University of Thessaly
Electrical and Computer Engineering

Machine Learning for Energy Consumption Analysis

Μηχανική Εκμάθηση για την Ανάλυση της Κατανάλωσης Ενέργειας

Author:
Anastasios Kachrimanis

Supervisors:
Dr Aspassia Daskalopulu
Dr Lefteri H. Tsoukalas

Volos, 2017

To my family and friends

Acknowledgments

I would first like to thank Dr Aspassia Daskalopulu and Dr Lefteri H. Tsoukalas for their continuous support, consulting and guidance. They consistently allowed this paper to be my own work, but steered me in the right direction whenever they thought I needed it.

I would also like to acknowledge Dr Elias Houstis and I am gratefully indebted to him for his irreplaceable help in order to acquire the necessary knowledge on the field of Data Science.

Finally, I want to express my profound gratitude to my parents and my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Rapid increase of energy consumption has raised concerns over supply difficulties, exhaustion of energy resources and heavy environmental impacts such as global warming, climate change and greenhouse gasses. Ecological, economic and policy reasons require the reduction of energy consumption in buildings. Overall worry for energy conservation has hiked up, with the attention brought to energy consumption buildings, notably the large public ones.
To achieve lower energy consumption and better energy efficiency, the role of the Building Energy Management Systems (BEMS) is significant. These systems can contribute to the continuous energy management and therefore to a better cost and energy saving performance. Hence, data collection and analysis for the implementation of prediction algorithms is critically important.
In this paper Principal Component Analysis (PCA) and K-means clustering are proposed for monitoring electricity consumption in buildings. Through PCA the correlations between independent variables (weather conditions) and energy consumption can be found, allowing us to separate different buildings into certain groups. By using K-means, we are able to evaluate the conclusions of the PCA. This study aims at detecting abnormal behaviours in consumption patterns as well as which independent variables are responsible for them, thus acquiring the knowledge to make the right decisions for better energy efficiency.

Περίληψη

Η ταχεία αύξηση της κατανάλωσης ενέργειας έχει δημιουργήσει ανησυχίες λόγω του προβλήματος εφοδιασμού, της εξάντλησης των ενεργειακών πόρων και των καίριων περιβαλλοντικών επιπτώσεων όπως η υπερθέρμανση του πλανήτη, η αλλαγή του κλίματος και του φαινομένου του θερμοκηπίου. Για οικολογικούς, οικονομικούς και πολιτικούς λόγους, απαιτείται η μείωση της κατανάλωσης ενέργειας κτιρίων. Η συλλογική συνείδηση για εξοικονόμηση ενέργειας έχει αυξηθεί, με την προσοχή να στρέφεται κυρίως στην ενεργειακή κατανάλωση των κτιρίων, ιδίως στα μεγαλύτερα, δημοσίας χρήσεως.
Για να επιτευχθεί χαμηλότερη κατανάλωση ενέργειας και καλύτερη ενεργειακή απόδοση, ο ρόλος των Συστημάτων Διαχείρισης Κτιρίων (BEMS) είναι άκρως σημαντικός. Τα συγκεκριμένα συστήματα μπορούν να συμβάλουν στη συνεχή ενεργειακή διαχείριση για καλύτερη απόδοση κόστους και εξοικονόμησης ενέργειας. Ως εκ τούτου, η συλλογή δεδομένων και η ανάλυση τους για την εφαρμογή αλγορίθμων πρόβλεψης είναι εξαιρετικά θεμελιώδης.
Σε αυτή την εργασία οι αλγόριθμοι Principal Component Analysis και K-means clustering προτείνονται για την παρακολούθηση της κατανάλωσης ηλεκτρικής ενέργειας στα κτίρια. Μέσω της ανάλυσης PCA μπορούν να βρεθούν και να υπολογισθούν οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών (μετεωρολογικές συνθήκες) και της κατανάλωσης ενέργειας, επιτρέποντάς μας να διαχωρίσουμε κτίρια σε διαφορετικές ομάδες. Χρησιμοποιώντας K-means, είμαστε σε θέση να αξιολογήσουμε τα συμπεράσματα της PCA ανάλυσης. Η μελέτη αυτή στοχεύει στην ανίχνευση μη φυσιολογικών συμπεριφορών που σχετίζονται με την κατανάλωση, αποσαφηνίζοντας ποιες ανεξάρτητες μεταβλητές ευθύνονται για αυτές, επιτρέποντας μας να αποκτήσουμε τις απαραίτητες γνώσεις για να ληφθούν οι σωστές αποφάσεις που θα μας οδηγήσουν σε καλύτερη ενεργειακή απόδοση.

