Modelling the energy demand of individual buildings in Switzerland

Author: Andy Gubser
Quartnerstrasse 2
CH-8882 Unterterzen
andy.gubser@stud.hslu.ch

Lecturer: Prof. Dr. Philipp Schütz
Lucerne School of Engineering and Architecture
Technikumstrasse 21
CH-6048 Horw
philipp.schuetz@hslu.ch

Co-Lecturer: David Suter
geoimpact AG
Heinrichstrasse 267
CH-8005 Zürich
david.suter@geoimpact.ch

Lucerne University of Applied Science and Arts
Master in Applied Information and Data Science (MScIDS)
Autumn Semester 2021

Heitenried, 23 December 2021

Management Summary

Climate change is in full swing, and to avoid the worst consequences, we must eliminate our greenhouse gas emissions. One of the most significant contributors is the building sector, and sustainable solutions are ready and competitive. However, most buildings are still heated with fossil fuels, and many of them urgently need energy retrofits. Geoimpact aims to change this and provide building owners and other stakeholders with fact-based information about sustainable alternatives. To improve their consulting, they need a reliable estimate of the heating energy consumption of each building.

This is the question we are dealing with in this master thesis. We have investigated which models achieve the highest accuracy in predicting the energy performance of individual buildings. Which target variables, which feature variables, and which regressors are appropriate. And whether we can increase accuracy by training multiple models on specific subsets rather than using one model for all buildings.

To this end, we merged energy consumption data from Biel, Geneva, and St. Gallen with data from our industry partner geoimpact and data from the Federal Register of Buildings and Dwellings (RBD) and Minergie. We trained models with different regressors and target variables. We evaluated their predictions with performance evaluation indicators such as the Coefficient of Variance (CV) and the Mean Absolute Percentage Error (MAPE). Thereby, we assessed the Multiple Linear Regressor (MLR), the Random Forest Regressor (RFR), the XGBoost Regressor (XGB), and the Support Vector Regressor (SVR) on the target variables of heating energy consumption (HEC), heating energy consumption per square meter of area (HEPI), and both log-transformed (log-HEC, log-HEPI).

We found that the RFR and XGB are the best fitting regressors and that it does not matter so much on the target variable. We get leverage when we divide the buildings into more homogeneous groups, such as residential buildings and old residential buildings, thus reducing the data variance. We have shown that this increases the accuracy of the models. This finding may be generalizable to other data science projects.

Based on our results, geoimpact can use the prototype for modelling heating energy consumption, which can be further improved with additional information (additional features and observations). In addition, more extensive studies with higher geographic coverage can be conducted. For this purpose, we recommend including more information on occupant behaviour, climate conditions, and building characteristics, such as the insulation quality of facades, roof, windows, and their solar orientation.

Table of Content

List of Tables

List of Illustrations

List of Abbreviations

AL	Applications and Light
ANN	Artificial Neural Network
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CECB	Cantonal Energy Certificate for Buildings
COP	Coefficient of Performance
Copula	D-vine copula quantile regression
CRISP-DM	Cross-Industry Standard Process for Data Mining
CV	Coefficient of Variation
DHW	Domestic Hot Water
DTR	Decision Tree Regressor
EGID	Building identifier (*Gebäudeidentifikator*)
EQM	Energy Quantification Methods
ERA	Energy Reference Area
FOEN	Federal Office for the Environment
GBR	Gradient Boosting Regressor
HEC	Heat Energy Consumption (unit: kWh/year)
HEPI	Heat Energy Performance Indicator (unit: kWh/m2/year)
HVAC	Heating, ventilation, and air conditioning
log-HEC	Log-transformed Heat Energy Consumption (unit: kWh/year)
log-HEPI	Log-transformed Heat Energy Performance Indicator (unit: kWh/m2/year)
MAE	Mean Average Error
MAPE	Mean Absolute Percentage Error
MLR	Multiple Linear Regressor
PEM	Performance Evaluation Measures
RBD	Federal Register of Buildings and Dwellings (*Eidgenössisches Gebäude- und Wohnungsregister, GWR*)
RC	Resistance-Capacitance models
RFR	Random Forest Regressor
RMSE	Root Mean Squared Error
SEP	Swiss Energy Planning
SFOE	Swiss Federal Office of Energy
SH	Space Heating
SIA	Schweizerischer Ingenieur- und Architektenverein
SLR	Simple Linear Regressor
SVR	Support Vector Regressor
XGB	XGBoost Regressor

List of data sources

Data Category	Data Source	Extraction Date
Heat Energy Consumption Biel	Energie Service Biel (ESB)	September 7th 2021
Heat Energy Consumption Geneva	Système d'information du territoire à Genève (SITG)[1]	June 5th, 2021
Heat Energy Consumption St. Gallen City	Umwelt und Energie, Stadt St. Gallen	June 18th, 2021
Building Characteristics	Federal Buildings and Housing Register (RBD)	December 11th, 2021
Building Characteristics	Minergie Schweiz	May 19th, 2021
Calculations of geoimpact	geoimpact	October 12th, 2021

[1] « Source : Système d'information du territoire à Genève (SITG), imprimé et/ou extrait en date du 5 juin 2021 », https://ge.ch/sitg/fiche/2177

1. Introduction

The United Nations Paris Agreement sets ambitious climate targets to avoid the most serious consequences of human-made climate change. Switzerland has decided to reduce its final energy consumption to below 1990 levels by 2030 and to achieve climate neutrality by 2050 (Federal Office for the Environment FOEN, 2020).

The building stock consumes about 100 TWh or about 45% of Switzerland's final energy demand (Swiss Federal Office of Energy SFOE, 2020a) and account for 25% of the country’s annual greenhouse gas emissions (Federal Office for the Environment FOEN, 2021b). 60% of the buildings use heating oil or natural gas for the processing of space heating or domestic hot water (Federal Office for the Environment FOEN, 2021a; Federal Statistical Office FSO, 2017; Swiss Federal Office of Energy SFOE, 2020b).

62% of residential buildings in Switzerland were built before 1980 (Federal Statistical Office FSO, n.d.). Their average consumption varies widely between 120 and 250 kWh/m2/year (Amt für Umwelt und Energie AUE, 2020; Binz et al., 2020; Federal Office for the Environment FOEN, 2021c; Gerster & Nietlisbach, 2014; Schweizerische Energie-Stiftung, n.d.; Umwelt und Energie UWE, 2015; Wick, 1982). With better energy insulation, these values could be reduced to 60 to 80 kWh/m2/year. (Binz et al., 2020; energie-environnement.ch, n.d.; Gloor, 2020). It is especially worthwhile to renovate roofs and replace windows in older houses. This alone can reduce the need for heating and cooling energy by 20 to 30% (Federal Laboratory for Materials Testing and Research, 2021).

Despite this great energy saving potential, only about 45% of the buildings built before 1980 have been retrofitted until 2014 (Konferenz Kantonaler Energiedirektoren EnDK, 2014). In 2020, there are still more than 1 million buildings are in urgent need of energetic retrofits (Swiss Federal Office of Energy SFOE, 2020). The retrofit rate – the proportion of buildings that are retrofitted in a year - is stagnating at around 1% per year. In other terms, it would require 100 years to retrofit the complete building stock, which is too slow to realize the energy transition (Federal Laboratory for Materials Testing and Research, 2021).

Another issue is the replacement of fossil heating systems. Energieforschung Stadt Zürich (2019) and Swiss Federal Office of Energy SFOE (2021) have found that, although modern heat pumps can compete with fossil heating systems in terms of price (Energie Schweiz AG, n.d.), more than 50% of fossil heating systems are replaced by another fossil heating system (Energieforschung Stadt Zürich, 2019; Swiss Federal Office of Energy SFOE, 2021). Energieforschung Stadt Zürich (2019) found that in urban areas, it is even 80%, and concludes that building owners underestimate the need of refurbishment and lack of planning. They wait until the heat system fails before retrofitting, and then under time pressure, they opt for the obvious one-to-one replacement.

To increase the retrofit rate and promote renewable energy sources in residential buildings, Switzerland addresses this challenge with the Energy Strategy 2050 (UVEK, n.d.). The Swiss Federal Office of Energy (2020b) envisages a halving of heat and electricity consumption in Swiss buildings by 2030 and the replacement of all fossil-fuel heating systems by 2050. The consumption of heat and electrical energy is to be reduced from 145 kWh/m2/year today to 60 kWh/m2/year.

The Swiss Federal Office of Energy (2020b) promotes its energy efficiency labels for buildings - the best known are the Cantonal Energy Certificate for Buildings (CECB) and Minergie. In this way, the SFOE promotes the determination of the energy efficiency of every building in Switzerland and makes this information available to the public by 2050. Thus, the energy efficiency of the building becomes a criterion for sale or rent.

1.1. Geoimpact and Swiss Energy Planning

There are private initiatives to promote the transparency of buildings energy performance and thus to increase the retrofit rate. The company geoimpact provides the platform Swiss Energy Planning (SEP). The platform presents daily updated building and infrastructure data from different sources. Thereby, synergies and potentials at the building and neighbourhood level are easily detected and can be quickly exploited. The goal is to equip the stakeholders and decision-makers in the energy transition with the right tools and information to master the challenges in the process toward a sustainable energy supply. (SwissEnergyPlanning SEP, n.d.)

Geoimpact works to capture the building's heat energy needs and to integrate them into SEP. With this paper, we address the estimation of the heat energy demand, that estimation allows for more efficient consulting and preparation of preliminary offer for heating system replacement. On a district or community level, the heat demand estimation is used for heat planning, e.g., for the planning of heat networks. (geoimpact AG, personal communication, 2021)

1.2. Research Question and Hypothesis

In this thesis, we model explicitly the heat energy demand of individual buildings in Switzerland. Thereby, heat energy includes the energy used for space heating (SH) and domestic hot water (DHW). To achieve this, we use data on measured energy consumption for the cities of Biel, Geneva, and St. Gallen, combined with data on building characteristics from RBD and Minergie, and some calculations from geoimpact.

Here, we investigate whether multiple models trained on specific subsets of buildings result in higher prediction accuracy compared to one single model for predicting the heat demand of all buildings. Having split the data into a training, a validation, and a test set. Thus, we can play around with the training and validation set and look for more homogeneous building classes, e.g., data of residential buildings, where we can potentially achieve higher accuracy. Each model is validated on Performance Evaluation Metrics.

Thus, we formulate our main research as follows:

Which set of models achieve the highest accuracy in predicting the energy performance of individual buildings?

To answer the question, multiple smaller steps are needed. Therefore, we separate the main research question into the following sub-questions regarding data, target variable, features, and model evaluation:

Target Variables

1) For which dimensions of energy consumption (Absolute Energy Consumption, per m2, log-transformed or not) do we achieve the highest accuracy?

Features

2) What relevant characteristics can be found in the literature for predicting the energy performance of buildings?
3) Which relevant features are important for the machine learning models developed here?

Model evaluation

4) Can we achieve higher prediction accuracy by using multiple models trained on specific subsets of buildings, compared to one single model for predicting the heat demand of all buildings?
5) Can we achieve a Coefficient of Variation (CV) below 25% (ASHRAE, 2005, p. 3)?

6) Can we achieve a Maximal Absolute Percentage Error (“MaxAPE”) below 15% (geoimpact AG, personal communication, December 16, 2021)?
7) When do we need higher geographic resolution?

1.3. Structure of the thesis

The remainder of this thesis is structured as follows: Chapter 2 provides theoretical foundations about Energy Quantification Methods (EQM), the Statistical Learning Algorithm used and Performance Evaluation Measures. Chapter 3 explains the design of this study and the methodology. In chapter 4, we present our results. Chapters 5 discusses and concludes.

2. Background of Estimating Building’s Heat Energy Consumption

This chapter provides a broad theoretical foundation to address the research question of estimating the heat energy consumption of individual buildings. Section 2.1 discusses the appropriate target measure of the model. Section 2.2 introduces the relevant literature of Energy Quantification Methods (EQM). Section 2.3 explains the Machine Learning Methods used and shows their advantages and drawbacks. Then, in section 2.4, we discuss how to evaluate the predictions with Performance Evaluation Measures.

2.1. Energy Consumption Discussion

We further discussed, how to define the examined heat energy consumption. Literature distinguishes three terms, 1) heat energy consumption (German: *Energieverbrauch*), 2) heat energy demand (*Energiebedarf*) and 3) useful heat (*Nutzwärme*).

While heat energy consumption 1) is determined from real meter reading data and billed by the energy supplier, the heat energy demand 2) is the amount of energy required to heat the building, under a standardized framework. The energy demand is mainly driven by the building's thermal insulation quality and the heating system and does not consider subjective habits of the occupants. Thus, heat energy consumption and heat energy demand can differ fundamentally, and we cannot derive one metric from the other (Ackermann, 2019).

