KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Energy Technology

Second cycle 30 credits

PV self-consumption: Regression models and data visualization

MARTOS TÓTH

Stockholm, Sweden 2022

KTH Industrial Engineering and Management

Master of Science Thesis

Department of Energy Technology

KTH 2022

PV self-consumption: Regression models and data visualization

TRITA-ITM-EX 2022:344

Martos Tóth

Approved	Examiner	Supervisor
2022-06-22	Hatef Madani	Nelson Sommerfeldt
	Commissioner	Contact person
	David Stoltz	David Stoltz

Abstract

In Sweden the installed capacity of the residential PV systems is increasing every year. The lack of feed-in-tariff-scheme makes the techno-economic optimization of the PV systems mainly based on the self-consumption. The calculation of this parameter involves hourly building loads and hourly PV generation. This data cannot be obtained easily from households. A predictive model based on already available data would be preferred and needed in this case. The already available machine learning models can be suitable and have been tested but the amount of literature in this topic is fairly low.

The machine learning models are using a dataset which includes real measurement data of building loads and simulated PV generation data and the calculated self-consumption data based on these two inputs. The simulation of PV generation can be based on Typical Meteorological Year (TMY) weather file or on measured weather data. The TMY file can be generated quicker and more easily, but it is only spatially matched to the building load, while the measured data is matched temporally and spatially. This thesis investigates if the usage of TMY file leads to any major impact on the performance of the regression models by comparing it to the measured weather file model. In this model the buildings are single-family houses from south Sweden region.

The different building types can have different load profiles which can affect the performance of the model. Because of the different load profiles, the effect of using TMY file may have more significant impact. This thesis also compares the impact of the TMY file usage in the case of multifamily houses and also compares the two building types by performance of the machine learning models.

The PV and battery prices are decreasing from year to year. The subsidies in Sweden offer a significant tax credit on battery investments with PV systems. This can make the batteries profitable. Lastly this thesis evaluates the performance of the machine learning models after adding the battery to the system for both TMY and measured data. Also, the optimal system is predicted based on the self-consumption, PV generation and battery size.

The models have high accuracy, the random forest model is above 0.9 R^2 for all cases. The results confirm that using the TMY file only leads to marginal errors, and it can be used for the training of the models. The battery model has promising results with above 0.9 R^2 for four models: random forest, k-NN, MLP and polynomial. The prediction of the optimal system model has promising results as well for the polynomial model with 18% error in predicted payback time compared to the reference.

Keywords: Self-Consumption, photovoltaics, battery, machine learning, solar energy, random forest, k-nearest neighbor, multi-layer perceptron, lasso regression, ridge regression, linear regression

Sammanfattning

I Sverige ökar den installerade kapaciteten för solcellsanläggningarna för bostäder varje år. Bristen på inmatningssystem gör att den tekniska ekonomiska optimeringen av solcellssystemen huvudsakligen bygger på egen konsumtion. Beräkningen av denna parameter omfattar byggnadsbelastningar per timme och PV-generering per timme. Dessa uppgifter kan inte lätt erhållas från hushållen. En prediktiv modell baserad på redan tillgängliga data skulle vara att föredra och behövas i detta fall. De redan tillgängliga maskininlärningsmodellerna kan vara lämpliga och redan testade men mängden litteratur i detta ämne är ganska låg.

Maskininlärningsmodellerna använder en datauppsättning som inkluderar verkliga mätdata från byggnader och simulerad PV-genereringsdata och den beräknade egenförbrukningsdata baserad på dessa två indata. Simuleringen av PV-generering kan baseras på väderfilen Typical Meteorological Year (TMY) eller på uppmätta väderdata. TMY-filen kan genereras snabbare och enklare, men den anpassas endast rumsligt till byggnadsbelastningen, medan uppmätta data är temporärt och rumsligt. Denna avhandling undersöker om användningen av TMY-fil leder till någon större påverkan på prestandan genom att jämföra den med den uppmätta väderfilsmodellen. I denna modell är byggnaderna småhus från södra Sverige.

De olika byggnadstyperna kan ha olika belastningsprofiler vilket kan påverka modellens prestanda. På grund av dessa olika belastningsprofiler kan effekten av att använda TMY-fil ha mer betydande inverkan. Den här avhandlingen jämför också effekten av TMY-filanvändningen i fallet med flerfamiljshus och jämför också de två byggnadstyperna efter prestanda för maskininlärningsmodellerna.

PV- och batteripriserna minskar från år till år. Subventionerna i Sverige ger en betydande skattelättnad på batteriinvesteringar med solcellssystem. Detta kan göra batterierna lönsamma. Slutligen utvärderar denna avhandling prestandan för maskininlärningsmodellerna efter att ha lagt till batteriet i systemet för både TMY och uppmätta data. Det optimala systemet förutsägs också baserat på egen förbrukning, årlig byggnadsbelastning, årlig PV-generering och batteristorlek.

Modellerna har hög noggrannhet, den slumpmässiga skogsmodellen är över 0,9 R2 för alla fall. Resultaten bekräftar att användningen av TMY-filen endast leder till marginella fel, och den kan användas för träning av modellerna. Batterimodellen har lovande resultat med över 0,9 R2 för fyra modeller: random skog, k-NN, MLP och polynom. Förutsägelsen av den optimala systemmodellen har också lovande resultat för polynommodellen med 18 % fel i förutspådd återbetalningstid jämfört med referensen.

Nyckelord: Egenanvändning, photovoltaics, batteri, maskininlärning, solenergi, random forest, k-nearest neighbors, multi-layer perceptron, lasso regression, ridge regression, linjär regression

Acknowledgement

This research is funded by the Swedish Energy Agency through the Design for Energy Effective Lifestyles program (Project Number 48103-1).

Thanks also goes to David Stoltz and Fredrik Balderud at Karlstad Energi for supplying the load data.

I am grateful for the support of my supervisor, Nelson Sommerfeldt. I appreciate your confidence in me and your mentorship throughout my thesis work. Thank you for your helpful advice and your availability every week.

I would also like to thank for the support of my family and friends during my thesis and studies.

Table of Contents

List of Figures

List of Tables

Nomenclature

CV: Cross-Validation
DHI: Diffuse Horizontal Irradiance
DNI: Direct Normal Irradiance
GHI: Global Horizontal Irradiance
k-NN: K-Nearest Neighbors
KPI: Key Performance Index
MAE: Mean Absolute Error
MBE: Mean Bias Error
ML: Machine Learning
MLP: Multi-layer Perceptron
PV: Photovoltaics
PVGIS: Photovoltaic Geographical Information System
RSD: Relative Standard Deviation
SAM: System Advisor Model
SC: Self-Consumption
SF: Solar Fraction
TMY: Typical Meteorological Year

1. Introduction

In Sweden, almost half of the installed grid-connected PV systems were roof-mounted residential systems in 2020. This category had the largest share in the installed PV capacity in the previous years. The installed capacity keeps increasing but the change in installed capacity of residential systems from 2019 to 2020 is relatively small compared to 2018-2019 where it is almost doubled [1]. This indicates a slowdown in adoption of PV systems in the residential sector. The electricity price stayed at similar levels from 2018 to 2020 [2] while the turnkey PV system prices keep decreasing from year to year [3]. The average crystalline silicon module efficiency is kept increasing as well [4] which means that the profitability of the PV systems keeps increasing.

To support the installation of the PV systems, there are subsidies which reduce the capital cost by giving tax credits. There is no feed-in-tariff scheme in Sweden [1], thus the business model of these systems is driven by SC. The SC is determined by hourly building load and hourly PV generation. So, to calculate the profitability of the PV system, the hourly building load data is needed. This data cannot be easily accessed therefore models were made to simplify the process. In the literature only two works have been done to solve this problem, one is by Galli [5] and the other is by McKenna [6]. Galli tested various machine learning (ML) models and McKenna used linear regression that can predict the SC from the annual building load and the annual PV generation. In this method the input parameters are easily obtainable, and this makes the process simpler for the customer. The training dataset for the ML models includes the SC, annual building load and annual PV generation. The SC is calculated from the measured hourly building load and the simulated hourly PV generation.

The regression model is using a training dataset that has one or more input variables and one output variable. During the training, the model is finding the best coefficients to fit the model to the samples with the smallest error. Then this model can be tested on new datapoints, during this process the model is predicting the output variables based on the input variables. The predicted and real output value can be compared, and various measures can be introduced to assess the performance of the model. The regression models' complexity has high impact on the performance. A simpler model may not represent every relationship between the input variables and the output variable. This results in underfitting. A complex model may represent relations in the input and output that do not exist and if it tested on new data, the model has low performance. This is called overfitting.

2. Problem statement and objectives

From weather stations hourly weather data can be acquired and matched spatially and temporally with the building load data, this weather data will be referenced as "measured". In the previous work of Galli [5], the training of the ML algorithms was done by using temporally and spatially mismatched weather data. This could lead to errors in predicting the SC with the models due to the variance in weather conditions in different years. Therefore, the building load will be matched spatially and temporally with the weather data. Then the models are retrained, and the performance of the regression models are evaluated and compared to the models in Galli's work [5]. Also, the performance of only spatially matched weather data is tested and further referenced as "TMY".

Different building types have different consumption profiles throughout the year. In Galli's work [5] single-family houses consumption data was used in the regression models. With the use of a different building types, the performance of the models can be showed and further used in techno-economic analysis. The models will be trained on the multifamily houses' consumption data with both measured and TMY weather data to show if using TMY data has any impact on the performance with this building type.

The battery prices have been decreasing year after year and the subsidies offer relatively high amount of cash back that makes the PV plus battery systems potentially profitable. The addition of the battery in the model can be done by recalculating the SC and then the regression models can be retrained with the new dataset. The payback time can be calculated also, and this could help potential buyers easily decide the optimal PV size and whether adding battery to the PV system is worth it or not. A battery model will be created and added to the PV system for both the single-family houses and multifamily houses. Then a new SC dataset is created, so the models can be retrained, and evaluated. After this the SC can be predicted from the annual building load, annual PV generation and battery size. A financial model will be also made to calculate the payback time of the system from the same inputs.

These objectives will contribute to have a robust regression model that can predict SC for residential buildings with an acceptable level of accuracy. The building loads are obtained from buildings in or near Karlstad, Sweden for all of the models, so they are only applicable in that region.

3. Methodology

Two types of weather data are generated, the measured and the TMY. The measured data will be preprocessed to have the same hourly datapoints as the building load and no missing values. It will also be split by years to match temporally the building load, and this will result in weather data from 2015 to 2021.The simulation of the PV generation is done by the PVWatts model [7] from the PySAM library. The hourly DNI values are calculated with the DIRINT model [8]. This model is a modified DISC model that uses the hourly zenith angle to have better performance in predicting the DNI values from the GHI. The DISC model is derived statistically from a large multi-climatic experimental data base [8].

The PV generation is assessed with multiple orientations to have a broader generation profile. This is necessary to obtain a large dataset which can be used as training data for the regression models. The different orientation is set by the tilt and azimuth angles. The simulation is run for the seven years of the measured data and for the TMY data separately.

The building loads are obtained from Karlstad Energi. The single-family house loads are the same as in [5] and it is already preprocessed to fill missing data gaps. It contains five years of measured loads from 2015 to 2019 for 108 buildings. The multifamily house data will be preprocessed by removing the buildings that have a significant number of missing values and other buildings missing values will be filled with previous or next hours of data if available or previous or next days, years if available.

The next step will be to calculate the SC. This can be done with the hourly PV generation and

building load. In each hour it checks if the PV generation is higher than the building load if it is than the building load is added to the hourly SC sum and if it is not, then the PV generation is added to this sum. After summing this value for the whole year and dividing it with the annual PV generation the annual SC is obtained. To have a larger dataset, the hourly PV value will be multiplied with a scaling factor. This linearly scaled by 10 points between the initial PV size which is 1 kW and the largest which annual PV generation is the same as the annual building load. The generation of the SC dataset is similar for both the single-family houses and multifamily houses.

A descriptive analysis is done on the annual building loads and the SC dataset to identify any outliers which could help improve the performance of the regression models. Then the SC dataset is used to train and test regression models. Eight different regression model are used: random forest, k-nearest neighbors, MLP, polynomial, ridge, lasso, linear and the McKenna model [6]. These models are implemented using the scikit-learn library in Python. The testing is done by using cross-validation, to better detect the overfitting of the regressions. The performance of the models is compared by mean absolute error (MAE) and R^2 (r-squared). The overfitting of the models is checked by the adjusted R^2 and the bias is by the mean bias error (MBE). This is done separately for single-family houses and multifamily houses.

For the battery model, the annual SC needs to be calculated for different battery sizes. To be able to compare the different PV sizes with different battery sizes, relative battery size is introduced which is calculated by dividing the battery size by the PV size. For the battery model a financial model is also added so the payback time of the different systems can be assessed. This model is calculating the payback time based on the predicted SC from the regressions, the PV size, the annual PV generation, and the battery size.

4. Modelling

In this chapter the steps of the modelling will be discussed. First the steps of the data processing are described. Then the battery model is described, after that the KPI's are introduced with which the regression models' performance can be quantified. The last section describes the financial model in detail.

