sensei

D7.1 Methods for the dynamic measurement and verification of energy savings

Beyond simple linear regression methods

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 847066.

Smart Energy Services to Improve the Energy Efficiency of the European Building Stock

Deliverable nº:	**D7.1**
Deliverable name:	**Methods for the dynamic measurement and verification of energy savings**
Version:	**1.0**
Release date:	**15/04/2021**
Dissemination level:	**Public**
Status:	**Submitted**
Author:	**HEBES – Sotiris PAPADELIS**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 847066.

Peer reviewed by:

Partner	Reviewer
IEECP	Filippos ANAGNOSTOPOULOS
CIMNE	Benedetto GRILLONE

Table of contents

Abbreviations and Acronyms

Acronym	Description
PPA	Power Purchase Agreement
ESC	Energy Supply Contract
EEM	Energy efficiency measure
M&V	Measurement and verification
P4P	Pay-for-Performance
BMS	Building Management System
IPMVP	International Performance Measurement and Verification Protocol
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
EVO	Efficiency Valuation Organization

Research Institute
Foreword for SENSEI report

April 2021

Investor perspective on better measuring mass building retrofits and payments for performance of energy efficiency

Author: Murray Birt
Senior ESG Strategist, DWS Group
Steering Committee of the Energy Efficiency Financial Institutions Group (EEFIG)
Co-chair of the Real Estate Paris Aligned Investing Initiative, Institutional Investors Group on Climate Change (IIGCC)

Foreword

Imagine if a power system operator could contract for energy efficiency savings in buildings that could help accelerate retirement of fossil fuel power generation and avoid or delay expensive power grid upgrades.

Imagine if a heating supply company could write a contract for deep thermal building retrofits as an alternative to importing fossil gas.

Imagine if the paradigm could be flipped from retrofits only being about energy cost savings to include new sources of revenue, helping create compelling propositions for consumers and energy service companies (ESCOs).

Imagine if the energy carbon reduction from government building retrofit programs could be more accurately measured instead of estimated or 'deemed'.

In fact, you don't have to imagine this future – just look across the Atlantic Ocean where multiple jurisdictions in the United States and in Canada are implementing these ideas.

From my personal perspective, I commend the work of the Horizon 2020 funded project SENSEI which aims to enable this vision to become a reality in Europe: Paying for Performance for energy efficiency.

Research Institute

Foreword for SENSEI report

April 2021

SENSEI has already produced a report[1] analysing how ten (10) US and Canadian jurisdictions are implementing pay for performance, as well as a report on how power supply companies and system operators might place a financial value on different types of energy efficiency retrofits, based on when and where different retrofits create value in the energy market.

DWS Group wrote a report[2] on the EU's Renovation Wave strategy which recommended that Europe develop a pay for performance strategy. As well, the private sector steering committee members of the Energy Efficiency Financial Institutions Group[3] also provided a similar recommendation to the European Commission. Better measurement is a foundation for linking energy markets with building renovation. Therefore, the SENSEI report that you have in your (virtual) hands, sets out methods for dynamic measurement and verification of energy savings. Although standards for measurement and verification (M&V) of energy savings already exist, they focus on high-level processes rather than specific tools and quantitative techniques. The energy efficiency meter developed in this report, named *eensight*, advances the state of play in M&V in terms of methods and toolkits using Machine Learning algorithms.

eensight is very relevant to the Renovation Wave Communication[4] which stated that the Commission will "*establish a trusted scheme for certifying energy efficiency meters in buildings that can measure actual energy performance improvements".*

Supporters of energy efficiency and stronger building renovations could examine how the new SENSEI tool could be piloted in Europe..

Important information – EMEA

DWS is the brand name of DWS Group GmbH & Co. KGaA and its subsidiaries under which they operate their business activities. The respective legal entities offering products or services under the DWS brand are specified in the respective contracts, sales materials and other product information documents. DWS, through DWS Group GmbH & Co. KGaA, its affiliated companies and its officers and employees (collectively "DWS") are communicating this document in good faith and on the following basis.

This document has been prepared without consideration of the investment needs, objectives or financial circumstances of any investor. Before making an investment decision, investors need to consider, with or without the assistance of an investment adviser, whether the investments and strategies described or provided by DWS Group, are appropriate, in light of their particular investment needs, objectives and financial circumstances. Furthermore, this document is for information/discussion purposes only and does not constitute an offer, recommendation or solicitation to conclude a transaction and should not be treated as giving investment advice.

The document was not produced, reviewed or edited by any research department within DWS and is not investment research. Therefore, laws and regulations relating to investment research do not apply to it. Any opinions expressed herein may differ from the opinions expressed by other legal entities of DWS or their departments including research departments.

[1] SENSEI June 2020 "Experience and lessons learned from pay-for-performance (P4P) pilots for energy efficiency" and SENSEI December 2020 "The Boundary Cases for the P4P Rates" https://senseih2020.eu/publicdeliverables/

[2] DWS May 2020 www.dws.com/insights/global-research-institute/green-healthy-buildings-as-economic-stimulus/

[3] www.eefig.eu

[4] https://ec.europa.eu/energy/sites/ener/files/eu_renovation_wave_strategy.pdf

Research Institute

Foreword for SENSEI report

April 2021

The information contained in this document does not constitute a financial analysis but qualifies as marketing communication. This marketing communication is neither subject to all legal provisions ensuring the impartiality of financial analysis nor to any prohibition on trading prior to the publication of financial analyses.

This document contains forward looking statements. Forward looking statements include, but are not limited to assumptions, estimates, projections, opinions, models and hypothetical performance analysis. The forward looking statements expressed constitute the author‘s judgment as of the date of this document. Forward looking statements involve significant elements of subjective judgments and analyses and changes thereto and/ or consideration of different or additional factors could have a material impact on the results indicated. Therefore, actual results may vary, perhaps materially, from the results contained herein. No representation or warranty is made by DWS as to the reasonableness or completeness of such forward looking statements or to any other financial information contained in this document. Past performance is not guarantee of future results.

We have gathered the information contained in this document from sources we believe to be reliable; but we do not guarantee the accuracy, completeness or fairness of such information. All third party data are copyrighted by and proprietary to the provider. DWS has no obligation to update, modify or amend this document or to otherwise notify the recipient in the event that any matter stated herein, or any opinion, projection, forecast or estimate set forth herein, changes or subsequently becomes inaccurate.

Investments are subject to various risks, including market fluctuations, regulatory change, possible delays in repayment and loss of income and principal invested. The value of investments can fall as well as rise and you might not get back the amount originally invested at any point in time. Furthermore, substantial fluctuations of the value of any investment are possible even over short periods of time. The terms of any investment will be exclusively subject to the detailed provisions, including risk considerations, contained in the offering documents. When making an investment decision, you should rely on the final documentation relating to any transaction.

No liability for any error or omission is accepted by DWS. Opinions and estimates may be changed without notice and involve a number of assumptions which may not prove valid. DWS or persons associated with it may (i) maintain a long or short position in securities referred to herein, or in related futures or options, and (ii) purchase or sell, make a market in, or engage in any other transaction involving such securities, and earn brokerage or other compensation.

DWS does not give taxation or legal advice. Prospective investors should seek advice from their own taxation agents and lawyers regarding the tax consequences on the purchase, ownership, disposal, redemption or transfer of the investments and strategies suggested by DWS. The relevant tax laws or regulations of the tax authorities may change at any time. DWS is not responsible for and has no obligation with respect to any tax implications on the investment suggested.

This document may not be reproduced or circulated without DWS written authority. The manner of circulation and distribution of this document may be restricted by law or regulation in certain countries, including the United States.

This document is not directed to, or intended for distribution to or use by, any person or entity who is a citizen or resident of or located in any locality, state, country or other jurisdiction, including the United States, where such distribution, publication, availability or use would be contrary to law or regulation or which would subject DWS to any registration or licensing requirement within such jurisdiction not currently met within such jurisdiction. Persons into whose possession this document may come are required to inform themselves of, and to observe, such restrictions.DWS Investment GmbH. As of: 14.04.2021

Issued in the UK by DWS Investments UK Limited which is authorised and regulated by the Financial Conduct Authority (Reference number 429806).© 2021 DWS Investments UK Limited. CRC 082561_1.0(04/2021)

Executive summary

This deliverable aims at contributing to the advancement of the automated measurement and verification (M&V) methods for energy efficiency. Automated M&V combines real-time data and predictive modelling methods so that to produce tools to understand the characteristics of a building's energy consumption, and provide continuous feedback on the most probable impact of an energy efficiency intervention.

The deliverable consists of two (2) parts. The first part summarizes the main aspects of an M&V process, including the data requirements, the metrics against which to evaluate a baseline energy consumption model, and the methods for the quantification of uncertainty commonly suggested by the relevant M&V standards. Then, it presents the state-of-play in terms of M&V 2.0 methods and reviews existing models for predictive baseline modelling.

The review carried out showed that there are only a very limited number of M&V frameworks that are ready to be tested and adopted by practitioners. Although the literature on M&V methods is extensive, the gap between: (a) presenting a methodology and its results and (b) offering the tools for practitioners to experiment with this methodology and integrate the parts that they find valuable is most often significant.

Furthermore, no M&V tool can be considered the best one a priori. The interaction between any tool and a specific building – as described by the dataset of its energy consumption – determines whether the tool is suitable for this building or, from the opposite point of view, whether the building is adequately predictable given the selected tool for M&V. Understanding the aforementioned interrelation requires:

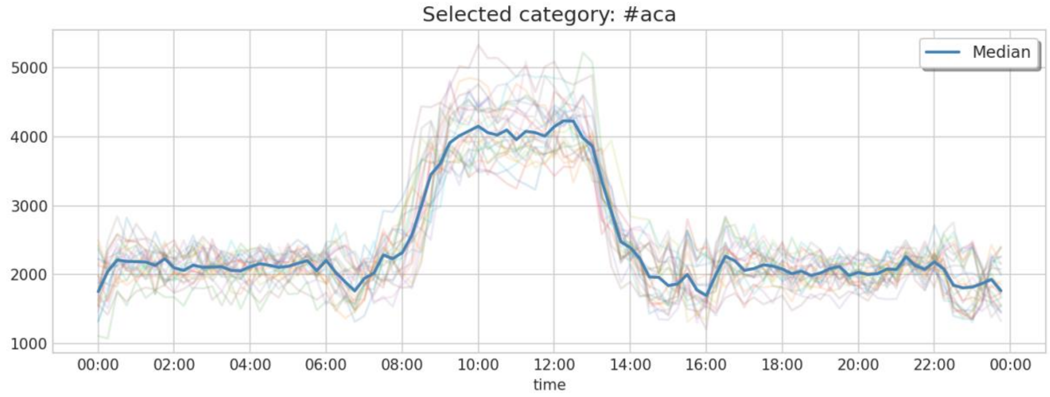
- Access to a diverse set of M&V models and workflows,
- Tools that automate the implementation and evaluation of experiments with different combinations of buildings and M&V model specifications, and
- Methods and indicators for categorizing buildings according to the characteristics of their energy consumption data.

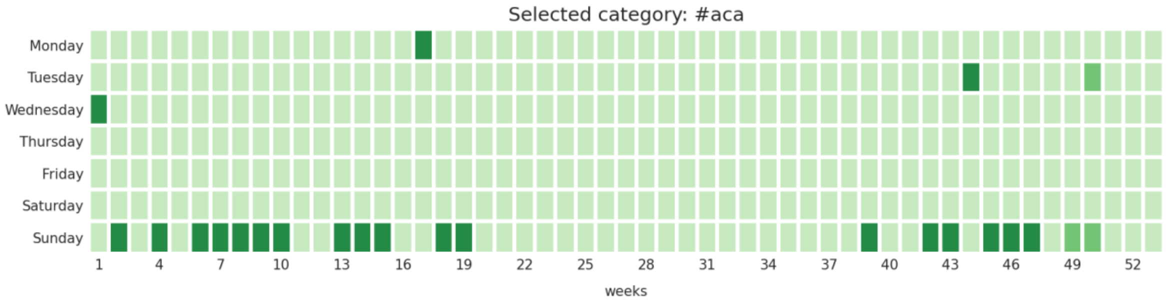
This deliverable contributes to the abovementioned requirements by developing new methods and tools to better understand the characteristics of a building's energy consumption and estimate the energy savings from an intervention, as well as providing all of its methods and tools as a reproducible open source project for practitioners to experiment and test on different datasets. SENSEI aspires to fuel more testing and more experimentation with the fundamental calculations of M&V.

The second part of the deliverable presents the details of the proposed M&V methodology, which comprises four (4) main components:

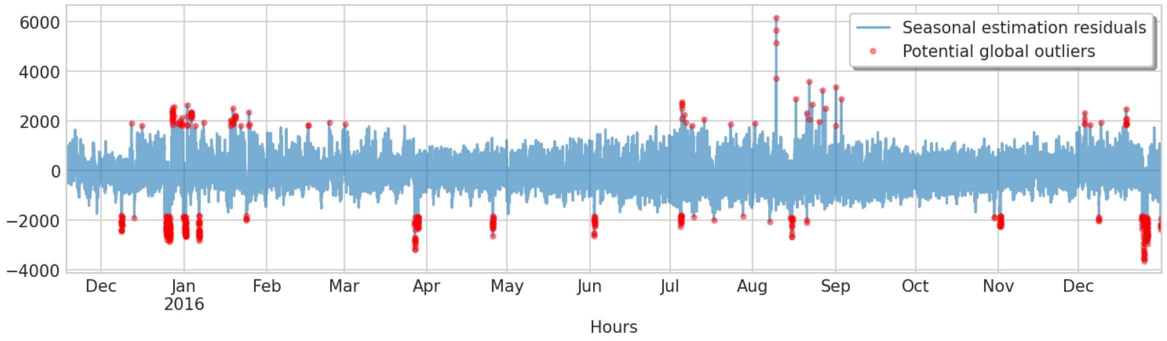
1. **Data exploration and visualization workflow**. It consists of a series of exploratory analysis steps that can be carried out so that a given dataset is better understood, as well as expectations for the corresponding building's consumption can be formed and, later, used for evaluating the patterns that the baseline energy consumption model uncovered. This part of the methodology is meant to guide the implementation of dashboards and interactive visualizations for communicating insights regarding a building energy consumption dataset.

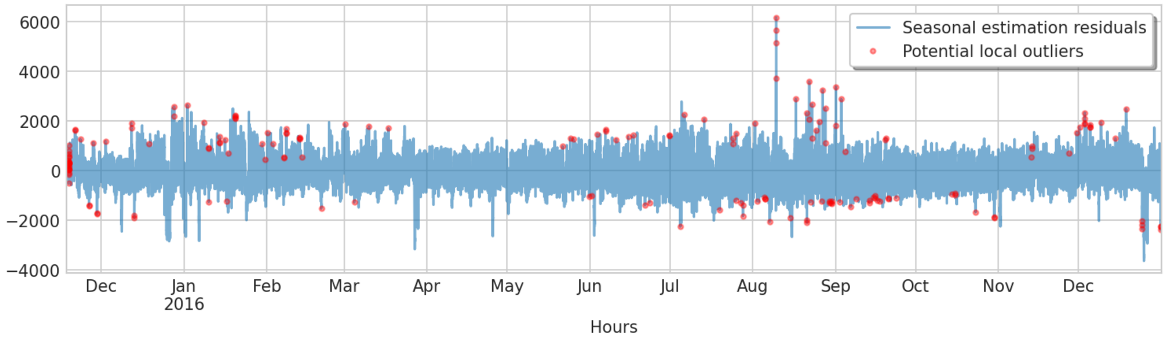
The main concept behind this step of the methodology is the interactive categorization of a building's energy consumption into similar shapes, and the exploration of the calendar distribution of the identified categories:





2. **Data preprocessing workflow**. SENSEI proposes a replicable workflow for M&V data preprocessing. The goal of the preprocessing stage is to validate the available data about a building, as well as to identify potential outliers. The outlier detection methodology is structured around the concept of identifying global and local indications of outlier existence in the data:





3. **Day typing workflow**. Instead of developing a methodology for the sake of developing a new methodology, we started from the widely heard argument that only simple M&V methodologies can gain acceptance for P4P applications. This requirement reflects the fact that no standard exists so that a model can be trusted as long as it adheres to it, irrespectively of whether the model is open source or based on a linear regression formulation that makes it possible to recreate in a spreadsheet.

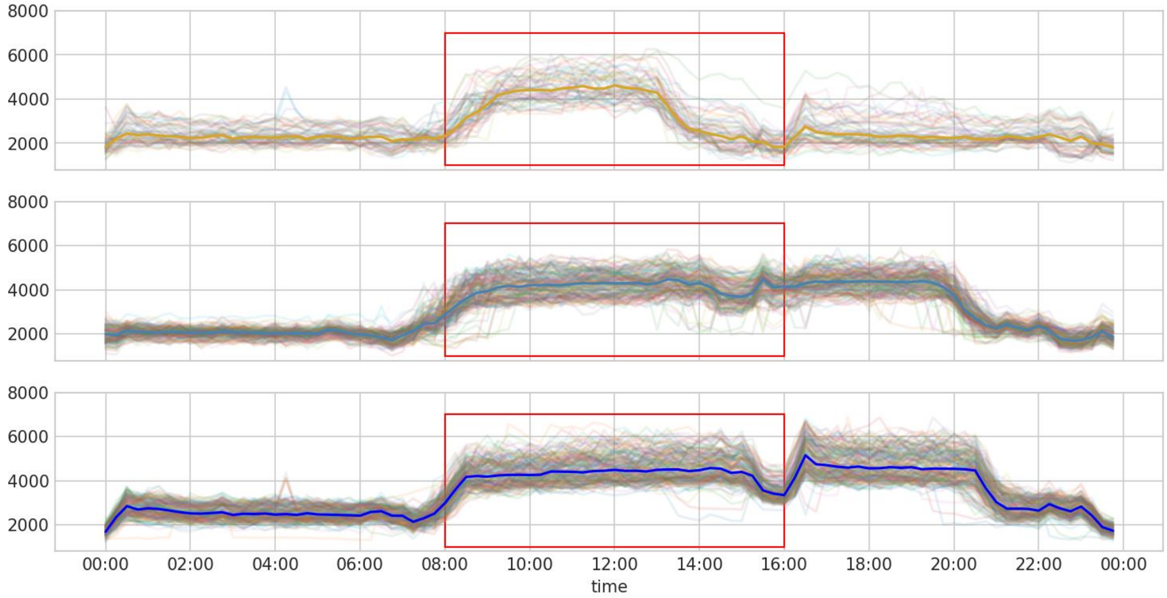
Based on the aforementioned, we worked backwards from devising a possible way of testing and validating a black box model to building an M&V methodology around it. To this end, we started from the concept of prototypes. Prototypes are a selection of representative instances from the data, such that a small set of them can describe adequately well the whole dataset.

In the case of energy consumption data, the prototypes can be found in the frequently recurring consumption profiles that define our expectations about the way a building's energy consumption varies according to the hour of the day, the day of the week, or the outdoor temperature. The use of prototypes brings forth two benefits:

(a) They provide a reference to judge any prediction and/or (possibly black box) model. One can use the prototype – from the set of prototypes – that is closest to a given observation, as a reference explanation for a prediction and monitor its difference with an audited model. The audited model may perform better or worse, but it should differ from the reference in a similar way both for historical data and for the counterfactual predictions.

(b) We can avoid making a priori assumptions regarding the daily, weekly and yearly similarities between the different observations in a building's energy consumption data. Instead, similarity to a prototype is an indicator of an opportunity to categorize different observations together, hence allowing predictive strength to be shared among the observations in each category – rather than just in each month or each month and its previous and subsequent months.

In practical terms, the day typing method aims at categorizing similar consumption profiles as long as it is possible to explain the association of each observation in the dataset with its category given only features that can be constructed without access to the actual consumption data.



4. **Workflow for baseline model application**. This step concerns training a predictive baseline model, evaluating its fitness and quantifying the uncertainty in its predictions. We have not carried out an extensive comparative test between different models and methodologies. Instead, we aimed at showcasing that the proposed methodology can deal well with buildings of different levels of predictability.

Every plot and result in this deliverable is completely reproducible. The relevant notebooks, as well as all the open source functionality that accompanies this deliverable can be found in the GitHub repository at https://github.com/hebes-io/eensight.

1 Introduction

1.1 The context for the deliverable

Renewable energy generation assets are financed based on the value of the energy they will produce. Typically, the owner of a generation asset secures a Power Purchase Agreement (PPA) or an Energy Supply Contract (ESC) for selling the generated energy, and this agreement can be used for attracting investment capital to purchase and install the necessary equipment. Based on this clear value proposition, a series of business model innovations have come to exist and offer solutions for the financing and deployment of renewable energy generation assets, the aggregation of the produced energy, and the securitization of the expected revenues.

In contrast, the energy efficiency services sector is lagging behind renewable energy generation in terms of demand for investments and business models for large-scale deployment. While a lot of work has taken place on measuring energy savings and formally covering risks and uncertainties at the *individual project level*, scaling energy efficiency up to project portfolio or programme level still faces challenges. One of the main barriers to market growth has been the uncertainty about the magnitude and persistence of the achieved efficiency improvements; renewable energy generation assets produce a measurable outcome, while energy efficiency can only be estimated through counterfactual analysis, which leads to increased uncertainty and potential for disputes. The heterogeneity in building structures, operating schedules and implemented energy efficiency measures (EEMs) only amplifies this uncertainty.

While there is no way of overcoming the need for counterfactual analysis, uncertainty can be mitigated when all the parties involved in the process of up-scaling energy efficiency can agree on the way this counterfactual analysis will be carried out. A main argument of the SENSEI project is that advanced measurement and verification (M&V), sometimes called *M&V 2.0*, can lay the foundation for energy efficiency project aggregation schemes and/or energy efficiency support programs by providing the insights that are necessary for all parties involved in up-scaling energy efficiency to correctly evaluate risks and expected benefits. M&V 2.0 combines real-time data and predictive modelling methods so that to produce: (a) tools to understand the characteristics of a building's energy consumption, and (b) continuous feedback on the most probable impact of an energy efficiency intervention.

Up-scaling energy efficiency to project portfolio or programme level would create an additional opportunity: the opportunity to valorise energy efficiency when it contributes to offsetting the need for new power generation plants or transmission network upgrades. In this framework, energy efficiency would be regarded as a load modifier. Load-modifiers are those resources that are not necessarily seen or optimized by the power or capacity market, but they modify the power system's load shape in ways that harmonize with the system operator's grid operations. To this end, the SENSEI project aims at preparing the ground for pilots that will act as a workbench for demonstrating that M&V 2.0 can indeed support the treatment of energy efficiency as a reliable load modifier. The pay-for-performance (P4P) concept has been adopted by SENSEI as a way to structure these pilots and steer the EEM selecting process towards measures that are beneficial for both building owners and the power system. As a consequence, all the transactions envisioned by SENSEI are defined by three (3) components:

1. **Performance indicators**. The transactions are structured around performance-based agreements that reward improved energy efficiency, while transparently allocating the costs and risks. When treating energy efficiency as a load-modifying resource, these indicators reflect the fact that the value of energy efficiency may vary according to the hour of the day and the season of the year.

2. **Compensation agreements**. The rules that dictate the proposed transactions aim at promoting the optimization and maintenance of the installed EEMs. In addition, the contractual arrangements aim at allocating the risks according to each party's ability to control or mitigate them, and at compensating for the uptake of the risks.

3. **Measurement and verification methods**. The role of the M&V process is to quantify and monitor the performance indicators. This deliverable contributes to the advancement of the existing M&V methods by developing and open sourcing tools to better understand the characteristics of a building's energy consumption and estimate the efficiency gains from an intervention.