The last metric, useful heat energy, 3), describes the energy output of a heating system used to heat the building. It is calculated by multiplying the drive energy with the heat system's efficiency. The drive energy is the energy used to operate the heating system (Kamoshida et al., 1990) and corresponds to energy consumption in our terminology. The efficiency ranges from 0.7 to 1 for gas and oil heating systems (Thermondo, 2019) and can exceed values of 4 for modern heat pumps. The efficiency of heat pumps is also known as coefficient of performance (COP). (Hoval Schweiz, n.d.). (Kamoshida et al., 1990; Thermondo, 2019).

In our study, we focus on 1) heat energy consumption. For two reasons: We have no information on the efficiency or age of the heating systems, a major issue, especially for district heat. Further, more than 90% of all examined buildings heat with gas or oil, and their efficiency is about 1. Consequently, the Heat Energy Consumption for these buildings is about equal to Useful Heat.

Thereby, we reduce complexity and error prunes due to 2) potentially miss leading energy demand values. 3) In theory, it would be ideal to model useful heat. However, with further information, we cannot evaluate the energy efficiency of district heated buildings. Since most buildings are heated with boilers, the efficiency of the heating systems does not vary that much and is approximately equal to 1. Consequently, the heat energy consumption is approximately the same as useful heat.

2.1.1. Functional form

Braulio-Gonzalo et al. (2021) investigated which target variable is the most appropriate for modelling total residential energy consumption, consisting of heat and electricity. They conclude that their proposed response variable Energy Consumption per Occupant, Area and Year outperforms the other better known response variables Energy Consumption per Year, Energy Consumption per Occupant and Year, Energy Consumption per Area and Year.

Since the occupant’s number in our study is not known, we discuss the potential target variables Heat Energy Consumption (HEC, kWh/year) and Heat Energy Performance Indicator (HEPI, unit: kWh/m2/year).

HEPI is usually calculated dividing the HEC through the energy requirement area (ERA). Simply speaking, the ERA describes the sum of actively heated or cooled areas per building. But, many specificities makes the variable hard to determine exactly; thus, this variable is usually unknown (Schluck et al., 2019; Schweizerischer Ingenieur- und Architekten-Verein SIA, 1982; Umwelt und Energie UWE, 2008). We approximate the ERA by the Gross Floor Area, the multiplication of the number of floors with the building floor area as recommended by Swiss Federal Office of Energy SFOE (2016).

We append the log-transformed HEC (log-HEC) and log-transformed HEPI (log-HEPI) to the list of our target variable candidates. Thereby, we expect to increase the model's accuracy (Wilhelm, 2020) by capturing multiplicative effects of the features on our target variable (Osmulski, 2018) and by reducing the relative importance of outliers in the target variable (ND, 2020).

Consequently, our target variable candidates are: HEC, HEPI, log-HEC and log-HEPI. Although, we train each model on different target variables, the predicted and the actual values are back-transform to HEC (unit: kWh/year) before the model evaluation.

2.1.2. Available datasets for building heat energy estimations

The Swiss Federal Office of Energy (2020b) has issued several energy efficiency labels for buildings in Switzerland. One of these labels is the Cantonal Building Energy Performance Certificates (*Gebäudeenergieausweis der Kantone,* CECB). The CECB is the most extensive database for buildings energy-related data in Switzerland. More than 50'000 buildings are recorded, including quality assessments of the building envelope and the building's overall energy efficiency (Swiss Federal Office of Energy SFOE, 2019; Verein GEAK-CECB-CECE, 2021).

Another large dataset is held by the Minergie Association, containing data for about 45'000 buildings in Switzerland. Minergie is a Swiss building standard for new and modernized buildings equipped with a high-quality building envelope and low-energy consumption (Minergie, 2021).

The use of such a database would be handy, however, while CECB's data is strictly protected by law (A. Husi, personal communication, April 12, 2021), the Minergie data only includes energy-efficient buildings. Another potential data source are property management companies, providers of heat contracting and billing services. But, these companies have the data not processed for extraction or restrict the release for business or data protection reasons. And many data is either faulty or outdated (Carisch et al., 2020).

2.2. Energy Quantification Methods (EQM)

Although, we model explicitly buildings HEC, in this chapter, we also discuss literature that considers Overall Energy Consumption consisting of heating, cooling and electricity. The literature of Energy Quantification Methods (EQM) divides the methods used into top-down and bottom-up approaches.

Bourdeau et al. (2019), Deb & Schlueter (2021), Foucquier et al. (2013), Li et al. (2014), Rabani et al. (2021), Swan & Ugursal (2009) and Zhao and Magoulès (2012) provides a sound founded understanding of EQM. They distinguish 1) top-down approaches, they model the energy consumption of a sample building stock and 2) bottom-up methods, they estimate the energy consumption of a sample of individual buildings.

2.2.1. Top-Down Approaches

Top-down methods analyse a whole sample building stock on pre-aggregated datasets such as CECB. The methods evaluate large-scale regional or urban retrofit measurements; thus, they are also named Urban Building Energy Modelling. Deb & Schlueter (2021) itemizes different applications of these approaches:

1) Energy benchmarking identifies the retrofit potential through clustering and dimension reduction.
2) Feature extraction identifies the most influential variables by regression and data mining.
3) Energy signature correlates energy consumption and weather data to gain insights into heating and cooling systems.

Advantages of these methods are the use of existing data sets; data-driven, theory-independent benchmarks; and the suitability for policymaking and large-scale planning. Disadvantages are the dependence on data quality and consistencies and the unsuitability for individual retrofitting of buildings. (Deb & Schlueter, 2021)

Researchers use 1) energy benchmarking techniques to assess building's energy-saving potential, mainly through clustering and dimension reductions. Geyer et al. (2017) and Gao & Malkawi (2014) presented hierarchical and k-means clustering approaches. Alternatively, (Park et al., 2016) proposed a procedure based on correlation analysis, decision-trees and analysis of variance (ANOVA). Liu et al. (2017) combined clustering approaches with decision-trees to benchmark hourly energy data against annual values. Re Cecconi et al. (2019) combined energy performance certificates with geolocated data. Streicher et al. (2019) utilized the CECB to classify buildings in Switzerland based on their SH consumption. They found that the ones with the highest demand of SH are rather old single houses located in the countryside, and the main predictor is ERA.

The main drivers of building's energy consumption are detected by 2) feature extraction through data mining and regression techniques. Tong et al. (2016) analyzed occupant's behavior with smart meter and demographic data. They found that household's overall energy demand highly correlates with its occupants' employment status and internet usage. McLoughlin et al. (2012) concluded that the buildings' typology, the bedrooms count as well as households age and composition, social class, and cooking habits all significantly impact buildings electricity consumption.

Energy signature methods 3) correlate high-resolution energy consumption data collected by smart meters with weather data to gain insights into the building's heating and cooling systems. Pasichnyi et al. (2019) analyzed the potential of different retrofitting scenarios of buildings by combining energy signature methods with EnergyPlus, a tool to simulate buildings energy flows. Westermann et al. (2020) categorized building's types and heat system through unsupervised methods.

2.2.2. Bottom-Up Approaches

The bottom-up approaches analyse the renovation strategies of individual buildings. Advantages of bottom-up approaches are the detailed evaluation of different retrofit measures for an individual building and the specific and accurate predictions. Disadvantages are their strong dependency on data quantity and quality. (Bourdeau et al., 2019; Deb & Schlueter, 2021; Foucquier et al. 2013; Li et al. 2014; Rabani et al. 2021; Swan & Ugursal, 2009; Zhao & Magoulès, 2012)

Bourdeau et al. (2019), Deb & Schlueter (2021), Foucquier et al. (2013), Li et al. (2014), Rabani et al. (2021), Swan & Ugursal (2009) and Zhao & Magoulès (2012) classify them as follows:

1) White-box models consider the physical constraints of the building.
2) Black-box models use statistical techniques to fit the features on buildings' energy demand.
3) Grey-box models combine white- and black-box methods.

Since 1) white-box models fully embed building's physical behavior in the modelling, they can explain the link between input and forecasted energy demand in full detail. Therefore, they require extensive prior knowledge and data about the building. Nevertheless, researchers often use white-box models as EQM simulation applications. Examples are EnergyPlus (Crawley et al., 2000, 2001; Wang et al., 2018) and IDA-ICE (Rabani et al., 2021). Another white-box application is *Tachion*. This software simulates the impact of different retrofit scenarios on the building's energy demand for SH, SC, Applications and

Light. It considers building properties such as ERA, construction year, number of floors, heated attic and basement, windows alignment and occupants behavior. (Solar Campus GmbH, 2007, 2020)

In contrast, 2) black box models look for purely data-driven relationships between input data and predicted energy demand without knowing the building physical conditions and energy flows. They use statistical learning and training data. Existing literature tends to favour artificial neural networks (ANN) and support vector regressions (SVR) (Holcomb et al., 2009; Wenninger & Wiethe, 2021). But (Wenninger & Wiethe, 2021) shows that Extreme Gradient Boosting (XGB) and Random Forest Regressor (RFR) and D-vine copula quantile regression (Copula) can achieve similar predictive accuracy. Thereby, all black-box methods outperformed the white-box method with almost 50% higher predictive power.

Finally, 3) grey-box models combine the advantages of white- and black-box models. They use simplified building dynamics models of white-box models (Resistance-Capacitance models, RC) with data fitting methods of black-box models. Examples are a platform for retrofitting simulations (Melillo et al., 2020; Schuetz et al., 2020, 2019) and the degree-day method (Al-Homoud, 2001; Berger & Worlitschek, 2019).

The next two approaches of Swiss Federal Office of Energy (2016) and Ecospeed Immo (2021) come close to the goal of geoimpact, but these two approaches are too simplified and less accurate to integrate into their software (geoimpact AG, personal communication, December 16, 2021).

Swiss Federal Office of Energy (2016) has developed the application Sonnendach.ch, which allows any building owner to retrieve the solar energy potential of their building. The application also provides estimates for SH and DHW consumption. These models are based on building type specific SIA guideline values for SH and DHW consumption per square meter. The ERA is approximated in a simplified way as multiplication of the number of floors by the building floor area. Now the SH consumption is estimated by multiplying by the corresponding SIA guideline value with the ERA, and adjust the result for the building's construction period, the heat system efficiency, and climatic conditions. To estimate the DHW consumption, the ERA is multiplied by the SIA guideline value for DHW per square meter.

Ecospeed Immo (2021) creates energy and greenhouse gas reports for several regions in Europe. They claim to achieve a prediction accuracy of more than 90% on a regional level (Ecospeed AG, 2019). In their model they use the ERA as recorded in the RBD if applicable, or calculate the ERA by multiplying the living space by a guideline value depending on the building category (Federal Office for the Environment FOEN, 2016). Similar to the calculation of Swiss Federal Office of Energy (2016), the SH consumption is estimated by the multiplication of the ERA by a specific energy parameter that depends on building's type and construction year. The resulting estimate is corrected for the heating system and climate conditions. The DHW consumption is modelled analogously to the calculation of SH, using a DHW specific energy parameter.

Wenninger and Wiethe (2021) note that there is a shortage in literature for data-driven models that predict HEC on a annual basis. Our study fits into the stated lack of studies, and we would classify this thesis as a black-box models study.

2.3. Machine Learning Methods

Wenninger and Wiethe (2021) has shown the algorithms RFR, XGB, SVR can achieve high prediction accuracy. SVR is even the most favoured method besides ANN. In addition, we use MLR to set a benchmark for the more complex models.

This section provides a basic understanding of these methods. We mainly refer to “An Introduction to Statistical Learning” and “The Elements of Statistical Learning”, the works of James et al. (2013) and Hastie et al. (2016).

2.3.1. Simple Linear Regressor (SLR)

In James et al. (2013, Chapter 3) linear regression models are discussed in detail. The basis of these algorithms is the Simple Linear Regressor (SLR), in which a feature variable is regressed on a target variable. This is usually illustrated by a scatter plot with the feature on the x-axis and the target on the y-axis (see Fig. 1). The SLR puts a straight line through the points so that the squared difference between the straight line and each data point is minimized (residual sum of squared errors).

The straight line is formulated mathematically as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{y}$ is the estimate of the target variable based on the feature variable x, $\hat{\beta}_0$ describes the estimated intercept, $\hat{\beta}_1$ the estimated slope of the fitting line. We can use this formula to predict the target variable $\hat{y}$ for given x.