4.1. Data processing

The data processing has a high impact on the performance of the regressions. The data can be noisy or have outliers that will change the trend of data. These will harm the goodness of the fit of the regression models. So, at every stage the data should be processed carefully. The three main data streams are the weather data, hourly building loads and PV generation which is calculated based on the weather data. All of the data processing is done in Python version 3.8 programming language.

4.1.1. Weather data

Two types of weather data are used in this study. The first is obtained from the Swedish Meteorological and Hydrological Institute (SMHI) station at Karlstad airport [9]. This weather data is spatially and temporally matched to the building loads. The second weather data is generated by Meteonorm 8.0 using the years 1996 to 2015 [10]. This climate data is only

spatially matched to the building load. These will be referenced further as “measured” and “TMY” accordingly.

For the measured weather data, the hourly values for the GHI, ambient dry bulb temperature and wind speed is downloaded. Only these parameters are downloaded since the PVWatts model in PySAM is only using these inputs plus the diffuse solar irradiance [7] which is not available from SMHI. Then measured weather data is filtered from 2015 to 2021, since the single-family house building loads are from 2015 to 2019 and the multifamily house loads are from 2019 to 2021. Most of the parameters had missing values in certain hours but not a significant amount, it was less than 1% for all the parameters. These missing values have been replaced first by previous or next hour if available then previous or next day if available and lastly by previous or next year. Only the GHI data had no missing values from 2015 to 2021. Then the DNI is calculated with the DIRINT model from the pvlib toolbox in Python. The DIRINT model uses GHI and hourly zenith angle to calculate the DNI. The zenith angle is calculated based on the latitude and longitude of the weather station with the solarposition function in the pvlib library. Then the DHI can be calculated by subtracting the DNI values multiplied by the cosine of the zenith angle from the GHI. After this all the weather data is available. To match it correctly to the building loads, the weather data is split for each year and seven EnergyPlusWeather (EPW) files are generated. The TMY weather data does not need any processing.

4.1.2. Building load

The single-family house loads have hourly loads for 108 buildings from 2015 to 2019 and are the same that Galli used in [5]. This does not need any processing, other than splitting it for each year to match the measured weather data temporally. The multifamily house loads were obtained from Karlstad Energi and contains hourly load profiles for 81 building from the southern region of Sweden. This data ranges from 2019 to early 2022. To only have whole years in the training data set, the values from 2022 are removed. Only four houses had significant amount of data missing, these are removed from the dataset leaving with 77 buildings, and other two had 21 hours missing from the 3 years of samples. These missing hours were filled with data from same hour in the previous day. All the other houses had no missing hourly values. To match the yearly measured weather data, the dataset is split into three years.

The monthly load pattern for five houses from the single-family house dataset is presented in Figure 1. All the buildings have similar load patterns with higher consumption during the winter months and lower consumption in the summer months, except for the building with 6.1 MWh annual building load. This house seems to have the same consumption in almost every month. The monthly load patterns for the multifamily houses can be seen in Figure 2. The building with 200.5 MWh annual load and the building with 20.2 MWh have similar load patterns as the single-family houses because the consumption of these houses is higher in the winter months than the summer. The building with 150.9 MWh annual load has higher consumption in winter, but the change is not significant compared to the summer months. The house with 62.1 MWh annual load has around the same consumption for every month. The building with 100.6 MWh has the strangest profile, since it has around 10,000 kWh consumption in every month except for October, November, and December where the monthly consumption drops to nearly zero kWh. The different load patterns indicate the usage of different electrical appliances such as air conditioning unit.

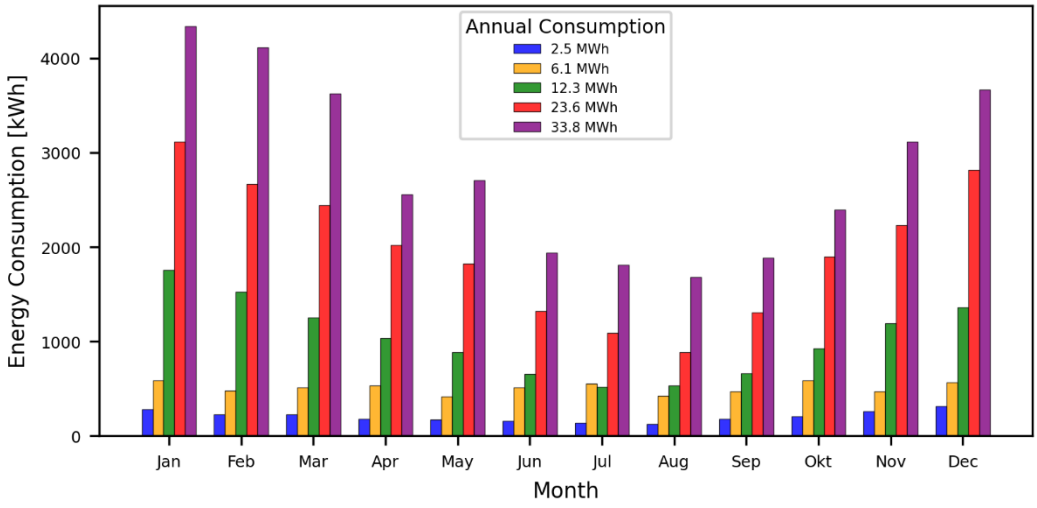

Figure 1: Monthly load pattern samples for single-family houses

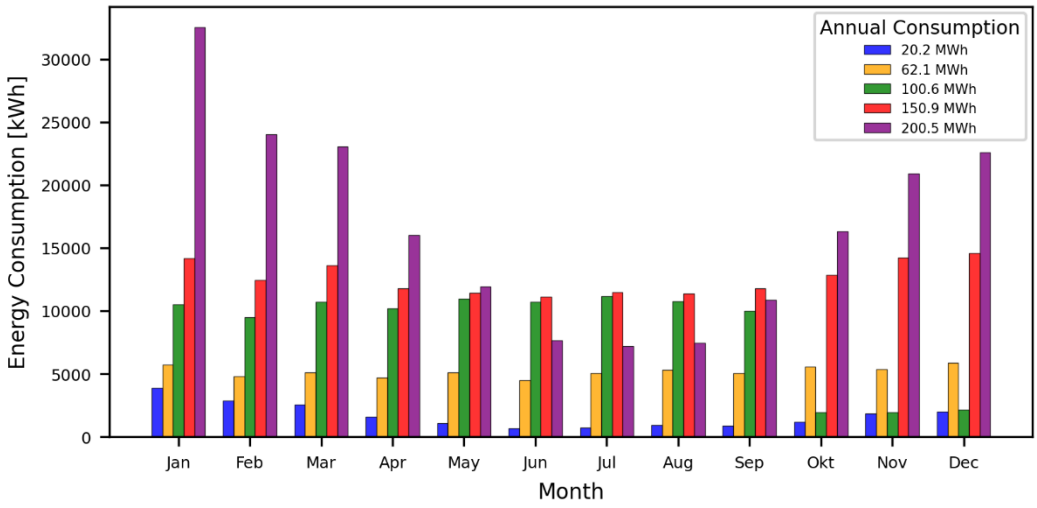

Figure 2: Monthly load pattern samples for multifamily houses

The change in annual building load for the single-family house dataset is presented in Figure 3. The five single-family house (denoted by sfh on the plot) has small changes in annual load except for sfh 3 where it increases gradually from around 12 MWh to around 20 MWh, from 2015 to 2019. A similar plot presents this data for the multifamily houses in Figure 4. The buildings in this sample have small changes from year to year except for the multifamily house 3 (denoted by mfh 3 on the plot), which has decreased its consumption from around 100 MWh in 2019 to around 25 MWh in 2020. This large change indicates that a high consumption appliance has taken away (e.g., a switch from electric to district heating) and/or the building was renovated in 2019.

The relatively high change in building loads from year to year does not mean that these buildings are outliers since each building considered to be a separate one for each year. So, these samples should not be removed from the dataset. However, these variations may have an impact on the SC variance and a descriptive analysis could help understand the uncertainty.

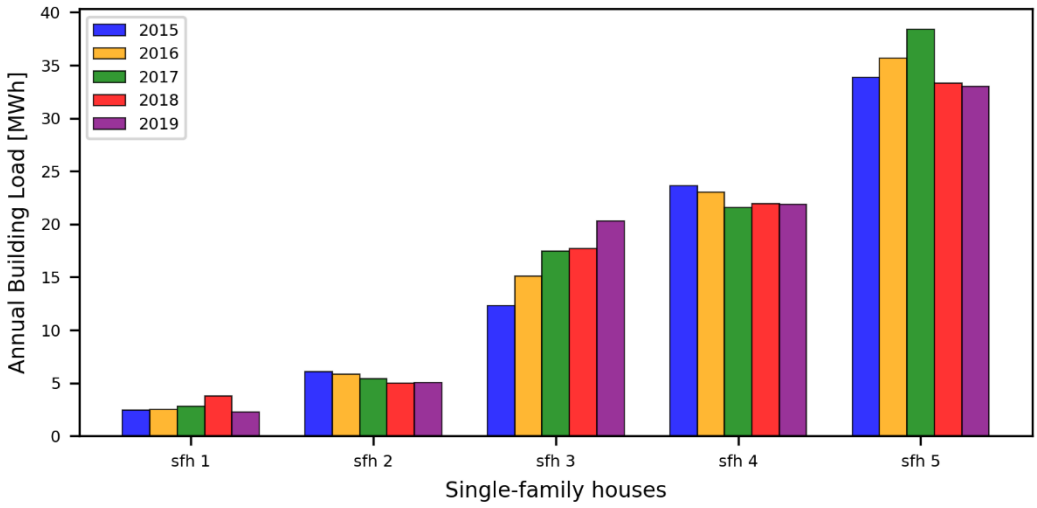

Figure 3: Sample of annual building loads for single-family houses

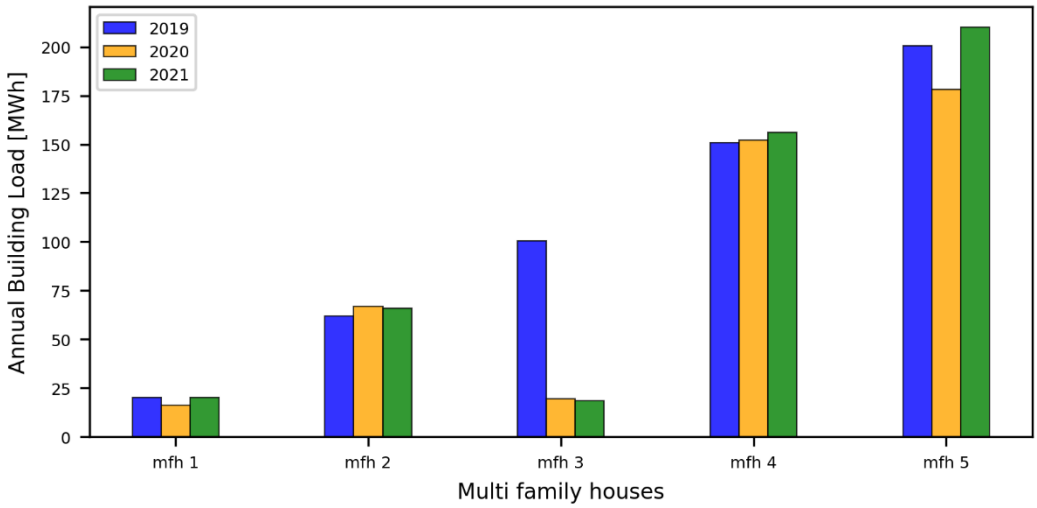

Figure 4: Sample of annual building loads for multifamily houses

4.1.3. PV generation

The hourly PV generation profiles are simulated with PySAM's PVWatts module [7]. To have a broader training dataset at the end, multiple orientations of PV systems is used. The tilt angle is changed between 0° and 90° with 15° increments while the azimuth angle is changed between 0° and 300° with 60° increments. This results in 42 generation profiles, the yields kWh/kW can be seen for each orientation for the TMY data in Figure 5. The simulation is done for seven years of measured weather data and for the TMY data. The PV size for the simulation is 1 kW_p and the PV array is fixed roof mounted. The efficiency of the inverter is set to 98% while the total system loss to 12%. These parameters are set according to Galli's study [5]. The monthly PV generation for the 30° tilt angle and 180° azimuth angle orientation from 2015 to 2021 is showed in Figure 6.