The aforementioned components are summarized in the following diagram:

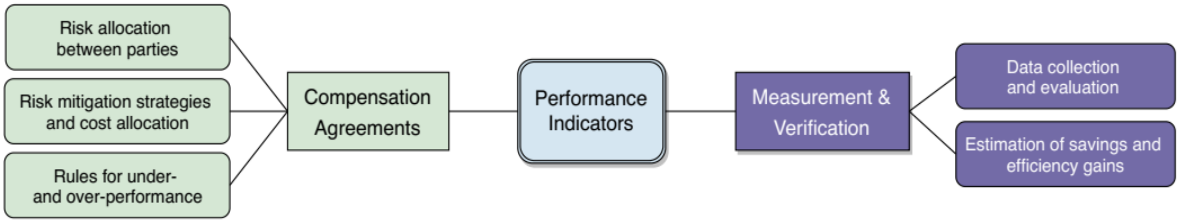

Risk allocation between parties
Risk mitigation strategies and cost allocation
Rules for under- and over-performance
Compensation Agreements
Performance Indicators
Measurement & Verification
Data collection and evaluation
Estimation of savings and efficiency gains

Figure 1.1: The components defining the SENSEI transactions

Part A: Background and state-of-play for M&V2.0

2 Background to Measurement and Verification

2.1 The concept of measurement and verification of energy savings

Measurement and verification (M&V) of energy savings is fundamentally an impact assessment problem, where the goal is to estimate the counterfactual – i.e. what would the energy consumption of a building have been had an energy efficiency intervention not occurred – using two sources of information. The first source is the past behaviour of the building's energy consumption. Energy consumption reflects events and operations that take place inside the building and, as a result, recurrent events and routine operations lead to daily, weekly and yearly seasonality in the consumption that can be exploited for estimating the counterfactual – to the extent that the seasonality remains unaffected by the intervention. The second source is the behaviour of other time series that were predictive of the building's energy consumption prior to the intervention. The most often utilized time series is the outdoor air temperature.

The relationship for the quantification of the energy savings from a retrofit project is commonly defined as:

$$\begin{aligned} \mathit{energy\ savings} = {} & \mathit{Counterfactual\ consumption\ based\ on\ baseline\ period\ data} \\ & - \mathit{Actual\ consumption\ during\ the\ reporting\ period} \\ & \pm \mathit{Routine\ adjustments} \\ & \pm \mathit{Non\text{-}routine\ adjustments} \end{aligned} \tag{2.1}$$

where:

Baseline period: The period of time prior to the intervention during which data is gathered so as to determine the relationship between energy consumption and the different independent variables that can predict it.

Reporting period: The period of time following the intervention during which data is gathered so as to calculate energy savings (avoided energy use).

Routine adjustments: Routine adjustments account for any energy-governing factors expected to change routinely during the reporting period, such as weather, operating hours or service levels (e.g. a new tenant requires different indoor conditions). The characterization "expected" means that the predictive model developed during the baseline period is able to adjust to such changes.

Non-routine Non-routine adjustments account for unexpected changes in energy use. The *adjustments:* fact that the changes are "unexpected" means that the driving factors of these changes were not included as independent variables in the baseline predictive model. By definition, they render the predictive model less relevant and require adjustments either to the model or to the baseline period energy data so as to reflect the same set of conditions as the ones observed during the post-intervention period.

The diagram below summarizes the concept of energy savings metering: the estimation of the energy savings by comparing the energy consumption after the implementation of an EEM (i.e. during the reporting period) to a baseline that represents what the consumption would have been without this measure (counterfactual).

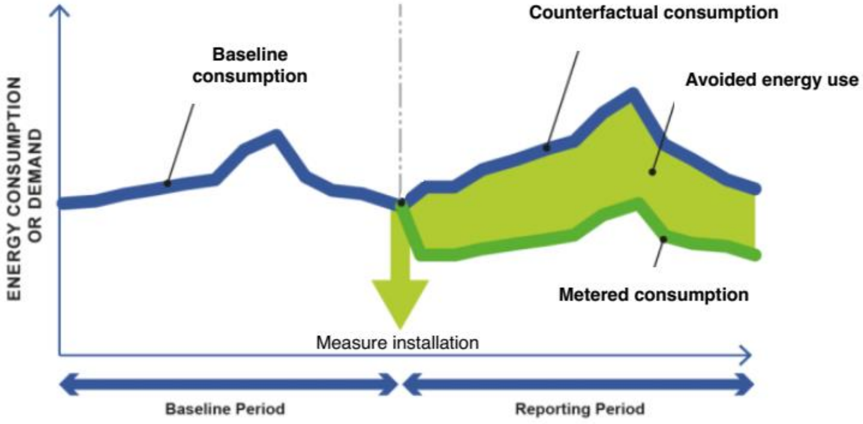


Figure 2.1: Estimation of energy savings from an energy efficiency measure (Adapted from IPMVP Generally Accepted M&V Principles, 2018)

2.2 The scope of M&V in SENSEI

2.2.1 Excepted data availability

It is assumed that energy consumption and outdoor air temperature is the only data that is available for M&V, since this data corresponds to the minimum set of data that may be available across all buildings in a portfolio and that would allow for a consistent analysis.

2.2.2 Relevant levels of energy savings estimation and monitoring

There are three (3) levels of M&V that are relevant for the SENSEI project:

1. **Single project monitoring**. Since the SENSEI project assumes that an energy retrofit project portfolio is derived by aggregating many projects that are designed and implemented by different ESCOs and through different energy performance contracts, it is necessary to have access to M&V services that can track the contribution of each individual project with respect to the performance of the overall portfolio.

2. **Project portfolio monitoring**. SENSEI is concerned with up-scaling energy efficiency to a project portfolio level. Accordingly, it is necessary to have access to M&V services that can track the performance of a portfolio as a whole.

3. **Program evaluation**. SENSEI promotes a paradigm where the evaluation of an energy efficiency program (from the perspective of the entity that compensates the projects) takes place in parallel with its implementation rather than after its completion.

2.2.3 Relevant IPMVP options

To quantify the impact from an efficiency upgrade intervention, the International Performance Measurement and Verification Protocol (IPMVP) proposes four (4) options:

Option A – Retrofit Isolation: Key Parameter Measurement. This option makes sense when the EEMs involve some parameters that are known with a high degree of certainty and other parameters that can be measured cost-effectively. An example is the case where the EEMs concern the building's HVAC systems and relevant data can be collected directly from a Building Management System (BMS).

Option B – Retrofit Isolation: All Parameter Measurement. This option is relevant when a given EEM's parameters are uncertain but can be measured in a cost effective way. Similarly to Option A, data can be collected either through the BMS or by using temporary meters.

Option C – Whole Facility: This option is applicable if the estimated project-level savings are large compared to the random or unexplained energy variations that occur at the whole-facility level, and if savings fluctuate over a seasonal or annual cycle. Option C may be the best approach if the EEMs cause complex, significant interactive effects.

Option D – Calibrated Simulation: For EEMs where it is prohibitive to meter all required parameters, one can utilize a simulation model that is calibrated to the actual pre-intervention data.

In a SENSEI pilot, it is expected that the actual EEMs to be implemented are selected by the respective ESCOs, and the M&V process does not require access to specific parts of the buildings' technical systems. As a consequence, Option C is predominantly the relevant M&V option. The viability of Option C is dependent on finding a predictive model that is able to explain a sufficient part of the metered energy consumption's variability.

2.3 Data requirements

Sufficient historical data is necessary for the development of reliable baseline predictive models. One indicator for data sufficiency is the *coverage factor*. The coverage factor refers to the range in the observed values of the model's independent variables during the baseline period. The collected data should cover the full range of the operating conditions of the building, and, ideally, predictions should concern operating conditions that the model has already seen (interpolation). Extrapolation refers to the case when the model predicts the energy consumption for values of the independent variables that are outside of the range used to train it.

ASHRAE Guideline 14 allows extrapolation for 10% above and 10% below of the baseline period outdoor air temperature range for models that use temperature as an independent variable. There are no definite criteria for dealing with seasonality, but when seasonal variations are significant, the data used for training the baseline model must provide enough information to capture these variations.

In any case, the available historical data must allow the analyst to identify the most frequent energy consumption patterns and answer at least the following questions:

- How does energy consumption vary across different hours of the day and days of the week?
- Are there specific hours and/or days when the building is probably not used?
- What part of the energy consumption is directly linked to the outdoor air temperature?
- Is there a start-up period for the building's operation (typically early in the morning) and is the sensitivity of the energy consumption to the outdoor air temperature during this period different than the one for the consumption during the rest of the day?
- Are there timing differences between weather changes and the associated energy use in the building?

2.4 Baseline model evaluation

Given adequate historical data, the next step is the development of the baseline predictive model. While the following chapters will discuss this step in much more detail, it is useful to note at this point that the functionality and modelling approach of an M&V predictive model generally aims at:

- Capturing both the daily and weekly patterns of energy consumption, as well as how these patterns change during the year;
- Estimating a flexible relationship between outdoor air temperature and energy consumption that reflects the impact of temperature on the demand for heating and cooling;
- Using historical energy usage data to screen buildings by determining if a building is a good fit for M&V analysis given the particular model at hand. This is done by calculating what level of energy savings would be required so that the model is able to distinguish them from the random or unexplained energy variations that occur at the whole-facility level.

The ASHRAE Guideline 14 proposes the use of the following quantitative metrics to evaluate the fitness of a baseline predictive model:

(1) The coefficient of variation of the root mean squared error ($CV(RMSE)$);

(2) The normalized mean bias error ($NMBE$).

The $\boldsymbol{CV(RMSE)}$ provides a quantification of the typical size of the error relative to the mean of the observations. The root mean squared error (RMSE) is calculated by:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.2}$$

where:

n The number of observations

y_i The consumption of the i^{th} observation ($i = 1,2, \dots, n$)

$\hat{y}_i$ The estimation for the i^{th} observation's consumption.

The $CV(RMSE)$ is calculated by:

$$\mathrm{CV(RMSE)} = \frac{1}{\bar{y}} \times \mathrm{RMSE} \times 100(\%) \tag{2.3}$$

where:

$\bar{y}$ The mean value of the observed consumption data.

The minimum ASHRAE Guideline 14 requirements for a baseline model's $CV(RMSE)$ are:

- For estimations in a reporting period that lasts less than 12 months: CV(RMSE) < 20%
- For estimations in a reporting period that lasts from 12 to 60 months: CV(RMSE) < 25%
- For estimations in a reporting period that lasts more than 60 months: CV(RMSE) < 30%

The $NMBE$ is computed by:

$$NMBE = \frac{1}{\bar{y}} \times \frac{\sum_{i=1}^{n}(y_i - \widehat{y_\imath})}{n} \times 100(\%) \tag{2.4}$$

The ASHRAE Guideline 14 requires that $|NMBE| \leq 0.5\%$.

A general rule is that the quantification of both $CV(RMSE)$ and $NMBE$ should utilize data that the baseline predictive model has not seen before. This is the typical application for cross-validation: a re-sampling procedure used to estimate how a predictive model's performance will generalize when applied on new data.

The most common method for cross-validation is the K-fold cross-validation. In K-fold cross-validation, the dataset is split into K smaller subsets (folds). For each of these subsets of data, the predictive model is trained using the remaining $K - 1$ subsets, and is evaluated on the subset that was left out. The predictive performance is computed by averaging over the results of all the subsets used for evaluation. This process is summarized in Figure 2.2.

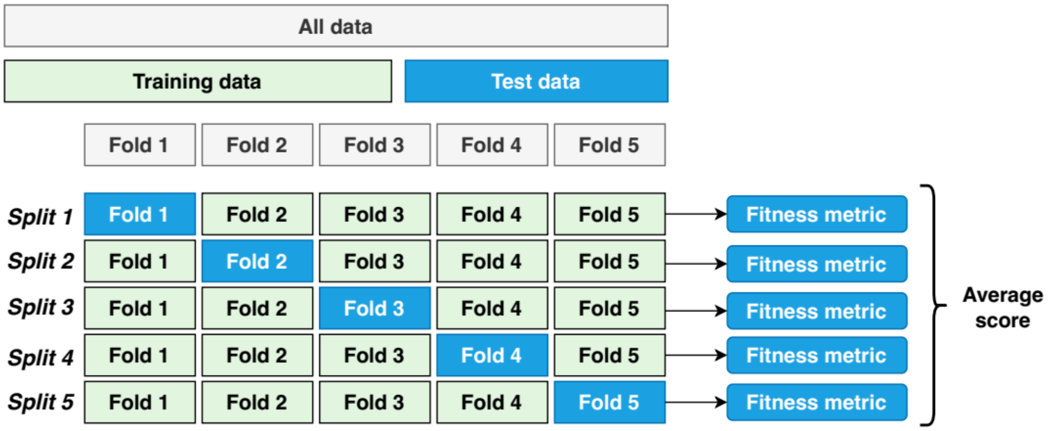


Figure 2.2: The standard K-fold cross-validation method

The fundamental assumptions behind the use of the K-fold cross-validation are that all observations come from the same generative process and that this generative process has no memory of past generated observations. However, M&V models operate on time-series data. Bergmeir, Hyndman and Bonsoo (2018)[5] have shown that although there are theoretical problems that invalidate the fundamental assumptions of the K-fold cross-validation when applied to time-series data, it can still be applied when the utilised predictive model leads to uncorrelated errors. This is achieved mainly when the predictors include lagged values of the response variable (energy consumption). In most cases, this is not aligned with the way baseline energy consumption models are constructed.

When the assumptions for the applicability of the K-fold cross-validation to time-series data do not hold, a rolling origin approach (sometimes referred to as the walk-forward cross-validation method) can be employed. In this case, the training set consists each time only of observations that occurred prior to the observations that form the test set. Thus, no future observations are used in constructing the forecast.

The following diagram illustrates this process:

[5] Bergmeir C., Hyndman R. J., Bonsoo Koo B. (2018) "A note on the validity of cross-validation for evaluating autoregressive time series prediction," Computational Statistics & Data Analysis, Vol. 120, pp. 70-83

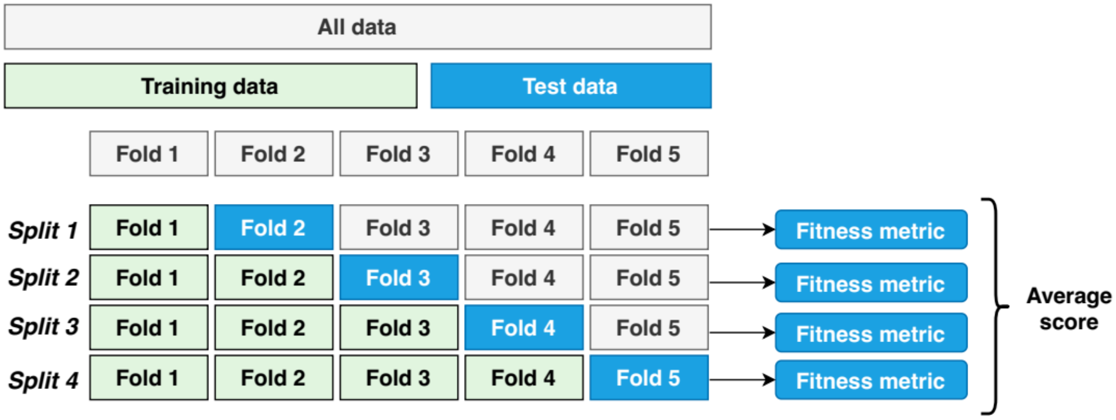


Figure 2.3: The walk-forward cross-validation method

2.5 Quantification of uncertainty

The expected savings must be greater than the uncertainty in the baseline model's predictions. ASHRAE Guideline 14 refers to the ratio of the uncertainty to the savings (i.e. the savings uncertainty as a percentage of the estimated savings) as the fractional savings uncertainty (FSU), and requires that $FSU < 0.5$ at a confidence level of 68%.

In addition, ASHRAE Guideline 14 provides a closed-form equation to quantify the FSU when the only input variable that is considered for the baseline model is the outside air temperature and the baseline consumption is estimated by a linear regression model using the ordinary least squares method:

$$FSU = 1.26 \cdot t_{1-\frac{a}{2},dof} \cdot \frac{CV(RMSE) \cdot \sqrt{\left(1+\frac{2}{n}\right)\frac{1}{m}}}{F} \tag{2.5}$$

where:

$t_{1-\frac{a}{2},dof}$ The t-statistic for confidence level a and degrees of freedom dof (the degrees of freedom of the baseline model is the total number of observations in the baseline period minus the number of explanatory variables in the model not including the intercept);

n The number of observations in the baseline period (used to estimate the baseline model);

m The number of observations in the reporting period;

F The savings fraction defined as the energy savings during the reporting period divided by the predicted consumption during that same period.

The relationship in (2.5) implies that the greater the $CV(RMSE)$, the greater the savings fraction F needs to be so as to adhere to the ASHRAE guidance.

When high resolution data is used (15-min, hourly or daily) the model's prediction residuals will be autocorrelated. For this case, ASHRAE Guideline 14 suggests the formula:

$$FSU = 1.26 \cdot t_{1-\frac{a}{2},dof} \cdot \frac{CV(RMSE)\sqrt{\frac{n}{n'}\left(1+\frac{2}{n'}\right)\frac{1}{m}}}{F} \tag{2.6}$$

where:

n' The effective number of observations in the baseline period after accounting for autocorrelation:

$$n' = n\frac{1-\rho}{1+\rho}$$

where ρ is the lag 1 autocorrelation coefficient of the baseline model errors.

However, if we adopt more flexible model representation and fitting approaches, we can no longer rely on closed form solutions for uncertainty quantification. Furthermore, existing literature[6] suggests that the standard methods for estimating the total savings uncertainty over the post-installation period tend to underestimate uncertainty. The tendency to underestimate the uncertainty is stronger for hourly models than for daily models.

[6] Samir Touzani, Jessica Granderson, David Jump, Derrick Rebello (2019) "Evaluation of methods to assess the uncertainty in estimated energy savings," Energy and Buildings, Volume 193, pp. 216-225

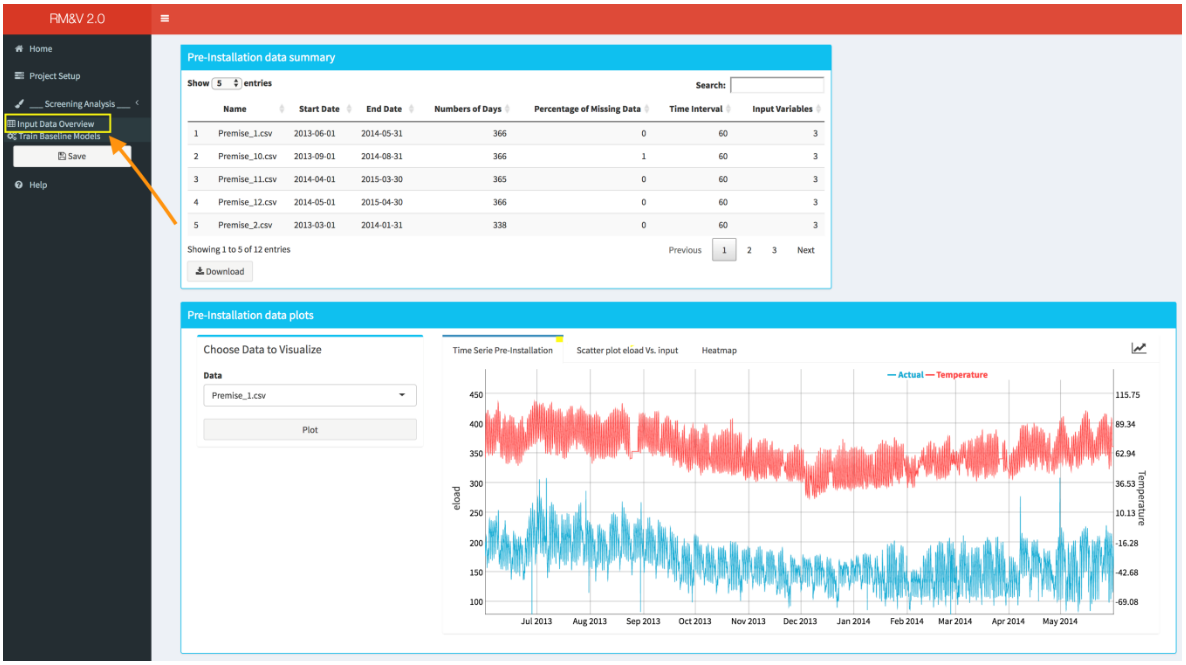
3 The state-of-play in M&V 2.0

3.1 Introduction

This chapter focuses on the state-of-play in terms of M&V 2.0 methods. The Efficiency Valuation Organization (EVO) white paper titled "IPMVP's Snapshot on Advanced Measurement & Verification"[7] already summarizes the different types of advanced M&V tools and presents five freely available M&V tools: ECAM, RMV2.0, OpenEEMeter, UT3 M&V Module and NMECR. We provide additional details for RMV2.0 and OpenEEMeter, as we will build on these details in the following chapters. Furthermore, we present results from competitions that reward modelling approaches for predicting energy consumption with high accuracy. These competitions tend to act as a means of crowdsourcing and benchmarking prediction models.

3.2 The LBNL RMV2.0 model

The LBNL RMV2.0[8] is an open-source package for performing M&V for commercial buildings developed by the Lawrence Berkeley National Laboratory. It is meant to run locally, and it provides a graphical user interface (GUI) that is browser-based (Figure 3.1).



Pre-Installation data summary

Show 5 entries   Search:

	Name	Start Date	End Date	Numbers of Days	Percentage of Missing Data	Time Interval	Input Variables
1	Premise_1.csv	2013-06-01	2014-05-31	366	0	60	3
2	Premise_10.csv	2013-09-01	2014-08-31	366	1	60	3
3	Premise_11.csv	2014-04-01	2015-03-30	365	0	60	3
4	Premise_12.csv	2014-05-01	2015-04-30	366	0	60	3
5	Premise_2.csv	2013-03-01	2014-01-31	338	0	60	3

Showing 1 to 5 of 12 entries

Figure 3.1: The graphical user interface of LBNL RMV2.0

[7] https://evo-world.org/en/news-media/evo-news/1175-evo-releases-a-white-paper-on-advanced-measurement-verification

[8] https://lbnl-eta.github.io/RMV2.0/

The user has two (2) options for selecting the type of the baseline predictive model:

(a) A Time of Week and Temperature (TOWT) model, and
(b) A Gradient Boosting Machine (GBM) model.

3.2.1 The TOWT model

The TOWT model[9] is a piecewise linear model where the energy consumption is predicted as a combination of two terms, one that relates the energy consumption to the time of the week and one that captures the piecewise-continuous effect of the outdoor air temperature.

The way electricity consumption varies with the outdoor temperature is generally a reliable indicator of building occupancy. Accordingly, the temperature effect is estimated separately for periods of the day with high and with low energy consumption in order to distinguish between occupied and unoccupied building periods. For this, the temperature data is split into up to seven (7) binned features. Bins with fewer than 20 hours are combined with the next closest bin by dropping the larger bin endpoint, except for the largest bin, where the lower endpoint is dropped. Assuming that the bins are created using N bin endpoints, the temperature features are constructed as follows:

- If the temperature T is greater than B_1, the first temperature feature is $T_1 = B_1$ and the algorithm proceeds to the next step. Otherwise, $T_1 = T$ and $T_i = 0$ for $i = 2, \dots, N$.
- For $i = 2, \dots, N$, if the temperature T is greater than B_i, then $T_i = B_i - B_{i-1}$ and the algorithm proceeds to the next i. Otherwise, $T_i = T - B_{i-1}$ and $T_j = 0$ for $j = i + 1, \dots, N$.
- If the temperature T is greater than B_N, then the last temperature feature is equal to $T - B_N$, and all other features are equal to zero.

Conceptually, the temperature model can be represented by the diagram below.