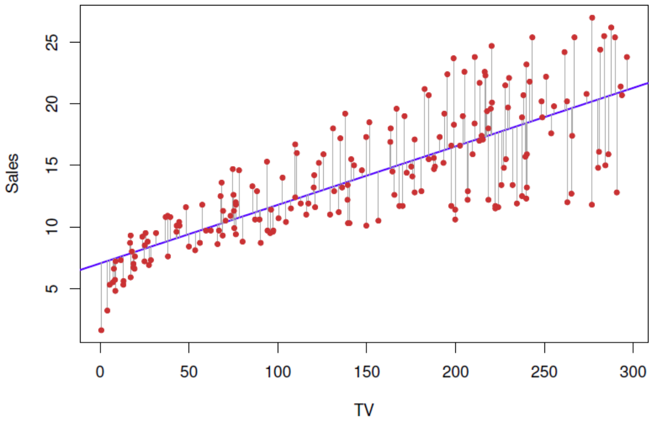

Figure 1: Graphical representation of the concept of multiple linear regression, illustrated in James et al. (2013, p. 62)

2.3.2. Multiple Linear Regression (MLR)

In contrast to SLR, the Multiple Linear Regressor estimates the target variable with several features. With two features and one target variable, the fitting line becomes a plane. We put this plane through the points and thereby we minimize the squared difference between the plane and the individual points.

The plane is formulated mathematically as: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \hat{\beta}_1 x_1 \times \ \hat{\beta}_2 x_2$.
In case of more feature variables, the plane turns into a hyperplane and can mathematically be expressed as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \hat{\beta}_1 x_1 \times \ \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

By fitting the plan through the points and minimizing the residual sum of squared errors, we get estimates of the coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ and we can use the previous formula to predict $\hat{y}$.

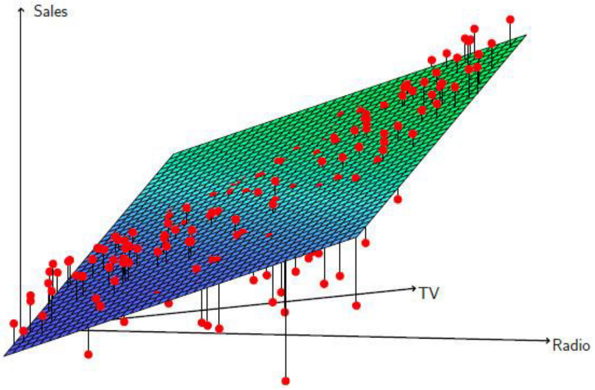

Figure 2: Graphical representation of the concept of multiple linear regression (James et al., 2013, p. 81)

The advantages of MLR are its interpretability, when the assumptions are met they also achieve high accuracy (James et al., 2013, p. 12). However, these assumptions are rarely met in real-world problems. The assumptions of linear models are violated if at least one of the following conditions occurs: 1) a non-linear relationship between target and feature variable, 2) correlated error terms, 3) non-constant variance of the error term, 4) outliers, 5) high-leverage observations and 6) collinearity (James et al., 2013).

2.3.3. Decision Tree Regression (DTR)

Decision tree regression, or regression tree for short, forms the basis for RFR as well as Gradient Boosting Regressions. Derived from James et al. (2013, Chapter 8.1), we briefly describe the main features of such regression trees. Fig. 3, on the left, shows a two-dimensional value space with two feature variables, one on each axis. The target variable R is in the third dimension. When we run the regression tree, the partition space is split into segments. The first partition is on the x-axis, where x_1 is equal to t_1. How and where the partitioning occurs is determined by the algorithm. Behind this logic is a mathematical concept called information entropy. It states that when partitioned correctly, each partition increases the amount of information about our data. The algorithm processes the information entropy and finds the optimal split of our data set. The second split is performed where X_2 is equal to t_2, but X_1 is less than t_1. The third split is performed when X_1 is equal to t_3. The fourth split is performed when X_2 is equal to t_4, but X_1 is greater than t_3.

Fig. 3, on the right, shows the corresponding decision tree that is drawn depending on the splits made by the algorithm. The algorithm stops if the next split cannot add additional information that exceeds a certain threshold, or if the next split would push the observations in a partition below the minimum required observations per partition. By adding splits, we add information to the system. The final leaves are called terminal leaves.

After splitting the value depending on the two feature variables, we now want to predict the target variable of each partition. We take the average of the target variables for each point in a partition, resulting in the values R_1, R_2, R_3, R_4, R_5. These values are assigned to each new point that falls within their terminal partition. For example, if a new observation falls in the partition x_1 less than t_1 and x_2 less than t_2, the decision tree predicts a R_1. The prediction procedure can also be retraced in fig. 3. Each new observation passes through the tests until it reaches the final sheet with the predicted values. (Eremenko, 2020).

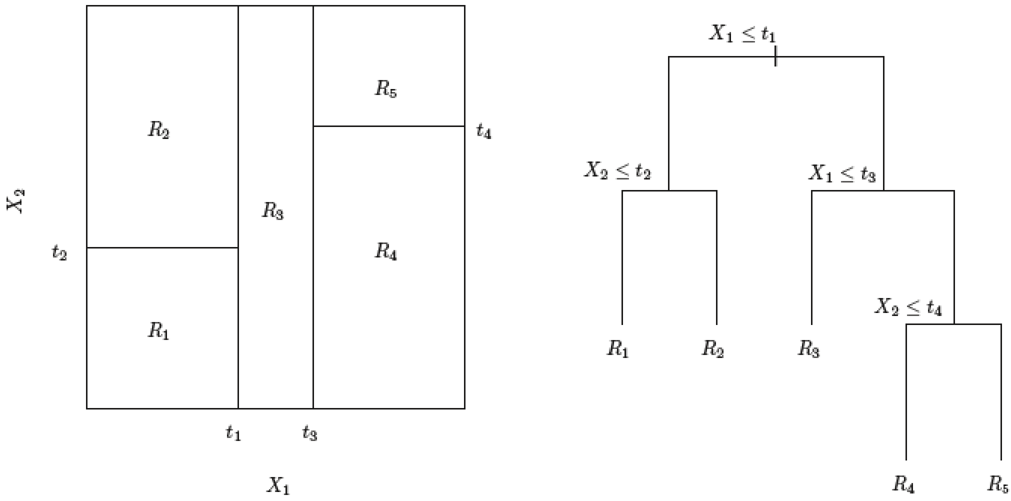

Figure 3: On the left is the two-dimensional value space with feature variables X_1 and X_2, on the right, is the associated decision tree, illustrated in: James et al. (2013, p. 308)

2.3.4. Random Forest Regression (RFR)

James et al. (2013, Chapter 8.2.2) explained the Random Forest Regressor. We summarize the main features. The RFR is an ensemble learning algorithm. The main idea behind ensemble methods is that by averaging of resulting predictions, their variance can be reduced, and the prediction accuracy increased. The algorithm uses a bootstrap procedure, which allows taking repeated samples from a single training set. The RFR constructs B regression trees using B bootstrap training sets. The algorithm however only considers a subset of the predictors for each split, thereby, it is ensured that in the presence of a strong predictor, not all decision trees make the same initial split. Consequently, the individual predicted values of the resulting trees are decorrelated among themselves. This further reduces the variance of the average prediction value and thus increases the predictive power even more compared.

The following illustration of (Bakshi, 2020) shows the interplay between individual decision trees.

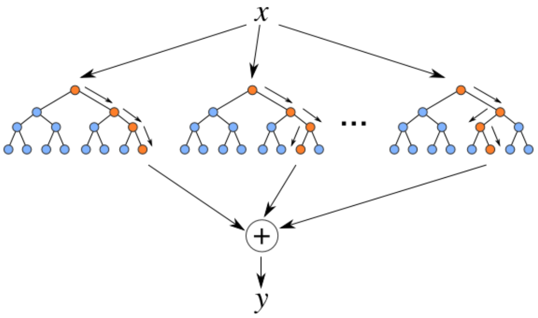

Figure 4: Graphical interpretation of RFR, illustrated in: Bakshi (2020)

The advantages of RFR are that they can handle large data sets, compute the importance of features, and are thus explainable and interpretable. They are flexible and can easily handle categorical variables and missing values. They also have high predictive power. The disadvantages of RFR are slow prediction speed, the inability to extrapolate or predict rare outcomes, i.e., the predicted values are within the bounds of the training data, limited control is given to the modeller. (Lisowski, 2019; Pradhan, 2019)

2.3.5. Gradient Boosting Regressor (GBR) and XGBoost Regressor (XGB)

Derived James et al. (2013, Chapter 8.2.3) we firstly explain the Gradient Boosting Regressor (GBR). And then how it differs from XGBoost (XGB) the implementation we use in our models.

Like RFR, the GBR combines multiple decision trees to increase predictive power. Unlike bootstrapping in RFR where the decision trees grow in parallel, in GBR they grow sequentially. The algorithm learns slowly. Each decision tree uses information from the trees that have grown before. The decision tree is fitted to the residuals of the previous model and not to the target variable. The new tree is inserted into the fitted functions to update the residuals. Each of these trees can be tiny, with only a few endpoints whose minimum number is determined by the algorithm.

"XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way." (XGBoost, n.d.)

In comparison with RFR, XGB has the advantages of higher computation speed, higher speed through parallelization. The disadvantage is difficult tuning and the tendency to overfit when the data is noisy.

2.3.6. Support Vector Regressor (SVR)

Support Vector Regression are explained in detail by Hastie et al. (2016, Chapter 12.3.6) and Hawkins (2015). We briefly explain the main characteristics of this algorithm. Unlike other EQM, SVR uses a kind of buffer zone - the so-called ϵ-insensitive tube - in which deviations between true and predicted values are not considered in the loss function (see fig. 5). SVR tries to find the narrowest tube in the centre of the area while minimizing prediction errors. It uses a symmetric loss function that penalizes positive and negative deviations equally (see fig. 6). The advantages of this algorithm are that its complexity is independent of the dimension of the input, its excellent generalization ability with high prediction accuracy, and its robustness to outliers.

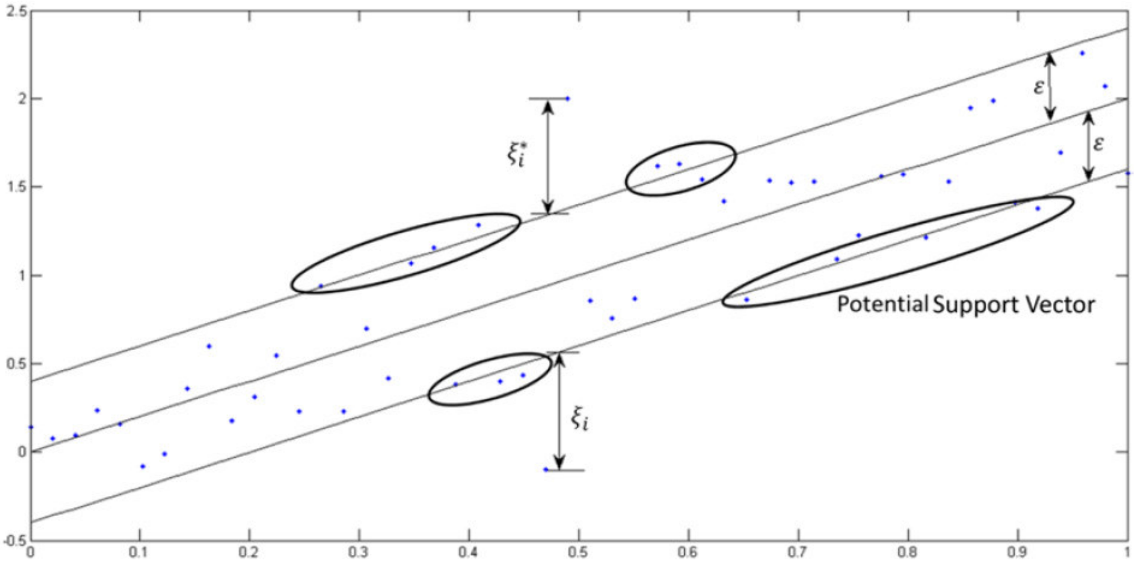

Figure 5: One-dimensional linear SVR

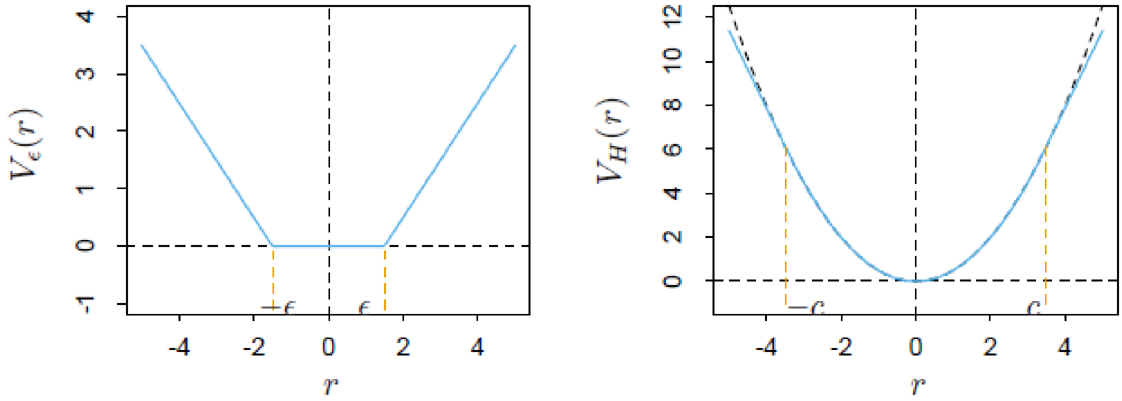

Figure 6: The left panel shows the ϵ-insensitive error function used by Support Vector Regression. The right panel shows the error function used in Huber's robust regression, illustrated in: Hastie et al. (2016, p. 435).