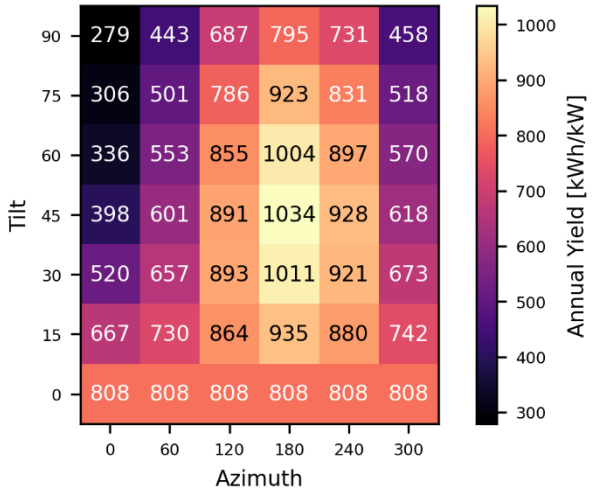

Figure 5: Annual PV yield by orientation for TMY data

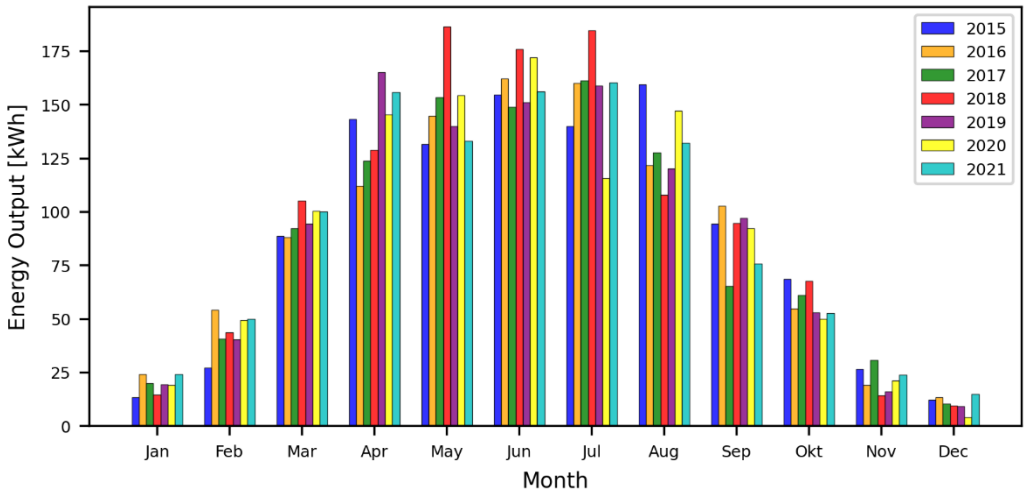

Figure 6: Monthly PV generation from 2015 to 2021

4.1.4. Self-consumption

The SC of a building is calculated by dividing the self-consumed PV generation with the total PV generation [11]. After all the data is preprocessed, the SC can be calculated. This is done for the matched seven years of building load with the PV generation and for the TMY based PV generation and the seven years of loads. The annual SC is calculated by summing up the hourly PV generation that is being used for supplying the load. This can be done by checking if the hourly PV generation is smaller than the hourly building load. If it is smaller, then the hourly PV generation is added to the sum. If it is not, then the hourly building load is added. After going through the year, the sum should be divided by the annual PV generation to get the annual SC. To have a larger dataset for training, the PV generation is scaled up based on the how much is the annual PV generation of the annual building load. This scaling factor is going from 10% to 100% with an increment of 10%. The 100% means that the annual PV generation equals to the annual building load. The previously shown 42 PV generation profiles with this scaling factor becomes 420 different profiles. This results in 226,800 data samples with the 540 single-family house profiles (each house is separated for each year, 5*108=540), and 97,020 data samples with the 231 multifamily house profiles (each house is separated for each year,

3*77=231).

The validation of the simulation results can be done by comparing the results to the SC values based on measurements in Sweden from 2018 published by Stridh [12]. The simulated and measured results can be seen in Figure 7. Both the simulated and the measured SC values are from 2018. The blue dots are the Swedish measurements, and the blue line is the average of them, those results are from [12]. The dotted line indicates the minimal and maximal values of SC for a certain solar fraction. Only few measured data point lies outside the min and max lines, which indicates that the simulation was done properly. The relatively small difference between the average of measurements and average of simulations, proves this as well.

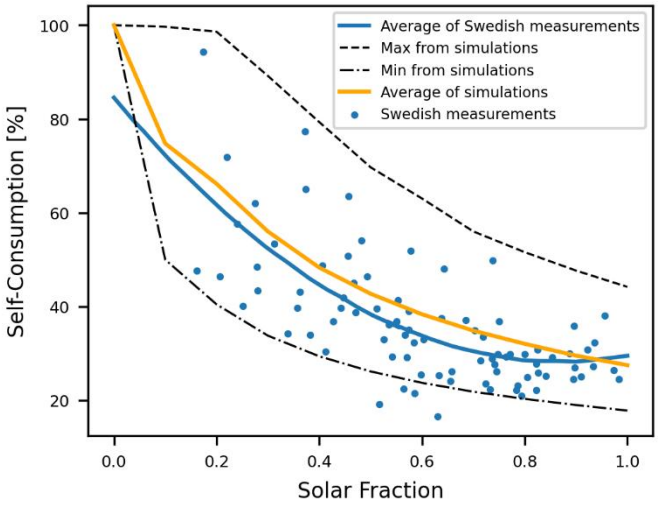

Figure 7: Simulated SC compared to measured, both from 2018

After the validation of the simulated SC values, the datasets for both single-family houses and multifamily houses can be used for training the regression models and the datasets do not have samples that will harm the performance of these models.

4.2. Battery model

There are seven parameters that are used in the battery model: charging efficiency, discharging efficiency, inverter efficiency, charging rate, depth of discharge (DOD), battery size and state of charge. The charging energy is higher than the actual added energy in the battery, this is the charging efficiency. The discharging efficiency is similar to this except the output energy to the consumer will be smaller than the discharged energy in the battery. The inverter efficiency is a loss during converting the DC current from the battery to AC, to use it in the household. The charging rate is the maximum energy that can be added to the battery per hour. The maximum DOD is the percentage of the battery capacity that can be used. This is implemented to increase the lifetime of the battery. The battery size or capacity is the amount of energy that it can store, and the state of charge represents how much energy is in the battery compared to its capacity.

The battery model is built based on the model by Han et.al. [13]. This model is simplistic but graphs the main characteristics of a battery system. The battery model should decide whether it can be charged from the surplus energy or discharged to serve the load. It also should return the SC value to cut off computation time. The battery starts from 100% charge state, which

means the available energy in the battery is its capacity. The battery has a maximum DOD to increase the lifetime, so the minimal charge state of the battery cannot be lower than this. The maximum hourly charge or discharge is represented by the charge rate. These values are offset by the charging and discharging efficiencies, and for supplying the load the current from the battery must go through an inverter to convert it to AC, so the inverter efficiency will lower the discharged energy. Multiple battery sizes to each PV size should be used to have a broader training data for the regression and the trends could be better visualized. To be able to compare the different PV plus battery sizes, relative battery sizes should be used similarly to [14] [15]. The relative battery sizes are between 0.0 kWh/kW and 2.0 kWh/kW, and it is calculated by dividing the capacity of the battery with the PV size. The battery model uses the same PV generation and building load dataset as for the only PV system.

To decide whether the battery should be charged or discharged, first the hourly load and PV generation should be compared. If the PV generation is higher than the load in the given hour and the battery state of charge is less than 100%, then the battery can be charged. The upper limit for the charged amount is the charge rate and the lower limit is the missing capacity if its lower than the charge rate. If the surplus PV generation is higher than the maximum chargeable amount, then the remainder energy will be sold to the grid. In these cases, the hourly building load is added to the hourly SC sum. If the PV generation is smaller than the load in the given hour and the battery state of charge is above 100-DOD level, then the battery can be discharged. The upper limit for the maximum dischargeable capacity is the charge rate or the capacity that would lead to the minimal state of charge. In this case the hourly PV generation and the supplied energy from the battery multiplied by the discharge and inverter efficiencies are added to the hourly SC values. The annual SC values can be calculated by summing up the hourly SC values and dividing it by the annual PV generation. After calculating the SC for all the battery sizes, the resulting dataset will have 1,134,000 samples for the single-family houses and 485,100 samples for the multifamily houses.

The study by Han et.al. [13] includes prices for a wide range of batteries from different countries. Considering this is a Li-ion battery with similar prices as the battery used by Campana et.al. [16], the parameters of this were used in this battery model. The parameters are listed in Table 1. The charging and discharging efficiencies are considered to be the same. The charge rate is calculated by dividing the battery capacity with the max charge-discharge power and the inverter efficiency is the same as in the PV model.

Table 1: The battery parameters

Parameter	**Value**
Max battery capacity [kWh]	210
DOD [%]	80
Charging efficiency [%]	94
Discharging efficiency [%]	94
Max charge-discharge power [kW]	50
Max charge-discharge rate [h]	4
Inverter efficiency [%]	98

4.3. Performance evaluation

The performance of the models assessed by the same metrics used by Galli [5], so the models can be truly compared. These two metrics are the mean absolute error (MAE) and the R^2 (r-squared). The MAE can be calculated by equation 1, where n is the number of samples, $\hat{y}$ is the predicted values and y is the measured ones. The R^2 can be calculated by equation 2, where $\bar{y}$ is the mean of the measured values. The R^2 can indicate the goodness of the fit of a regression model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{2}$$

For further analysis the adjusted R^2 and the MBE is used. If the difference between the adjusted R^2 and the R^2 is relatively large, then the model is likely overfitting. If the MBE is relatively high, then the model is biased and it is underfitting. The adjusted R^2 is calculated by equation 3, where N is the number of samples and p is the number of variables. The MBE is calculated by equation 4.

$$R^2{}_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{3}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i - y_i \tag{4}$$

4.4. Financial model

The financial model is used with the battery model to calculate the profitability of the different systems. The KPI for this is the payback time which is calculated by dividing the investment cost (CAPEX) of the systems with the income from the systems. The profits from the systems are relative to the system which has no PV and no battery installed. The model uses a nonlinear cost function for the PV system published by Sommerfeldt [17], this can be seen in equation 3. The PV_{size} has unit of kW, and the resulting cost is in SEK. The profit of the system can be calculated by multiplying the SC with the PV generation and the electricity price. Another income source is when the surplus energy is sold to the grid this can be calculated by multiplying the 1-SC with the PV generation and the electricity sell back price. After this the payback can be calculated.

$$C_{PV} = \left(8.6 + 15.9 * PV_{size}{}^{-0.55}\right) * PV_{size} * 1000 \tag{3}$$

The electricity price was taken from the Eon website [18] at 2022.04.11 and it is 1.4 SEK/kWh. The sellback price of the electricity is 0.34 SEK/kWh [1]. The subsidies that are being used in Sweden are presented by the IEA (International Energy Agency) in [1]. Currently there is 50% tax credit on the battery investment cost and 15% tax credit on the PV system cost. These two subsidies are used in the model by reducing the investment of the battery and PV by the according value. Only the battery price is in US$, so this is converted to SEK based on exchange rate on Bloomberg [19]. The exchange rate on 2022.04.11 8:18 UTC+2 was 9.4559 US$/SEK, so the battery price becomes 4730 SEK.

5. Results

This chapter will present the results in two sections. The first is for the single-family houses, this building type was used in Galli's work [5] and a comparison is made between the different regression models. The second is for the multifamily houses, this building type has different consumption profile than the previous one. The difference in using measured and TMY climate data is tested, so the significance of using spatially and temporally matched weather data can be showed.

5.1. Single-family houses

This building type as mentioned before was used in Galli's work [5] with regression models trained on only TMY weather data from several locations in Sweden. This section is providing an analysis on the used dataset for training of the ML models, after that a performance comparison between the different models is presented. Lastly the addition of batteries to the PV system is analyzed following with choosing the optimal system based on the payback time.

5.1.1. Descriptive statistics

The purpose of this section is to identify the characteristics of the datasets and to describe their properties through statistical analysis. Also, outliers can be detected in the datasets which can harm the performance of the ML models.

The visual representation of the dataset that is used to train the regression models for single-family houses can be seen in Figure 8. It shows the calculated self-consumption as a function of building load and PV generation. The annual building load is on the x-axis, while the color of the datapoints corresponds to the annual PV generation. The dataset has few samples in the region where the annual building load is greater than 25 MWh. This indicates that the regression models will have a larger error in this region for predicting SC. The datapoints with similar color have similar PV generation, if both the building load is relatively small (below 10 MWh) the datapoints have similar colors and relatively large variance in SC. This means that for similar inputs the dataset has different SC values, so the regression models may have larger error in this region. Similarly, when the building load is higher 10-25 MWh, but the PV generation is relatively low, the variance of the SC is relatively high. This means if the solar fraction (annual PV generation divided by annual building load) of the inputs is relatively small the model will likely have larger errors. And this is also true if the solar fraction is very close to zero, since in the dataset the lowest solar fraction values are around 0.1. The high variance of SC can be caused by buildings with similar annual load have different consumption profiles due to distinct electrical appliances. So, this relatively high variance of SC in certain regions does not imply that the dataset has some outliers. These conclusions should be treated with some consideration due to the resolution of the color bar.