[9] J. L. Mathieu, P. N. Price, S. Kiliccote and M. A. Piette (2011) "Quantifying Changes in Building Electricity Use, With Application to Demand Response," in IEEE Transactions on Smart Grid, vol. 2, no. 3, pp. 507-518

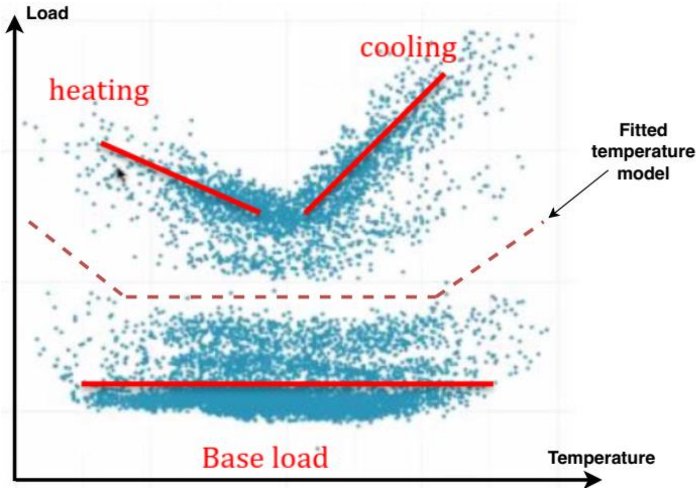


Figure 3.2: The general form of a piecewise-linear temperature model

Subsequently, each week is divided into 168 hourly time-of-week intervals starting from Monday: interval 1 is from midnight to 1 A.M. on Monday morning, interval 2 is from 1 A.M. to 2 A.M. and so on. If more than the 65% of the data points that correspond to a specific time-of-week are above the fitted curve, the corresponding hour is flagged as "Occupied", otherwise it is flagged as "Unoccupied."

Finally, the predicted consumption for the occupied hours is given by:

$$\hat{y}_{occ}(t) = a^{occ}_{TOW[t]} + \sum_{j} \beta^{occ}_{j} T_j(t) + \varepsilon_t \tag{3.1}$$

where:

$\hat{y}_{occ}$ The predicted consumption for the occupied hours

TOW The time of week index: $TOW \in [1,2, \dots ,168]$
The notation $TOW[t]$ means the value of index TOW at time t

$a^{occ}_{TOW[t]}$ The effect of the time-of-week on the energy consumption during the occupied hours

β^{occ}_{j} The coefficient for the j^{th} bin of the temperature when the building is occupied

$T_j(t)$ The transformed temperature feature that corresponds to bin j

ε_t The noise of the regression model

In contrast, the predicted consumption for the unoccupied hours is given by:

$$\hat{y}_{u}(t) = a^{u}_{TOW[t]} + \beta_u T(t) + \varepsilon_t \tag{3.2}$$

where:

$\hat{y}_u$ The predicted consumption for the unoccupied hours

$a^u_{TOW[t]}$ The effect of the time-of-week on the energy consumption during the unoccupied hours

β_u The coefficient of the linear term for the outdoor temperature

$T(t)$ The outdoor temperature at time t

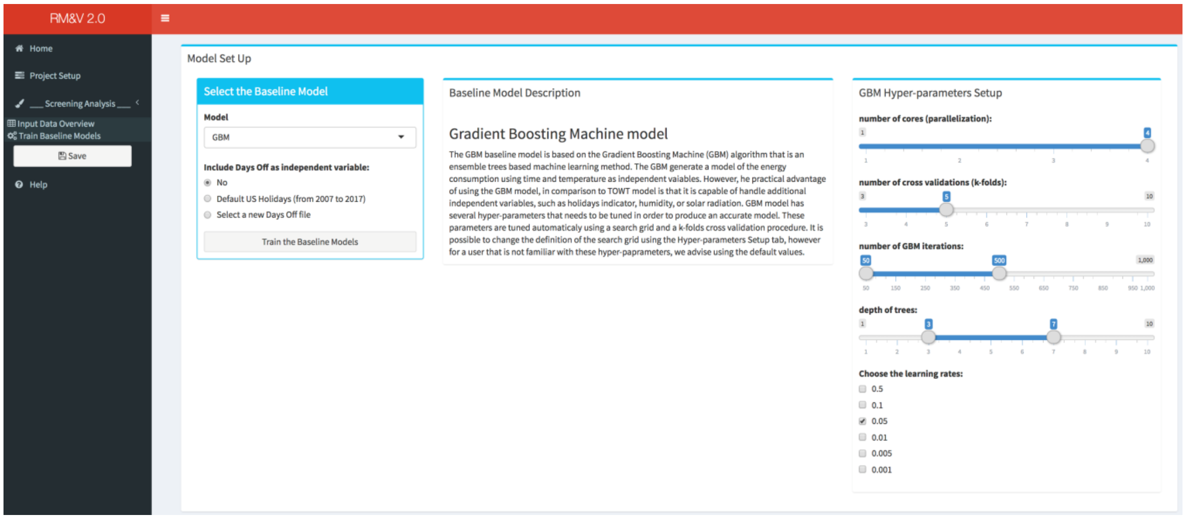
ε_t The noise of the regression model

The way temperature is treated by the TOWT model implies an expectation that the relationship between energy consumption and outdoor temperature is linear when the building is not occupied (i.e. the building's HVAC systems operate at or near the dead-band), whereas it is more flexible when the building is occupied (usually U-shaped).

In addition, a weighting factor can be added to give more statistical weight to days that are nearby to the day being predicted. The time scale of the weighting function (i.e., the number of days that are nearby to the predicted day) is the only hyper-parameter of the TOWT model.

3.2.2 The GBM model

The practical advantage of using a GBM model[10], compared to the TOWT model, is that it is capable of handling additional independent variables, such as holiday indicators, humidity, or solar radiation. The disadvantage of such models, however, is that their structure is hidden. Furthermore, the GBM model has several hyper-parameters that need to be tuned in order to produce an accurate model (Figure 3.3).



Model Set Up

Select the Baseline Model

Model: GBM

Include Days Off as independent variable:

● No

○ Default US Holidays (from 2007 to 2017)

○ Select a new Days Off file

Baseline Model Description

Gradient Boosting Machine model

The GBM baseline model is based on the Gradient Boosting Machine (GBM) algorithm that is an ensemble trees based machine learning method. The GBM generate a model of the energy consumption using time and temperature as independent vaiables. However, he practical advantage of using the GBM model, in comparison to TOWT model is that it is capable of handle additional independent variables, such as holidays indicator, humidity, or solar radiation. GBM model has several hyper-parameters that needs to be tuned in order to produce an accurate model. These parameters are tuned automaticaly using a search grid and a k-folds cross validation procedure. It is possible to change the definition of the search grid using the Hyper-parameters Setup tab, however for a user that is not familiar with these hyper-papranmeters, we advise using the default values.

GBM Hyper-parameters Setup

number of cores (parallelization): 1 – 4

number of cross validations (k-folds): 3 – 5 (range 3 to 10)

number of GBM iterations: 50 – 500 (range 50 to 1,000)

depth of trees: 3 – 7 (range 1 to 10)

Choose the learning rates:

☐ 0.5

☐ 0.1

☑ 0.05

☐ 0.01

☐ 0.005

☐ 0.001

Figure 3.3: Selection of values for the GBM model's hyper-parameters

[10] Touzani, S., Granderson, J. and Fernandes, S., 2018 "Gradient boosting machine for modelling the energy consumption of commercial buildings," Energy and Buildings, 158, pp. 1533-1543

Gradient boosting produces a prediction model in the form of an ensemble of many weak prediction models, most often decision trees (Figure 3.4). Each model in the ensemble is not very accurate, but all of them together create a powerful model.

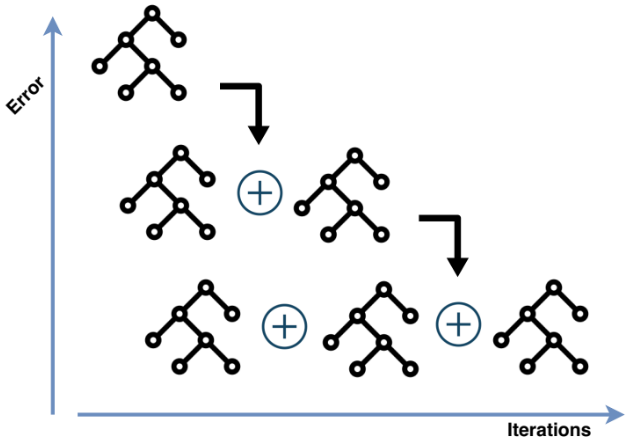


Figure 3.4: Conceptual representation of gradient boosting

GBM models support a modelling workflow where first the model is applied on the available data and then, based on the results, the most important inputs are highlighted, and the strongest interactions are identified. This means that we rely on the models' predictive capability to identify which of the included features are most impactful, as well as how these features interact with each other.

3.3 The OpenEEMeter model

The OpenEEMeter model[11] is an open source implementation of the CalTRACK set of methods[12]. The CalTRACK methods focus specifically on calculating weather-normalized metered energy savings for determining payments under pay-for-performance (P4P) programs for residential buildings. In terms of its modelling approach, CalTRACK builds on the aforementioned TOWT model. It assumes that the energy consumption of a building follows a daily and a weekly pattern, as well as that these patterns change from month to month. Accordingly, instead of using a single baseline model for estimating the counterfactual energy consumption during all times of the year, up to twelve (12) separate models may be used for a particular building – one for predicting the counterfactual in each calendar month. In particular, each model is fitted using:

- Data from the same calendar month in the 365 days prior to the intervention date;
- Data from the previous and subsequent calendar months in the 365 days prior to the intervention date. These data points are given a weight of 0.5 when fitting the model.

[11] http://eemeter.openee.io/

[12] https://www.caltrack.org/

Implicitly, behind CalTRACK's design decision of building a monthly model by also utilizing data from the previous and subsequent calendar months, there is the assumption that: (a) there is information in the previous and subsequent months that is relevant for the monthly model (i.e. learning something about these months would tell us something about the month under study), and (b) only the previous and subsequent months are relevant for the monthly model.

A CalTRACK model can be summarised as follows:

Dependent variable: Energy consumption per hour

Independent variables:

- Seven (or fewer) temperature features
- 168 binary dummy variables indicating the time-of-week. Each week is divided into 7x24 = 168 hourly time-of-week intervals starting from Monday: interval 1 is from midnight to 1 A.M. on Monday morning, interval 2 is from 1 A.M. to 2 A.M. and so on.
- An occupancy binary variable interacting with the temperature and time-of-week variables.

3.4 Competitions

3.4.1 The ASHRAE - Great Energy Predictor III competition

In 2019, ASHRAE hosted the Great Energy Predictor III machine learning competition on the Kaggle platform[13]. The competitors were provided with over 20 million points of training data from 2,380 energy meters collected for 1,448 buildings from 16 sources. The competition's overall objective was to find the most accurate modelling solutions for the prediction of over 41 million private and public test data points. Miller et al. (2020)[14] provide a summary of the competition and its results.

3.4.2 Power Laws: Forecasting Energy Consumption

In 2018, Schneider Electric hosted a competition with the objective of forecasting building energy consumption from little data: (a) historical building consumption data, (b) historical weather data and weather forecast for one or a few places geographically close to the building, (c) calendar information, identifying working and off days, and (d) meta-data about the building, e.g., whether it is an office space, a restaurant, etc. More than 200 building sites were considered.

[13] https://www.kaggle.com/c/ashrae-energy-prediction

[14] Clayton Miller et al. (2020) "The ASHRAE Great Energy Predictor III competition: Overview and results," Science and Technology for the Built Environment, DOI: 10.1080/23744731.2020.1795514

3.4.3 Useful observations from the competition results

The competitions showcased that:

- Gradient boosted decision tree models have consistently greater predictive capability compared to alternative modelling approaches.
- Machine learning models can overfit when the dataset includes outliers and discords. A discord is a consumption profile that is very different to all the other profiles in the dataset. In some cases, a discord is the result of outliers in metered data. In other cases, the respective profiles correspond to national holidays and the dataset does not contain enough instances of these holidays. Potential outliers and discords should be removed from the data that is used for training the predictive model – but not from the data that is used for evaluating the model.
- The inclusion of smoothed and lagged values for the outdoor temperature improves the predictive capability of a model. Buildings have different thermal characteristics (envelope U-values and heat capacities) that can introduce a delay between outdoor temperature changes and the resulting changes in heating and cooling energy loads.
- Creating ensembles of different models improves predictive accuracy. Ensemble methods include methods that combine forecasts from different models and methods that combine forecasts from models trained with different subsets of the available data.

Part B: Overview and technical details of proposed methodology

4 Exploratory analysis of the demo datasets

For the presentation of the different steps that compose the proposed methodology for M&V, we have made use of two (2) open datasets for energy building consumption:

Demo building #1 The building with $\mathrm{SiteId} = 50$ that is found in the open dataset of building electricity consumption that Schneider Electric has made available on their Data Exchange[15] and was used in the forecasting energy consumption competition described in **Section 3.4.2**. The granularity of the data is 15 minutes.

Demo building #2 The building with name $\mathrm{FOX\_OFFICE\_ROWENA}$ from the dataset of Building Data Genome Project 2, which is an open data set made up of 3,053 energy meters from 1,636 buildings[16]. The granularity of the data is 1 hour.

Both datasets include data for two (2) consecutive years. For model development, training and fine-tuning, we assume that data is available only for the 1st year. The data for the 2nd year is used only for evaluating the model's performance.

The electricity consumption of the demo building #1 (DB1 hereafter) is presented in Figure 4.1 and the electricity consumption of the demo building #2 (DB2 hereafter) in Figure 4.2.

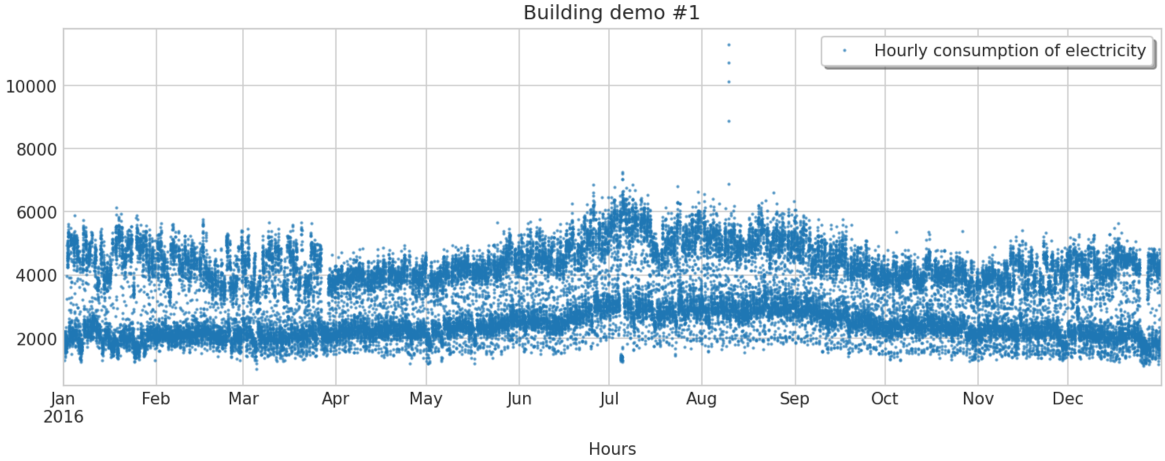


Figure 4.1: The electricity consumption of the demo building #1

[15] https://shop.exchange.se.com/apps/54008/forecasting-building-energy-consumption

[16] https://github.com/buds-lab/building-data-genome-project-2

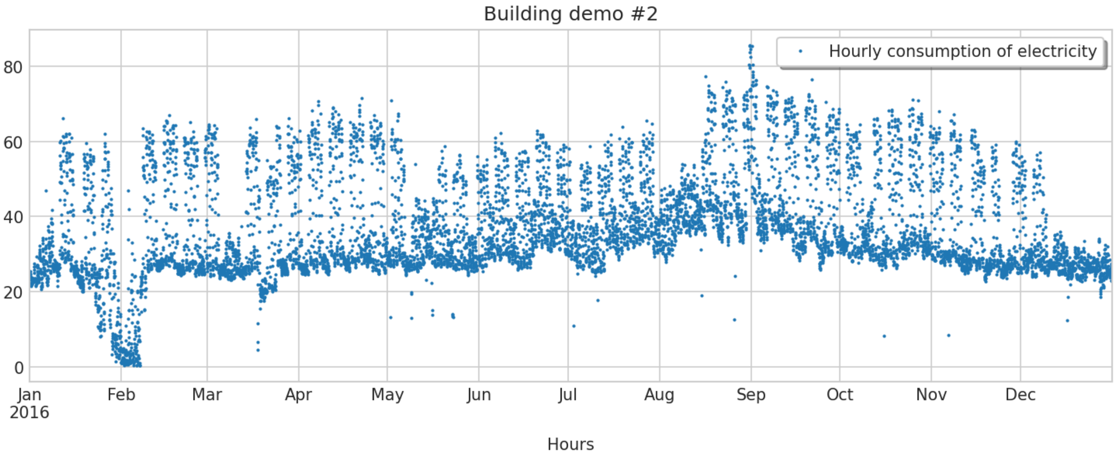


Figure 4.2: The electricity consumption of the demo building #2

In this chapter, we present a series of exploratory analysis steps that can be carried out so that a given dataset is better understood, as well as expectations for the corresponding building's behaviour are formed and, later, used for evaluating the patterns that the baseline energy consumption model uncovered.

4.1 Profiling the daily consumption of a building

Ignoring the differences due to the different seasons of the year, Figure 4.3 presents the median electricity consumption per hour of the day and day of the week for DB1. The plot indicates that, for this building, Sundays are on average different from the other days of the week.

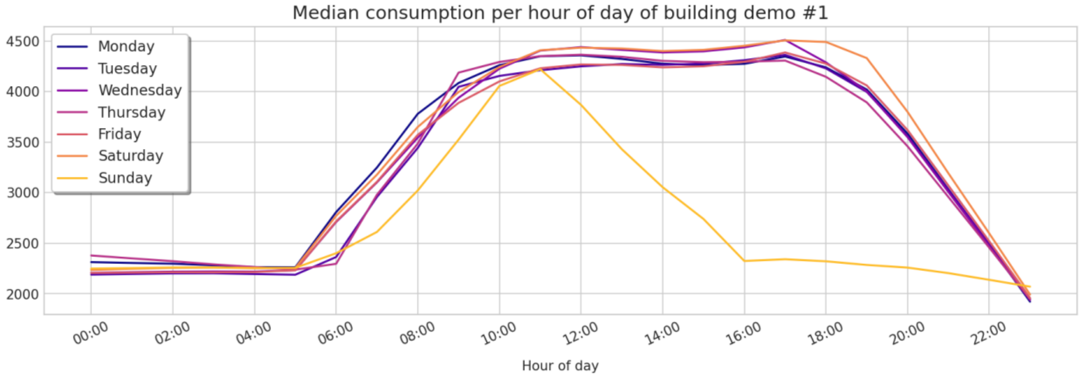


Figure 4.3: The median electricity consumption per hour of day and way of week for building demo #1

Similarly, the plot of Figure 4.4 presents the median electricity consumption per hour of the day and day of the week for DB2. The plot indicates that, for this building, there is a clear difference between weekdays and weekends.

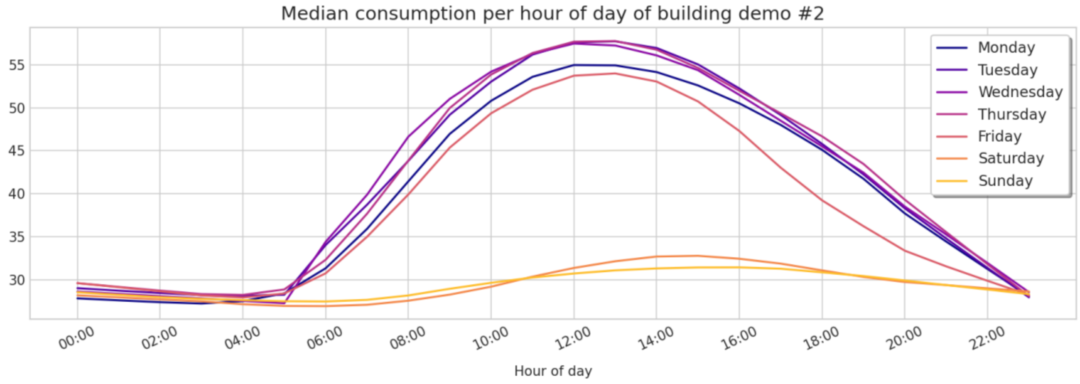


Figure 4.4: The median electricity consumption per hour of day and way of week for building demo #2

The proposed approach for exploratory analysis of a building's energy consumption data utilizes the *symbolic aggregate approximation* (SAX)[17]. The SAX approximation of a time series transforms the original data into symbolic words. The algorithm begins by normalizing the data (so that its mean is zero and its standard deviation is one), breaks it into equally sized segments, and computes the average of each segment. Finally, the algorithm assigns an ordinal value to each segment's average so that all the regions defined by the ordinal values have approximately the same probability.

SAX requires two (2) user-defined parameters: (a) the number of the ordinal values to use, and (b) the number of segments. The number of segments determines the amount of information retained by the SAX approximation. A small number of segments could lead to a very compact representation that may summarize the data too much (filtering out too much information), whereas a very large number of segments could fail in filtering out the noise in the data.

Since there is no effective way to select the SAX parameters in an automated way, SAX is better suited for the interactive and visualization-driven part of the proposed workflow. In contrast, the *day typing approach* that is proposed by this deliverable later on is meant to be utilized in an automated and parameter-light way. The method can be applied on complete daily profiles or on any number of non-overlapping daily intervals, such as for instance: 00:00-08:00, 08:00-16:00 and 16:00-00:00.

If we apply a SAX approximation with five (5) ordinal values and three (3) segments, i.e. one segment per 8 hours, on all the daily consumption profiles of the DB1, we get the categories and the number of days in each one of them as summarized in Figure 4.5 below.

[17] Lin J., Keogh E., Wei L., Lonardi S. (2007) "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15(107)

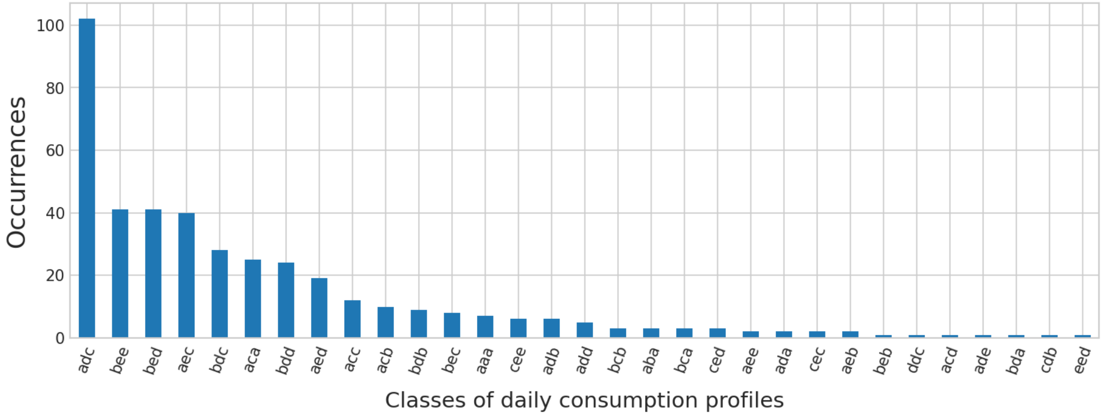


Figure 4.5: The SAX-based categorization of all the daily consumption profiles in demo building #1

Having access to the categories of Figure 4.5, we can interactively filter the daily energy consumption profiles by selecting a category and visualizing the corresponding daily profile subset (Figure 4.6).