Advantages of SVR are its robustness to outliers, its easy implementation, its generalization capability, and prediction accuracy. The disadvantages are that they tend to underperform with large and noise data.

2.4. Performance Evaluation Measures (PEM)

Table 1 shows the measures the models are evaluated on. Thereby, true values are abbreviated with the letter A, predicted values with the letter F.

Table 1: Performance Evaluation Measures (Wenninger & Wiethe, 2021)

Performance evaluation measure	Equation	Unit, value range	Best value
Coefficient of Variation (CV)	$CV(A,F) = \frac{\sqrt{\frac{1}{N}\sum_{i=i}^{N}(F_i - A_i)^2}}{\frac{1}{N}\sum_{n=i}^{N} F_i}$	$-, (0, \infty)$[2]	0
Mean absolute percentage error (MAPE)	$MAPE(A,F) = \frac{100\%}{n}\sum_{i-1}^{n}\left\lvert\frac{A_i - F_i}{A_i}\right\rvert$	$\%, [0, \infty]$	0
Root-mean-squared error (RMSE)	$RMSE(A,F) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}$	$\frac{kWh}{m_2 a}, [0, \infty]$	0
Mean absolute error (MAE)	$MAE(A,F) = \frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$	$\left(\frac{kWh}{m_2 a}\right)^2, [0, \infty]$	0

ASHRAE (1997), the American Society of Heating, Refrigerating and Air-Conditioning Engineers, suggests that researchers evaluate their models by their accuracy, sensitivity, versatility, speed and cost, reproducibility, and ease of use. Therefore, the predictive model preferably was explainable and fast as a linear regression with high prediction power. However, this may not be possible.

ASHRAE (2005, p. 3) recommends a coefficient of variation (CV) of less than 25% as a measure of sufficient accuracy. Another relevant evaluation measure is the Mean Percentage Error (MAPE). ECOSPEED AG (2019) claims to achieve a MAPE below 10% with their regional model. Both the CV and the MAPE have the advantage of being unitless and independent of the scale of the target variable. We have demonstrated scale independence in Appendix 7.1.

We evaluate all models based on HEC measured in kWh/year. The target variables HEPI, log-HEC and log-HEPI are recalculated to HEC before the predictions are evaluated with the true values.

[2] The coefficient of variance is negative when $F_i < 0$, however, that is not the case in this study.

3. Methodology

To address the research question and benchmark the different subsets and EQM, a suitable methodology and study design is necessary. The study design of Cross-Industry Standard Process for Data Mining (CRISP-DM) is considered as the guideline for big data analysis and was initially presented by Wirth and Hipp (2000). The original six steps are 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Modelling, 5) Evaluation, and 6) Deployment. The idea of this process is that iteration between the different phases is the rule rather than the exception. We do not go through the steps once and have the problem solved. Often the whole process is a data investigation and with each iteration we gain more knowledge that can be applied in the next iteration (Provost & Fawcett, 2013).

In contrast to the original process, we allow in our study for more interaction between the stages. Therefore, our study design is slightly modified and similar to the one presented by Wenninger and Wiethe (2021), shown in fig. 7. Each step is explained in detail below.

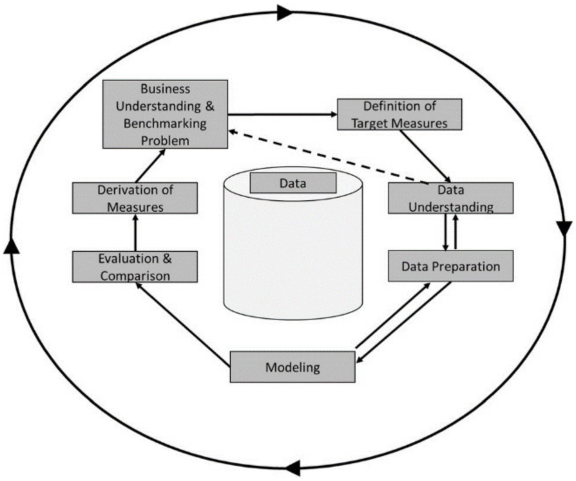

Figure 7: Our Methodology derived from and illustrated by Wenninger and Wiethe (2021)

3.1. Business Understanding & Benchmarking Problem

Like Wenninger & Wiethe (2021), we extended the first phase to include the benchmarking problem. In this phase, we study the industry partner's problem respectively our research questions and gather domain-specific knowledge about EQM (see section 2.2) and PEM (see section 2.4).

3.2. Data Understanding

We collect data and examine its strengths and limitations. In this phase, we dive deep into the data and learn about the structures and correlations with the business problem. To do so, we inspect the data by univariate and multivariate descriptive statistics and plots (section 4.1).

In this study, we use data on the measured energy consumption of individual buildings in the cities of Biel, Geneva, and St. Gallen, combined with data on building characteristics from RBD and Minergie and in-house calculations from geoimpact.

Data source	Number of buildings	Population ratio
Census	1,800,000	
Geneva	15,000	0.83%
St. Gallen	15,000	0.83%
Biel	9,000	0.005%
Survey[3]	*80*	*0.0045%*

Table 2: Overview of available data

The georeferenced dataset of Biel includes the measured consumption data of gas, electricity, and water for about 9’000 individual buildings. The data is provided by Energy Service Biel (Extraction date: September 7th, 2021).

The data of Geneva includes georeferenced data for measured HEC including SH and DHW per square meter and year for about 15’000 buildings over the years 2011 to 2021. The data is public available on https://ge.ch/sitg/fiche/2177. (Extraction date: 5th June 2021) [4]

The data of the St. Gallen city contains georeferenced building and consumption data of electricity, water, district heat and gas for about 15'000 buildings over the years 2017 to 2019. Additional datasets include details of the city's heating and hot water systems such as system type, heating capacity, boiler's year of construction. The data is provided by the Environment and Energy Office of St. Gallen (Extraction date: June 18th, 2021).

Each of the data sets of Biel, Geneva and St. Gallen include the unique building identifier (EGID). This identifier enables to connect the data to the other datasets of building characteristics and the calculations from geoimpact.

The Federal Buildings and Housing Register (RBD) provides the main properties of the buildings considered in this project. The RBD distinguishes four different building typologies: 1) purely residential buildings (single family and multifamily houses), 2) residential buildings with ancillary uses, 3) buildings mainly used for non-residential purposes, and 4) buildings without residential use.
The recordings further include the building's primary data such as an address, location coordinates, construction year, renovation year (if applicable), number of flats, number of rooms, floor space.

The Minergie association defines the Swiss standards for efficient and comfortable buildings. Their georeferenced data include modelled energy performance indexes and Minergie label information for more than 45,000 buildings. In our model, we only use the information, whether a building is Minergie certificated.

We extend the data with the following estimates of geoimpact:

- Building renovation pressure: Renovation pressure of the building due to past renovations in comparison with similar buildings and relating to geographical location (0 low, 1 high).
- Building renovation rate neighborhood: number of renovations in neighborhood relating to number of buildings.
- Population density: number of inhabitants per hectare.
- Population growth: Net population growth in one hectare.

[3] We did not use data collected as part of the survey we prepared in the preliminary study. The data includes 80 buildings with information on their thermal energy and electricity (based on utility bills), primary building characteristics, and, at best, household self-generated electricity and electric car consumption. The questionannaire can be found in section 7.5.

[4] «Source: Système d'information du territoire à Genève (SITG), imprimé et/ou extrait en date du 5 juin 2021

3.3. Data Cleaning and Preparation

With a sound understanding, we prepare the datasets of the cities of Biel, Geneva and St. Gallen as well as the other RBD and Minergie data and combine them with the master data of Geoimpact. We gain first insights by inspecting univariate and multivariate descriptive statistics and plots. We prepare the data by removing duplicates, handling outliers and missing values. Thereby, we mainly analyzed outliers in the target variable, and we decided to drop buildings which HEPI undercut 10 kWh/m2/year and exceeds 300 kWh/m2/year.

The Geoimpact dataset contains information as to the buildings of the cities Biel, Geneva and St. Gallen with building locations and features originally retrieved from the RBD, as well as Geoimpact in house calculations. This dataset serves as the master dataset and the other datasets are linked to it. As an additional variable, the information whether the building boiler is used for the provision of both SH and DHW was also obtained from the RBD. From the Minergie dataset, only the EGID were extracted to identify which buildings in the master dataset are Minergie certified.

The data of Biel contains the consumption of annual electricity, gas, and water as well as the identifiers EGID of 9’000 buildings. For Geneva, the ERA and EPI for individual energy carriers are available for 18’000 buildings over 10 years. We reconstruct the measured heating demand by multiplying the ERA by the EPI. For St. Gallen city we got comprehensive data on building characteristics, actual consumption of electricity, water, gas and heat, and modeled energy consumption. However, we are interested in only measured energy consumption values of individual sites, i.e., sites that are self-heated. The resulting data frame for each city exists of only two columns: one for HEC and one for the unique building identifier EGID that links the data to the other data sets.

Table 3 shows the variables we used in our study in the following groups: 1) Building Attributes, 2) Heat Systems, 3) Energetic Insulation, 4) Location and 5) Target Variable.

The considered energy sources, gas, oil, and district heating can provide SH and DHW, then, a central heating system is installed, and we cannot distinguish which part of the energy sources is used for which energy usage. Since both sources have different drivers, DHW depends on mainly on occupants’ behaviour, SH mainly on building characteristics (Federal Office for the Environment FOEN, n.d.), we want to identify buildings that use such a central heating system. Therefore, we created the variable "Use of boiler for SH and DHW”.

We also create the variable Buildings Gross Floor Area, by multiplying the number of floors with the buildings footprint area, to approximate the Buildings Energy Reference Area. This value is used in the denominator of the target variable HEPI.

3.4. Modelling and Evaluation

In the contexts of this study, we have rejected the idea of a single model that accurately predicts the HEC of each building, instead, we model multiple models each estimating the HEC of a subset of buildings. First, we model the estimate the different target variables and evaluate them using the Performance Evaluation Measure. We select the preferred estimator and then proceed to analyse more homogeneous building subgroups. By dividing buildings into homogeneous subsets, we expect to increase our predictive power. A subset includes buildings in a single city, e.g., Biel, Geneva or St. Gallen, or residential buildings only, or buildings built before 1980, or some combination thereof.

Table 3: Input parameter for EQM in our study

Category	**Variables**	**Value**	**Data Source**
Building Attributes	Buildings Construction Year	Year	RBD
Building Attributes	Buildings Renovation Year	Year	RBD
Building Attributes	Number of Floors	-	RBD
Building Attributes	Number ofApartments	-	RBD
Building Attributes	Buildings Gross Floor Area (Approx. ERA)	m2	Own Calculation (RBD)
Building Attributes	Buildings Footprint Area	m2	RBD
Building Attributes	Buildings Volume	m3	Geoimpact
Building Attributes	Building Use	12 categories including multi-Family (73%), Single-Family (12%), Others (15%)	Geoimpact
Heat System	Heating Type	8 categories including gas (70%), oil (23%), district heating (4%), Others (3%)	RBD
Heat System	Use of boiler for SH and DHW	Yes, No	Own Calculation (RBD)
Energetic insulation	Minergie-Standard	Yes, No	Minergie
Energetic insulation	Building Renovation Pressure	%	Geoimpact
Location	Neighborhoods Renovation Rate	%	Geoimpact
Location	Population Density	# Inhabitants per hectare	Geoimpact
Location	Population Growth	Net population growth per % per hectare	Geoimpact
Location	Altitude	Meters above sea level	Geoimpact
Target Variable 1	Heat Energy Consumption (HEC)	kWh/a	Biel, Geneva, St. Gallen
Target Variable 2	Heat Energy Performance Indicator (HEPI)	kWh/m_2a	Biel, Geneva, St. Gallen, RBD

Wenninger and Wiethe (2021) has shown the algorithms RFR, XGB, SVR can achieve high prediction accuracy. SVR is even the most favoured method besides ANN. In addition, we use MLR to set a benchmark for the more complex models. We use MLR, RFR, XGB and SVR in our study. With this selection, we examine a wide range of methods, from simple methods such as MLR to complex methods such as SVR.