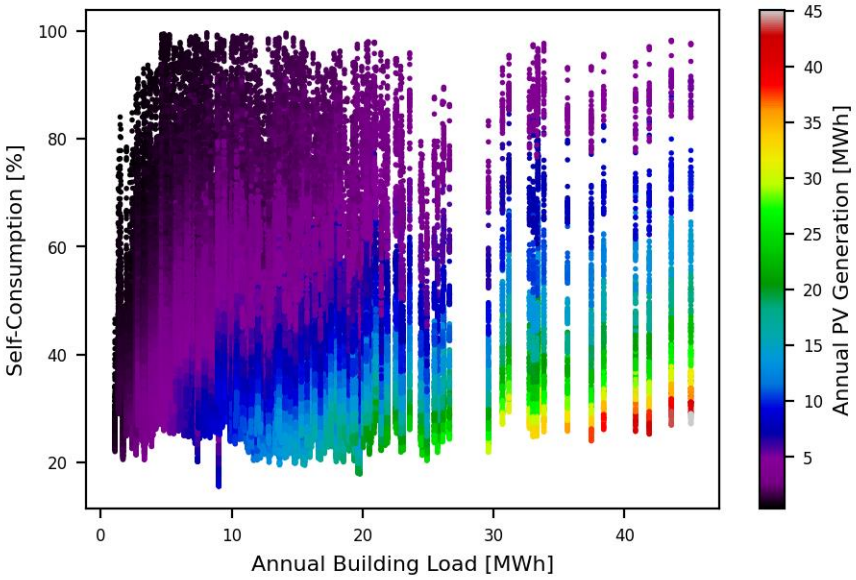

Figure 8: The calculated self-consumption for single-family houses

The relative standard deviation (RSD) of the annual building is showed in Figure 9. This figure depicts the standard deviation of annual building load from 2015 to 2019 divided by the mean of annual building loads for these years for each single-family house. Most of the buildings - around 75% - are below 10% of standard deviation and only four buildings from the 108 are above 30%, with 48% as the highest. This does not mean that those four buildings are outliers and should not be included in the training dataset, since as mentioned before in the 4.1.2 Building load section the buildings are separated for every year. So, this will not affect the performance of the regression models. However, the prediction of the optimal PV system is based on SC and building load as well, so it will be affected. Around 25% of the buildings have a variation in annual building load of above 10%. This much change from year to year would change the prediction for SC significantly and also the prediction for optimal PV system size. So, for this task the latest consumption data should be used and undergoing renovation or purchase of new high consumption electrical devices (e.g. heat pump, electric vehicle) should be considered.

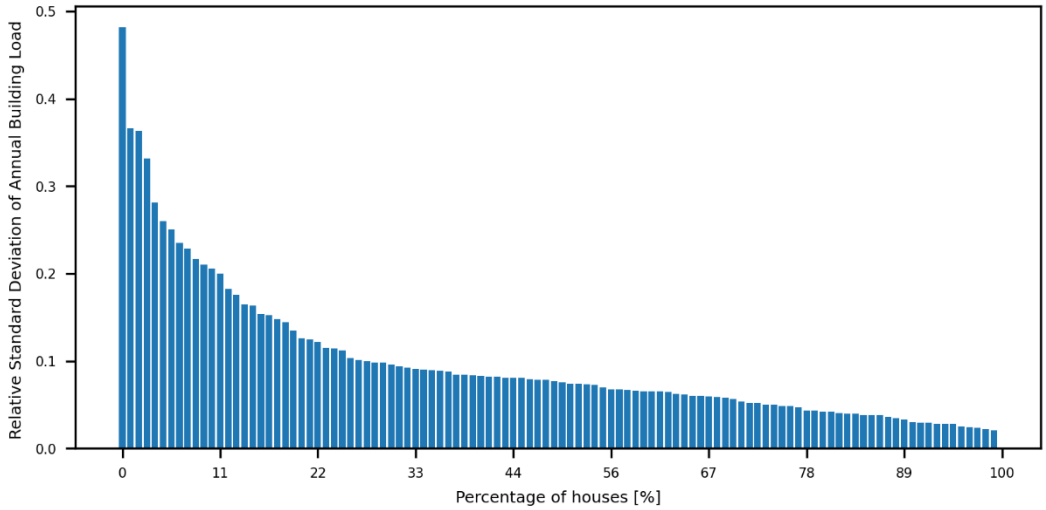

Figure 9: Relative standard deviation of the annual building load for single-family houses

The RSD of the annual SC is showed in Figure 10. The standard deviation for the annual SC is calculated based on the annual SC from 2015 to 2019 for the 108 single-family houses. Then this value is divided by the mean of the annual SC for these five years and depicted on the figure below. The blue color indicates the measured data, and the orange indicates the TMY data. A major part of the dataset (around 95%) is below 10% of RSD, and about 50% of the houses is below 5% of RSD. This indicates that for most of the buildings, the change in annual SC from year to year is relatively small. The RSD for 6 buildings out of 108 is larger than 10% and the highest value is around 28%. The statistics mentioned above are both true for the measured and TMY data. The two dataset has marginal differences in RSD. This could imply that the usage of TMY data will not have a significant impact on the training of the models compared to using measured data. And as mentioned before the relatively high variance in annual building load will lead to relatively high variance in annual SC, so for techno-economic analysis the year of evaluation should be chosen carefully.

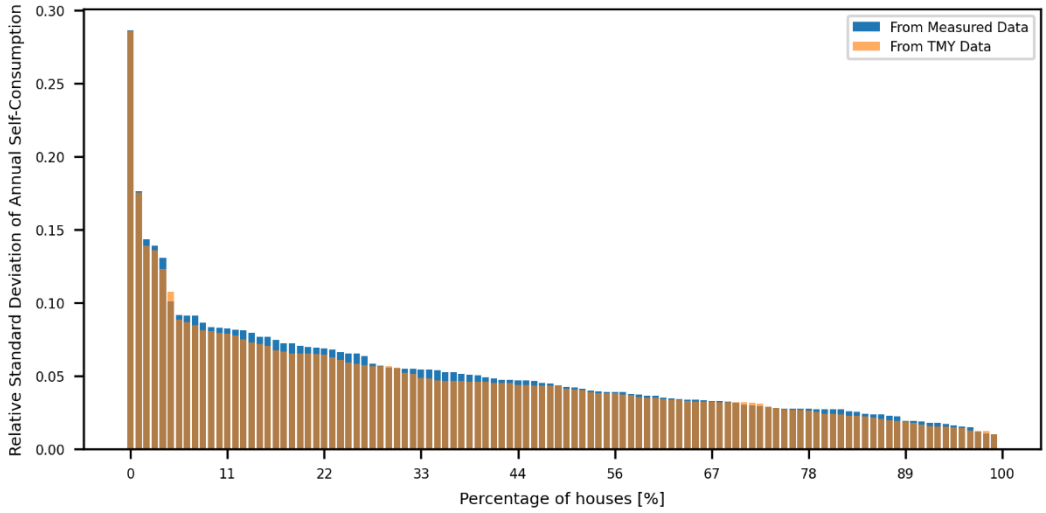

Figure 10: Relative standard deviation of the annual self-consumption for single-family houses

5.1.2. Performance comparison

The comparison is made between three models' performance. The first model is from Galli's study [5] which uses TMY data for five locations from PVGIS-SARAH database [20]. The second model is using measured data from SMHI station at Karlstad airport [9]. Therefore, this model uses a spatially and temporally matched weather data to the building load. To further investigate the impact on using just spatially matched weather data TMY data for Karlstad is used in the third model generated by Meteonorm 8.0 [10]. These models will be refenced as "Galli models", "Measured" and "TMY" further on.

The same regression models and building loads are used as in Galli's study to be able to fully compare their performance. The mean absolute error (MAE) of the models can be seen in Table 2. The table shows the MAE values for the three models and the relative difference between the Measured and Galli models in the 4th column and the relative difference between the TMY and Galli models in the last column. The random forest (Rnd Forest in Table 2) model performed the best in all the models based on the lowest MAE value. The difference between the different models is 0.0% for both the Measured model and the TMY model. For most regression models the diff value is positive for both Measured and TMY model except for Linear and McKenna regression models. So, those models underperform compared to the Galli models. This was expected because the Measured and TMY models were trained on a dataset which has fewer entries than the Galli models' training dataset [21] [22] [23]. The Measured and TMY model uses 226k datapoints for training while the Galli model 1.08M. The largest difference is in the k-NN regression with 20.0% difference, but the MAE value is still relatively small. Comparing the Measured and TMY models, the difference is marginal for all the regressions.

Table 2: MAE of Measured, TMY, and Galli regressions for single-family houses

Model	Gal	Meas	Diff	TMY	Diff
Rnd Forest	0.015	0.015	0.0%	0.015	0.0%
k-NN	0.025	0.030	20.0%	0.030	20.0%
MLP	0.038	0.042	10.5%	0.042	10.5%
Polynomial	0.038	0.042	10.5%	0.041	7.9%
Ridge	0.038	0.042	10.5%	0.042	10.5%
Lasso	0.058	0.058	0.0%	0.058	0.0%
Linear	0.074	0.071	-4.1%	0.072	-2.7%
McKenna	0.073	0.069	-5.5%	0.069	-5.5%

The coefficient of determination (R^2) of the models can be seen in Table 3. This table shows the results similarly as in Table 2. The random forest (Rnd forest in Table 3) regression performed the best from all the models because it has the highest R^2 for all the models. The Lasso, Linear and McKenna regression have the lowest values, all of them are below 0.65. The other models have performed well, every has a value above 0.85. Only the random forest and the k-NN model has R^2 greater than 0.9. These models can be considered as well-trained models. All the regressions underperformed compared to the Galli model's regressions, except for the McKenna model. This was expected as mentioned before for the MAE values, since the Measured and TMY models using a smaller dataset. The biggest difference is in the Lasso regression in the Measured model with a value of -6.4%. The Measured and TMY values have a marginal difference across all of the regressions, similarly to the MAE values. The measured and TMY models' adj. R^2 and MBE is presented in Table A1 in Appendix A. The adj. R^2 and the R^2 values have no or marginal differences, which indicates that the models have no

overfitting issues. The MBE values for the models are acceptably small (less than 0.001), so the regressions have no bias, and they are not underfitting.

Table 3: R^2 of Measured, TMY, and Galli regressions for single-family houses

Model	Gal	Meas	Diff	TMY	Diff
Rnd Forest	0.985	0.977	-0.8%	0.980	-0.5%
k-NN	0.956	0.923	-3.5%	0.928	-2.9%
MLP	0.907	0.875	-3.5%	0.880	-3.0%
Polynomial	0.907	0.875	-3.5%	0.885	-2.4%
Ridge	0.897	0.857	-4.5%	0.868	-3.2%
Lasso	0.670	0.627	-6.4%	0.648	-3.3%
Linear	0.641	0.627	-2.2%	0.640	-0.2%
McKenna	0.646	0.647	0.2%	0.658	1.9%

The results above showed similar values in both MAE and R^2 for the Measured and TMY models, but this does not mean that using the TMY climate data has no impact on the regression models. Since the models above were trained and tested on their respective datasets. So, to test the impact of using only spatially matched weather data to the load data, the TMY model should be tested on the Measured dataset. The obtained values from this evaluation can be seen in Table 4. This table shows the MAE and R^2 values for the regressions that were trained on the TMY dataset and tested on the Measured dataset in the "Abs" columns. The "Diff" column is showing the difference between the "Abs" column values and the MAE and R^2 values for the TMY model presented in Table 2 and in Table 3.

Again, the random forest (Rnd forest in Table 4) has the biggest R^2 and the smallest MAE, indicating that this model performed the best. The Linear regression has the lowest R^2 and the biggest MAE, similar to the previous evaluations. All the MAE values have a positive difference meaning that they are bigger than the model that used TMY for training and testing. The largest difference is between the random forest models with 40%, but the MAE value is still relatively small. The reason for the large differences is unknown. The smallest difference was between the Lasso regressions with 1.7%, the reason for this is unknown as well. The other models have around 5% difference except for the k-NN regression which has 6.7%.

Similarly, to the MAE values, almost all the regression models performed worse. Only the Lasso regression performed better with 15.1% increase in $R^{2,}$ and the Ridge regression showed no change with 0.0% difference. The better performance of the Lasso regression is unexpected and the reason for this is unknown. The highest decrease in R^2 can be seen with the Linear model with -3.0% change. The worse performance in MAE and R^2 for almost all of the ML models suggest that using only spatially matched weather data to the load can have an impact on the regression models performance. And the usage of measured data which matches spatially and temporally to the building load may have a better performance.

Table 4: TMY trained models on Measured data for single-family houses

	MAE		R^2	
Model	**Abs**	**Diff**	**Abs**	**Diff**
Rnd Forest	0.021	40.0%	0.967	-1.3%
k-NN	0.032	6.7%	0.923	-0.5%
MLP	0.044	4.8%	0.867	-1.5%
Polynomial	0.044	4.8%	0.869	-1.8%
Ridge	0.044	4.8%	0.868	0.0%
Lasso	0.059	1.7%	0.746	15.1%
Linear	0.074	4.2%	0.621	-3.0%
McKenna	0.072	4.3%	0.641	-2.6%

5.1.3. Battery model

The battery model's regressions are trained on two different datasets. One is the measured and the other is the TMY. The performance of both is evaluated and can be compared. With the comparison the difference in performance of the two methods can be measured.

The MAE of the two methods can be seen in Table 5. The best performing model for both method is the random forest regression. The second-best performing model has double the MAE as the random forest. The worst performing models are the Linear and McKenna regression. There is no difference between the two methods for most of the regressions except for the random forest and k-NN regressions. But these differences are relatively small as well. The random forest has the highest with 5.6% difference.