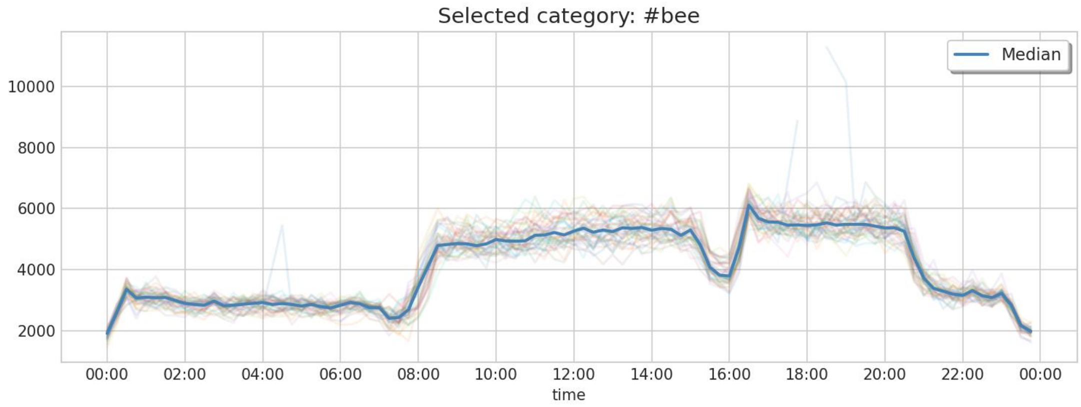


Figure 4.6: Daily consumption profiles of building demo #1 according to selected category

In addition, we can visualize calendars that highlight the days that correspond to each category. The plot in Figure 4.7 suggests that the category of Figure 4.6 includes all days but Sundays, during the June-August period.

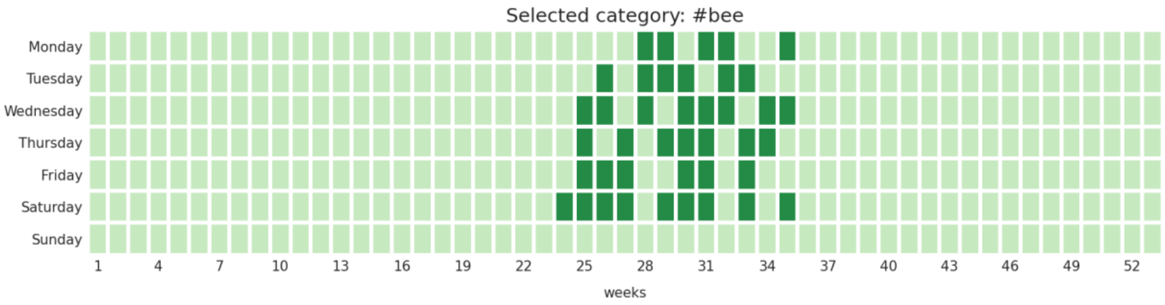


Figure 4.7: Yearly distribution of the selected category

As another example, Figure 4.8 depicts the daily consumption profiles and the yearly distribution for one of the smaller categories found through the SAX approximation. The plot suggests that the consumption profiles in this category correspond predominantly to Sundays, and exclude the May –September period too.

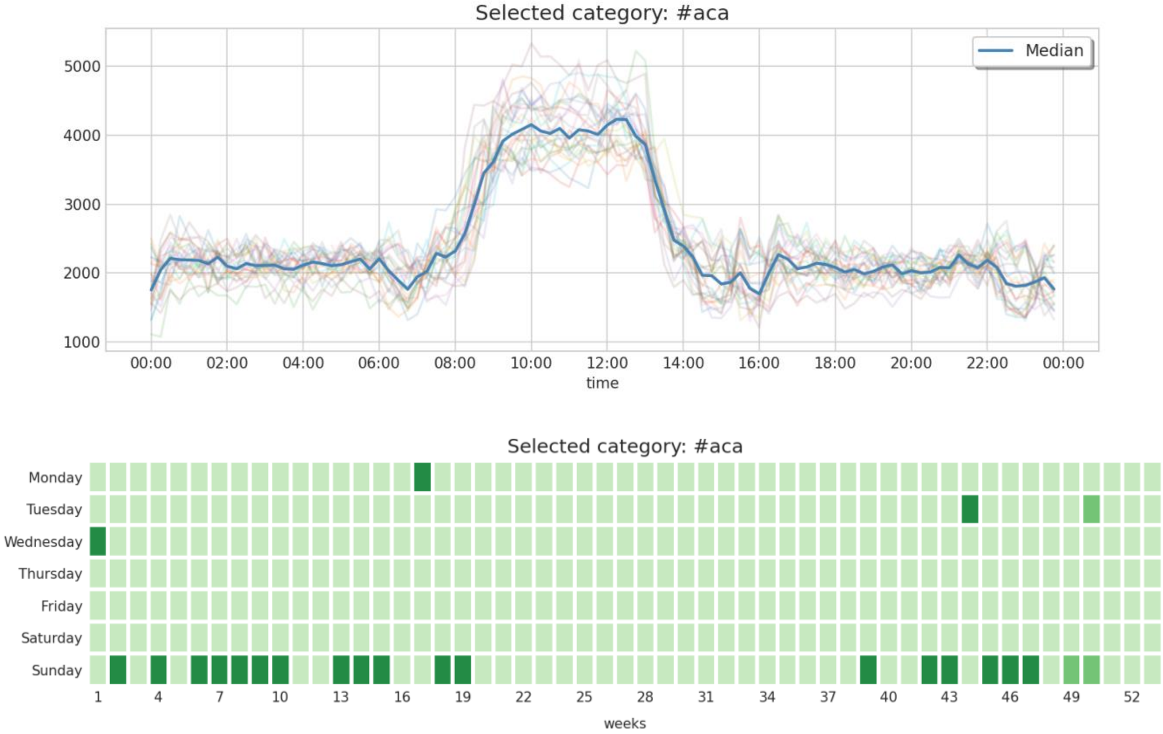


Figure 4.8: Daily consumption profiles and yearly distribution of another category from building demo #1

A similar approximation applied on all the daily consumption profiles of the DB2 leads to the categories and the number of days in each one of them as presented in Figure 4.9.

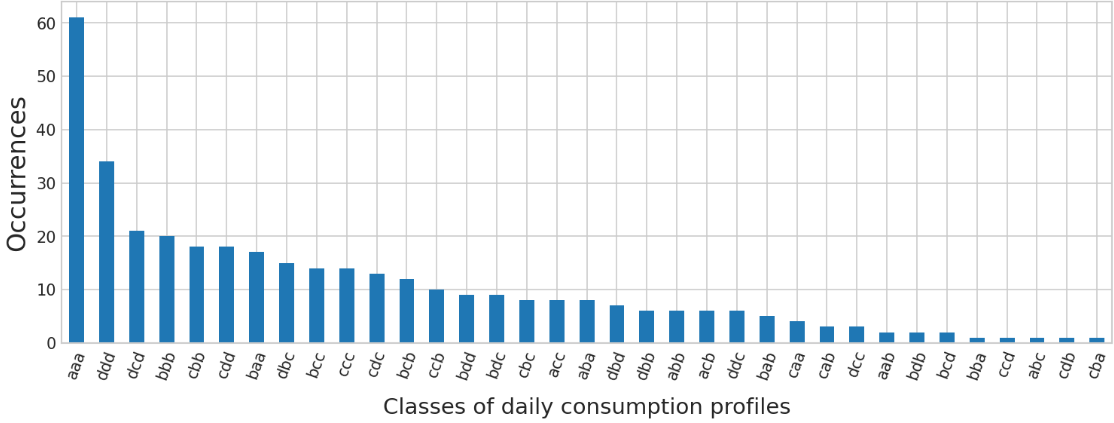


Figure 4.9: The SAX-based categorization of all the daily consumption profiles in demo building #2

Again, we can visualize the daily consumption profiles and the yearly distribution of any selected category (Figure 4.10).

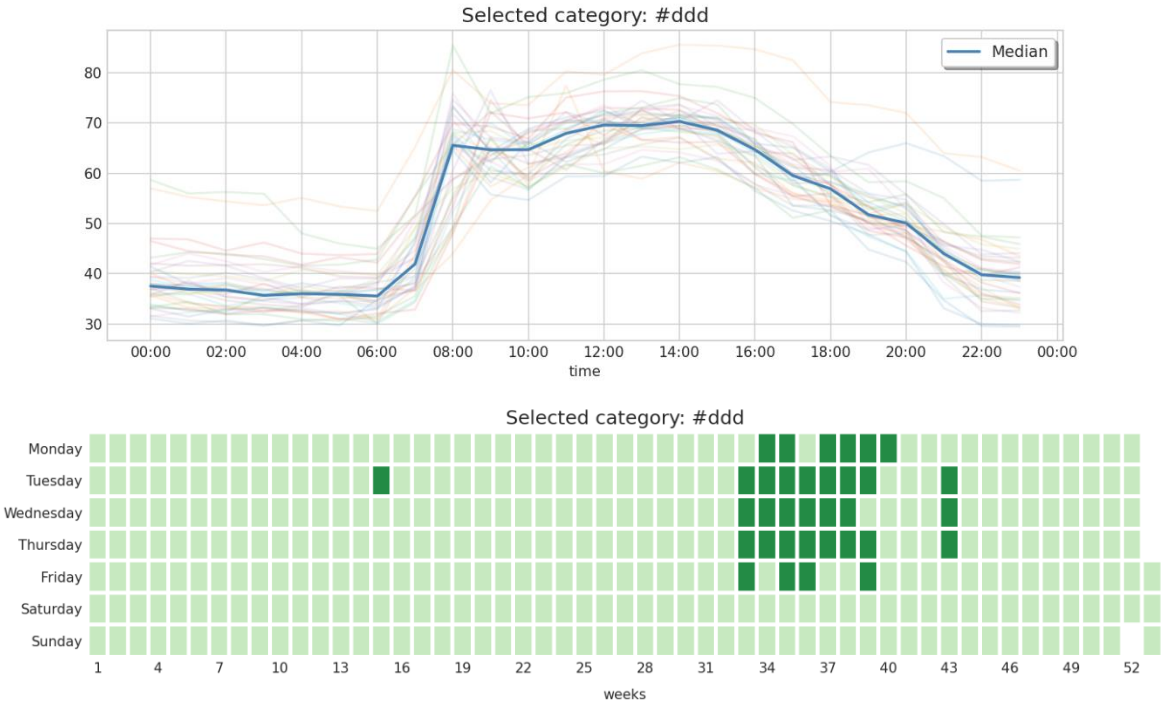


Figure 4.10: Daily consumption profiles and yearly distribution of a selected category from building demo #2

If the selected SAX configuration (number of ordinal values and segments) leads to a large number of categories, hierarchical clustering can be applied to dynamically group the categories into a user-defined number of aggregated ones. As an example, the dendrogram of Figure 4.11 presents the way aggregated categories can be created by merging similar categories together.

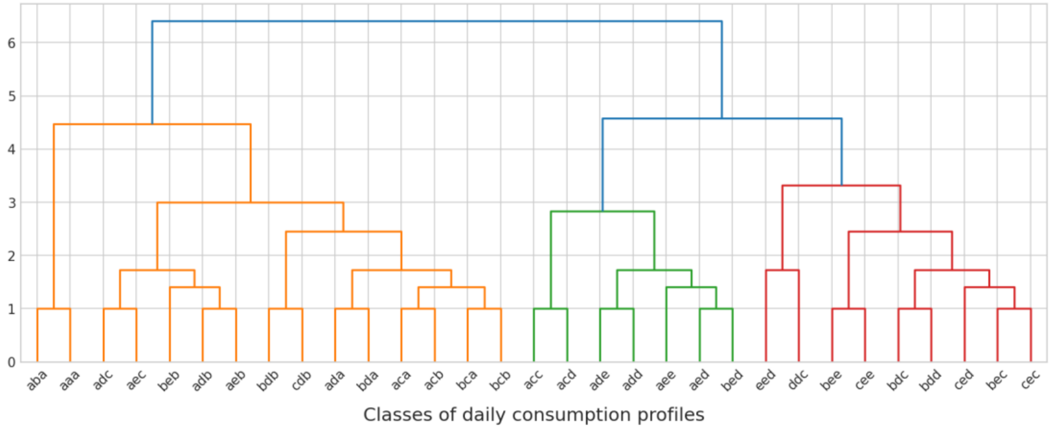


Figure 4.11: Dendrogram of aggregated categories for demo building #1

If we proceed to reduce the number of categories of DB1 to six (6), we get the categories and the number of days in each one of them as summarized in Figure 4.12 below.

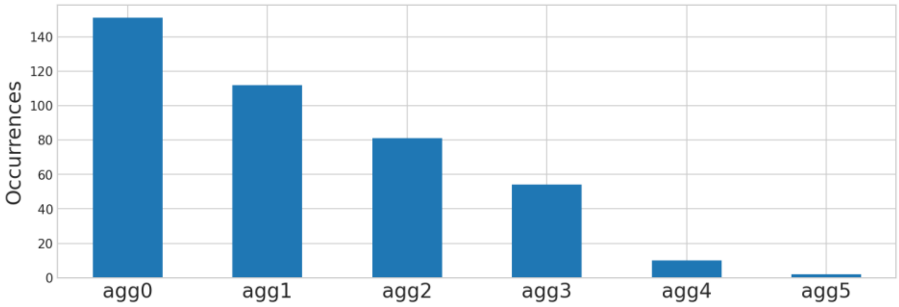


Figure 4.12: Aggregated categorization of all the daily consumption profiles of demo building #1

Similarly to what was mentioned above, one may filter the daily energy consumption profiles by selecting a category and visualizing the corresponding daily profile subset. The plots in Figure 4.13 suggest that:

(a) The 1st largest aggregated category includes all days except Sundays and excludes the summer period;

(b) The consumption profiles in the 2nd largest aggregated category correspond mainly to the summer period;

(c) The consumption profiles in the 4th largest aggregated category correspond predominantly to Sundays.

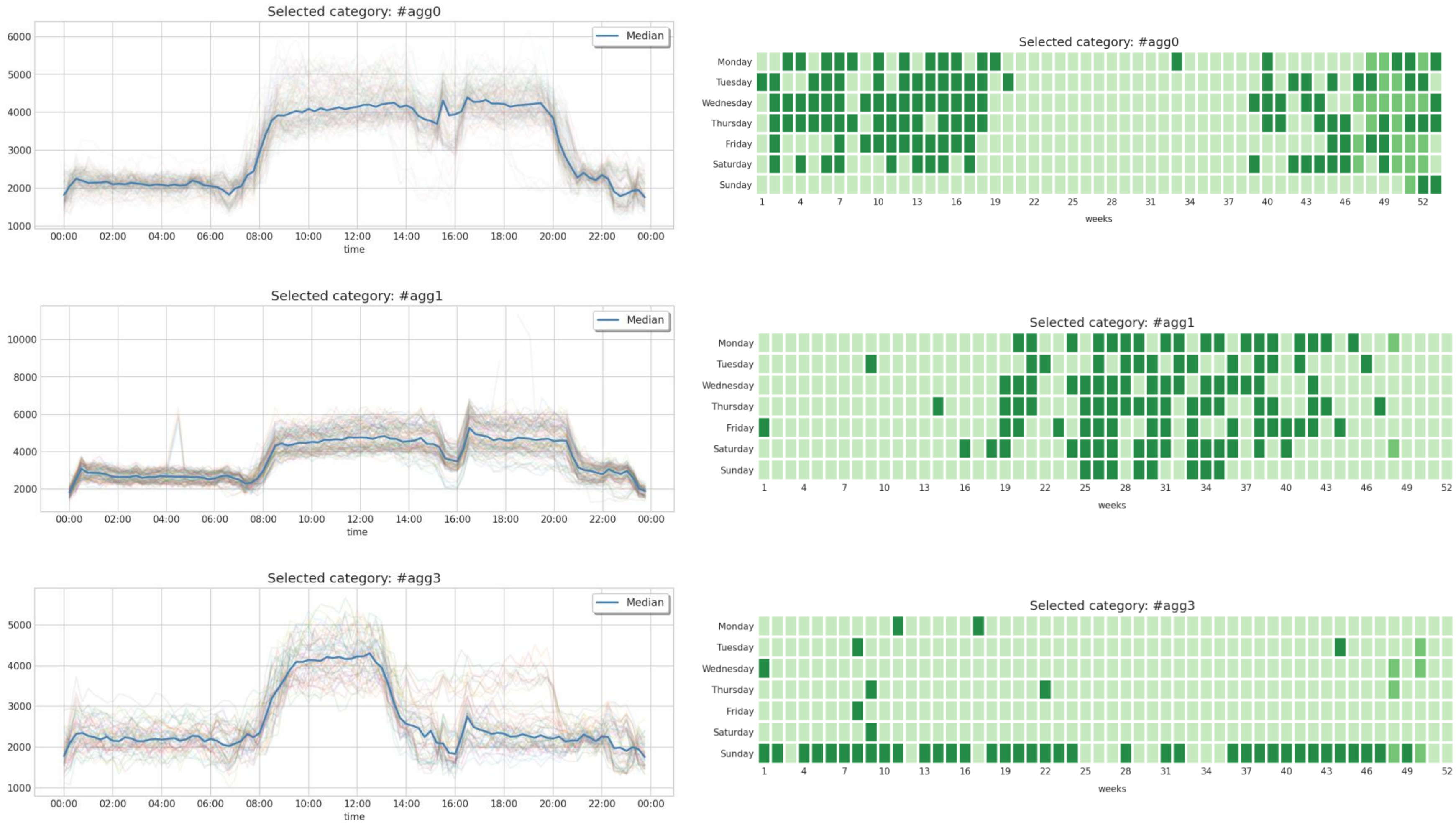


Figure 4.13: Daily consumption profiles of demo building #1 according to selected aggregated category

4.2 Exploring the impact of outdoor air temperature on daily consumption

The plot of Figure 4.14 presents the outdoor air temperature alongside with the electricity consumption of DB1, while the plot in Figure 4.15 presents the outdoor air temperature alongside with the electricity consumption of DB2.

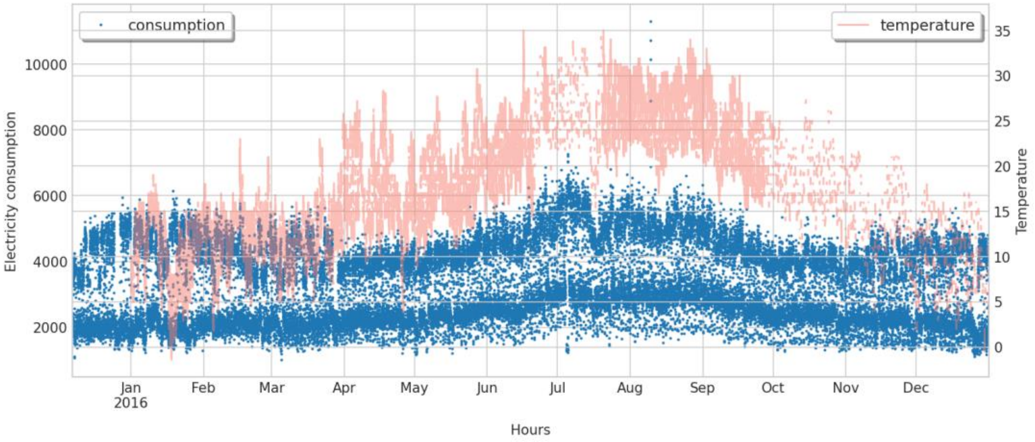


Figure 4.14: The outdoor air temperature data for the demo building #1

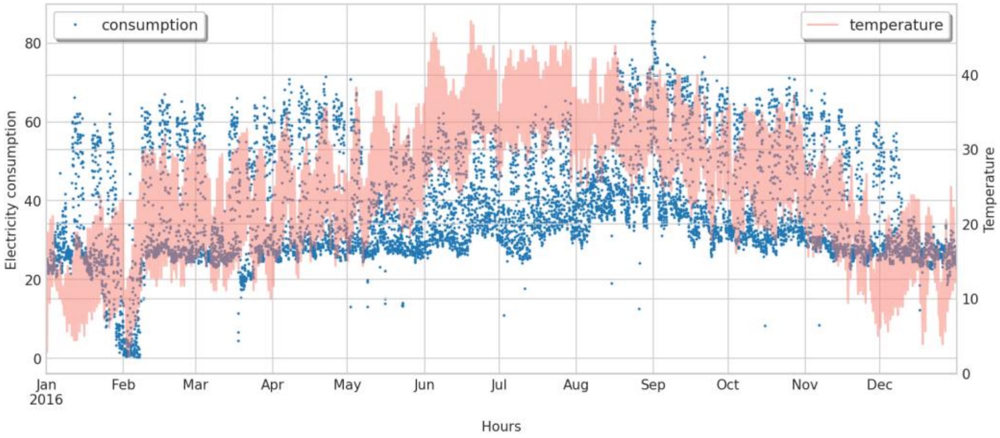


Figure 4.15: The outdoor air temperature data for the demo building #2

An alternative, and probably more informative, way to visualize the relationship between outdoor temperature and energy consumption is to drop the time information and plot consumption as a function of temperature as presented in Figure 4.16 for DB1.

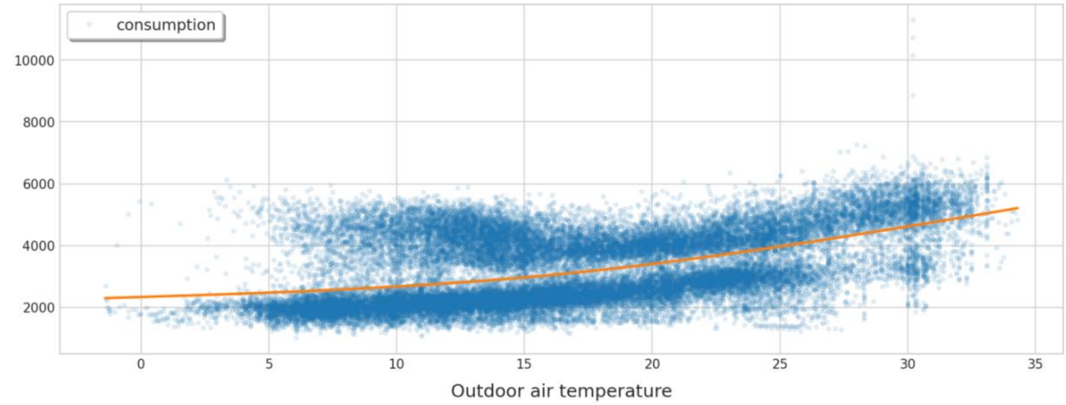


Figure 4.16: Energy consumption as a function of the outdoor air temperature for demo building #1

The continuous curve in the plot of Figure 4.16 has been derived by fitting on the data a natural[18] cubic spline model with three (3) degrees of freedom. It can be seen that the curve splits the dataset into two subsets. As it was discussed in the Section 3.2.1, the TOWT model utilizes this (or a similar) curve to distinguish between hours when the building is (most likely) occupied and hours during which the building is (most likely) unoccupied (Figure 4.17).

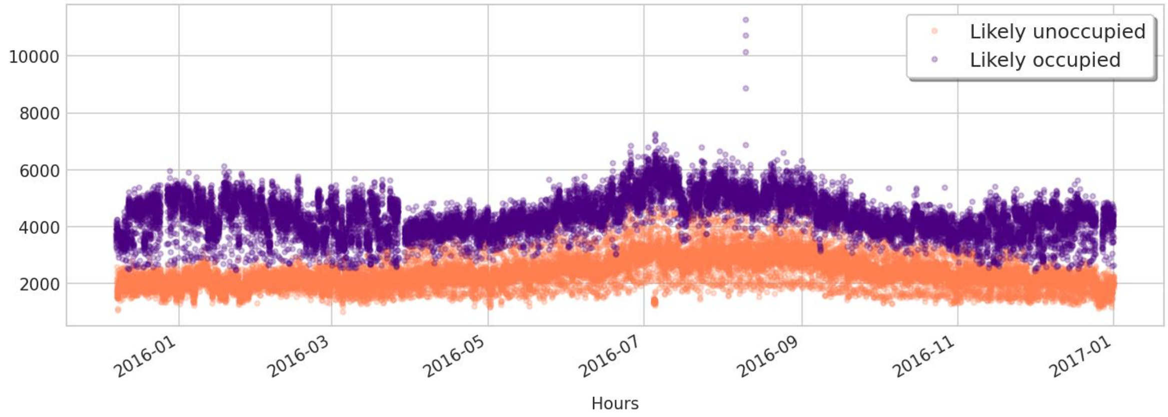


Figure 4.17: Distinguishing between occupied and unoccupied hours in demo building #1

Figure 4.18 shows the distribution of the likely unoccupied and likely occupied hours in the dataset. The plot indicates that the DB1 is generally occupied or in use from 08:00 to 20:00.