In this stage, we modelled the four regressor candidates MLR, RFR, XGB and SVR on the four target variables HEC, HEPI, log HEC, log HEPI. The phase consists of the following three phases:

- Exploration Phase (on validation set)

In the procedure of the modelling phase, we first split the data into an exploratory set – including training and validation data - and a test set. The training and validation data allows us to tinker (Kozyrkov, 2019). We try several regressors on different subsets and different PEM. In total, we trained and evaluated more than 1000 model settings. These models are evaluated using PEM (see Section 2.4) When we have our favorite model setting including pre-processing pipeline, features, and the subsets that we want the regressor and target variable candidates to test on, we proceed with the next phase.

- Evaluate the best fitting Regressor (test set)

Based on our favorite model setting, we look for the best fitting regressor and the best-fitting target variable. Therefore, we model each of the four regressors on each of the four target variable candidates. We do this once for all buildings in the data, and once for residential buildings only. In total, we train and evaluate 32 different model settings (4 Regressors X 4 Target Variables X 2 Subsets). We analyze the result tables and choose the favorite regressor and the favorite target variables to proceed.

- Evaluate the best fitting Subset (test set)

Having a regressor candidate, we divide the data set into more homogenous subsets to potentially improve the prediction power of our model. These subsets are identified by previous diagnosis plots and include all buildings, residential buildings, residential buildings built before 1980 and their combination with the canton attribute.

We train these subsets and for each we evaluate the prediction accuracy with PEM. Having found the best-fitting regressor, target variable for each subset, we are ready to deploy.

3.5. Deployment

In the final phase, we discuss our models and findings with geoimpact. We derive implications for the integration and further development of the algorithms and the integration into SEP. We agreed to share the project via a private repository on GitHub (https://github.com/andygubser/building-energy-modelling). If any questions arise, we will stay in contact after the submission of this thesis.

4. Results

Section 4.1 provides descriptive analysis and shows the distribution of the variables across the three cities of Biel, Geneva, and St. Gallen. Section 4.2 evaluates the best fitting regressor and target variables. Section 4.3 evaluates these models on subsets.

4.1. Descriptive Analysis

In this section, we provide insight into the distribution of variables. Note that, as mentioned in Section 3.2, we only consider buildings with HEPI values between 10 and 300 kWh/m2/year. The plots differ the cities by its cantons, thus, BE is for Biel, GE is for Geneva, and SG for St. Gallen City. Table 4 shows the distribution of variables in the train set. It is noticeable that the data includes both very small buildings (8 m3 volume) and very large buildings (167 apartments, almost 1 million m3 volume).

Table 4: Descriptive analysis of training data

No	variable	count	min	25%	50%	75%	max
1	Building Construction Year	11176	1560	1919	1965	1989	2019
2	Building Renovation Year	6593	1954	1993	2010	2016	2021
3	Number of floors	11209	1	3	4	6	31
4	Number of apartments	11209	0	2	7	14	167
5	Building Gross Floor Area	11209	27	480	1’010	1’904	224’232
6	Building Footprint Area	11209	24	146	229	320	62’996
7	Building Volume	10554	7.68	1’285	2’733	5’208	977’068
8	Minergie Standard	11209	0.0	0.0	0.0	0.0	1.0
9	Renovation Pressure	11207	0.0	0.0904	0.270	0.559	0.993
10	Renovation Rate	11207	0.002298	0.01667	0.02236	0.0328	0.0505
11	Population Density	10760	3	56	116	223	5212
12	Population Growth	10755	-84.4	-3.23	-0.43	1.75	451
13	Altitude	11209	348.1	399	422	443	874
14	HEC	11209	1000	54’135	117’124	225’611	12’095’064
15	HEPI	11209	10	90	119	147	300

Figure 8 shows the building characteristics per city. Most were built before 1980, but the buildings in Geneva are not as old as the buildings in St. Gallen or Biel, and they are larger (more floors, more apartments, larger ERA, and larger building footprint). The building mentioned with the 167 apartments is also located in Geneva. About 90% of the buildings are residential, e.g., single- or multi-family houses. With the exception that Biel has more single-family homes than St. Gallen, the two cities are similar in terms of building characteristics.

Figure 9 shows the distribution of energetic insulation variables. Buildings in Geneva are in average better insulated than the buildings in Biel and St. Gallen. Figure 10 shows that the almost every building in Biel is heated with gas (>90%), in St. Gallen about 85% use gas, about 15% district heating, in Geneva about 60% use gas, about 30% oil. Further, in Geneva almost every building has a boiler installed for DHW and SH.

Figure 11 shows the distribution of location variables. St. Gallen has the highest altitude. The city is about 700 meters above sea level. While St. Gallen and Biel have about the same population density, it is much higher in Geneva. The renovation rate in Geneva is the highest, followed by St. Gallen.

Figure 12 shows the distribution of the target variables HEC and HEPI at the city level. Geneva has the highest HEC value of the three, probably mainly due to its larger buildings, and the highest HEPI value. This contradicts the intuition of Geneva's younger and more energy-insulated building stock. The average building in Switzerland consumes about 145 kWh/m2/year (Swiss Federal Office of Energy

SFOE, 2020a). This is about the same as the average building in Geneva. However, the HEPI values for Biel and St. Gallen seem to be unexplainably low.

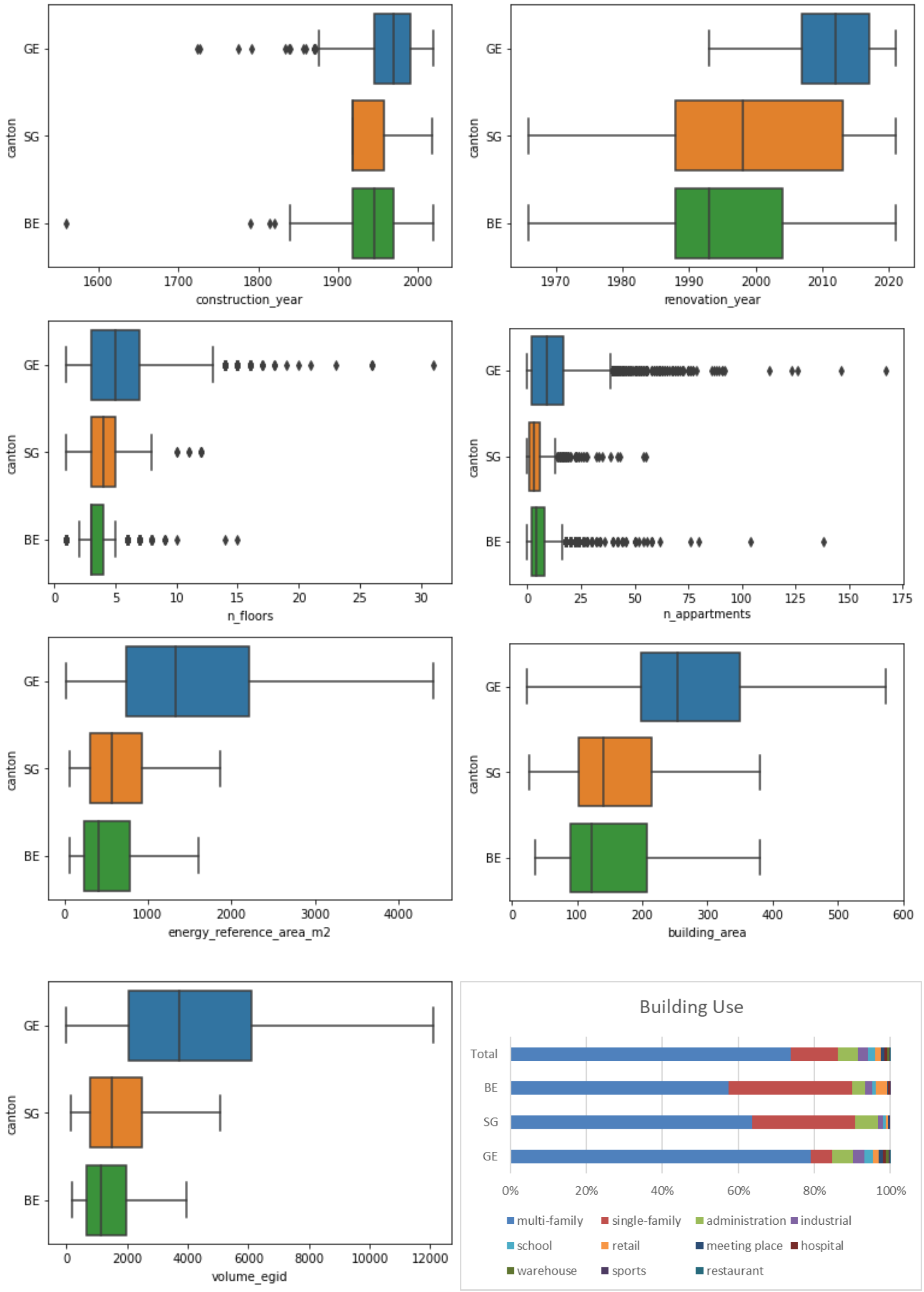

Figure 8: Distribution Plots for Building Attributes

Figure Glossary, from top left to bottom right: Building Construction Year, Building Renovation Year, Number of Floors, Number of Apartments, Building Gross Floor Area (approx.. ERA), Building Footprint Area (Building Area), Building Volume, Building Use.

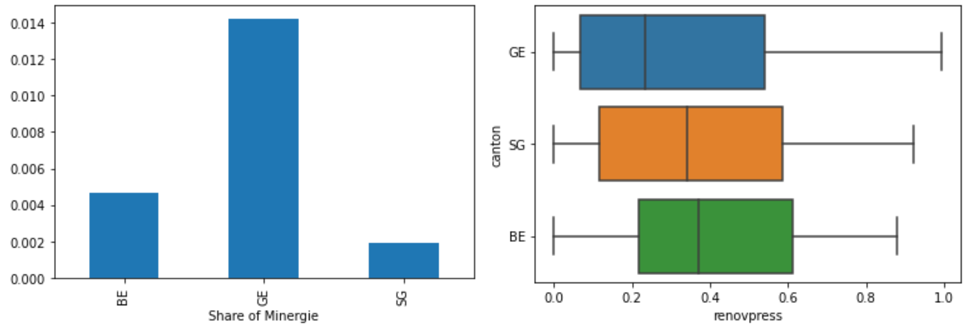

Figure 9: Distribution of Energetic Insulation Variables

Figure Glossary: left panel: Share of Minergie certificated Buildings per City;
right panel: Renovation Pressure

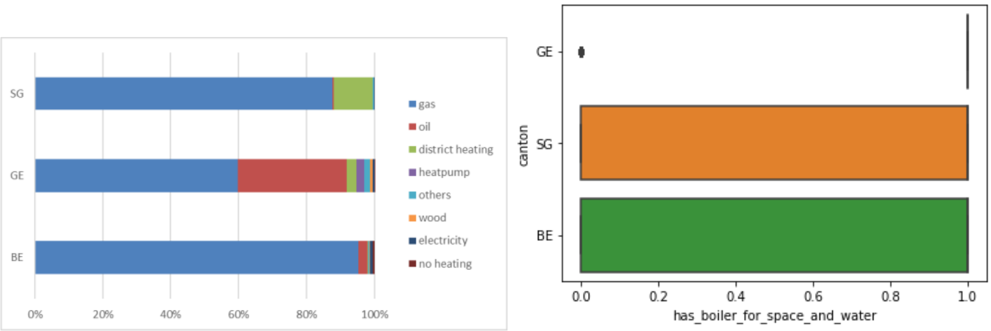

Figure 10: Distribution Plots of Heat System Variables

Figure glossary: left panel shows heat system types per city; right panel shows whether a boiler is installed for SH and DHW