Table 5: MAE of Measured, TMY regressions for single-family houses battery model

Model	**Meas**	**TMY**	**Diff**
Rnd Forest	0.017	0.018	-5.6%
k-NN	0.035	0.036	-2.8%
MLP	0.046	0.046	0.0%
Polynomial	0.046	0.046	0.0%
Ridge	0.046	0.046	0.0%
Lasso	0.073	0.073	0.0%
Linear	0.093	0.093	0.0%
McKenna	0.092	0.092	0.0%

The R^2 of the two methods can be seen in Table 6. The random forest regression has the highest score, and the Linear regression has the lowest. The random forest, k-NN, MLP, Polynomial and Ridge models have relatively high R^2 values with above 0.89. The difference between the two methods is below 1% for all the models. The adj. R^2 and the MBE values for these models are in Table A2 in Appendix A. The adj. R^2 and the R^2 values have marginal differences thus the models are not overfitting. The MBE values are close to zero (less than 0.001), so the models are not biased.

Table 6: R^2 of Measured, TMY regressions for single-family houses battery model

Model	Meas	TMY	Diff
Rnd Forest	0.983	0.982	0.1%
k-NN	0.939	0.937	0.2%
MLP	0.910	0.907	0.3%
Polynomial	0.909	0.904	0.6%
Ridge	0.899	0.894	0.6%
Lasso	0.663	0.663	0.0%
Linear	0.632	0.630	0.3%
McKenna	0.639	0.636	0.5%

To truly test if only using spatially matched weather data to the building load has impact on the regressions' performance, the measured data is tested on the regression models that are trained on the TMY dataset. The results from this evaluation can be seen in Table 7. The difference in MAE for all of the models is positive and relatively small the largest is 5.6% for the k-NN model. The positive value means that it is underperformed. The difference in R^2 values is below 1% except for the Lasso regression where it is 16.4%. The reason for this large difference is unknown. Every model underperformed except for the Lasso and Ridge regressions.

Table 7: TMY trained models on Measured data for single-family houses battery model

	MAE		R^2	
Model	**Abs**	**Diff**	**Abs**	**Diff**
Rnd Forest	0.018	0.0%	0.974	-0.8%
k-NN	0.038	5.6%	0.935	-0.2%
MLP	0.046	0.0%	0.906	-0.1%
Polynomial	0.047	2.2%	0.903	-0.1%
Ridge	0.047	2.2%	0.903	1.0%
Lasso	0.073	0.0%	0.772	16.4%
Linear	0.095	2.2%	0.627	-0.5%
McKenna	0.094	2.2%	0.634	-0.3%

5.1.4. Optimal system

The polynomial and the random forest regressions will be evaluated for the battery model. The polynomial model is relatively simplistic compared to the random forest with relatively high level of accuracy and it has been used by Karlstad Energi for commercial purposes, and the random forest model had the best performance from the regressions. A building is chosen from the dataset, and SC values are predicted based on the annual building load and annual PV generation. The annual PV generation is chosen based on the solar fraction of the system. It is ranging from 0.1 to 1.0 solar fraction with an increment of 0.1. This will give 10 SC values. The chosen building has the calculated SC values in the dataset and the models' results can be compared to these. The tilt angle of PV system and the azimuth angle is chosen to be 30˚ and 180˚ accordingly.

The building with 15.1 MWh annual building load is used for the evaluation. The results in Figure 8 suggested that between 10 MWh and 20 MWh annual building load has sufficient amount of data compared to annual building loads of 25 MWh or higher where very few entries are in the dataset. And if the annual building load less than 10 MWh, the datapoints seemed to have larger variance in SC. The reference dataset is shown in Figure 11. The right figure shows the SC for different relative battery sizes and solar fraction values for the building with 15.1 MWh annual building load. The relative battery size is the battery capacity (kWh) divided by

the PV size (kW). The left figure shows the calculated payback time based on the SC, PV size, annual PV generation and battery size. The plotted data contains 10 solar fraction and 5 relative battery sizes resulting in 50 datapoints, the colors of these have been interpolated with matplotlib bessel interpolation [24] to visualize the results better.

The SC values for a given solar fraction tends to increase with the increase of relative battery size. It is expected since larger battery capacity can store more surplus PV generation which can be used in the hours of no insolation. For a given relative battery size, the SC tends to increase with decreasing solar fraction values. This is expected as well because the lower solar fraction means lower PV generation thus the surplus energy will be lower, and the self-consumed part will be higher. The lowest SC is at 1.0 solar fraction and 0.0 kWh/kW relative battery size. The highest SC is at 0.1 solar fraction and 2.0 kWh/kW relative battery size.

The payback time has similar trends as the SC. It increases below 0.6 solar fraction with increasing relative battery sizes. However, above 0.6 solar fraction it has a local minimum around 0.5 kWh/kW relative battery size. This is more visible above 0.9 solar fraction. For a given relative battery size the payback time has a local minimum around 0.2-0.4 solar fraction for all relative battery sizes. The optimal system based on the reference datapoints has 13.7 years payback time and 0.839 SC with 2.71 kW PV size (0.16 solar fraction) and 0.0 kWh battery size. This indicates that with subsidies on the battery, it still has a relatively high investment cost compared to the additional income that it produces.

The polynomial model results are plotted in Figure 12. The polynomial SC values have similar trends as described previously regarding the SC values from the reference dataset. The scales of the two figures' color bar are the same for both the payback and SC, so they can be compared. It seems that the SC values are being underpredicted because at low solar fraction values (below 0.2) the datapoints have darker red color compared to orange color in Figure 12. The values in this region are close to 1.0 SC in the reference datapoints but in the polynomial regression datapoints this region's values are around 0.8 SC. And around 0.6 solar fraction and 1.5 kWh/kW – 2.0 kWh/kW relative battery size the reference values are around 0.7 SC while the polynomial values are around 0.5-0.6 SC.

This underprediction of SC results in a higher payback time in most cases. This can be seen by the reference plot's dominant color is dark and light blue in most regions and light green above 0.6 solar fraction. On the polynomial plot, the dominant color is mainly green. The highest differences seem to be in the region where the solar fraction is greater than 0.6 and the relative battery size is greater than 1.0 kWh/kW. A similar trend can be seen on reference plot as on the polynomial, where above 0.6 solar fraction the payback time has a local minimum between 0.5-1.0 kWh/kW relative battery size. Below 0.6 solar fraction the payback has monotonic increase with increasing relative battery sizes. The optimal system has a payback time of 14.9 years and 0.682 SC with 3.36 kW PV size (0.2 solar fraction) and 0.0 kWh battery capacity. The worst system has 20.5 years of payback time with 16.8 kW PV size and 33.6 kWh battery capacity which is the largest in both PV and battery size.

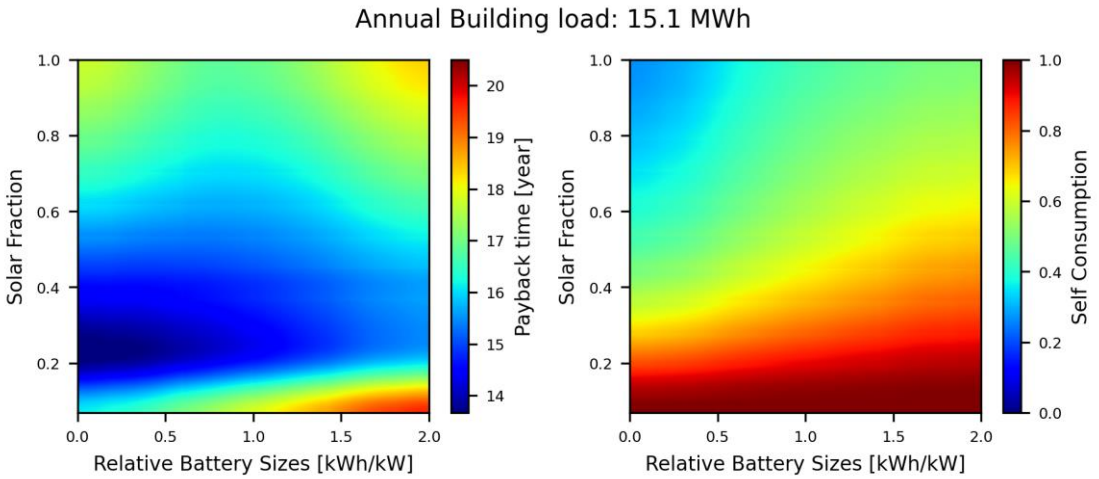

Figure 11: The payback time (left figure) from the reference and the SC (right figure) for single-family houses

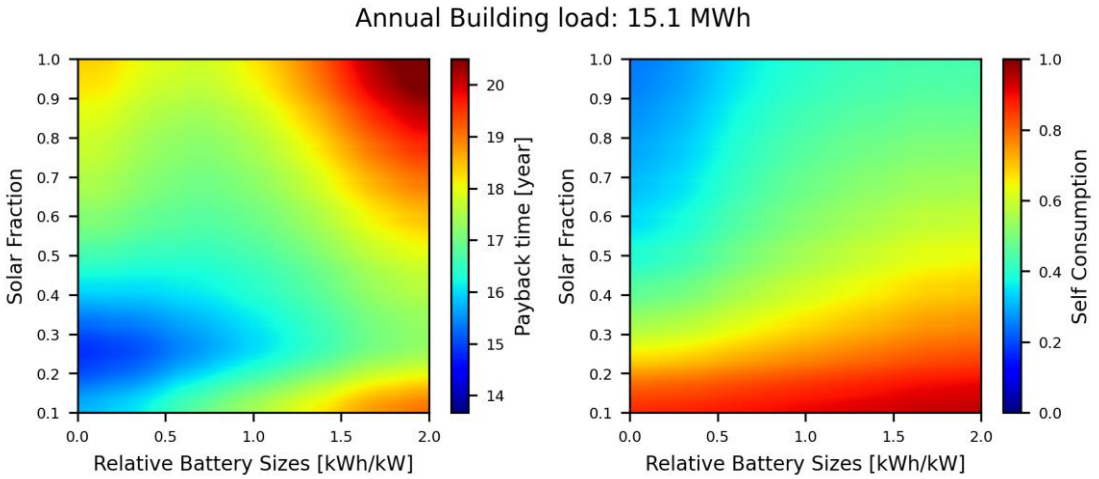

Figure 12: The payback time (left figure) from the polynomial regression and the SC (right figure) for single-family houses

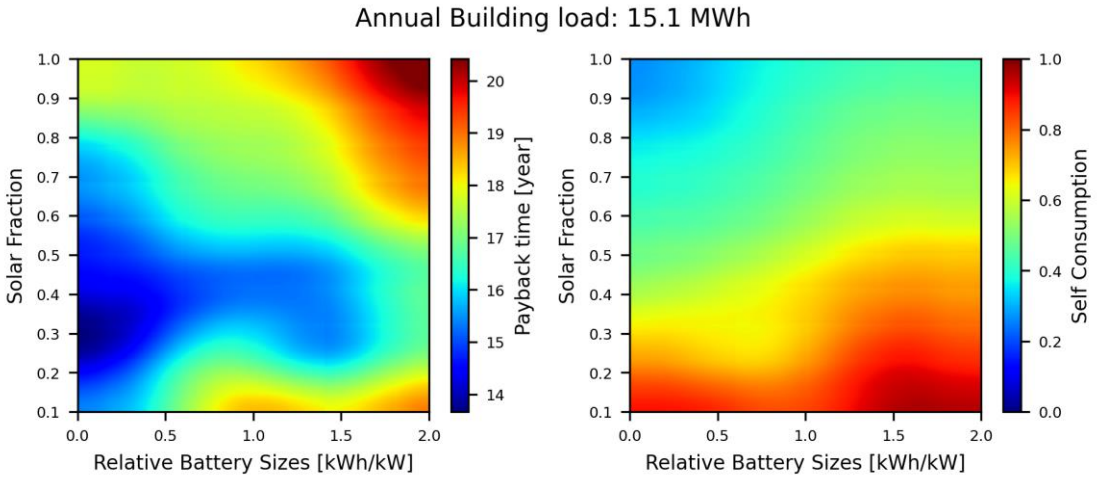

Figure 13: The payback time (left figure) from the random forest regression and the SC (right figure) for single-family houses

The results from the random forest model can be seen in Figure 13. The color bar has the same range as in Figure 12 and in Figure 11. The SC values are closer to the reference above 0.3 solar fraction than the values from the polynomial model. However, the SC has trends around 0.3 and below, that are not shown in the reference. Such as the SC has a local minimum around 0.5-1.0 kWh/kW relative battery size and a local maximum around 1.5 kWh/kW relative battery size for a given solar fraction. While the SC on the reference has a monotonic increase with increasing battery sizes. These new trends can indicate that the random forest model is too complex, and it is overfitting but the similar adj. R^2 and R^2 values suggest that the model is not overfitting. This should be further investigated in the future.

The payback time also seems to be closer to the reference than the polynomial values because the blue color became more dominant on the plot. The new trends on the SC values propagated to the payback time plot and created new trends below 0.3 solar fraction. The payback has a local minimum in this region which was not observed in the reference plot. The local maximum above 0.6 solar fraction disappeared and the payback time has a monotonic increase in this region. This also could be an indication of the overfitting of the model. The optimal system from this model has 13.8 years payback time and 0.658 SC with 5.04 kW PV size (0.3 solar fraction) and 0.0 kWh battery size.