[18] The term *natural* means that the second derivatives of the spline polynomial are set equal to zero at the endpoints of the interval of interpolation. This forces the spline to be a straight line outside of the interval.

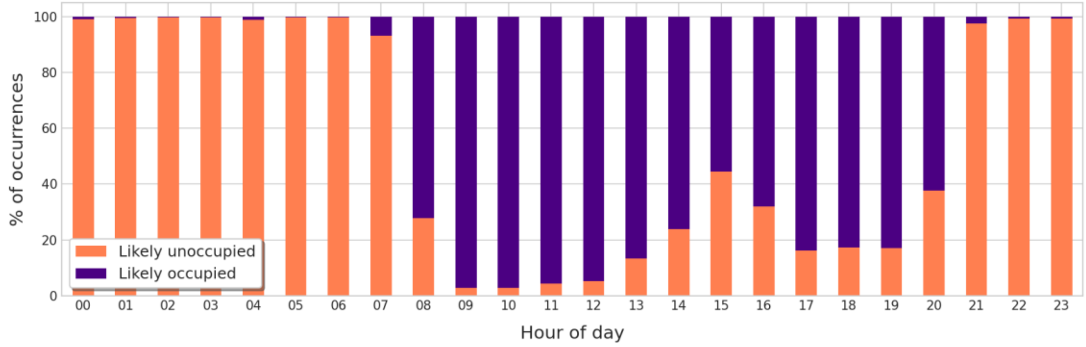


Figure 4.18: The distribution of the high load and low load hours in demo building #1

A supplementary way to visualize the impact of outdoor air temperature on a building's energy consumption is by selectively fixing both the hour of the day and the day of the week. Indicatively, the plot in Figure 4.19 below visualizes the electricity consumption for four (4) specific days of the week (Mondays, Wednesdays, Saturdays and Sundays) and for the 14:00 and 22:00 hours each day for the case of DB1.

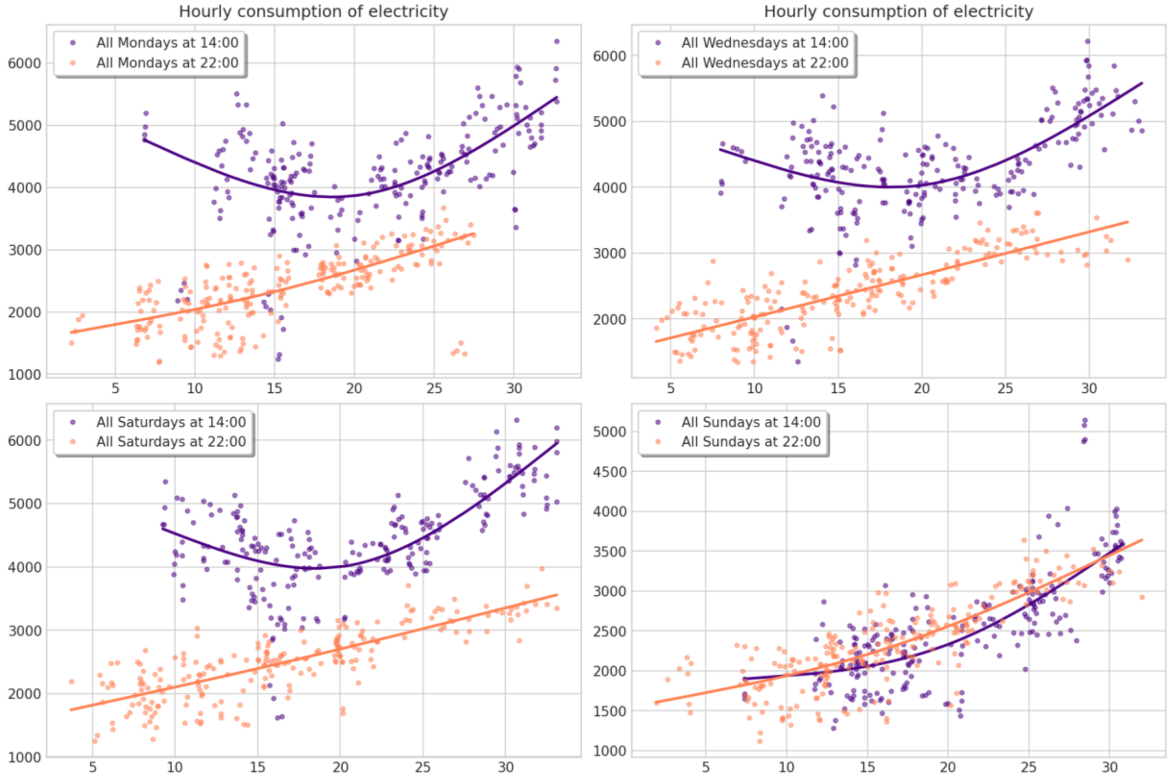


Figure 4.19: The effect of the outdoor air temperature for different hours and days of the week

The plots of Figure 4.19 can serve as a basis for another way of distinguishing between hours when the building is most likely occupied and hours during which the building is most likely unoccupied. In particular, we can estimate one temperature-consumption curve for each combination of hour of the day and day of the week and, then, cluster all the curves according to the similarity of their coefficients[19]. The correspondence of the hours of the day and days of the week to their clusters is depicted in Figure 4.20. The plot indicates that the relationship between electricity consumption and temperature during the hours from 08:00 to 20:00 for all days expect Sundays and from 09:00 to 13:00 for Sundays is different compared to the relationship during the remaining hours and days, which implies differences in occupancy as well.

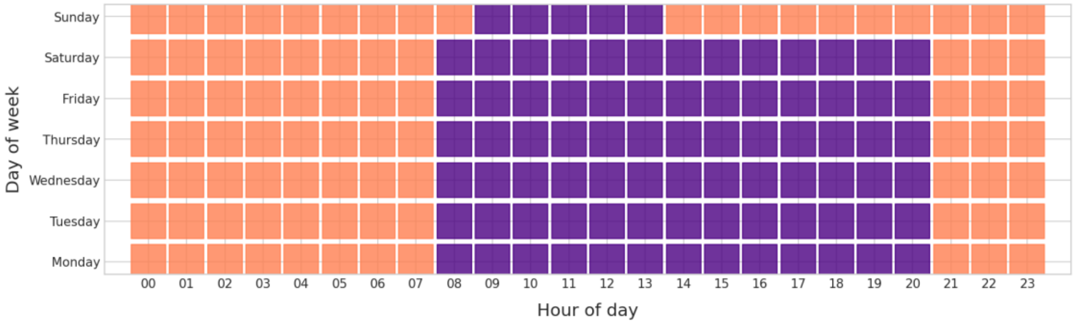


Figure 4.20: The temperature-consumption clusters per hour of day and day of week (2 clusters)

If we aim for three (3) instead of two (2) clusters, we get the correspondence between clusters and the combinations of hours of the day and days of the week of Figure 4.21. The plot indicates that the relationship between electricity consumption and temperature during the start-up and turn-down hours (08:00 and 20:00, respectively) is generally different from the relationship during occupied and unoccupied hours.

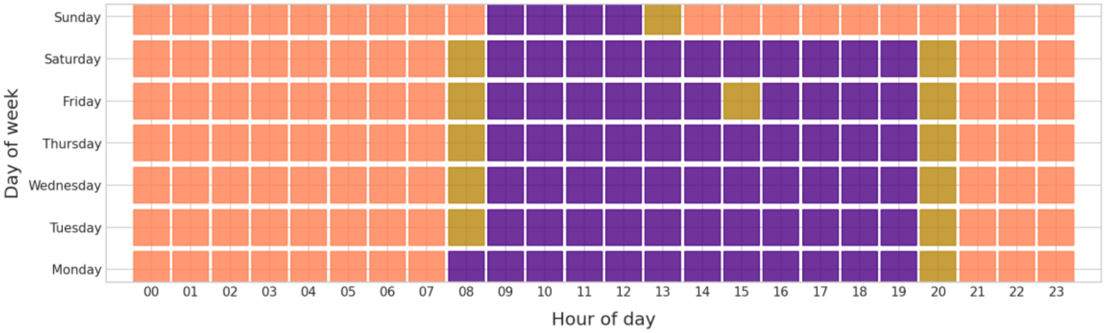


Figure 4.21: The temperature-consumption clusters per hour of day and day of week (3 clusters)

The same process can be applied on the data of DB2 with less satisfying results. The energy consumption as a function of the outdoor air temperature is presented in Figure 4.22. The distinction between the likely unoccupied and the likely occupied hours is less clear than for DB1.

[19] Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003) "Unsupervised Curve Clustering Using B-Splines," Scandinavian Journal of Statistics 30, no. 3, pp. 581-95

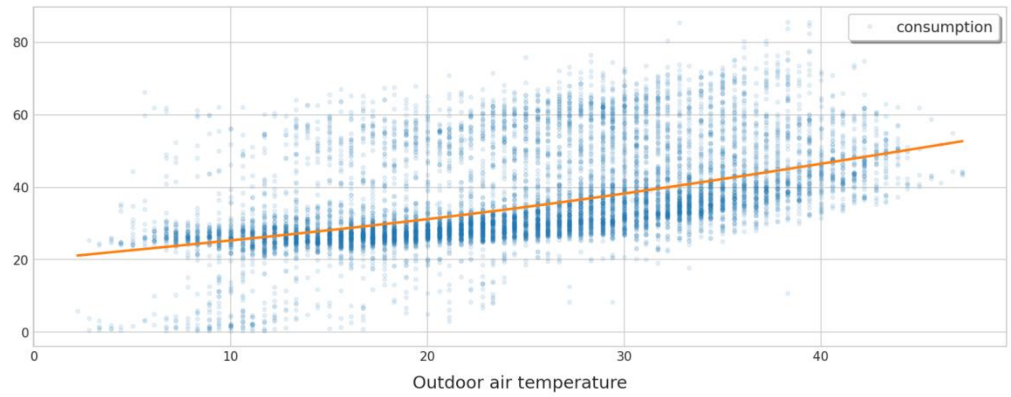


Figure 4.22: Energy consumption as a function of the outdoor air temperature for demo building #2

The distinction between the hours when the building is (most likely) occupied and the hours during which the building is (most likely) unoccupied are presented in Figure 4.23 and Figure 4.24.

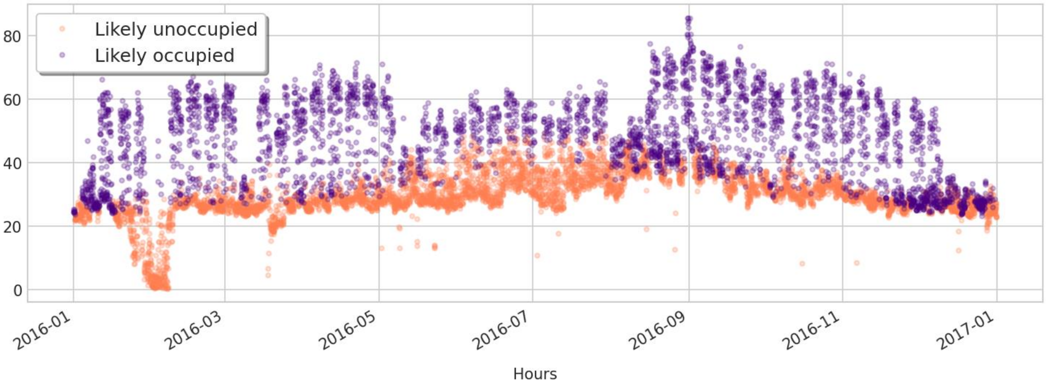


Figure 4.23: Distinguishing between occupied and unoccupied hours in demo building #2

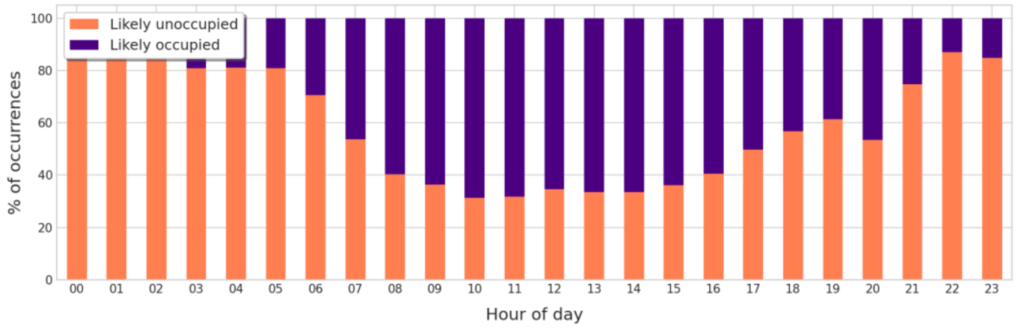


Figure 4.24: The distribution of the high load and low load hours in demo building #2

Although the results of Figure 4.24 are aligned with the results of Figure 4.4, they do not make it easier to distinguish between occupied and unoccupied hours in DB2. Accordingly, and since the methodology proposed in this deliverable aims at targeting as many types of buildings as possible, we do not make use of this approach for formulating the temperature-related component of the proposed M&V model. Instead we rely on a non-linear (gradient boosted tree model) to capture the interactions between the outdoor temperature, the day of the week and the hour of the day. The interactions that the model has actually identified can be uncovered through a process that is similar to the way the plots of Figure 4.19 were generated: selectively fixing both the hour of the day and the day of the week, and summarizing the predictions of the model for different temperature values.

5 The SENSEI workflow for data preprocessing

5.1 The goal of the preprocessing pipeline

The data preprocessing pipeline that is presented in this chapter assumes that energy consumption and outdoor air temperature is the only data that is available for M&V. We consider an energy consumption and outdoor air temperature dataset validated if:

(1) **There are no duplicate values in the dataset's timestamps**. Duplicate timestamps are treated separately for energy consumption and for temperature data. In both cases, if the range of the energy consumption or temperature values that share a timestamp is short – according to a user-defined threshold – they are replaced by their average. Otherwise, they are treated as missing values.

(2) **There are no missing values in the dataset's timestamps**. If there are missing timestamps, they are added and the respective data is treated as missing values.

(3) **Potential outliers are identified and marked**. Outlier detection is carried out separately for energy consumption and for temperature data.

(4) **There is enough data available for the energy consumption of the building under study**. Baseline energy consumption data must cover at least one full year before any energy efficiency intervention. In addition, and adopting the data requirements of the CalTRACK set of methods, data must be available for over 90% of hours in each calendar month – ***after excluding the potential outliers***.

(5) **There are no missing values in the outdoor air temperature data**. If temperature data is missing, the missing values are imputed. The outdoor air temperature changes smoothly from one hour to the next, so interpolating over a 6-hour window around a missing observation is a sensible approach for imputation. This is in line with CalTRACK's requirement that temperature data may not be missing for more than six (6) consecutive hours.

(6) **There are no missing values in the energy consumption data**. Missing values for energy consumption data do not pose a problem when training the predictive baseline model but they may lead to daily consumption profiles that are mistakenly regarded as unusual when we search for common and uncommon patterns in the data. Accordingly, the proposed workflow imputes the energy consumption data for the identification of patterns in the daily consumption, but ***does not include*** the imputed values in the dataset that is used for the predictive model's training.

The steps of the preprocessing pipeline are summarized in Figure 5.1 below.

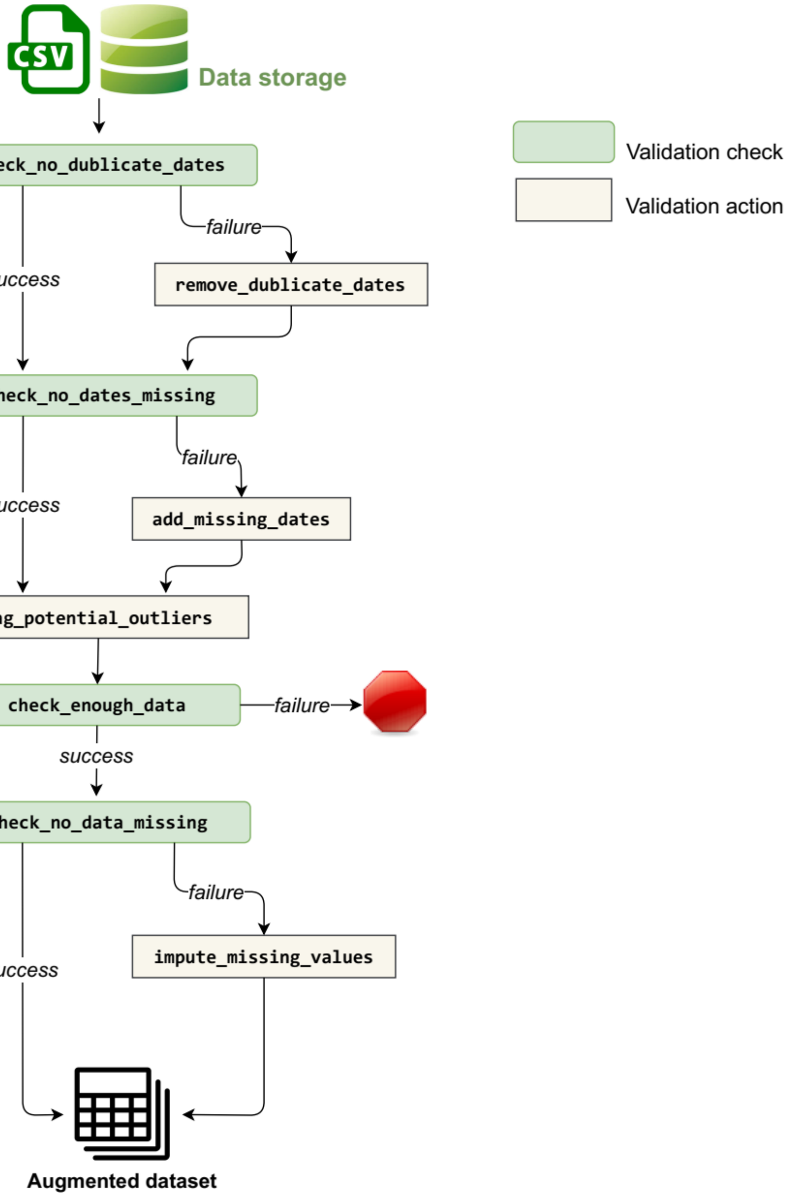


Figure 5.1: The steps for screening and validating the available data

5.2 Outlier identification

The proposed approach for outlier identification is outlined next:

Step 1: Global filter. The first step screens for non-physically plausible values as well as unlikely values in the data. For power consumption data, negative and zero values are filtered out. For both consumption and temperature data, values that are at least 10 times larger than the median value are also removed. The threshold of ten times the median value aims at removing the most extreme outliers. Furthermore, consecutive occurrences of constant values are filtered out as well.

Step 2: Seasonal filter. The second step captures the seasonal cycle of the data through a trend and seasonality decomposition approach that utilizes a Fourier series expansion of the form:

$$y(t) = a + bt + \sum_{n=1}^{N} \left(\alpha_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{5.1}$$

where:

a The offset of the linear trend

b The slope of the linear trend

t The day since a pre-specified epoch. For hourly data, t will take decimal number values.

N Parameter that controls the flexibility of the expansion. Suggested values are[20] $N = 4$ for daily seasonality, $N = 10$ for yearly seasonality

P The length of the seasonality: $P = 1$ for daily seasonality, $P = 365.25$ for yearly seasonality. For energy consumption data, we fit a different daily seasonality component for each day of the week.

α_n, b_n Regression coefficients for the Fourier series expansion terms.

The reason for applying seasonal decomposition before outlier identification can be seen in Figure 5.2. The upper panel shows the multimodal distribution of the power consumption of the DB1, alongside with a Normal distribution that has been fitted on the data, while the lower panel corresponds to the same aspects of DB2.

Since seasonality leads to multimodal distributions, methods that rely on the assumption that the data follows a Normal distribution – such as simple three-sigma rules, the Grubbs test[21] or the Extreme Studentized Deviate (ESD) test[22] – should be used ***only after*** a seasonal filter has been applied to the data.

[20] Taylor S. J. and Letham B. (2018) "Forecasting at scale," The American Statistician 72(1), pp. 37-45

[21] Frank E. Grubbs (1969) "Procedures for detecting outlying observations in samples," Technometrics, 11(1), pp. 1-21

[22] Bernard Rosner (1975) "On the detection of many outliers," Technometrics, 17(2), pp. 221-227

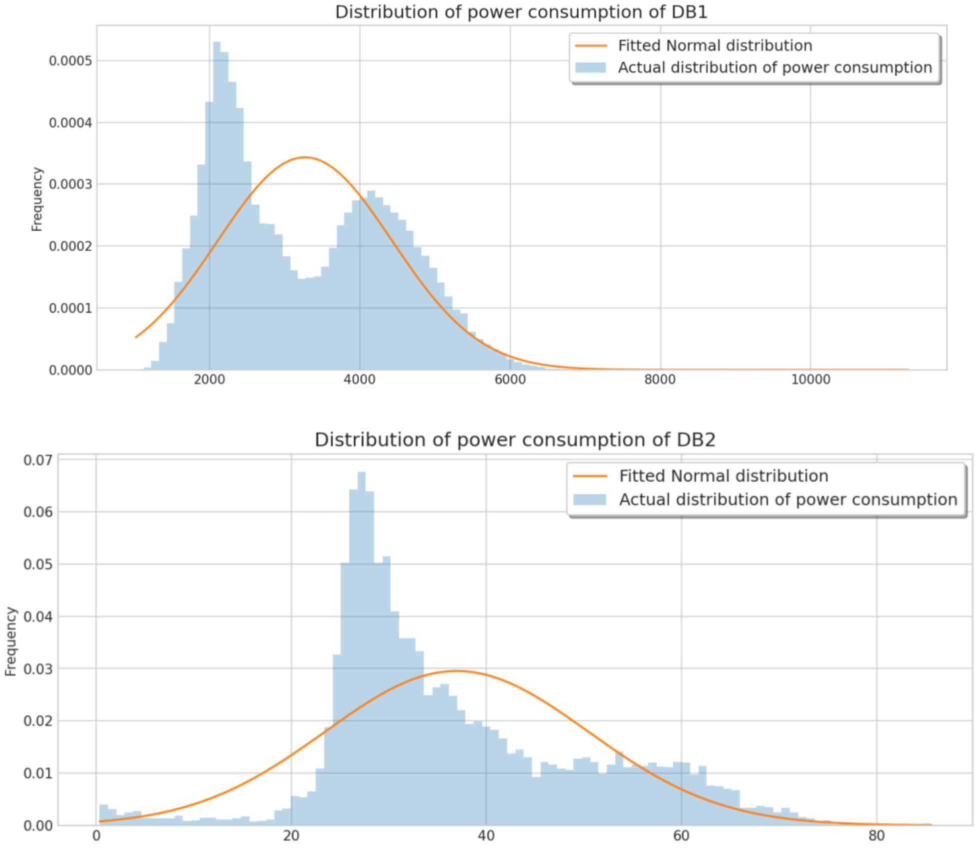


Figure 5.2: The multimodal distribution of the power consumption

The plot in Figure 5.3 shows the actual and the predicted power consumption of the DB1 for the first two (2) months of 2016, when applying the aforementioned modelling approach.

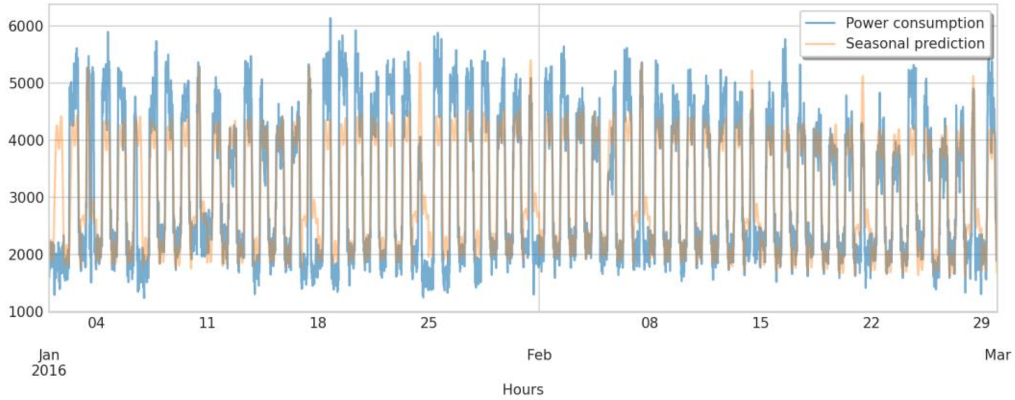


Figure 5.3: The actual and predicted power consumption of the demo building #1

The plot in Figure 5.4 presents the distribution of the residuals when subtracting the actual from the predicted power consumption for DB1. The distribution of the residuals resembles a Student's t distribution and, hence, it is easier to work with for detecting outliers.