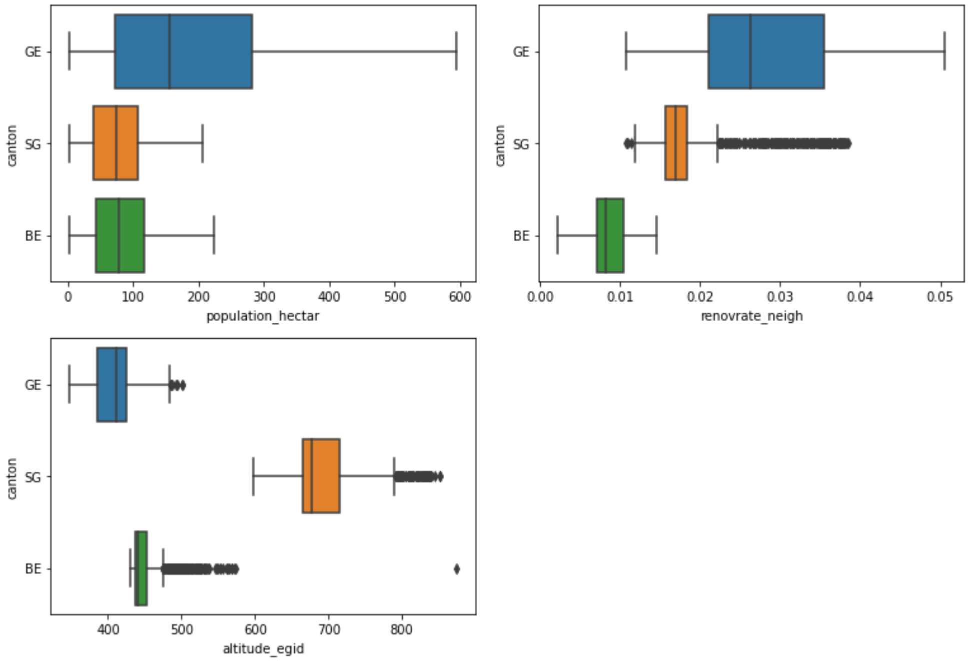

Figure 11: Distribution Plots of Location Variables

Figure Glossary: top left: Population Density per City, top right: Renovation Rate in the Neighborhood, bottom left: Altitude (meters above sea level)

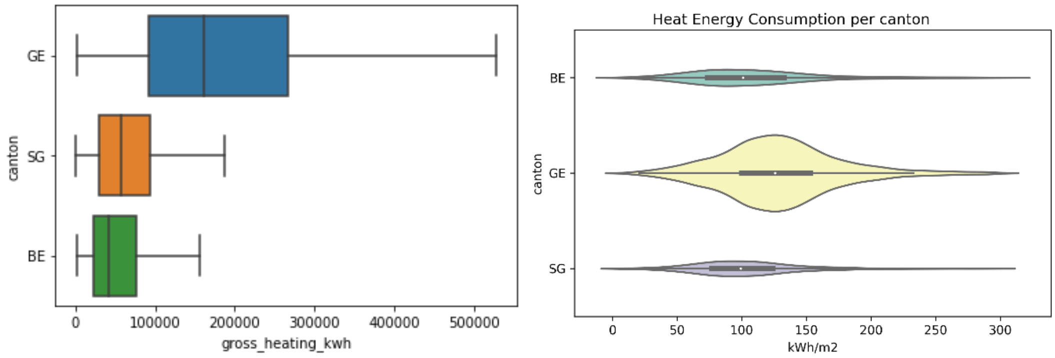

Figure 12: Distribution of Target Variables

Figure Glossary: left panel: HEC (kWh/year); right panel: HEPI (kWh/m2/year)

4.2. Evaluate the best-fitting regressor and target variable

We evaluate the regressors and target variables by running each EQM (MLR, RFR, XGB, SVR) on all buildings in the test set (full details in Appendix 7.2). Table 9 summarizes the results sorted by lowest CV. Each model is assigned a number for the corresponding model setting (column “No”). The column "ntest/ntrain" indicates how many buildings were used to train and how many to evaluate the model. The regressors are trained and tested on different target variables. But for performance evaluation, each predicted and true value is retransformed to HEC. Thus, we can compare not only the scale-independent measures CV and MAPE, but also MAE and RMSE.

Overall, the prediction accuracy of the EQM is low, but the goal in this analysis step is to identify the best fitting combinations of regressor and target variable. The best-fitting model setting is No 1&2 with the regressors XGB and RFR with target variable HEPI. These models achieve the lowest CV (0.68, 0.77) and low MAPE (0.32, both). In contrast, the regressors MLR and SVR only achieve the lowest CV of 0.94 and 0.99 in their best-fitted model setting (No 6&9).

Comparisons of models No 1&4, 2&3, as highlighted in greens, show that the log-transformation of the target variable HEPI does necessary lead to higher accuracy (lower CV and lower MAPE). However, model No 5&12, show that the log-transformation can significantly lower CV and MAPE of a XGB with target variable HEC.

Table 5: Validation of the best-fitting regressors and target variable on all buildings

No	Regressor	Target	n test/n train	CV	MAPE	MAE	RMSE
1	XGB	HEPI	4804/11209	0.68	0.32	44468	124745
2	RFR	HEPI	4804/11209	0.77	0.32	44198	143365
3	RFR	log HEPI	4804/11209	0.81	0.28	41945	138124
4	XGB	log HEPI	4804/11209	0.82	0.30	44098	141321
5	XGB	log HEC	4804/11209	0.84	0.30	44780	144803
6	MLR	HEC	4804/11209	0.94	0.51	55193	169818
7	RFR	HEC	4804/11209	0.97	0.32	45199	178920
8	RFR	log HEC	4804/11209	0.99	0.29	43335	168438
9	SVR	log HEPI	4804/11209	0.99	0.31	45242	174802
10	MLR	log HEPI	4804/11209	1.04	0.33	50428	183996
11	SVR	log HEC	4804/11209	1.07	0.34	46852	181436
12	XGB	HEC	4804/11209	1.09	0.33	46269	198352
13	SVR	HEPI	4804/11209	1.19	0.35	54225	223196
14	MLR	HEPI	4804/11209	1.29	0.36	55200	240076
15	SVR	HEC	4804/11209	2.34	1.38	120380	273813
16	MLR	log HEC	4804/11209	8.66	0.50	94207	1584390

Table 10 shows the results from fitting each EQM on residential buildings in the test set, sorted by lowest CV. As before, predicted, and true value are retransformed to HEC before the model is evaluated. Thus, all the four metrics can be compared across models.

While the SVR performed rather weak in the analysis on all buildings (see Table 9), with the restriction to only residential buildings, the regressor the lowest CV values of all regressors (No 1&2). The regressors XGB and RFR that performed best in the previous analysis, also do well in the analysis of residential buildings and the RFR achieve an only slightly higher CV than SVR (No 4&5, 3&7).

To summarize, both RFR and XGB are promising regressors and achieve similar accuracy on the one scenario including all buildings and on the other scenario only including residential buildings. However, since the RFR performs slightly better on the residential building's subset, and we will continue the analysis on this data, we decide to proceed with the RFR and not with XGB. Further, modelling these regressors on log-transformed target variables do not lower the resulting CV or MAPE. Therefore, we proceed the subset analysis with the target variables HEC and HEPI and the regressor RFR.

Table 6: Validation of the best-fitting regressors and target variable on residential buildings

No	Regressor	Target	ntest/ntrain	CV	MAPE	MAE	RMSE
1	SVR	HEPI	4054/9458	0.67	0.29	35027	108855
2	SVR	log HEPI	4054/9458	0.69	0.25	30993	108447
3	RFR	HEPI	4054/9458	0.69	0.25	29374	112356
4	XGB	HEPI	4054/9458	0.76	0.25	31292	122802
5	XGB	log HEPI	4054/9458	0.76	0.24	30099	118986
6	MLR	HEC	4054/9458	0.77	0.32	34812	123317
7	RFR	log HEPI	4054/9458	0.78	0.23	29173	121633
8	RFR	HEC	4054/9458	0.84	0.25	29823	135455
9	XGB	HEC	4054/9458	0.88	0.26	31820	141009
10	XGB	log HEC	4054/9458	0.88	0.24	31600	137475
11	RFR	log HEC	4054/9458	0.90	0.23	29281	138867
12	MLR	log HEPI	4054/9458	1.00	0.28	36781	155580
13	SVR	log HEC	4054/9458	1.00	0.26	32179	154775
14	MLR	HEPI	4054/9458	1.40	0.30	39386	223325
15	SVR	HEC	4054/9458	1.79	1.32	101406	210101
16	MLR	log HEC	4054/9458	7.77	0.36	84431	1536653

4.3. Evaluate prediction performance of building subsets

Having found our favorite regressor, RFR (see 4.2), we now run the regressor on different subsets (full details in Appendix 7.3), one including all buildings (Table 11), one including only residential buildings (Table 12) and one including only residential buildings built before 1980 (Table 13). Further, these buildings can be restricted only one of the three cities, as indicated by the column City.

The tables 11-13 show the results of fitting the RFR on the test set including all buildings (Table 11), residential buildings (Table 12) and residential buildings built before 1980. These subsets can be further restricted with a specific city. The RFR is fitted on the target variables HEC and HEPI.

Table 12, No 2&3, shows the results for all the three cities. We are also in the same scenario as before in section 4.3; therefore, we achieve similar accuracy levels. By restricting to buildings of the St. Gallen city, we can slightly increase our accuracy for the HEC target. However, other restrictions to buildings lead to weaker prediction results. Based on these results, we cannot evaluate the value added of splitting buildings into more homogenous groups.

Table 7: Validation of the prediction accuracy on all buildings (RFR)

No	City	Target	ntest/ntrain	CV	MAPE	MAE	RMSE
1	St.Gallen	HEC	948/1547	0.73	0.35	23547	60831
2	All three	HEC	6864/11209	0.77	0.31	41314	138411
3	All three	HEPI	6864/11209	0.77	0.30	40953	140030
4	Geneva	HEPI	4866/7945	0.83	0.25	48156	180221
5	Geneva	HEC	4866/7945	0.94	0.26	48977	202373
6	St.Gallen	HEPI	948/1547	0.95	0.35	26857	83181
7	Biel	HEPI	1051/1715	1.43	0.46	33039	107650
8	Biel	HEC	1051/1715	1.46	0.47	34435	110370

Therefore, we take another subset and analyze the accuracy of RFR for residential buildings only; the results are shown in Table 13, where No 4&7, show the repeated scenario analyzed in section 4.3.

Then, we take another subset and study the prediction power on residential buildings that were built before 1980. To keep the discussion simple, the CV and MAPE of the tables 11-13 are shown in the fig. 13 in dependence of their split.

Table 8: Validation of the prediction accuracy on residential buildings (RFR)

No	City	Target	ntest/ntrain	CV	MAPE	MAE	RMSE
1	St.Gallen	HEPI	558/1301	0.44	0.35	18312	31879
2	St.Gallen	HEC	558/1301	0.47	0.34	18462	33332
3	Geneva	HEPI	2854/6659	0.51	0.19	33534	105015
4	All three	HEPI	4054/9458	0.69	0.25	29331	111509
5	Geneva	HEC	2854/6659	0.83	0.19	34663	166533
6	Biel	HEC	642/1497	0.83	0.44	22465	49250
7	All three	HEC	4054/9458	0.83	0.25	29503	134208
8	Biel	HEPI	642/1497	0.87	0.45	23904	54286

Table 9: Validation of the prediction accuracy on residential buildings built before 1980 (RFR)

No	City	Target	ntest/ntrain	CV	MAPE	MAE	RMSE
	Geneva	HEC	1542/3596	0.23	0.19	30351	48872
	Geneva	HEPI	1542/3596	0.23	0.18	30251	49063
	All three	HEPI	2531/5903	0.32	0.30	28020	50823
	All three	HEC	2531/5903	0.38	0.30	29220	60057
	St.Gallen	HEPI	483/1127	0.40	0.30	16121	26211
	St.Gallen	HEC	483/1127	0.40	0.31	16346	26390
	Biel	HEPI	506/1178	0.52	0.45	18232	32221
	Biel	HEC	506/1178	0.61	0.47	19324	36193

Figure 13 shows the effects of the individual data splits on CV and MAPE. The effects of taking a subset of the buildings on CV are massive, while the effects on MAPE are negligible. This can be explained mainly by the fact that with each sub setting split we reduce the variance between the populations. However, since most of the buildings are or the ‘average’ building is residential and were built before 1980 (fig. 8), the MAPE is only weakly affected.
The CV of each city and with each target variable is reduced with each further data split. Thereby, the CV for HEC or HEPI run parallel, but mainly with a lower CV for HEPI. Except for Geneva where the CV for HEC increases relative to the CV for HEC when focusing on residential buildings, the difference decreases again with the further split when only the old buildings remain in the dataset. This is probably because there are relatively many new and large buildings in Geneva (fig. 8).

Finally, we conclude that dividing the data into more homogeneous groups has an impact on CV but not on MAPE and the division has led to a conversion of the target variables.