The optimal system for each model is presented in Table 8. The difference in payback time for the polynomial model compared to the reference is 1.2 years around 8.8%. The random forest model has marginal difference compared to the reference. However, the SC value for the random forest is 21.6% smaller and the optimal PV size is 86% higher. This causes the payback times to have almost the same value. For the polynomial model the SC is 18.7% smaller and the PV size is 24% higher. All the models have 0.0 kWh battery size as optimal. Based on these results and on the similar trends that the polynomial plots showed compared to the reference, the polynomial model seems to be a better regression to predict the SC and the payback time.

Table 8: Optimal systems based on payback time for the single-family houses

	Payback time	**SC**	**PV size [kW]**	**Battery size [kWh]**
Reference	13.7	0.839	2.71	0.0
Polynomial	14.9	0.682	3.36	0.0
Random forest	13.8	0.658	5.04	0.0

The previous table showed that the optimal system PV size can differ significantly from the reference value with each regression, so to directly compare the performance of them the reference data's optimal solar fraction should be chosen, and the models should be evaluated at that point. The results at 0.16 solar fraction from the models is presented in Table 9. The prediction from the random forest is 10.6% lower than the reference and the polynomial model underpredicted it by 11.8%. According to these results the random forest is better at predicting the SC than the polynomial.

Table 9: Values from the models at 0.16 solar fraction for single-family houses

	Payback time	**SC**	**PV size [kW]**	**Battery size [kWh]**
Reference	13.7	0.839	2.71	0.0
Polynomial	15.0	0.740	2.71	0.0
Random forest	14.9	0.750	2.71	0.0

5.2. Multifamily houses

This section follows a similar way for presentation of the results as in the Single-family houses section. First the statistics of the datasets is showed, and the main characteristics are discussed, then in the second subsection the performance of measured and TMY weather data is compared. Lastly results from the battery model are presented.

5.2.1. Descriptive statistics

The multifamily houses dataset has 77 building and three years of data for them from 2019 to 2021, since the buildings in each year are separated then it results in 231 building loads. The calculated SC values for this dataset are presented in Figure 14. It shows the calculated self-consumption in a function of building load and PV generation just as in Figure 8 for single-family houses. The plot shows that the dataset has few entries in the region where the annual building load is greater than 200 MWh. The regression models trained on this dataset will likely have the largest error in this region. The lower solar fraction regions seem to have a larger variance in SC compared to higher solar fraction regions, similarly as in Figure 8. The highest variance in SC occurs in low annual PV generation and low annual building load datapoints. But the resolution of the colors can be misleading.

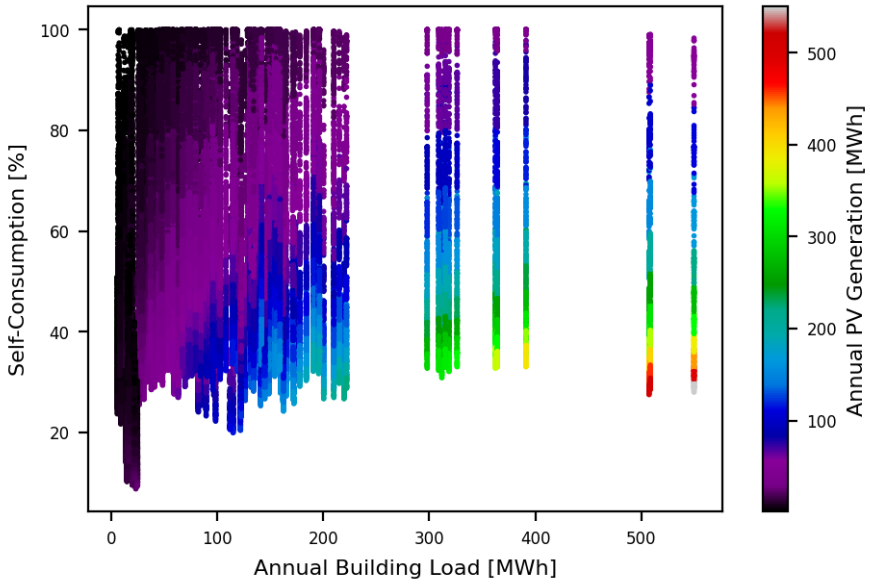

Figure 14: The calculated self-consumption for multifamily houses

The RSD of the annual building load for the multifamily houses is shown in Figure 15. The relative standard deviation is calculated the same way as for single-family houses. Most of the houses have smaller than 10% RSD. Three houses have relatively higher RSD compared to the others, one is around 50%, another is around 100% and the highest is around 120%. This means that from year to year some building annual load can change significantly. This does not mean that these three buildings are outliers and should be removed. Since the buildings are separated for each year. This will not influence the performance of the regression models. But it indicates that the evaluation of choosing the optimal PV system should be using the latest annual building load data.

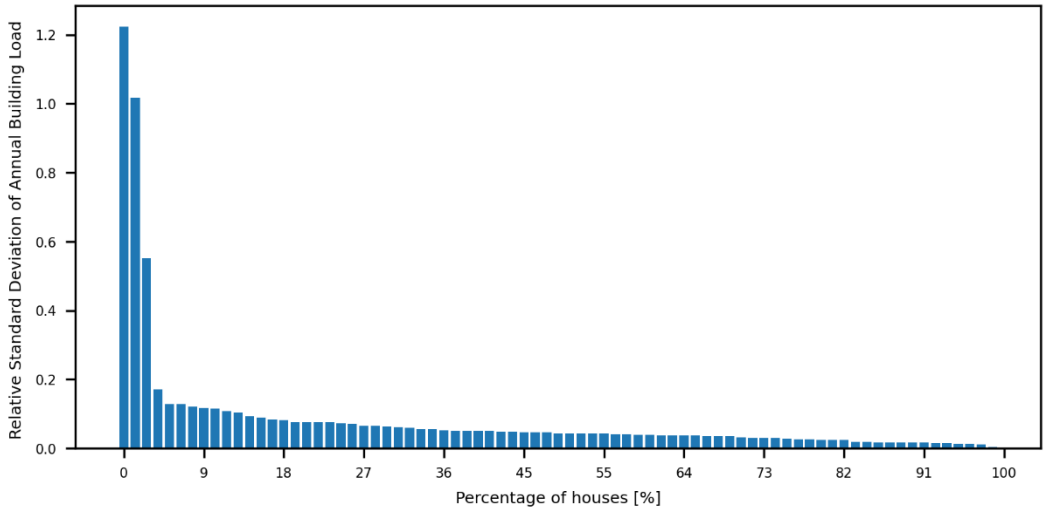

Figure 15: Relative standard deviation of the annual building load for multifamily houses

The RSD of the SC for multifamily houses is shown in Figure 16. Based on the measured data (shown in blue color), around 90% of the multifamily houses has 5% or lower RSD which is much higher than for single-family houses which was 50%. Almost all of the houses have around 12% or lower RSD with one exception that has around 26%. The TMY data (shown in orange) has very similar statistics and the difference between the two is only few percentage points at most. This indicates that the regression models will have similar performance independently from the used training data. For both data the RSD is marginal for 50% of the houses.

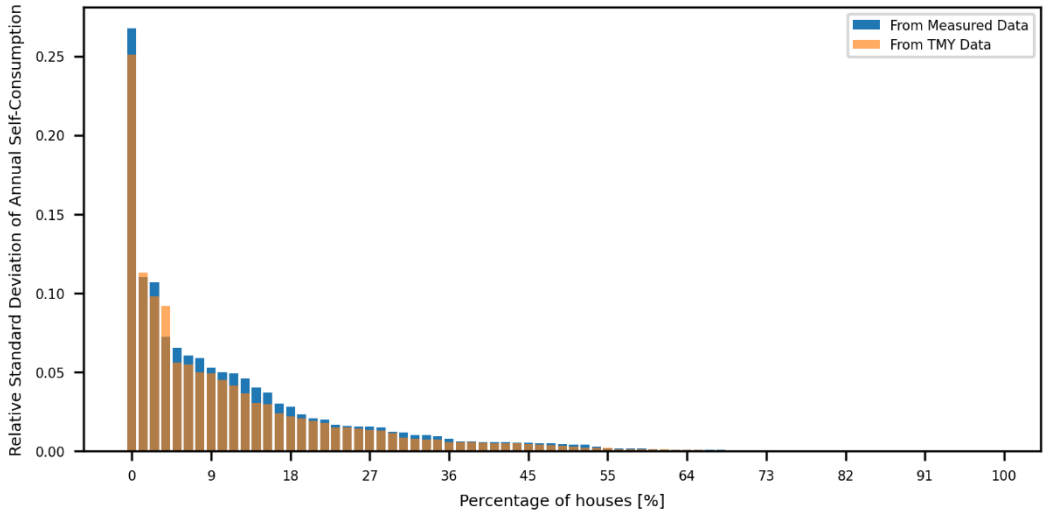

Figure 16: Relative standard deviation of the annual self-consumption for multifamily houses

5.2.2. Performance comparison

The comparison is made between two models the first is using the spatially and temporally matched weather data to the building load, further referenced as “measured”. The other model is using only spatially matched weather data to the building load, further referenced as TMY. The TMY and measured data is generated the same way as for the single-family houses.

The same regressions are used and the same metrics as in the performance evaluation for the single-family houses' regressions. The MAE values for the two model and the relative difference between them can be seen in Table 10. The random forest (Rnd Forest in Table 10) has the smallest MAE for both models and the Linear model has the highest MAE for both models. The difference between the models' MAE is marginal for all the regressions. The highest is 3.0% difference for the MLP model. All of the other has 0.0% difference except for the Lasso and Ridge models. This result was expected based on the low difference in RSD of SC as shown in Figure 16. Comparing these results to the results from the single-family house in Table 2, it can be seen that all the regression has higher MAE in the case of multifamily houses. In the case of k-NN regression is more than double and for all the other regressions it increased significantly except for the random forest.

Table 10: MAE of Measured, TMY regressions for multifamily houses

Model	**Meas**	**TMY**	**Diff**
Rnd Forest	0.018	0.018	0.0%
k-NN	0.066	0.066	0.0%
MLP	0.065	0.067	-3.0%
Polynomial	0.073	0.073	0.0%
Ridge	0.068	0.069	-1.4%
Lasso	0.082	0.083	-1.2%
Linear	0.125	0.125	0.0%
McKenna	0.119	0.119	0.0%

The R^2 values for the two models are presented in Table 11. The best performing model again is the random forest (Rnd forest in Table 11) for both the measured and TMY models. The worst performing is the Linear regression in both cases. The regressions have marginal differences for both models just as for the MAE above. The largest difference is below 1.0%. Comparing to the results in Table 3, the regressions from the single-family houses overperformed in almost every case expect the random forest which showed small improvements. This highlights the robustness of this regression. The highest difference was in the Linear regression with around 30% decrease for both models. Surprisingly the k-NN model had a 15% decrease in performance but according to Shokrzade et al. [25] the performance of the k-NN model can be jeopardized by big datasets that are noisier compared to smaller ones. And in this case the multifamily house dataset is around half of the single-family house data. For the other models this result shows that the sample size of the training dataset can have significant impact on the performance of the regressions. Both measured and TMY models have marginal differences in adj. R^2 and R^2 as it can be seen in Table B1 in Appendix B. The MBE values are relatively small as well (less than 0.0013).

Table 11: R^2 of Measured, TMY regressions for multifamily houses

Model	**Meas**	**TMY**	**Diff**
Rnd Forest	0.985	0.983	0.2%
k-NN	0.789	0.788	0.1%
MLP	0.828	0.821	0.9%
Polynomial	0.673	0.674	-0.1%
Ridge	0.749	0.748	0.1%
Lasso	0.578	0.575	0.5%
Linear	0.424	0.425	-0.2%
McKenna	0.474	0.474	0.0%

The performance of the regressions trained on the TMY dataset and tested on measured data is presented in Table 12. This test can show if using only spatially matched climate data to the

load data can affect the performance of the models for multifamily houses. The obtained results are unexpected since for almost every regression the R^2 values improved compared to results shown in Table 11. Only the Linear and McKenna regressions' R^2 has decreased. The MAE values are similar, other than the random forest, Linear and McKenna regressions all the models have been improved. Some changes are significant, for example the k-NN model's MAE decreased with 60.6% and the R^2 increased with 22.7%. The reason for this is unknown.

Table 12: TMY trained models on Measured data for multifamily houses

	MAE		R^2	
Model	**Abs**	**Diff**	**Abs**	**Diff**
Rnd Forest	0.021	16.7%	0.984	0.1%
k-NN	0.026	-60.6%	0.967	22.7%
MLP	0.065	-3.0%	0.821	0.0%
Polynomial	0.067	-8.2%	0.804	19.3%
Ridge	0.068	-1.4%	0.800	7.0%
Lasso	0.082	-1.2%	0.720	25.2%
Linear	0.127	1.6%	0.423	-0.5%
McKenna	0.121	1.7%	0.472	-0.4%

5.2.3. Battery model

Two methods are tested here as for the single-family house case. Both for the measured and TMY dataset the regressions are trained and evaluated. The performances are compared and to see if using TMY weather data has any impact on the performance the TMY trained models are tested on the measured dataset.