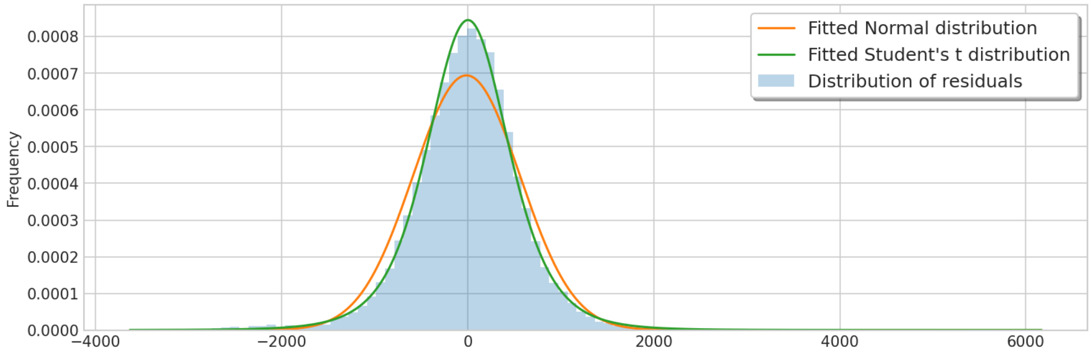


Figure 5.4: The benefit from applying seasonal decomposition before outlier detection

Step 3: Global outlier detection. The third step of the outlier detection process identifies observations in the available dataset as potential outliers if the value of their corresponding residuals lies outside the range defined by:

$$[mean^{\,all} - c \times scale^{\,all},\ \ mean^{\,all} + c \times scale^{\,all}] \tag{5.2}$$

where:

$mean^{\,all}$ The mean of a Student's t distribution fitted on all the residual values

$scale^{\,all}$ The scale of a Student's t distribution fitted on all the residual values

c User defined parameter (suggested value is 4).

The plot in Figure 5.5 shows the potential outliers identified in the seasonal estimation residuals of the DB1 using the global outlier detection approach, while the plot in Figure 5.6 shows the potential outliers identified in power consumption for January 2016.

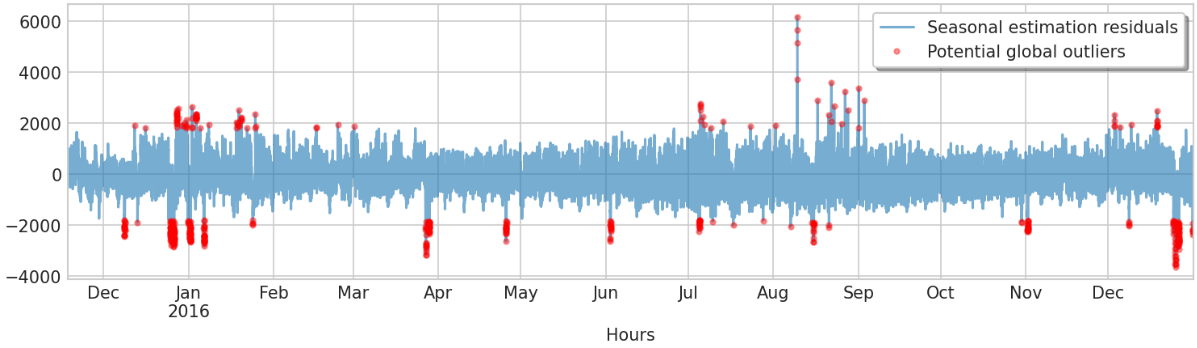


Figure 5.5: The potential outliers identified using the global outlier detection approach

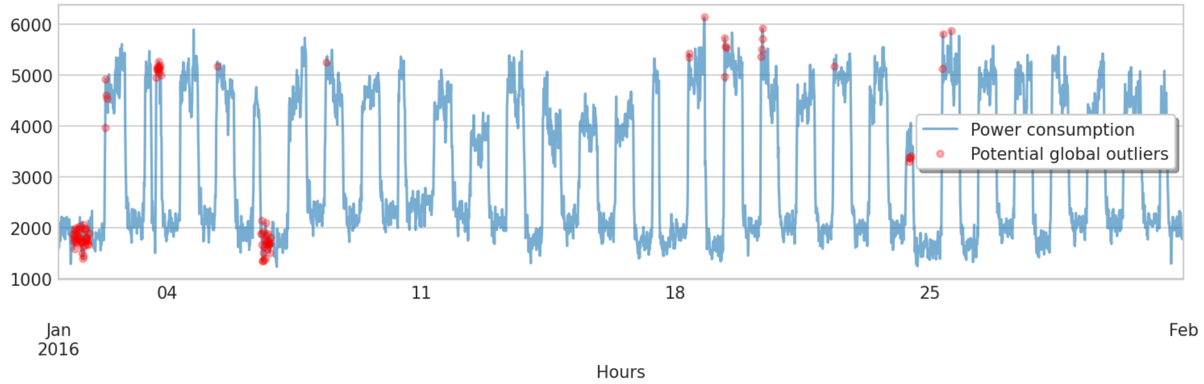


Figure 5.6: The potential outliers identified for January 2016 using the global outlier detection approach

Step 4: Local outlier detection. The final step of the outlier detection process retains from the outliers identified in the previous step only those that can be characterised as outliers when we also compare their values with the observations in the same day of the year.

The rationale for this approach can be explained by looking at the plot in Figure 5.7, which shows the actual and the predicted power consumption during the first two (2) weeks of 2016 in the dataset of DB1. An important observation from the plot is that the distance from the seasonal model's predictions is not by itself enough for detecting outliers when the whole day is misrepresented by the model (here a holiday is treated as a normal day).

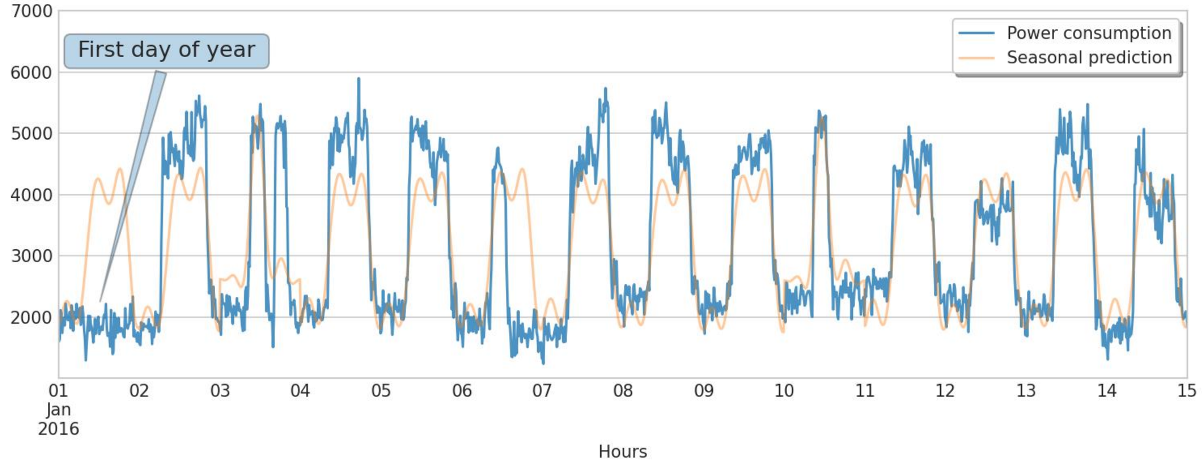


Figure 5.7: The actual and predicted power consumption during the first two weeks of 2016

Accordingly, the observations in the available dataset are marked as potential outliers if the value of their corresponding residuals lies outside the range defined by:

$$[median^{day} - c \times mad^{day},\ \ median^{day} + c \times mad^{day}] \tag{5.3}$$

where:

$median^{day}$ The median of all the residual values in the corresponding day

mad^{day} The median absolute deviation of all the residual values in the corresponding day

c User defined parameter (suggested value is 4).

This step is parameterised by the minimum number of observations that must be available for any given day so that to take the daily statistics into account. If the number of the available observations is lower than this threshold, only the global outlier detection results are considered.

The plot of Figure 5.8 shows the potential outliers identified in the power consumption dataset of the DB1 using the local outlier detection approach.

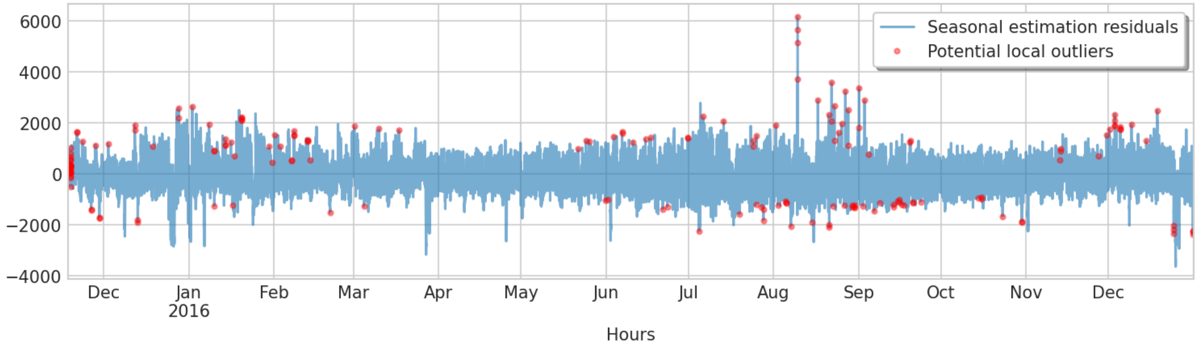


Figure 5.8: The potential outliers identified using the local outlier detection approach

For an observation to be marked as an outlier, both global and local results must agree. The plot of Figure 5.9 shows the potential outliers for the DB1 when combining the global and local results.

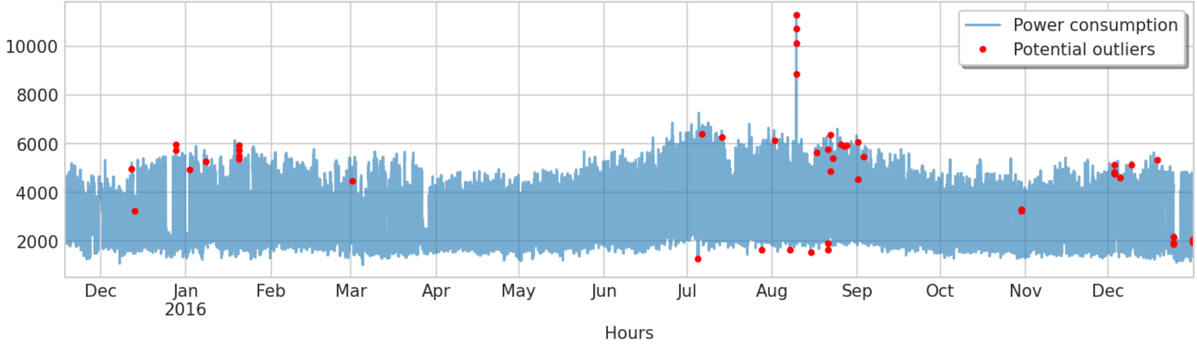


Figure 5.9: The potential outliers identified in the power consumption dataset

6 The SENSEI workflow for day typing

All the predictive models for estimating the baseline energy consumption of a building work by exploiting the daily, weekly and yearly seasonality of the consumption data. Some models make the assumption that this seasonality can be identified across the whole dataset, whereas others assume that each calendar month is so distinct from the others that it should have its own monthly model. In contrast, the proposed approach makes no a priori assumptions regarding the daily, weekly and yearly similarities between the different observations in the dataset. Instead, it relies on the day typing stage to distinguish all the days in the available dataset into different categories according to the shape and scale of their energy consumption profiles.

The goal of the day typing stage is to exploit similarities in the energy consumption profile of a building so as to increase the predictive accuracy of the model for the baseline energy consumption and decrease risk of overfitting during the baseline consumption estimation. The day typing method categorizes a building's daily or sub-daily consumption profiles according to their similarity to a small number of recurring patterns (prototypes) that can be found in the consumption time series. The key idea behind the proposed method is to identify a small number of recurring patterns that are very dissimilar to each other and can be used as points of reference for comparing all the remaining consumption profiles found in the dataset.

6.1 Definitions

The proposed methodology for day typing builds upon the distance and matrix profile data structures[23]. First, some necessary definitions:

Time series A time series T of length n is a time-ordered sequence of real numbers $[x_1, x_2, \dots, x_n]$.

Subsequence A subsequence $T_{i,m}$ of the time series T is a contiguous subset of T starting at position i with length $m < n$:

$$T_{i,m} = \left[x_i, x_{i+1,\dots,}x_{i+m-1}\right]$$

Distance profile A distance profile $D \in \mathbb{R}^{n-m+1}$ between a time series T of length n and a query time series $T^{(q)}$ of length m is another time series that stores the (normalized) Euclidean distance between $T^{(q)}$ and each possible subsequence $T_{i,m},\ i \in [1,2,\dots,n-m+1]$ of T. By definition, the distance

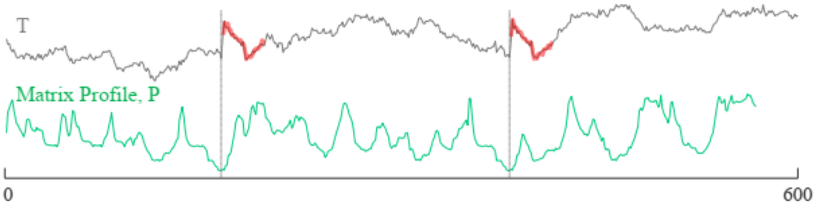
[23] Chin-Chia Michael Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh (2016) "Matrix profile I: All pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets," in 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, pp. 1317–1322

profile takes values close to zero at the location of the subsequences $T_{i,m}$ that are very similar to $T^{(q)}$.

Nearest Neighbour The nearest neighbour of a subsequence $T_{i,m}$ is the subsequence that has the smallest distance from it (the subsequence that is most similar to it).

Matrix profile A matrix profile $P \in \mathbb{R}^{n-m+1}$ of a time series T is another time series that stores at each position i the distance between the subsequence $T_{i,m}$ and its nearest neighbour. If a specific subsequence has a matrix profile value far greater than zero, it is unlike any other subsequence in the dataset (a discord), whereas if it has a close to zero value, it is a repeated pattern (a motif).

The following plot[24] depicts the matrix profile of an example time series. Low values correspond to repeated patterns.



Matrix profile index A matrix profile index of a time series T is a vector that stores at each position i the index of the nearest neighbour of the subsequence $T_{i,m}$. In other words, it stores the starting position of the subsequence that is most similar to $T_{i,m}$.

6.2 The matrix profile as a measure of energy consumption predictability

If we calculate the coefficient of variation (i.e. the standard deviation divided by the mean) of the energy consumption of DB1 and DB2, we get similar results: 0.35 and 0.36. In this sense, one could argue that these two buildings are equally easy or difficult to predict. However, a different way of evaluating the predictability of these two buildings is by looking at the matrix profile values of their energy consumption.

The diagram in Figure 6.1 presents the matrix profile values of all the subsequences in the consumption data of DB1 that: (a) have length m that corresponds to one day (this means that $m = 24$ for hourly data or $m = 96$ for 15-min data), and (b) start at 00:00 hours, so that to only compare subsequences that span a full day's period. The plot of Figure 6.1 corresponds to the matrix profile of

[24] Chin-Chia Michael Yeh, Towards a Near Universal Time Series Data Mining Tool: Introducing the Matrix Profile, University of California, Riverside, USA, 2018

a very well behaved building. Most values are low and only a few peaks are present. This implies that a lot of daily consumption profiles are similar to each other, and only a few of them are very different. The coefficient of variation of the matrix profile values is 0.21.

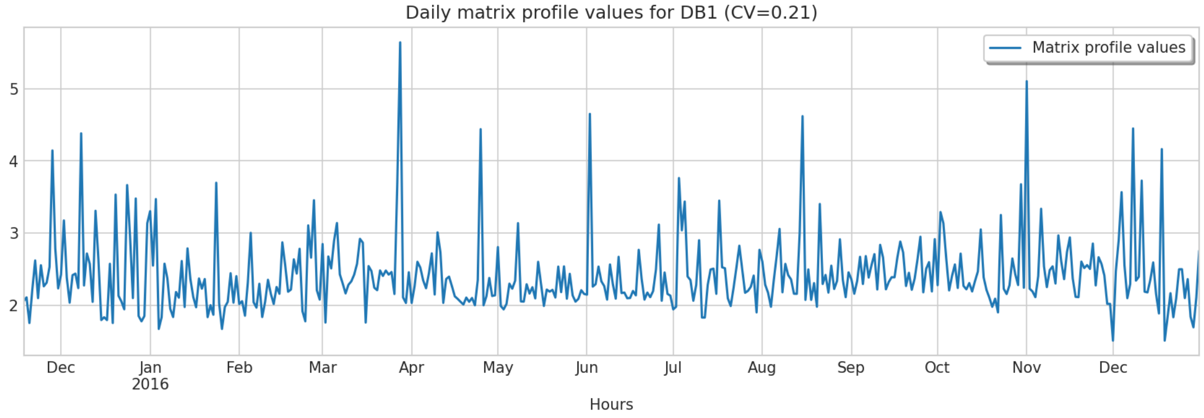


Figure 6.1: The daily matrix profile for the demo building #1

In contrast, the plot of Figure 6.2 indicates a much more challenging building to work with. It is the matrix profile values of the consumption data of DB2. The coefficient of variation of the matrix profile values is 0.65, which is three time higher than the one for DB2.

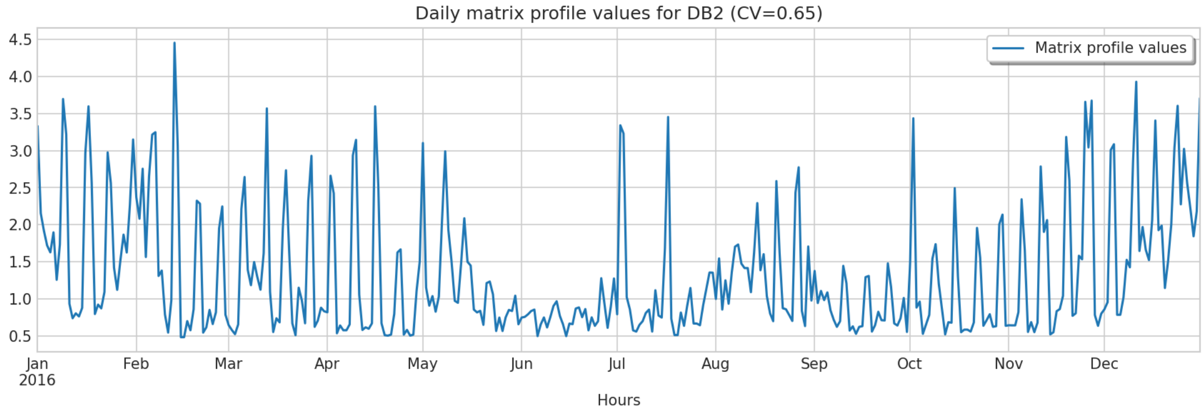


Figure 6.2: The daily matrix profile for the demo building #2

6.3 Categorization of a building’s consumption profiles

In this subsection, we present a method for the categorization of a building’s consumption profiles according to their similarity to a small number of recurring patterns that can be found in the available dataset. The key idea behind the proposed method is to identify a small number of recurring patterns that are very dissimilar to each other and can be used as points of reference for comparing all the remaining consumption profiles found in the dataset.

However, categorizing load profiles is useful for exploratory purposes but not for prediction. Accordingly, the proposed method aims at finding a way to rearrange the calendar structure of the building's data consumption so that similarity in load profiles can be translated into a similarity measure for time-based features that are only available during prediction time, such as the day of the week and the month of the year. The quality of the similarity measure can be evaluated, so that if it is low – i.e. the method fails to identify consistent rules for mapping recurrent profiles to specific times of the year – the practitioners can fall back to a different M&V approach that uses the existing calendar structure of the data.

The method can be applied on complete daily profiles or on any number of non-overlapping daily intervals. For demonstration purposes, the electricity consumption of the DB1 is split the into three (3) non-overlapping 8-hour intervals: 00:00-08:00, 08:00-16:00 and 16:00-00:00, and then the proposed method is applied on each and every interval, while for DB2, the method is applied on the complete daily profiles.

6.3.1 Demo building #1

Identify a small number of recurring patterns that are very dissimilar to each other

The daily consumption profiles of the DB1 are presented in the plot of Figure 6.3. It is possible to visually spot at least two different daily patterns: one where the high-load period spans from 09:00 to 20:00, and one where the high-load period spans from 09:00 to 13:00.

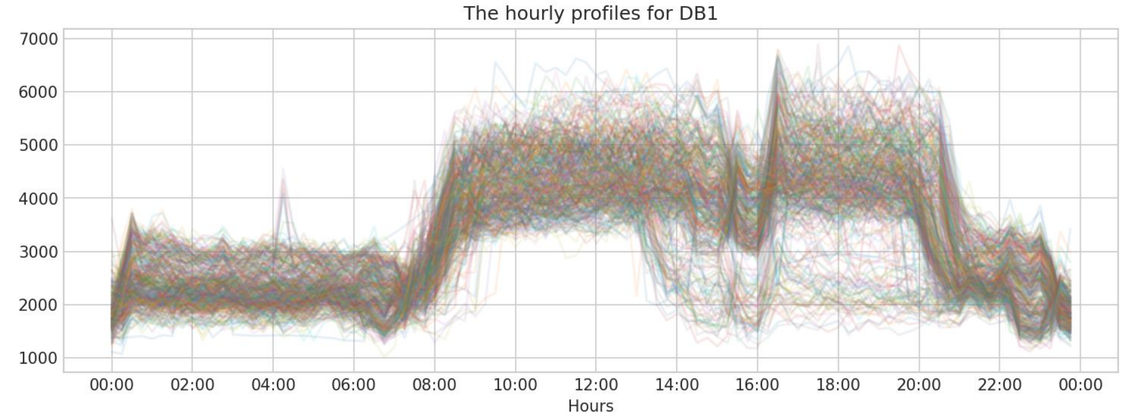


Figure 6.3: The consumption profiles of the demo building #1

The day typing process begins by calculating the matrix profile of the energy consumption data for a subsequence length m that corresponds to 8 hours, and isolating the matrix profile values of all the subsequences that start at 00:00 hours. Then, the reference patterns are identified through the sequence outlined next:

Step 1: Find the minimum value in the matrix profile. The proposed approach builds on the fact that values of the matrix profile that are close to zero indicate a repeated pattern. The subsequence with the smallest value in the matrix profile of all the subsequences that start at 00:00 hours is presented in Figure 6.4. This is the 1st selected reference pattern (prototype).

