2018.
[10] Kai Nino Streicher, Pierreves Padey, David Parra, Meinrad C Bürer, Stefan Schneider, and Martin K Patel. Analysis of space heating demand in the swiss residential building stock: Element-based bottom-up model of archetype buildings. Energy and Buildings, 184:300–322, 2019.
[11] Kai Nino Streicher, David Parra, Meinrad C Buerer, and Martin K Patel. Techno-economic potential of large-scale energy retrofit in the swiss residential building stock. Energy Procedia, 122:121–126, 2017.
[12] Bundesamt für Statistik. Eidgenössisches Gebäude- und Wohnungsregister Merkmalskatalog, 2015.
[13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.
[14] Manuel Koller and Werner A Stahel. Sharpening wald-type inference in robust regression for small samples. Computational Statistics & Data Analysis, 55(8):2504–2515, 2011.