The MAE of two methods is similar, the largest difference is in the random forest model with 13.3% as it can be seen in Table 13. The values for most of the model is relatively small, only the Linear and McKenna models are above 0.1. All the models' MAE has increased compared to the single-family house case except for the k-NN and random forest regressions.

Table 13: MAE of Measured, TMY regressions for multifamily houses battery model

Model	**Meas**	**TMY**	**Diff**
Rnd Forest	0.013	0.015	-13.3%
k-NN	0.030	0.031	-3.2%
MLP	0.060	0.059	1.7%
Polynomial	0.065	0.066	-1.5%
Ridge	0.065	0.066	-1.5%
Lasso	0.087	0.086	1.2%
Linear	0.127	0.124	2.4%
McKenna	0.121	0.120	0.8%

The R^2 values for the models are listed in Table 14. Only the random forest and the k-NN models are above 0.9, the rest is below 0.85. The Linear and McKenna regressions have a relatively high performance drop compared to the single-family house case. Their values are below 0.5 whereas in the single-family case it was above 0.6. For all the models the difference between the two methods is below 4.0% and for the random forest, k-NN, MLP and Polynomial it is below 1.0%. The adj. R^2 and MBE values for the models can be seen in Table B2 in Appendix B. These values suggest that the models are not overfitting and not biased (less than 0.00011).

Table 14: R^2 of Measured, TMY regressions for multifamily houses battery model

Model	Meas	TMY	Diff
Rnd Forest	0.991	0.987	0.4%
k-NN	0.951	0.947	0.4%
MLP	0.847	0.843	0.5%
Polynomial	0.811	0.803	1.0%
Ridge	0.763	0.750	1.7%
Lasso	0.547	0.528	3.6%
Linear	0.432	0.427	1.2%
McKenna	0.474	0.459	3.3%

The performance of the TMY trained models tested on the measured data is presented in Table 15. The random forest and Lasso regression show no change in MAE. The k-NN, Polynomial and Ridge models has better performance with decreased MAE and the rest has underperformed. The reason for this is unknown. The largest difference is in the MLP model with 10.2%. All the models' R^2 value are increased except for the MLP. The Lasso model has the largest increase with 33.1%.

Table 15: TMY trained models on Measured data for multifamily houses battery model

	MAE		R^2	
Model	**Abs**	**Diff**	**Abs**	**Diff**
Rnd Forest	0.015	0.0%	0.988	0.1%
k-NN	0.030	-3.2%	0.955	0.8%
MLP	0.065	10.2%	0.815	-3.3%
Polynomial	0.064	-3.0%	0.809	0.7%
Ridge	0.065	-1.5%	0.808	7.7%
Lasso	0.086	0.0%	0.703	33.1%
Linear	0.127	2.4%	0.430	0.7%
McKenna	0.123	2.5%	0.463	0.9%

5.2.4. Optimal system

The polynomial and the random forest models will be evaluated here for similar reasons as for the single-family houses case. A reference dataset is going to be used to be able to compare the performance of the regressions. A building is chosen from the dataset and the SC is predicted in the same as in the single-family house case.

The chosen building has 150.9 MWh annual building load. The selection was based on the trends that can be seen in Figure 14. Firstly above 200 MWh annual building load the dataset has fewer samples so the prediction will likely have lower performance, than below 100 MWh annual building load the variance of the SC is relatively high. The results for the three models will be plotted on similar figures as for the single-family houses. The relative battery sizes will be on the x-axis while the solar fraction on the y-axis and the color of datapoint will indicate the payback time and the SC.

The plots for the reference dataset can be seen in Figure 17. Both the SC and payback time has monotonic increase for any given solar fraction with the increase of relative battery sizes. This is different from single-family houses because there was a local minimum in the payback time if the solar fraction was above 0.6. The optimal system is in the relatively low solar fraction and smallest relative battery size region, similarly to the single-family houses. The optimal system based on the datapoints from the reference has 7.8 years payback time and 0.923 SC

with 34.47 kW PV size (0.21 solar fraction) and 0.0 kWh battery size.

The results for the polynomial model presented in Figure 18 and its color bars have the same ranges as in Figure 17. The polynomial model seems to underpredict the SC values compared to the reference in all datapoints, similarly to the single-family houses. The payback is similar to the reference, except the lower solar fraction (below 0.2) regions. Here the reference has higher payback times and the global maximum as well, while the polynomial model has its global maximum at 1.0 solar fraction and 2.0 kWh/kW relative battery size. The optimal system for the polynomial model has 8.5 years payback time and 0.819 SC with 33.53 kW PV size (0.2 solar fraction) and 0.0 kWh battery capacity.

The results from the random forest model are shown in Figure 19 and the color bars have the same ranges as in Figure 18 and in Figure 17. The SC plot shows a clear trend of a local minimum in SC for any given solar fraction. This is not visible on the reference plot, and it can be caused by the overfitting of the random forest model. The payback times seem to match the reference values well in certain regions, except when the solar fraction is below 0.2. The optimal system for the random forest has 8.1 years payback time and 0.997 SC with 16.76 kW PV size (0.1 solar fraction) and 0.0 kWh battery capacity.

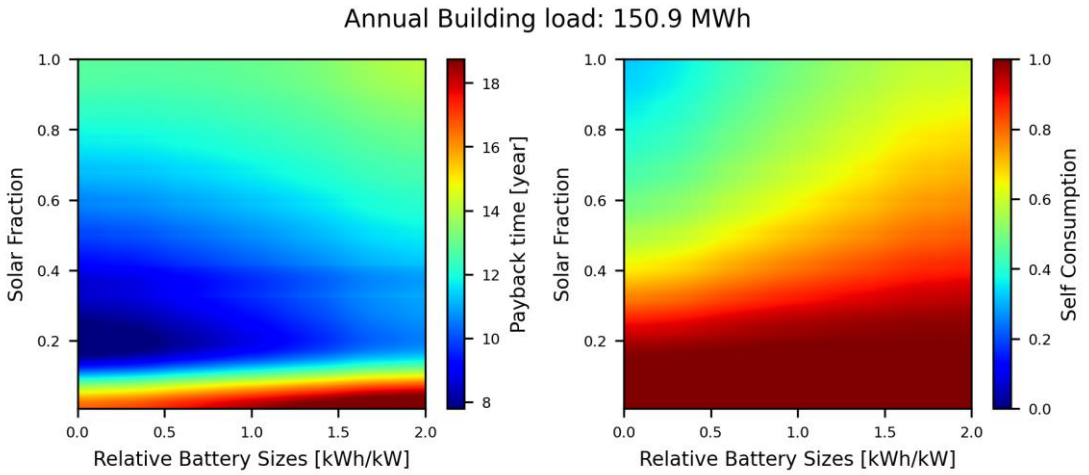

Figure 17: The payback time (left figure) from the reference model and the SC (right figure) for multifamily houses

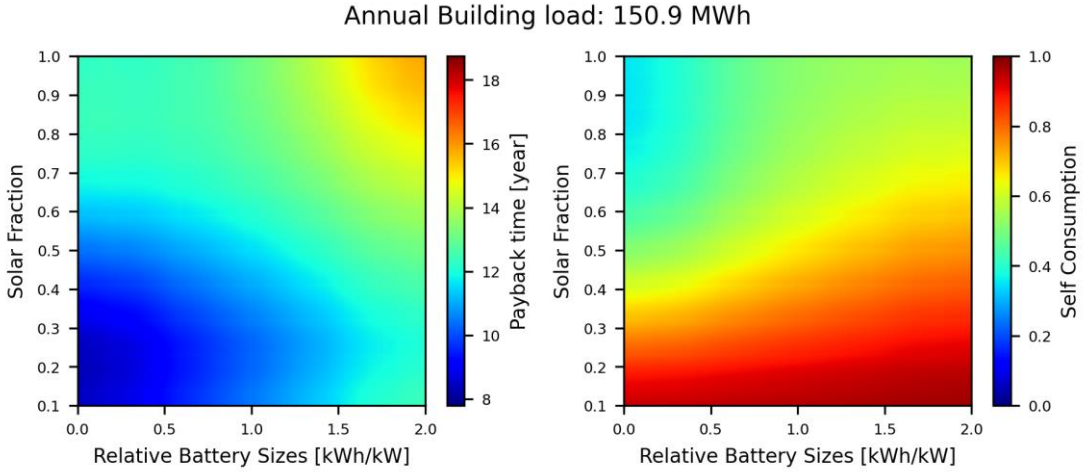

Figure 18: The payback time (left figure) from the polynomial regression and the SC (right figure) for multifamily houses

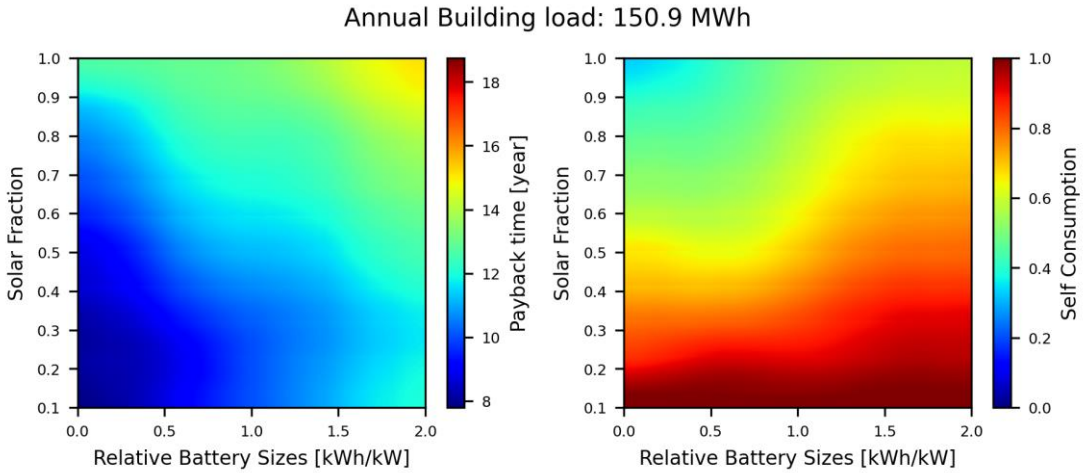

Figure 19: The payback time (left figure) from the random forest regression and the SC (right figure) for multifamily houses

The optimal systems' parameters are presented in Table 16. The polynomial model overpredicted the payback time with 0.7 year (around 9% difference). The random forest overpredicted as well with 0.3 year (around 3.8% difference). The SC value was underpredicted by the polynomial model by around 11.2% and overpredicted by the random forest model by around 8%. The biggest difference is in optimal PV size, the polynomial model underpredicted it by around 2.7%, while the random forest underpredicted it by 51.4%. This shows the new trends that the random forest model showed in the SC, propagated to the payback time and the PV size of the optimal system was highly affected. Based on these results, the better prediction model for the optimal system is the polynomial regression.

Table 16: Optimal systems based on payback time for the multifamily houses

	Payback time	**SC**	**PV size [kW]**	**Battery size [kWh]**
Reference	7.8	0.923	34.47	0.0
Polynomial	8.5	0.819	33.53	0.0
Random forest	8.1	0.997	16.76	0.0

The variation in predicted optimal PV size is high for the random forest regression which should have higher accuracy compared to the polynomial regression based on R^2 values. To directly compare the models' performance, the reference data's optimal solar fraction should be chosen. The prediction of the models at a 0.21 solar fraction is presented in Table 17. The polynomial underpredicted the reference SC value by 11.9% while the random forest predicted almost the same SC value as the reference with only 0.7% difference. For the polynomial model this led to a 0.7 years higher payback. The payback time for the random forest regression is the same as for the reference due to the marginal difference in SC. According to these results the random forest model has better performance.

Table 17: Values from the models at 0.21 solar fraction for multifamily houses

	Payback time	**SC**	**PV size [kW]**	**Battery size [kWh]**
Reference	7.8	0.923	34.47	0.0
Polynomial	8.5	0.813	34.47	0.0
Random forest	7.8	0.917	34.47	0.0

6. Discussion

The descriptive statistics for both the single-family house and multifamily house dataset showed that some regions contain only few samples, and this could influence the prediction when it is used for building loads in such regions. This could be improved by adding more buildings into the dataset which annual building loads are in this region.

The models from Galli's work had better performance compared to both TMY and Measured models for almost all regressions. The regressions in Galli's work used a training dataset that is almost five times larger than the training dataset of the TMY and Measured models. Multiple articles suggest that the size of the training dataset can have an impact on the performance of the regressions and the performance is better with larger sample size [21] [22] [23]. This is also visible when the results from the single-family house and multifamily house case is compared. The mean absolute error for all regressions significantly increased except for random forest. This increase is more than double for the k-NN model. The R^2 values lowered for almost all the models by a significant amount, the linear lowered by 30% and the k-NN by 15%. The explanation for this is the same that is mentioned before because the multifamily house dataset size is around half of the single-family house dataset.