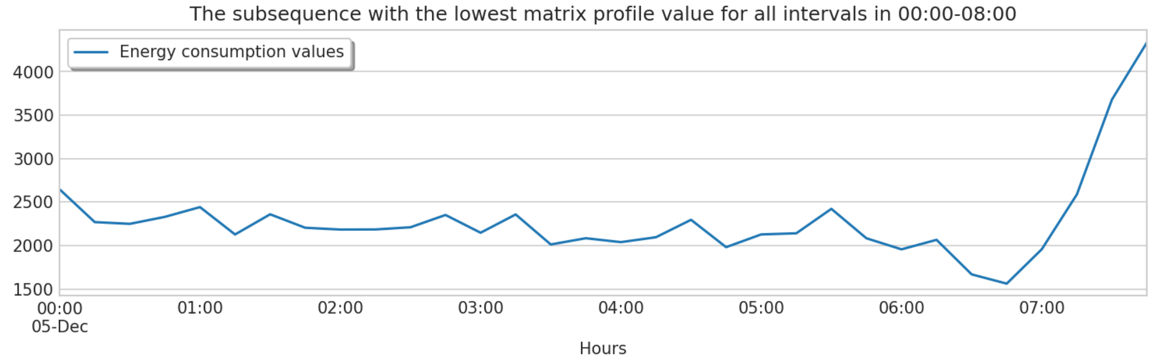


Figure 6.4: The 00:00-08:00 subsequence with the lowest matrix profile value

Step 2: Compute the distance of the selected reference pattern with respect to all the subsequences in the time interval under study. The result measures how similar each 00:00-08:00 subsequence in the time series is to the currently selected pattern (i.e. the one depicted in Figure 6.4 above). The fact that we only care about subsequences that start at 00:00 hours protects us from trivial matching. Trivial matching renders pattern detection meaningless[25], so when comparing two subsequences they should not match if they overlap. We store the distances from the first selected pattern into an array (denoted as $dist\_mp$).

Step 3: Compute a *relative* matrix profile by dividing the original matrix profile data with $dist\_mp$. The intuition is that if a subsequence has a very close nearest neighbour (i.e. a very low matrix profile value), but is very different from the selected reference pattern, then its value in the relative matrix profile should be also very low since when we compute the relative matrix we divide an already small number by a large one (Zhu et al. 2020)[26].

Step 4: Find the 8-hour subsequence that corresponds to the minimum value in the relative matrix profile. This is the 2nd selected pattern.

[25] Keogh E. and Lin J. (2005) "Clustering of time-series subsequences is meaningless: implications for previous and future research," Knowledge and Information Systems, vol. 8, no. 2, pp. 154–177

[26] Zhu, Y., Gharghabi, S., Silva, D.F. et al. (2020) "The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code," Data Mining and Knowledge Discovery 34, pp. 949-979, https://doi.org/10.1007/s10618-019-00668-6

Step 5: Compute the distance of the new reference pattern with respect to the consumption time series, and update the $dist\_mp$ array with the element-wise minimum between $dist\_mp$ and the aforementioned distance. In other words, we use $dist\_mp$ to store the distance between every subsequence and its closest match among all the patterns selected so far.

Step 6: Recalculate the relative matrix profile by dividing the original matrix profile data with the updated $dist\_mp$ array, and repeat from Step 4.

In order to filter out redundant patterns, we want to keep minimizing at every step of this iterative process the squared maximum mean discrepancy (MMD) between the distribution of the energy consumption in the selected prototypes and the distribution of the energy consumption in all data. The closer the squared MMD is to zero, the better the distribution of the prototypes fits the data.

The squared MMD is calculated as:

$$MMD^2 = \frac{1}{m(m-1)}\sum_{i,j=1}^{m} k(z_i, z_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(z_i, x_j) + \frac{1}{n(n-1)}\sum_{i,j=1}^{n} k(x_i, x_j) \tag{6.1}$$

where:

m The number of all prototypes selected after each iteration

n The number of all daily profiles

z A matrix with all prototypes

x A matrix with all daily profiles

$k(\cdot)$ A kernel function. We use the radial basis function kernel[27].

The plot in Figure 6.5 shows the progression of the squared MMD as we add more prototypes. The first observation of the plot corresponds to two (2) prototypes already selected (we cannot calculate the MMD with only one prototype). We stop the prototype selection process the first time the MMD drops below 0.2.

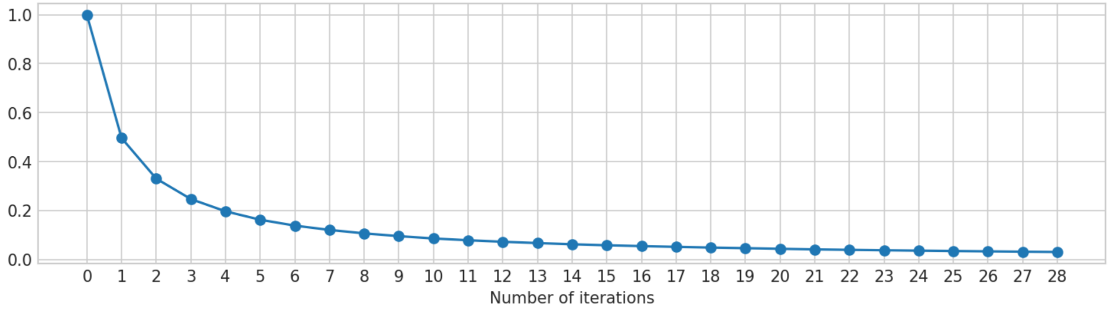


Figure 6.5: The progression of the squared MMD as we add more prototypes

[27] https://en.wikipedia.org/wiki/Radial_basis_function_kernel

The reference pattern of Figure 6.4 and the additional patterns selected by the aforementioned process are presented in Figure 6.6 below.

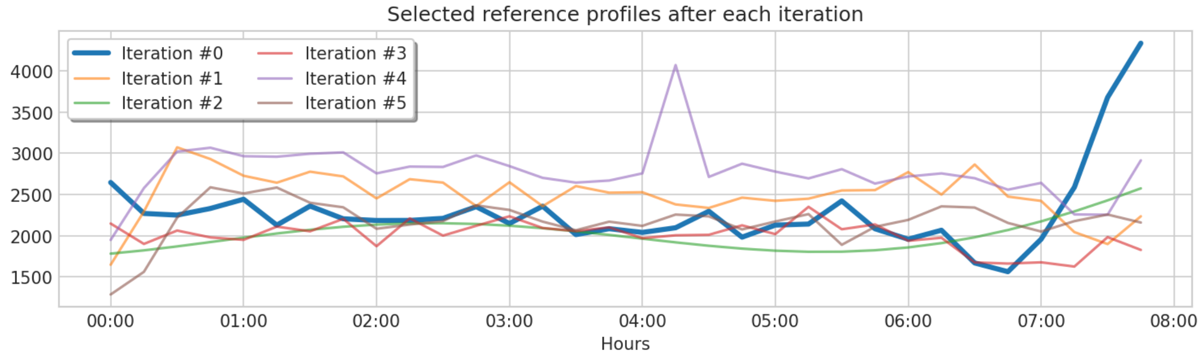


Figure 6.6: The selected reference profiles for the 00:00-08:00 interval after each iteration

Translate profile similarity to temporal proximity

Since consumption data is only available during the training of the predictive model, our goal is to translate the information about consumption profile similarity into information that can be constructed using daily-level data that is only available during prediction time:

- Day of the week: An ordinal feature taking values from 0 (Monday) to 6 (Sunday);
- Month of the year: An ordinal feature taking values from 1 to 12.

This implies that daily profile categories are useful only if they are predictable. In other words, they are useful if it is possible to explain the association of each observation in the dataset with its category given only features that can be constructed without access to the actual consumption data.

To this end, the proposed methodology makes use of *distance metric learning*. Distance metric learning aims at automatically constructing task-specific distance metrics from (weakly) supervised data. In this case, the derived distance metric is just a function that gets the aforementioned time-based features as inputs and returns the similarity in the respective load shapes. The underlying algorithm[28] has access to a set of positive and negative pairs of daily profiles, and its goal is to learn a distance metric that puts positive pairs close together and negative pairs far away. The construction of these pairs takes place at the same time that the prototypes are selected; positive pairs are constructed from profiles that are both similar to a given prototype, whereas negative pairs include one profile that is similar to a given prototype and one that is not.

According to the standard approach for model evaluation, part of the pairs is used for training the distance metric learning algorithm and part is used for evaluating its performance on unseen data.

[28] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon (2007) “Information-theoretic metric learning,” In Proceedings of the 24th international conference on Machine learning (ICML '07), pp. 209–216

For the evaluation, we test whether given the time information of two observations (day of week and month of year), the model can accurately predict whether their load profiles are similar or not. A practitioner can treat it as a binary classification problem, and if the accuracy is close or lower than 0.70, a different than the proposed modelling approach can be pursued. The balanced accuracy score for the DB1 was 0.93.

The learned distance function is utilized for categorizing the daily profiles into different clusters. However, clustering is ambiguous; different parameters of a clustering algorithm lead to different clusters. Accordingly, the proposed approach identifies the parameters of the clustering algorithm as part of the prediction problem. In other words, the clustering parameters are treated as hyperparameters to be optimized for optimal prediction accuracy.

If we manipulate the parameters of the clustering algorithm so that to produce three (3) clusters, we get the clusters of Figure 6.7. The red border denotes the interval over which the categorization has taken place.

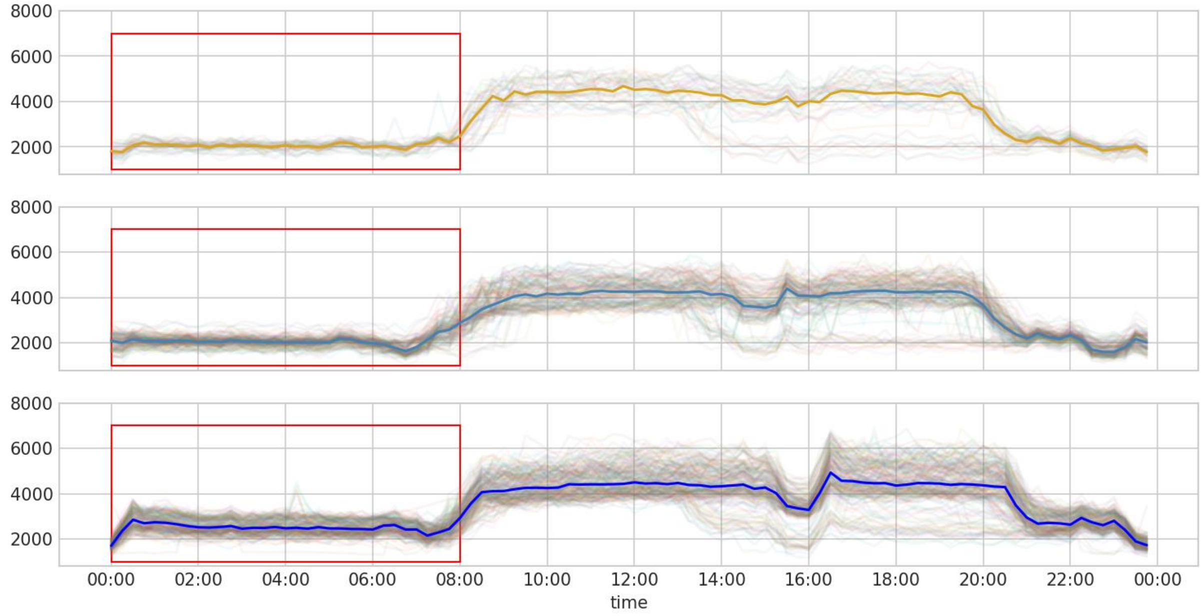


Figure 6.7: Categories into which the profiles of the 00:00-08:00 interval can be distinguished

In addition, the plot in Figure 6.8 shows how these three (3) categories are distributed throughout the year.

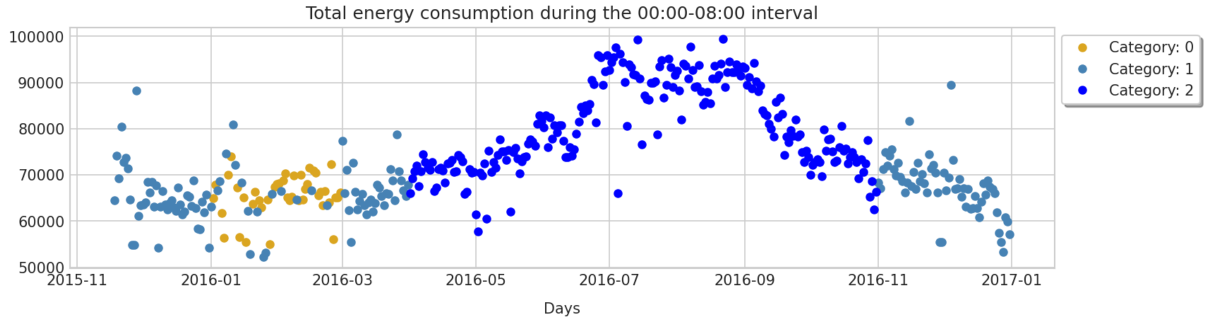


Figure 6.8: Yearly distribution of the categories of the 00:00-08:00 interval

From Figure 6.8 it can be seen that the proposed methodology focuses on load shape similarity rather than load level similarity. The prediction stage is responsible for capturing how the month, the day of the week, the hour of the day and the outdoor temperature can explain the variability of the consumption in each category.

By iteratively applying the prototype selection process already outlined for the 00:00-08:00 case, we identify the reference patterns for the 08:00-16:00 interval as presented in Figure 6.9.

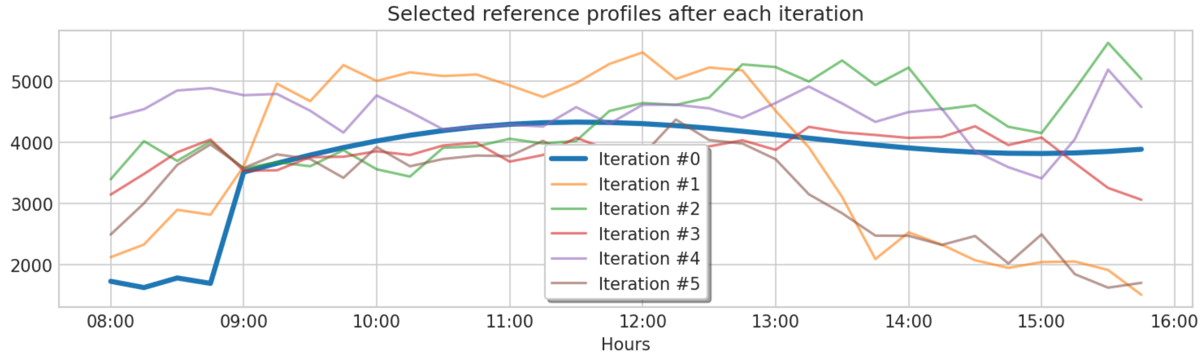


Figure 6.9: The selected reference profiles for the 08:00-16:00 interval after each iteration

If we again aim at three (3) clusters, the categorization of all consumption profiles for the 08:00-16:00 interval is presented in Figure 6.10 below.

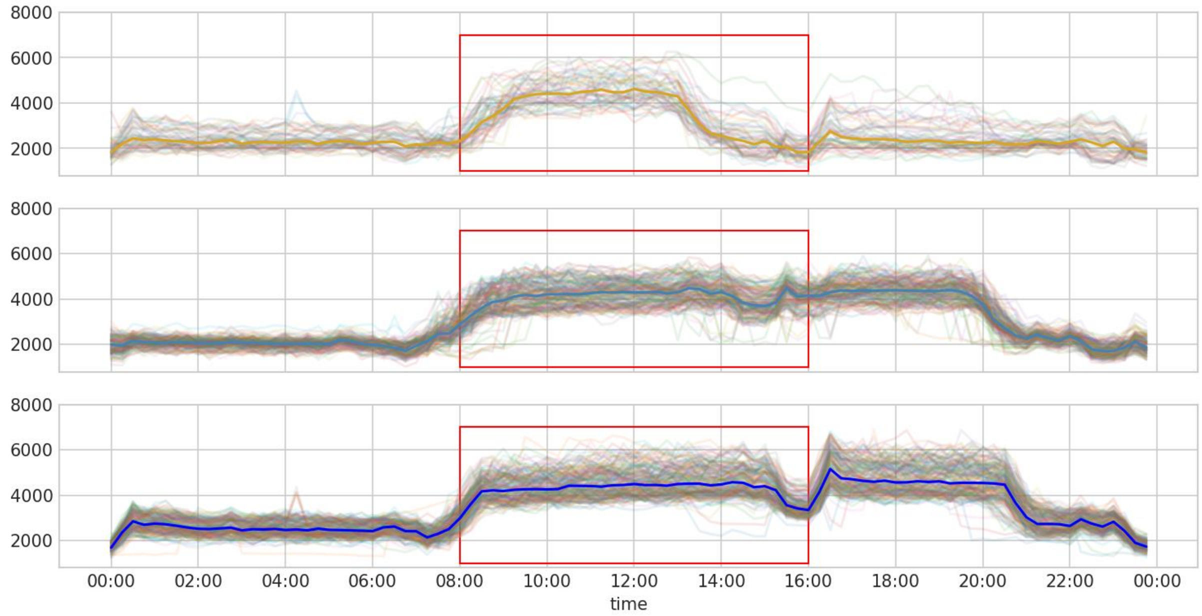


Figure 6.10: Categories into which the profiles of the 08:00-16:00 interval can be distinguished

The plot in Figure 6.11 shows how these three (3) categories are distributed throughout the year.

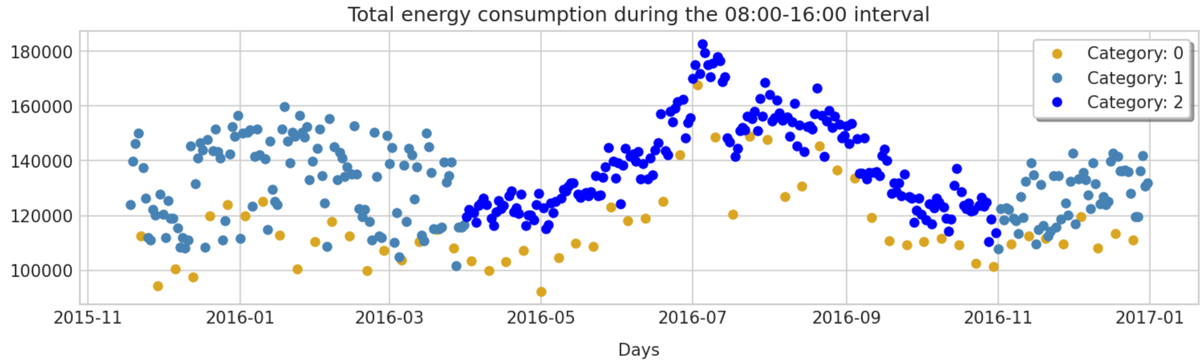


Figure 6.11: Yearly distribution of the categories of the 08:00-16:00 interval

Finally, if we again aim at four (4) clusters, the categorization of all consumption profiles for the 16:00-00:00 interval is presented in Figure 6.12 below.

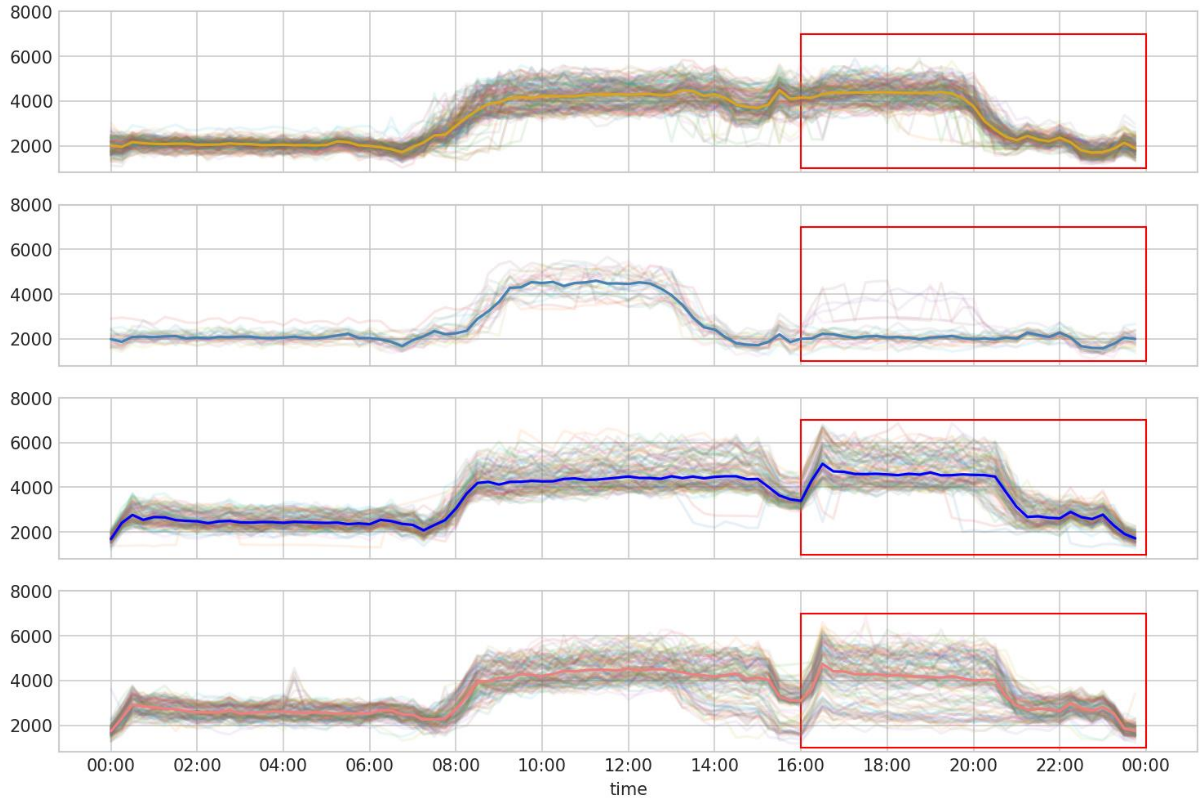


Figure 6.12: Categories into which the profiles of the 16:00-00:00 interval can be distinguished

The plot in Figure 6.13 shows how these four (4) categories are distributed throughout the year.

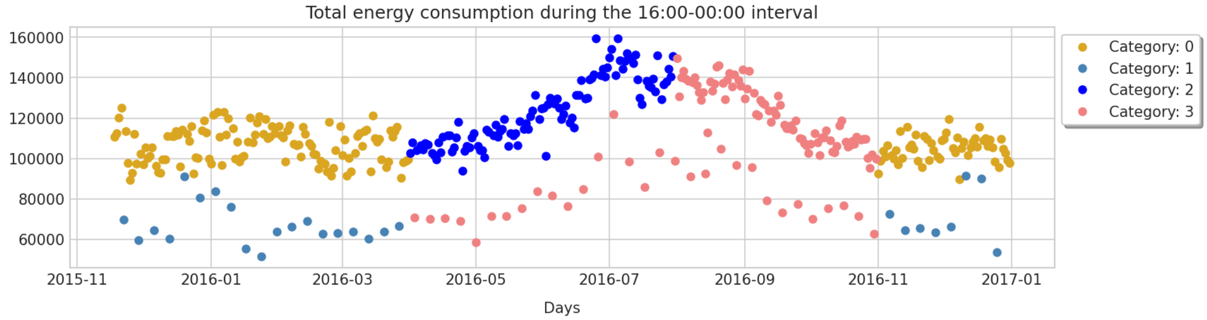


Figure 6.13: Yearly distribution of the categories of the 16:00-00:00 interval

6.3.2 Demo building #2

Identify a small number of recurring patterns that are very dissimilar to each other

The daily consumption profiles of the DB2 are presented in the plot of Figure 6.14. It is possible to visually spot at least two different daily patterns: one where the high-load period spans from 08:00 to 16:00 and one where the energy consumption is more or less flat.