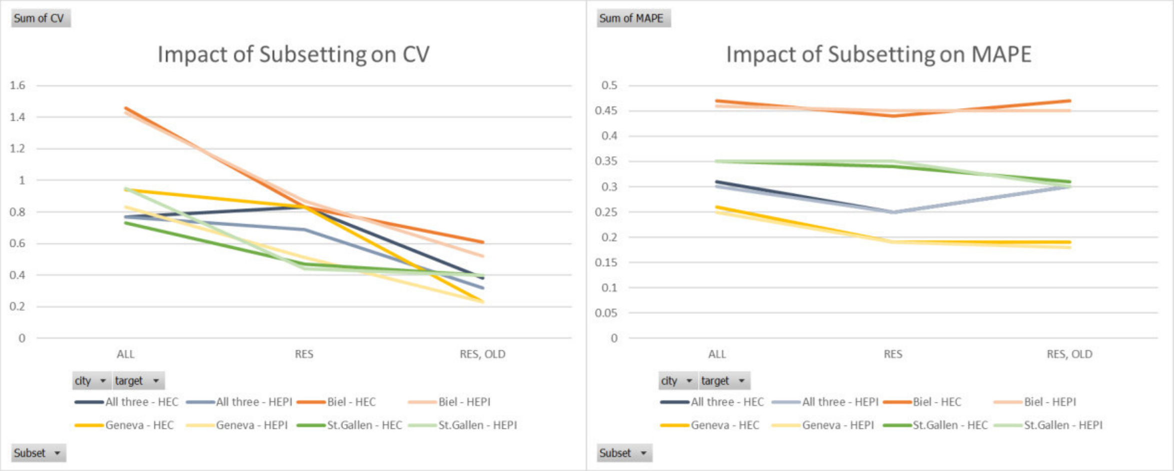

Figure 13: Impact of Taking Subsets, left panel: CV, right panel: MAPE

5. Discussion and Conclusion

In this chapter we discuss the research questions and conclude this study. In the introduction we stated the main research as follows.

Which set of models achieve the highest accuracy in predicting the energy performance of individual buildings?

1) What are the relevant characteristics for predicting the energy performance of buildings?
    a. What does literature say?
    b. What do we find in the data?

The studies in chapter 2.3 described the following key drivers of building's heat or overall energy consumption, clustered into the following groups (Braulio-Gonzalo et al., 2021):

- Household typology: ERA, building type
- Construction features: Building Age
- Technical building systems: heat system and heat system efficiency, building envelope, facade, roof, windows, solar orientation
- Socio-economic profile of occupants and households' composition
- Climatic and local conditions

The main feature we find in data are approximates for building size such as gross floor area (our approximation of ERA), the number of apartments and floors, buildings volume, as well as climatic and local conditions approximated by altitude, building's energetic insulation quality approximated by the renovation pressure ratio, and the buildings use type (single-, multifamily or others). The resulting features of our models are found in Appendix 7.4.

2) What are the most promising regressors for estimating heat energy consumption?
   Candidates: MLR, RFR, XGB, SVM

Section 4.2 shows that RFR achieve high accuracy for heating energy consumption in HEC or in HEPI, followed by XGB. MLR and SVR are less suitable.

The RFR also performs well also in comparison with other studies such as Wenninger and Wiethe (2021). On residential buildings, they achieved CVs between 0.33 and 0.35 with ANN, Copula, XGB, RFR and SVR. In our setting the RFR achieved at best CVs between 0.23 and 0.77 depending on the building sample.

3) How do we need to transform the target variable heat energy consumption to achieve the highest accuracy?
   Candidates: HEC, HEPI, log-HEC, log-HEPI

Further, section 4.2 shows that log-transformed target variables does not necessarily lead to improvements in accuracy. Therefore, we have proceeded the analysis without them, and pursued with Absolute Heat Energy Consumption (kWh) and Heat Energy Consumption per square meter (HEPI). Section 4.3 has shown that making the data more homogenous by sub setting decreases the CV on both targets' kWh and HEPI, with a slightly lower CV for HEPI. Because of these results, we do not recommend one target variable over the other.

4) Can we achieve higher prediction accuracy by using multiple models trained on specific subsets of buildings, compared to one single model for predicting the heat energy consumption of all buildings?
    a. Can we achieve a CV below 25% (ASHRAE, 2005, p. 3)?
    b. Can we achieve a Maximum Percentage Error of less than 15% (geoimpact AG, personal communication, December 16, 2021)?

Section 4.3 analyzed the effects on performance of subdividing into homogeneous populations, such as residential buildings and residential buildings built before 1980. Each subdivision improved the accuracy for both targets regarding CV, while MAPE remained constant. Thus, we can improve our prediction accuracy regarding CV by subdividing the data, but this is not possible for MAPE. We obtained the lowest CV of 0.23 for residential buildings in Geneva built before 1980 (see Table 14). However, this low CV was only achieved for this subgroup.

We also achieved the lowest MAPE below 0.2 for buildings in Geneva. Thus, our MAPE is already higher than the expected maximum percentage error of geoimpact. To achieve our accuracy, we need additional information, such as additional features for the energy condition of a building or the behaviour of the occupants. We also found that the energy efficiency of buildings in Biel and St. Gallen is lower than that of buildings in Geneva, even though there are more new buildings there. This could indicate that more accurate heat consumption data or more normalization steps are needed, such as correcting the heating degree days or the energy demand area.

The data show a large variation in building energy consumption between cities. Either the energy consumption really varies that much between cities, or our data is flawed or incomplete, or we need further information on our datasets, i.e., more characteristics and more observations from other regions of Switzerland. In any case, we need more data to scale the model on a nationwide level.

To conclude, our study modelled the buildings heat energy consumption on different sub samples. Thereby the RFR and the XGB were the best fitting regressors, the transformation of our target variable was not that significant. We get the big leverage when we reduce the variance within the data by sampling the buildings into more homogeneous groups, such as residential buildings and old residential buildings. We have shown that taking samples can increase the accuracy of the models and this finding may be generalizable to other data science projects.

6. Bibliography

Ackermann, T. (2019). *Energiebedarf versus Energieverbrauch oder Theorie versus Realität* (p. 26). Fachhochschule Bielefeld Campus Minden Institut für Bauphysik und Baukonstruktion. https://www.hausundgrund.de/sites/default/files/downloads/fh-bielefelduntersuchungenergiebedarfversusenergieverbrauch12112019.pdf

Al-Homoud, M. S. (2001). Computer-aided building energy analysis techniques. *Building and Environment*, *36*(4), 421–433. https://doi.org/10.1016/S0360-1323(00)00026-3

Ali, U., Shamsi, M. H., Hoare, C., Mangina, E., & O'Donnell, J. (2021). Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis. *Energy and Buildings*, *246*, 111073. https://doi.org/10.1016/j.enbuild.2021.111073

Amt für Umwelt und Energie AUE. (2020). *Energiebedarfsdaten Wohnen und Betriebe Kanton Bern Kurzdokumentation*. https://files.be.ch/bve/agi/geoportal/gbd/dm/dmrpe_be/RPE_Berechnung_Energiebedarf_EBBE_Kurzfassung_2020.PDF

ASHRAE. (1997). Energy Estimating and Modeling Methods. In *1997 ASHRAE handbook: Fundamentals.* (p. 30.2). ASHRAE.

ASHRAE. (2005). *ASHRAE's GUIDELINE 14-2002 FOR MEASUREMENT OF ENERGY AND DEMAND SAVINGS: HOW TO DETERMINE WHAT WAS REALLY SAVED BY THE RETROFIT.* 13.

Bakshi, C. (2020, June 9). *Random Forest Regression*. Medium. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Berger, M., & Worlitschek, J. (2019). The link between climate and thermal energy demand on national level: A case study on Switzerland. *Energy and Buildings*, *202*, 109372. https://doi.org/10.1016/j.enbuild.2019.109372

Binz, A., Bichsel, J., Geissler, A., Hall, M., Huber, H., Ragonesi, M., Steinke, G., & Weickgenannt, B. (2020). *Neubau. Energieeffizientes Bauen.* Faktor Verlag AG. https://pubdb.bfe.admin.ch/de/publication/download/7353

Bourdeau, M., Zhai, X. qiang, Nefzaoui, E., Guo, X., & Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society*, *48*, 101533. https://doi.org/10.1016/j.scs.2019.101533

Braulio-Gonzalo, M., Bovea, M. D., Jorge-Ortiz, A., & Juan, P. (2021). Which is the best-fit response variable for modelling the energy consumption of households? An analysis based on survey data. *Energy*, *231*, 120835. https://doi.org/10.1016/j.energy.2021.120835

Carisch, A., Eifert, M., Janes, P., Schütz, P., & Suter, D. (2020). *Gebäudepass «Cloud»- Whitepaper*.

Crawley, D. B., Lawrie, L. K., Winkelmann, F. C., Buhl, W. F., Huang, Y. J., Pedersen, C. O., Strand, R. K., Liesen, R. J., Fisher, D. E., & Witte, M. J. (2001). EnergyPlus: Creating a new-generation building energy simulation program. *Energy and Buildings*, *33*(4), 319–331.

Crawley, D. B., Pedersen, C. O., Lawrie, L. K., & Winkelmann, F. C. (2000). Energy plus: Energy simulation program. *ASHRAE Journal*, *42*(4), 49–56. Scopus.

Deb, C., & Schlueter, A. (2021). Review of data-driven energy modelling techniques for building retrofit. *Renewable and Sustainable Energy Reviews*, *144*, 110990. https://doi.org/10.1016/j.rser.2021.110990

Ecospeed AG. (2019). *ECOSPEED Region*. https://www.ecospeed.ch/region/en/

Ecospeed Immo. (2021). *GEBÄUDESCHARFE ENERGIEDATEN FÜR KANTONE UND GEMEINDEN. Bedienungsanleitung & Methodik*.

Energie Schweiz AG. (n.d.). *Heizkostenrechner*. Erneuerbar heizen. Retrieved December 14, 2021, from https://erneuerbarheizen.ch/heizkostenrechner/

energie-environnement.ch. (n.d.). *Wärmebedarf und GEAK - Fachstellen für Energie und Umwelt der Kantone Bern, Freiburg, Genf, Jura, Neuenburg, Waadt und Wallis*. Retrieved December 15, 2021, from https://www.energie-umwelt.ch/haus/renovation-und-heizung/gebaeudeplanung/waermebedarf-und-geak

Energieforschung Stadt Zürich. (2019). *Heizungsersatz: Vergleich ausgewählter Städte und Gemeinden* (Zwischenbericht Nr. 55). Energieforschung Stadt Zürich.

https://www.econcept.ch/media/projects/downloads/2019/11/201911_FP-2.8.1_Staedtevergleich_Modul_A_und_B_EFZ_Layout_Ber_gjsuRk6.pdf

Federal Laboratory for Materials Testing and Research. (2021, August 24). *Energy renovation: First sort, then refurbish*. https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-84812.html

Federal Office for the Environment FOEN. (2016). *Methodik zur Berechnung der kantonalen CO2-Emissionen im Gebäudebereich auf Basis des Gebäude- und Wohnungsregisters (GWR)*. 27.

Federal Office for the Environment FOEN. (2020, December 11). *Klimaschutz: Fünf Jahre Pariser Übereinkommen*. https://www.bafu.admin.ch/bafu/de/home/dokumentation/medienmitteilungen/anzeige-nsb-unter-medienmitteilungen.msg-id-81567.html

Federal Office for the Environment FOEN. (2021a). *Kenngrössen zur Entwicklung der Treibhausgasemissionen in der Schweiz 1990–2019*. 70.

Federal Office for the Environment FOEN. (2021b, April 12). *Greenhouse gas emissions from buildings*. https://www.bafu.admin.ch/bafu/en/home/themen/thema-klima/klima--daten--indikatoren-und-karten/daten--treibhausgasemissionen-der-schweiz/treibhausgasinventar/treibhausgasemissionen-der-gebaeude.html

Federal Office for the Environment FOEN. (2021c). *Kantonale Energiekennzahlen und CO2-Emissionen im Gebäudebereich*. 119.

Federal Statistical Office FSO. (n.d.). *Bauperiode*. Retrieved December 15, 2021, from https://www.bfs.admin.ch/bfs/de/home/statistiken/bau-wohnungswesen/gebaeude/periode.html

Federal Statistical Office FSO. (2017). *Energy field*. https://www.bfs.admin.ch/bfs/en/home/statistiken/bau-wohnungswesen/gebaeude/energiebereich.html

Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews*, *23*, 272–288. https://doi.org/10.1016/j.rser.2013.03.004

Gao, X., & Malkawi, A. (2014). A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, *84*, 607–616. https://doi.org/10.1016/j.enbuild.2014.08.030

geoimpact AG. (2021). *Die Energieverbrauchsschätzung wird benötigt für • eine effizientere Beratung sowie eine automatisierte Erstellung einer Richtofferte bei einen Heizungsersatz • Wärmeplanung auf Quartier oder Gemeindeebene, zB zur Planung von Wärmeverbünden* [Personal communication].

geoimpact AG. (2021, December 16). *A partner of Geoimpact requires estimates of a building's energy performance with a maximum percentage absolute error of 15%.* [Personal communication].