The random forest regression showed the best accuracy in all of the performance evaluations. This model is the most complex and that makes it hard to use it in commercial cases. The parameters of this regression can be saved in a file which can be shared to use it on any computer. However, this file takes up around 2 gigabyte storage space after compression. The runtime can take minutes to load this file into memory and then make prediction which makes this model inefficient. On the other hand, the polynomial regression has the balance between runtime and accuracy.

The random forest model shows signs of overfitting in the prediction of the optimal system because new trends appear on the plot which are not observable on the reference dataset. However, the adj. R^2 values have no difference from the R^2 meaning that there should be no overfitting issue with the regression. This problem should be further investigated to better understand this regression model.

The model used by McKenna is linear regression. This simple model performed well with 0.757 R^2. The more complex models in this thesis performed better, the random forest for the single-family house case reached 0.980 R^2. However, the random forest model showed signs of overfitting when the optimal system case was evaluated. New trends were visible on the SC plot predicted from the random forest model compared to the reference SC plot. The simpler polynomial model had no such issues, and it grasped the main trends from the reference dataset better. This suggests that using a more complex model is not always favorable and issues with overfitting should be considered.

The number of prediction models for the SC is fairly low in the literature. But the relevance of these tools is high because the techno-economic analysis on PV systems is based on SC. And predicting the optimal PV system based on easily available input data such as annual building load is important to be able to spread the PV technology widely. This will help to achieve the 7^{th} Sustainable Development Goal (SDG), Affordable and Clean Energy. Because the

additional renewable energy generation in the households would require less grid energy which is not carbon-free. Also, this will help to achieve the 13th SDG, Climate Action. Due to reducing the consumed grid energy and CO_2e emissions.

7. Conclusion

The models from Galli's work used both spatially and temporally mismatched weather data to the building load. The comparison of Galli's models with the TMY and Measured models showed relatively high impact on the performance. The mean absolute error for the k-NN model had 20% difference and 10.5% for the MLP, polynomial and ridge regressions. The highest difference in R^2 values was 6.4% for the lasso regression. The results from the single-family house case showed that the usage of only spatially matched weather data to the building load would have no significant impact on the performance of the regression models compared to using spatially and temporally matched data. But the usage of only spatially matched weather data will likely lower the performance of the regressions. The relatively low difference between the methods is most likely due to the variance of the SC between the TMY and measured datasets is marginal. In the case of multifamily houses, the performance of almost all of the regressions are increased when the TMY trained models were evaluated on the measured data. The mean absolute error lowered for all models except for the random forest, linear and Mckenna regressions. The R^2 increased for all models except for the linear and McKenna regressions. The reason for this is unknown but it is most likely due to the multifamily house's dataset has fewer samples. And these regression models perform better on a larger dataset.

The battery model of the single-family house case indicated that these ML models should be treated with caution because the random forest model showed signs of overfitting. And this resulted in an overprediction in optimal PV size by 86%. The polynomial model had similar trends as the reference and seemed to have a good fit on the dataset. This model predicted 8.8% higher payback time, 18.7% lower SC and 24% higher optimal PV size than the reference value. When the models were directly compared at the same solar fraction value, the random forest had marginal differences in SC compared to the reference dataset. And the polynomial model underpredicted the reference value by 7.2%. The random forest model for the multifamily house case had similar overfitting problems and, in this case, the optimal PV size is underpredicted by 51.4%. The polynomial model had similar trends as the reference dataset, only for small solar fraction values had big differences. This model predicted 9% higher payback time, 11.2% lower SC and 2.7% lower optimal PV size than the reference values. At the direct comparison the random forest model overpredicted the reference value by 6.4% while the polynomial model underpredicted it by 7.0%. Based on the results the polynomial model seemed to have better performance for determining the optimal system size.

The random forest and the k-NN model had above 0.9 R^2 value for the single-family house case, if the battery model was added then also the MLP and polynomial regressions had an R^2 value higher than 0.9. In the case of multifamily houses only the random forest model was above this criterion. If the battery model is added the k-NN regression is satisfied this. The adjusted R^2 and MBE values showed that the models are not overfitting, and they are not biased. To conclude, this study showed that the SC can be predicted with already available ML technologies with a reasonable level of accuracy.

8. Future work

The building load dataset in some load regions has very few samples. To build a robust and accurate regression model, this dataset could be broadened especially in those insufficient regions. The models should be tested when these regions are removed since they may have some effect on the performance.

The building load dataset only contained houses from southern Sweden with similar weather conditions, the effect of different insolation profiles on the models could be investigated and the usage of large scope of different latitudes as another input for the model.

The PVWatts model that is used to generate the hourly PV generation is the simplest model for this purpose in the PySam library. It only uses few weather parameters, but high variety of parameters are available from weather stations and can be included in the weather file and used in a more complex PV model. For Sweden the amount of clouds, precipitation amount and the snow depth can have relatively high impact on the PV generation profile thus probably higher difference in regression models' performance when using TMY weather data compared to measured data.

The results from the optimal system suggested that the random forest has overfitting issues because of the new trends that appeared on the plot compared to the reference plot. But the adjusted R^2 and the R^2 values had no difference for this model, so the random forest model should not be overfitting. This contradiction could be investigated by refining the training dataset or evaluating the model with varying hyperparameters (such as depth for random forest).

The empirical validation of the models is also important, and it should be done by measurements in households with PV installations. Also, smaller time steps can be investigated since the self-consumption can differ with different timesteps and the new smart meters will have measurements in every 15 minutes.

References

[1] IEA, "National Survey Report of PV Power Applications in Sweden," Swedish Energy Agency, 2020.

[2] "Statista," [Online]. Available: https://www.statista.com/statistics/418124/electricity-prices-for-households-in-sweden/. [Accessed 11. 04. 2022].

[3] J. Lindahl, D. Lingfors, A. Elmqvist and I. Mignon, "Economic analysis of the early market of centralized photovoltaic parks in Sweden," *Renewable Energy,* vol. 185, pp. 1192-1208, 2022.

[4] V. Benda and L. Cerna, "A Note on Limits and Trends in PV Cells and Modules," *Applied Sciences,* vol. 12, p. Article 3363, 2022.

[5] F. Galli, "Predicting PV self-consumption in villas with machine learning," M.S. Thesis, KTH, Stockholm, 2021.

[6] E. McKenna, J. Pless and S. J. Darby, "Solar photovoltaic self-consumption in the UK residential sector: New estimates from a smart grid demonstration project," *Energy Policy,* vol. 118, pp. 482-491, 2018.

[7] A. P. Dobos, "PVWatts Version 5 Manual," National Renewable Energy Laboratory, Denver, 2014.

[8] R. Perez, P. Ineichen, E. Maxwell, R. Seals and A. Zelenka, "Dynamic global-to-direct irradiance conversion models," *ASHRAE Transactions,* vol. 1, no. 98, pp. 354-369, 1992.

[9] "SMHI," [Online]. Available: https://www.smhi.se/data. [Accessed 30. 01. 2022].

[10] J. Remund, S. Müller, M. Schmutz and P. Graf, "Meteonorm Version 8," in *EUPVSEC*, Online conference, 2020.

[11] R. Luthander, J. Widen, D. Nilsson and J. Palm, "Photovoltaic self-consumption in buildings: A review," *Applied Energy,* vol. 142, pp. 80-94, 2015.

[12] B. Stridh, "Utvärdering av egenanvändning av solel i Sverige (Evaluation of self-consumption of PV electricity in Sweden)," Swedish Energy Agency, 2020.

[13] X. Han, J. Garrison and G. Hug, "Techno-economic analysis of PV-battery systems in Switzerland," *Renewable and Sustainable Energy Reviews,* vol. 158, p. Article 112028, 2022.

[14] A. Chaianong, A. Bangviwat, C. Menke, B. Breitschopf and W. Eichhammer, "Customer economics of residential PVebattery systems in Thailand," *Renewable*

Energy, vol. 146, pp. 297-308, 2020.

[15] Y. Li, W. Gao and Y. Ruan, "Performance investigation of grid-connected residential PV-battery system focusing on enhancing self-consumption and peak shaving in Kyushu, Japan," *Renewable Energy*, vol. 127, pp. 514-523, 2018.

[16] P. E. Campana, L. Cioccolanti, B. Francois, J. Jurasz and Y. Zhang, "Li-ion batteries for peak shaving, price arbitrage, and photovoltaic self-consumption in commercial buildings: A Monte Carlo Analysis," *Energy Conversion and Management*, vol. 234, p. Article 113889, 2021.

[17] N. Sommerfeldt, "Solar PV in prosumer energy systems," Phd thesis, School of Industrial Engineering and Management, KTH Royal Institue of Technology, Stockholm, Sweden, 2019.

[18] Eon, "Electricity price," [Online]. Available: https://www.eon.se/el/elpriser/aktuella. [Accessed 11. 04. 2022].

[19] "Bloomberg," [Online]. Available: https://www.bloomberg.com/quote/USDSEK:CUR. [Accessed 11. 04. 2022].

[20] A. G. Amillo, T. Huld and M. Richard, "A New Database of Global and Direct Solar Radiation Using the Eastern Meteosat Satellite, Models and Validation," *Remote Sens.*, vol. 6, pp. 8165-8189, 2014.

[21] D. D. Moghaddam, O. Rahmati, M. Panahi, J. Tiefenbacher, H. Darabi, A. Haghizadeh, A. T. Haghighi, O. A. Nalivan and D. T. Bui, "The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers," *CATENA*, vol. 187, no. 104421, 2020.

[22] A. Bailly, C. Blanc, E. Francis, T. Guillotin, F. Jamal, B. Wakim and P. Roy, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Computer Methods and Programs in Biomedicine*, vol. 213, no. 106504, 2022.

[23] A. F. Al-Anazi and I. D. Gates, "Support vector regression to predict porosity and permeability: Effect of sample size," *Computers & Geosciences*, vol. 39, pp. 64-76, 2012.

[24] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, pp. 90-95, 2007.

[25] A. Shokrzade, M. Ramezani, T. F. Akhlaghian and M. M. Abdulla, "A novel extreme learning machine based knn classification method for dealing with big data," *Expert Systems with Applications*, vol. 183, p. Article 115293, 2021.

Appendix A

Table A1: MBE and adj. R^2 values for measured and TMY model for single-family houses

	MBE		adj. R^2	
Model	**Meas**	**TMY**	**Meas**	**TMY**
Rnd Forest	-3.2E-04	-5.5E-05	0.977	0.980
k-NN	7.5E-04	1.1E-03	0.923	0.928
MLP	-1.0E-03	-1.4E-03	0.875	0.880
Polynomial	-3.8E-09	-8.5E-08	0.875	0.884
Ridge	-3.8E-07	4.0E-07	0.856	0.868
Lasso	1.4E-07	-2.5E-07	0.626	0.648
Linear	-1.7E-07	-6.2E-07	0.627	0.640
McKenna	5.7E-07	1.2E-06	0.647	0.658

Table A2: MBE and adj. R^2 values for measured and TMY model for single-family houses battery model

	MBE		adj. R^2	
Model	**Meas**	**TMY**	**Meas**	**TMY**
Rnd Forest	-3.2E-04	-5.5E-05	0.977	0.980
k-NN	7.5E-04	1.1E-03	0.923	0.928
MLP	-1.0E-03	-1.4E-03	0.875	0.880
Polynomial	-3.8E-09	-8.5E-08	0.875	0.884
Ridge	-3.8E-07	4.0E-07	0.856	0.868
Lasso	1.4E-07	-2.5E-07	0.626	0.648
Linear	-1.7E-07	-6.2E-07	0.627	0.640
McKenna	5.7E-07	1.2E-06	0.647	0.658

Appendix B

Table B1: MBE and adj. R^2 values for measured and TMY model for multifamily houses

	MBE		adj. R^2	
Model	**Meas**	**TMY**	**Meas**	**TMY**
Rnd Forest	-2.0E-04	2.5E-05	0.985	0.983
k-NN	-8.7E-04	-1.5E-03	0.789	0.787
MLP	4.8E-04	5.2E-05	0.828	0.821
Polynomial	1.3E-03	1.3E-03	0.672	0.673
Ridge	1.2E-06	-2.0E-06	0.749	0.747
Lasso	3.2E-07	-1.4E-06	0.577	0.574
Linear	-3.5E-07	-4.5E-07	0.424	0.425
McKenna	-3.9E-06	-4.5E-07	0.473	0.474

Table B2: MBE and adj. R^2 values for measured and TMY model for multifamily houses battery model

	MBE		adj. R^2	
Model	**Meas**	**TMY**	**Meas**	**TMY**
Rnd Forest	-1.1E-04	8.2E-05	0.991	0.987
k-NN	6.0E-05	-8.3E-05	0.951	0.947
MLP	9.9E-04	-5.9E-04	0.847	0.843
Polynomial	5.1E-06	9.7E-06	0.810	0.802
Ridge	7.3E-06	1.2E-05	0.763	0.750
Lasso	9.1E-07	3.8E-07	0.546	0.528
Linear	-9.8E-07	-1.0E-07	0.432	0.427
McKenna	-5.7E-07	-5.2E-07	0.474	0.459

TRITA – ITM-EX 2022:344