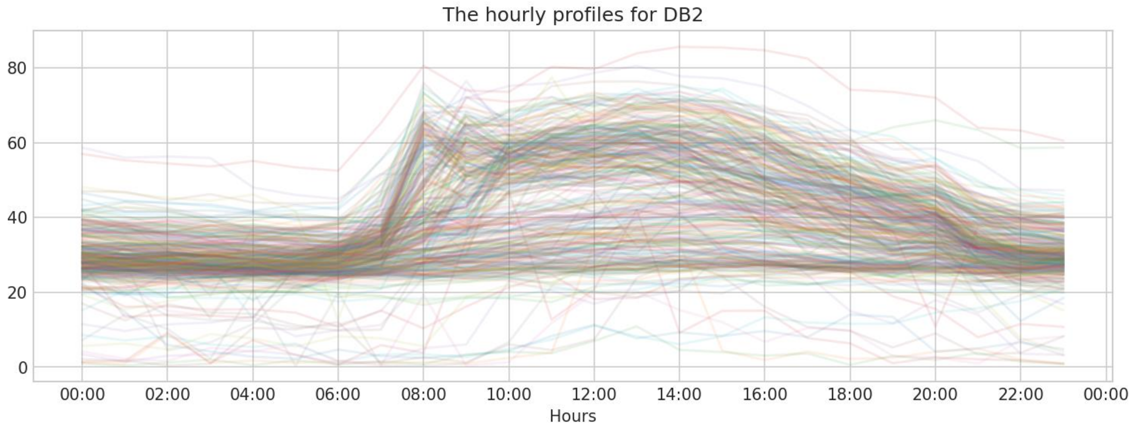


Figure 6.14: The consumption profiles of the demo building #2

The plot in Figure 6.15 shows the selected prototypes.

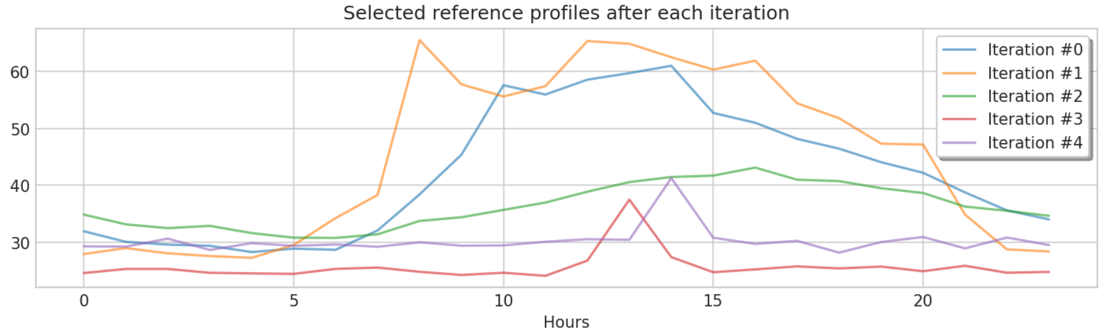


Figure 6.15: The selected prototypes for demo building #2

If we aim at ten (10) clusters, the categorization of all consumption profiles of DB2 is presented in Figure 6.16 below.

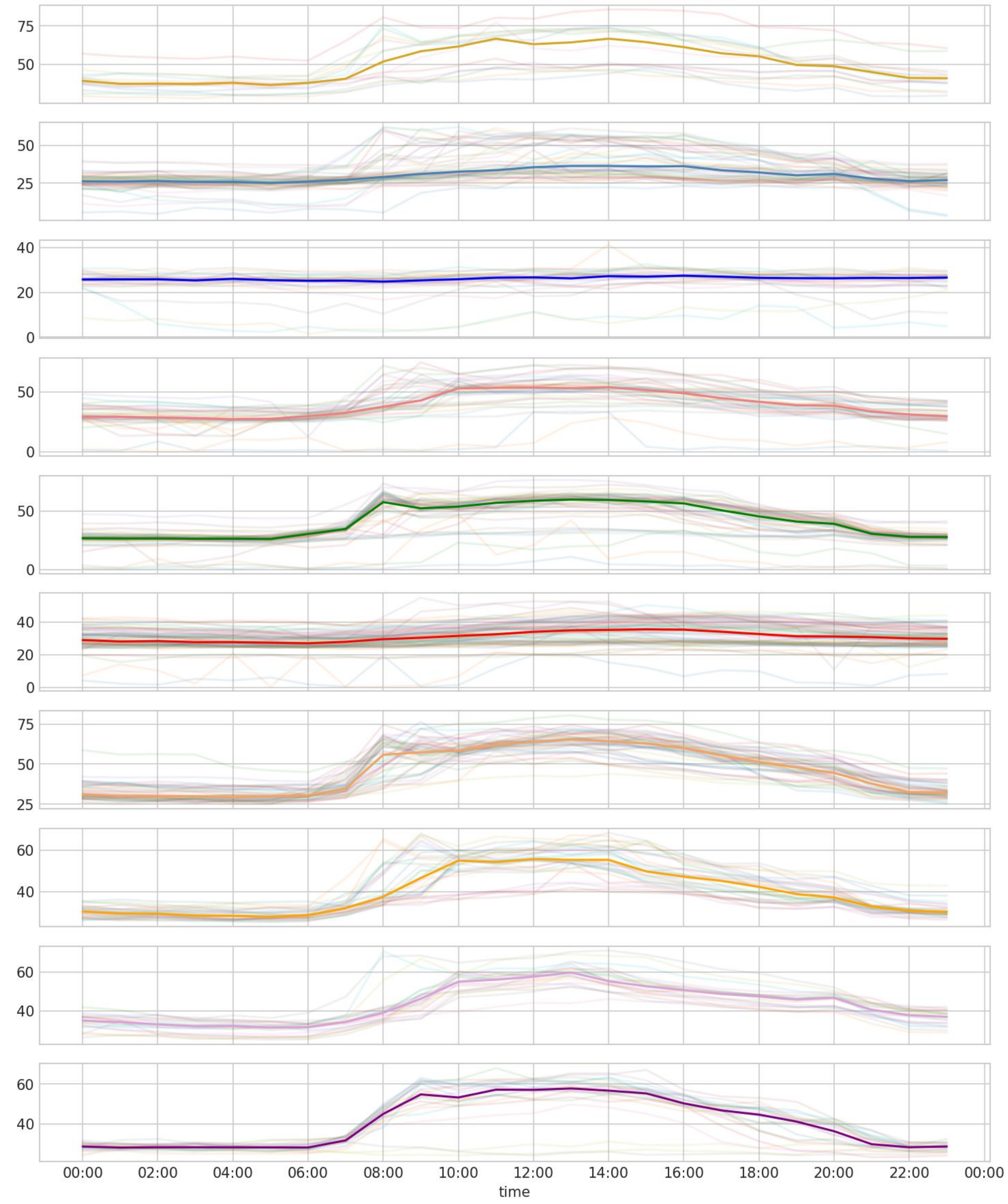


Figure 6.16: Categories into which the profiles of the demo building #2 can be distinguished

Finally, plot in Figure 6.13 shows how these ten (10) categories are distributed throughout the year.

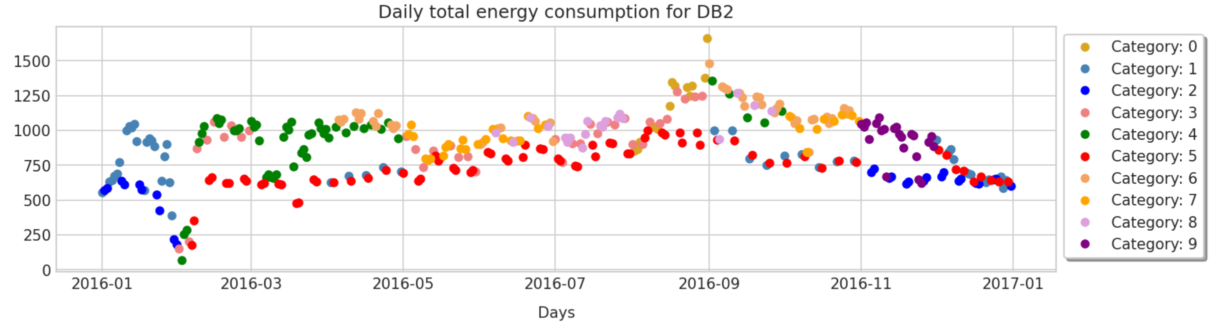


Figure 6.17: Yearly distribution of the categories of demo building #2

7 The SENSEI workflow for baseline model development

7.1 The formulation of the predictive model

The proposed model for M&V is a composition of two (2) distinct ones:

- A clustering model that associates an observation with a daily profile category given only daily-level data:
  - Day of the week: An ordinal feature taking values from 0 (Monday) to 6 (Sunday);
  - Month of the year: An ordinal feature taking values from 1 to 12.
- A regression model that aims at predicting the hourly energy consumption given the daily profile category and:
  - Day of the week: An ordinal feature taking values from 0 (Monday) to 6 (Sunday);
  - Hour of day: An ordinal feature taking values from 0 to 23;
  - The current temperature.

The regression model is a gradient boosted decision tree one. The parameters of the clustering algorithm are defined in such a way that the regression model minimizes its error on a hold-out / testing dataset. In other words, clustering is task-specific and aims at improving predictive capability. The distance function that the clustering algorithm employs is the one constructed during the day typing stage.

7.2 The evaluation of the predictive model

The quantification of the fitness metrics utilizes a cross-validation approach. The following table summarizes the results when applying the model of the data of DB1.

	$\boldsymbol{CV(RMSE)}$	$\boldsymbol{NMBE}$
00:00 - 08:00	12.3%	-0.38%
08:00 - 16:00	13%	0.15%
16:00 - 00:00	15%	0.06%
ASHRAE Guideline 14 requirements	$< 20\%$	$< \pm 0.5\%$

Moreover, the plot in Figure 7.1 shows the relation between actual and predicted consumption of DB1 for the 08:00-16:00 interval.

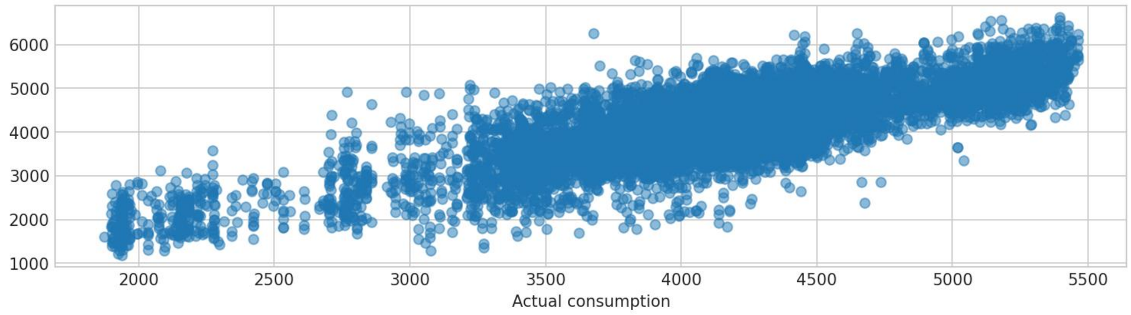


Figure 7.1: The relation between actual and predicted consumption of demo building #1

The evaluation results for the DB2 are:

	$CV(RMSE)$	$NMBE$
00:00 - 23:00	22.8%	-0.08%
ASHRAE Guideline 14 requirements	$< 20\%$	$< \pm 0.5\%$

The predicted and actual consumption for DB2 is depicted in Figure 7.2.

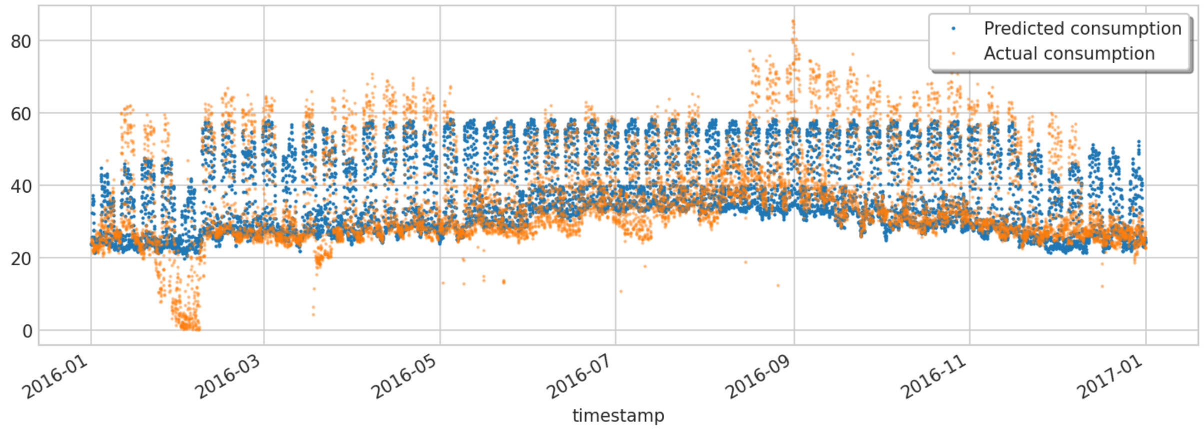


Figure 7.2: The predicted and actual consumption for demo building #2 for 2016

Since the results for the DB2 are marginally acceptable, we can evaluate its generalization capability by testing it on the dataset for 2017. For what is worth, the results are slightly better:

	$CV(RMSE)$	$NMBE$
00:00 - 23:00	18%	-0.05%
ASHRAE Guideline 14 requirements	$< 20\%$	$< \pm 0.5\%$

The plot of Figure 7.3 shows the relation between actual and predicted consumption of DB2 for the 2017, while Figure 7.4 shows the predicted and actual consumption for the same year.

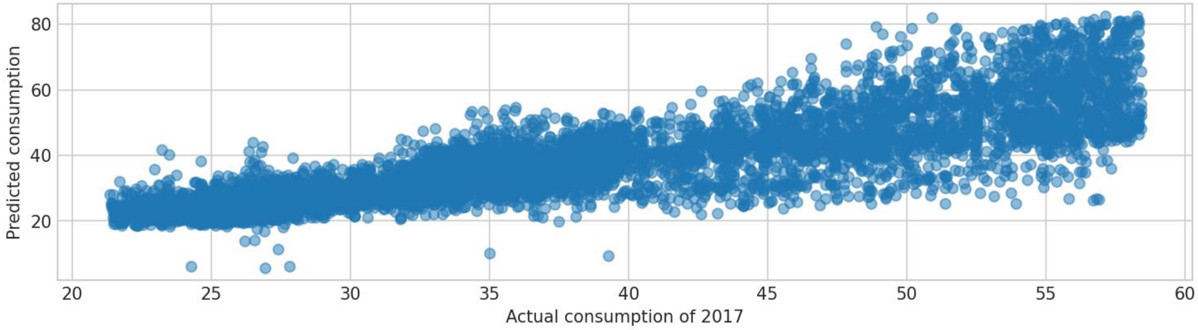


Figure 7.3: The relation between actual and predicted consumption of demo building #2

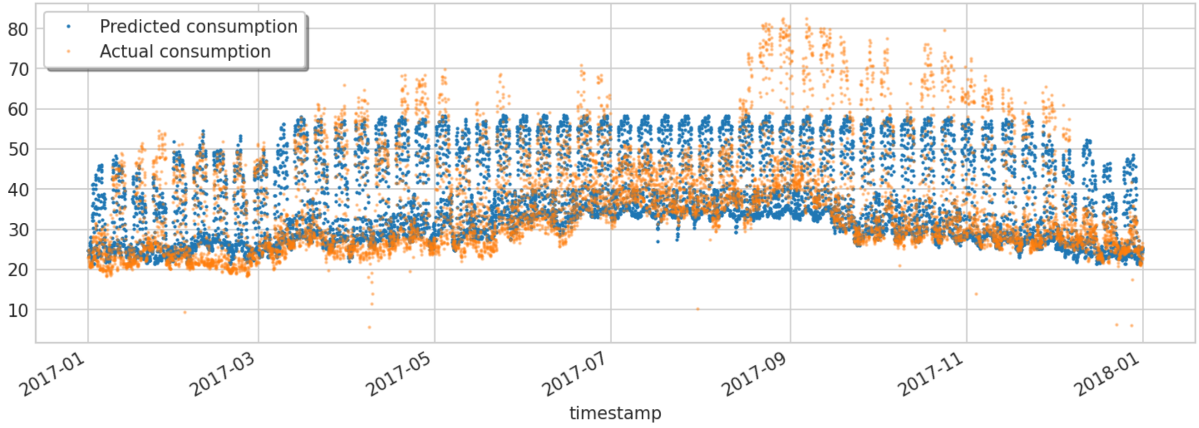


Figure 7.4: The predicted and actual consumption for demo building #2 for 2017

7.3 Constructing uncertainty intervals

Given a feature matrix X_{n+1} for a new test data point, our goal is to construct a prediction interval $C(X_{n+1})$ that is likely to contain the true value of the unknown response Y_{n+1}. If the estimation for Y_{n+1} is $\hat{Y}_{n+1}$, and the desired miscoverage rate is a (or alternatively, the desired confidence level is $1-a$), this goal can be stated as:

$$\text{find } C(X_{n+1}):\ P\{Y_{n+1} \in C(X_{n+1})\} \geq 1-a \tag{7.1}$$

An obvious, but naive, approach for constructing prediction intervals is to assume that the in-sample errors ($e_i = |Y_i - \hat{Y}_i|$ for $i = 1, \ldots, n$) can be treated as a reliable representation of the model's predictive uncertainty. However, using the errors of predictions on data that the model has already used for its training is an unreliable approach since the distribution of these errors is often biased downwards, i.e. the errors on the training data points are typically smaller than the errors on previously unseen test points.

An alternative approach for constructing prediction intervals is to split the training dataset into a part that is used only for the training of the predictive model (called the *proper training dataset*) and a part that is used for calculating the out-of-sample errors (the *calibration dataset*). Schematically, this approach (commonly referred to as *split conformal prediction*) is summarized in Figure 7.5 below.

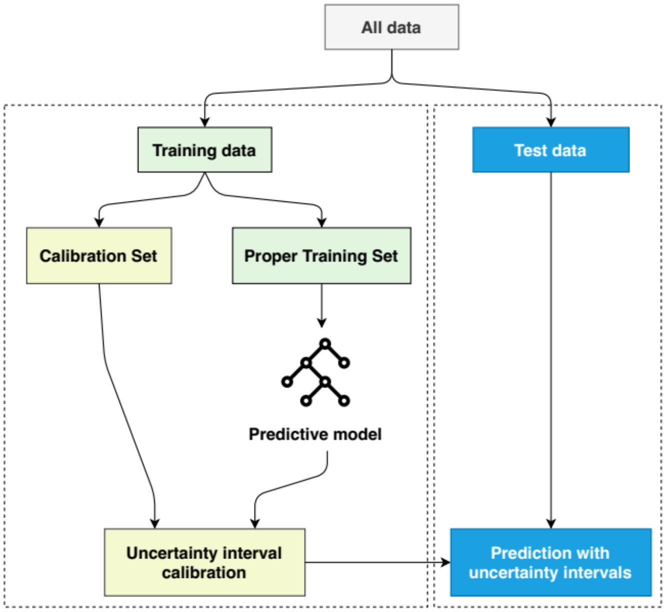


Figure 7.5: Outline of the split conformal prediction approach

The relevant algorithm includes the following steps[29]:

(1) Randomly split the indices of the training dataset $\{1, \ldots, n\}$ into two subsets: the proper training set I_1 and the calibration set I_2. The size of the calibration set m should only correspond to a small portion of the training set (i.e. $m \ll n$), as in the opposite case the removal of these examples will result in a significant reduction to the predictive ability of the underlying model and, consequently, to wider prediction intervals. A common choice is $m = 0.25 \times n$.

(2) Train the predictive model using the proper training dataset and calculate the prediction errors using the calibration dataset:

$$e_i = \left|Y_i - \hat{Y}_i\right| \text{ for } i \in I_2$$

These errors represent one of the ways to measure the nonconformity score of each

[29] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002) "Inductive confidence machines for regression," In Lecture notes in computer science: Vol. 2430, Proceedings of the 13th European conference on machine learning, pp. 334–356

instance $(X_i, Y_i),\ i \in I_2$. The nonconformity score reflects the strangeness of each instance compared to the rest of the dataset.

(3) Construct the prediction intervals as:

$$C(X_{n+1}) = \left[\hat{Y}_{n+1} - q_e\left(\frac{a}{2}\right),\ \hat{Y}_{n+1} + q_e\left(1 - \frac{a}{2}\right)\right] \tag{7.2}$$

where $q_e(i)$ is the $i\%$ quantile of the errors.

If the underlying predictive model has a skewed error distribution, using the absolute value of the errors for calibration should be avoided since it is possible for one of the boundaries to become overly optimistic, while the other becomes overly pessimistic[30]. To approximate the skewed error distribution of the underlying model, one can use a nonconformity measure based on the signed error of the model:

$$e_i = Y_i - \hat{Y}_i \text{ for } i \in I_2$$

In this case, we could define the prediction intervals at a given confidence level as:

$$C(X_{n+1}) = \left[\hat{Y}_{n+1} + q_e\left(\frac{a}{2}\right),\ \hat{Y}_{n+1} + q_e\left(1 - \frac{a}{2}\right)\right] \tag{7.3}$$

The approach that is utilized by the SENSEI model embeds the aforementioned steps into the cross-validation process. During cross-validation, the model is fitted onto a part of the available dataset (the training part) and it is evaluated on another (the evaluation part). All the errors in all the evaluation parts are aggregated and an empirical cumulative probability distribution is fitted on them. This function provides the quantile values in (7.2) and (7.3).

The plot of Figure 7.6 shows the actual consumption of DB1 during July of 2016 along with the uncertainty intervals for 95% and 99% confidence level.

[30] Linusson H., Johansson U., Löfström T. (2014) "Signed-Error Conformal Regression," In: Tseng V.S., Ho T.B., Zhou ZH., Chen A.L.P., Kao HY. (eds) Advances in Knowledge Discovery and Data Mining, PAKDD 2014, Lecture Notes in Computer Science, vol 8443, Springer

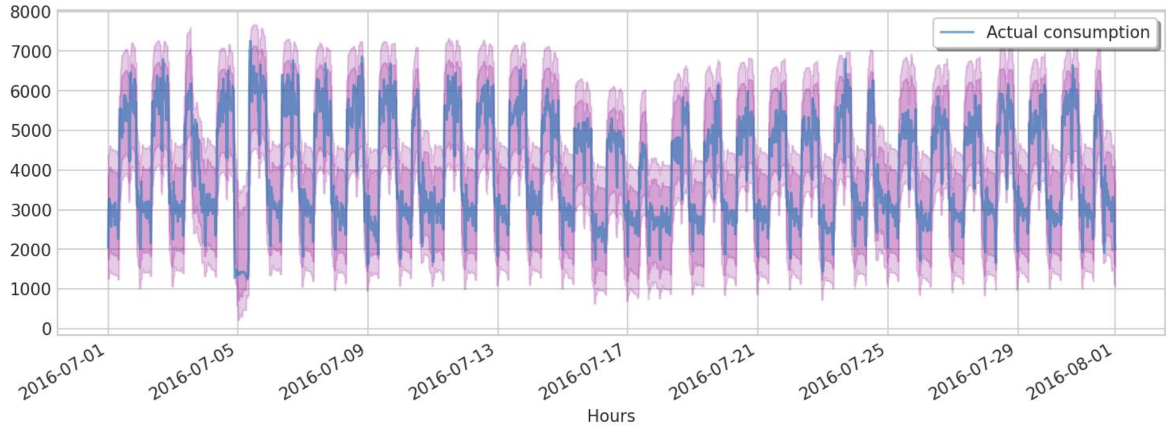


Figure 7.6: Actual consumption and uncertainty intervals for July 2016

8 Conclusions

This deliverable presented a workflow for M&V of energy savings that we believe can be useful to practitioners and help advance the state of play in terms of methods and toolkits. The presented workflow is modular so that each stage can be integrated with completely different predictive models than the one employed here.

Although standards for M&V of energy savings already exist, they focus on high-level processes rather than specific tools and techniques; devising methods and criteria for evaluating concrete M&V tool chains is work that still remains to be done. Until then, and even as a way to accelerate the emergence of the necessary initiatives, we need to experiment with the fundamental calculations of M&V and the ways to understand and deal with its uncertainty. SENSEI aspires to fuel more testing and more experimentation in the field.

Every plot and every result that was presented in this deliverable is completely reproducible. The relevant notebooks, as well as all the open source functionality that accompanies this deliverable can be found in the GitHub repository at https://github.com/hebes-io/eensight.