Gerster, S. A., & Nietlisbach, A. (2014). *Sinkende Energie- kennzahl von Wohnbauten*. 2.

Geyer, P., Schlüter, A., & Cisar, S. (2017). Application of clustering for the development of retrofit strategies for large building stocks. *Advanced Engineering Informatics*, *31*, 32–47. https://doi.org/10.1016/j.aei.2016.02.001

Gloor, R. (2020, February 22). *Energieverbrauch von Gebäuden*. energie.ch. https://energie.ch/heizenergieverbrauch/

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd edition). Springer.

Hawkins, J. (2015). *Support Vector Regression*. 14. https://doi.org/10.1007/978-1-4302-5990-9_4

Holcomb, D., Li, W., & Seshia, S. A. (2009). Algorithms for green buildings: Learning-based techniques for energy prediction and fault diagnosis. *Google Scholar, UCB/EECS-2009-138*.

Hoval Schweiz. (n.d.). *Wärmepumpe effizient betreiben – mit diesen Tricks klappt es | Hoval Schweiz*. Retrieved December 19, 2021, from https://www.hoval.ch/de_CH/W%C3%A4rmepumpe-effizient-betreiben-%E2%80%93-mit-diesen-Tricks-klappt-es/drei-tipps-effiziente-waermepumpe

Husi, A. (2021, April 12). *GEAK data for modelling of buildings energy demand* [Personal communication].

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kamoshida, J., Hirata, Y., Isshiki, N., Katayama, K., & Sato, K. (1990). Thermodynamic Analysis of Resorption Heat Pump Cycle Using Water-Multicomponent Salt Mixture. In T. Saito (Ed.), *Heat Pumps* (pp. 545–554). Pergamon. https://doi.org/10.1016/B978-0-08-040193-5.50064-X

Konferenz Kantonaler Energiedirektoren EnDK. (2014). *Energieverbrauch von Gebäuden*.

Kozyrkov, C. (2019, August 9). *The most powerful idea in data science*. https://towardsdatascience.com/the-most-powerful-idea-in-data-science-78b9cd451e72

Li, W., Zhou, Y., Cetin, K., Eom, J., Wang, Y., Chen, G., & Zhang, X. (2017). Modeling urban building energy use: A review of modeling approaches and procedures. *Energy (Oxford)*, *141*, Article PNNL-SA-129914. https://doi.org/10.1016/J.ENERGY.2017.11.071

Li, Z., Han, Y., & Xu, P. (2014). Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy*, *124*, 325–334. https://doi.org/10.1016/j.apenergy.2014.03.020

Lisowski, E. (2019). XGBoost and Random Forest® with Bayesian Optimisation. *KDnuggets*. https://www.kdnuggets.com/xgboost-and-random-forest-with-bayesian-optimisation.html/

Liu, J., Chen, H., Liu, J., Li, Z., Huang, R., Xing, L., Wang, J., & Li, G. (2017). An energy performance evaluation methodology for individual office building with dynamic energy benchmarks using limited information. *Applied Energy*, *206*, 193–205. https://doi.org/10.1016/j.apenergy.2017.08.153

McLoughlin, F., Duffy, A., & Conlon, M. (2012). Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, *48*, 240–248. https://doi.org/10.1016/j.enbuild.2012.01.037

Melillo, A., Durrer, R., Worlitschek, J., & Schuetz, P. (2020). First results of remote building characterisation based on smart meter measurement data. *Energy*, *200*, 117525. https://doi.org/10.1016/j.energy.2020.117525

Minergie. (2021). *Produktreglement Minergie v2021.1*. https://www.minergie.ch/media/201223_produktreglement_minergie_p_a_v2021.1_de.pdf

ND, D. (2020, February 6). *regression—Log-Transforming target var for training a Random Forest Regressor*. Cross Validated. https://stats.stackexchange.com/questions/447863/log-transforming-target-var-for-training-a-random-forest-regressor

Osmulski, R. (2018, March 9). *Why take the log of a continuous target variable?* Medium. https://towardsdatascience.com/why-take-the-log-of-a-continuous-target-variable-1ca0069ee935

Park, H. S., Lee, M., Kang, H., Hong, T., & Jeong, J. (2016). Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Applied Energy*, *173*, 225–237. https://doi.org/10.1016/j.apenergy.2016.04.035

Pasichnyi, O., Levihn, F., Shahrokni, H., Wallin, J., & Kordas, O. (2019). Data-driven strategic planning of building energy retrofitting: The case of Stockholm. *Journal of Cleaner Production*, *233*, 546–560. https://doi.org/10.1016/j.jclepro.2019.05.373

Pradhan, D. (2019). *What are the advantages and disadvantages for a random forest algorithm?* Quora. https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a-random-forest-algorithm

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* (1st edition). O'Reilly Media.

Rabani, M., Bayera Madessa, H., & Nord, N. (2021). Achieving zero-energy building performance with thermal and visual comfort enhancement through optimization of fenestration, envelope, shading device, and energy supply system. *Sustainable Energy Technologies and Assessments*, *44*, 101020. https://doi.org/10.1016/j.seta.2021.101020

Re Cecconi, F., Moretti, N., & Tagliabue, L. C. (2019). Application of artificial neutral network and geographic information system to evaluate retrofit potential in public school buildings. *Renewable and Sustainable Energy Reviews*, *110*, 266–277. https://doi.org/10.1016/j.rser.2019.04.073

Schluck, T., Streicher, K. N., & Mennel, S. (2019). Statistical modelling of the energy reference area based on the Swiss building stock. *Journal of Physics: Conference Series*, *1343*, 012031. https://doi.org/10.1088/1742-6596/1343/1/012031

Schuetz, P., Melillo, A., Businger, F., Durrer, R., Frehner, S., Gwerder, D., & Worlitschek, J. (2020). Automated modelling of residential buildings and heating systems based on smart grid monitoring data. *Energy and Buildings*, *229*, 110453. https://doi.org/10.1016/j.enbuild.2020.110453

Schuetz, P., Scoccia, R., Gwerder, D., Waser, R., Sturzenegger, D., Elguezabal, P., Arregi, B., Sivieri, A., Aprile, M., & Worlitschek, J. (2019). Fast Simulation Platform for Retrofitting Measures in Residential Heating. In D. Johansson, H. Bagge, & Å. Wahlström (Eds.), *Cold Climate HVAC 2018* (pp. 713–723). Springer International Publishing. https://doi.org/10.1007/978-3-030-00662-4_60

Schweizerische Energie-Stiftung. (n.d.). *Gebäudestandards*. Gebäudestandards in der Schweiz: Eine Übersicht. Retrieved December 15, 2021, from https://www.energiestiftung.ch/energieeffizienz-gebaeudestandards.html

Schweizerischer Ingenieur- und Architekten-Verein SIA. (1982). *Energiekennzahl*. 20.

Solar Campus GmbH. (2007). *Tachion Framework*. http://www.solarcampus.ch/

Solar Campus GmbH. (2020). *Benutzerdokumentation für die Gebäude-Energiesimulation*.

Streicher, K. N., Padey, P., Parra, D., Bürer, M. C., Schneider, S., & Patel, M. K. (2019). 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. https://doi.org/10.1016/j.enbuild.2018.12.011

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*(8), 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Swiss Federal Office of Energy SFOE. (2016, February 1). *Beilage "Solarkataster Schweiz: Abschätzung des Wärme- und Brauchwarmwasserbedarfs."* https://docplayer.org/33618265-Beilage-solarkataster-schweiz-abschaetzung-des-waerme-und-brauchwarmwasserbedarfs.html

Swiss Federal Office of Energy SFOE. (2019). *GAPxPLORE: Energy Performance Gap in existing, new, and renovated buildings Learning from large-scale datasets*. https://www.minergie.ch/media/2019_gapxplore_langversion.pdf

Swiss Federal Office of Energy SFOE. (2020). *Das Gebäudeprogramm Jahresbericht 2020*. https://www.dasgebaeudeprogramm.ch/media/filer_public/03/a6/03a60b89-4d45-487b-bf88-b1d421521a23/bfe_gebaudeprogrammjahresbericht_de_210805_final.pdf

Swiss Federal Office of Energy SFOE. (2020a). *Gebäudepark 2050—Vision des BFE*. https://pubdb.bfe.admin.ch/de/publication/download/8985

Swiss Federal Office of Energy SFOE. (2020b). *Analyse des schweizerischen Energieverbrauchs 2000–2019—Auswertung nach Verwendungszwecken*. https://www.bfe.admin.ch/bfe/en/home/versorgung/statistik-und-geodaten/energiestatistiken/energieverbrauch-nach-verwendungszweck.exturl.html/aHR0cHM6Ly9wdWJkYi5iZmUuYWRtaW4uY2gvZGUvcHVibGljYX/Rpb24vZG93bmxvYWQvMTAyNjA=.html

Swiss Federal Office of Energy SFOE. (2021). *Heizsysteme: Entwicklung der Marktanteile 2007-2020: Aktualisierung 2021*.

SwissEnergyPlanning SEP. (n.d.). *Swiss Energy Planning | Energieplanung | Schweiz*. SwissEnergyPlanning. Retrieved May 31, 2021, from https://www.swissenergyplanning.ch

Thermondo. (2019, December 5). *Wirkungsgrad der Heizung – wichtige Kennzahl für die Effizienz des Heizgeräts*. Wirkungsgrad der Heizung. https://www.thermondo.de/info/rat/vergleich/wirkungsgrad-der-heizung/

Tong, X., Li, R., Li, F., & Kang, C. (2016). Cross-domain feature selection and coding for household energy behavior. *Energy*, *107*, 9–16. https://doi.org/10.1016/j.energy.2016.03.135

Umwelt und Energie UWE. (2008, dez). *Beurteilen Sie den Energieverbrauch Ihres Gebäudes!* https://uwe.lu.ch/-/media/UWE/Dokumente/publikationen/Publikationen_01_A_bis_F/Energieverbrauch_beurteilen.pdf

Umwelt und Energie UWE. (2015). *Energiespiegel—Methodik und Diskussion*. https://uwe.lu.ch/-/media/UWE/Dokumente/Themen/Energie/Energiespiegel/Methoden_Energiespiegel.pdf?la=de-CH

UVEK, E. D. für U., Verkehr, Energie und Kommunikation. (n.d.). *Energiestrategie 2050*. Retrieved December 14, 2021, from https://www.uvek.admin.ch/uvek/de/home/energie/energiestrategie-2050.html

Verein GEAK-CECB-CECE. (2021). *Was ist der GEAK / GEAK*. https://www.geak.ch/der-geak/was-ist-der-geak/

Wang, D., Landolt, J., Mavromatidis, G., Orehounig, K., & Carmeliet, J. (2018). CESAR: A bottom-up building stock modelling tool for Switzerland to address sustainable energy transformation strategies. *Energy and Buildings*, *169*, 9–26. https://doi.org/10.1016/j.enbuild.2018.03.020

Wenninger, S., & Wiethe, C. (2021). Benchmarking Energy Quantification Methods to Predict Heating Energy Performance of Residential Buildings in Germany. *Business & Information Systems Engineering*, *63*(3), 223–242. https://doi.org/10.1007/s12599-021-00691-2

Westermann, P., Deb, C., Schlueter, A., & Evins, R. (2020). Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Applied Energy*, *264*, 114715. https://doi.org/10.1016/j.apenergy.2020.114715

Wick, B. (1982). *Energie im Mehrfamilienhaus: Verbrauchswerte und Sparpotential* [Text/html,application/pdf]. https://doi.org/10.5169/SEALS-74749

Wilhelm, F. (2020, May 4). *Honey, I shrunk the target variable*. Florian Wilhelm’s Blog. https://florianwilhelm.info/2020/05/honey_i_shrunk_the_target_variable/

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process modell for data mining. *Undefined*. https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard-process-modell-for-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24

XGBoost. (n.d.). *XGBoost Documentation—Xgboost 1.5.1 documentation*. Retrieved December 21, 2021, from https://xgboost.readthedocs.io/en/stable/

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049

Master of Science (MScIDS) in Applied Information and Data Science

Declaration of originality

The following declaration of originality must be signed by hand and included at the end of the Master's Thesis:

"The undersigned hereby declares that he or she

- wrote the work in question independently and without the help of any third party,
- has provided all the sources and cited the literature used,
- will protect the confidentiality interests of the client and respect the copyright regulations of Lucerne University of Applied Sciences and Arts."

Date and signature

23.12.2021 A. Cubser

Lucerne University of Applied Sciences and Arts
Master of Science (MScIDS) in Applied Information and Data Science
Lucerne, 30 September 2019