Contents

List of Figures

List of Tables

1 Introduction

When working on energy efficiency, especially when focusing on building efficiency, there are many of factors that affect final energy consumption [15]. Relations can be found between the equipment installed in the building and the weather (building acclimatization) but also with each building's occupants.

Nowadays, building automation has evolved sensibly beyond control systems for HVAC and lighting. Albeit the idea of smart buildings is present for quite a while now, with the emergence of IoT and a new generation of intelligent edge devices, pioneering inventions are now regarded as out-of-date. Constructions of all sizes and functions, residential or commercial, are progressively linked to smart ecosystems. These systems monitor and handle much more than climate control and lighting. Intelligent devices aggregate point data, combine, analyze, and issue data to edge computing where advanced analytic and forecasting engines [2] will allow new levels of control, while significantly upgrading the overall energy efficiency. These new systems are smarter, self-learning and innovative and make reducing energy consumption even easier and more valuable.

Building Energy Management Systems (BEMS) respond to climatic conditions, the operation of the building and the human interference. These systems enable building owners and occupants to monitor, maintain and manage electrical and electromechanical functions within a construction. BEMS development usually involves the installation of sensors, software, a network and a cloud-based data store. BEMS functions include administration of heating, ventilation and air conditioning systems, as well as lighting, security and safety operations. Apart from these, the greater inclination of a Building Energy Management System is its proven ability to reduce energy consumption. Typically, a BEMS favors connection with the user, enabling the operatives to program systems and keep the preferable conditions steady, alert for any surfacing anomalies, and execute pre-programmed algorithms [9], akin to a programmable logic controller (PLC) used in industrial settings.

The factors influencing building energy consumption can be separated in seven categories [26]:

- Building characteristics
- User characteristics
- Climate
- Building occupants' behaviour and activities
- Building services
- Indoor environmental quality requirements
- Social and economic factors

These seven influencing sources give a general idea of where to start to find the consumption behaviour causes and where to devote time to find relationships.

In this study, intelligent data-analysis methods, such as Principal Component Analysis (PCA) and K-means Clustering, are proposed for modelling and managing daily electricity consumption in buildings.

PCA is a projection technique which creates a representation of the dependencies between variables in a lower dimension space (orthogonal components). Thus, the technique congregates relationships among variables in this new subspace containing correlated information, while non correlated information falls in the residual space. The main purpose of a PCA implementation is the analysis of the data in order to identify the variables responsible for such large variations and their patterns.

Main advantages of the proposed approach to our data, which we suppose to derive from a BEMS are summarized below:

- Monitoring (sensor errors, abnormal user behaviours, excessive values, etc.).
- Finding abnormal consumption causes in the original monitored variables.
- Finding relations involving consumption and the other independent variables (weather conditions, year of construction, etc.).
- Detecting data errors or missing data.
- Separation of the buildings into groups with the same characteristics.

K-means clustering aims to divide a set of observations into k groups (clusters) in which each observation belongs to the group (cluster) with the nearest mean. Through this implementation we are able to crawl the data blocks based on their characteristics. In addition, K-means has been applied before and after PCA implementation to our data in order to identify the discrepancies of these two proposed case studies.

Section 2 describes a typical BEMS and its architecture, while in section 3 our data is presented. In the next two sections, Principal Component Analysis and K-means clustering are outlined, and the proposed methodologies are implemented in sections 6 and 7 respectively. Afterwards, conclusions and future work are presented in section 8.

2 Building Energy Management Systems (BEMS)

2.1 Management of Energy Consumption

The rapid proliferation of energy usage has raised concerns over supply difficulties, exhaustion of energy resources and heavy environmental impacts (global warming, climate change, etc.). Environmental, economic and policy reasons require the reduction of energy consumption in buildings.

Overall worry for energy conservation has hiked up, with the attention brought to energy consumption buildings, notably the large public ones.

The European Directive for n-ZEB (nearly - Zero Energy buildings) [19] demands the minimization of operation cost for heating, cooling and lighting systems. However, the proper operation of the systems and the required comfort level should be constantly monitored and adjusted to evaluate the appropriate operation. In large buildings, like a complex of firms, several systems from different suppliers are installed. Proper communication of a centralized management system, with the various installed information exchange systems and higher control commands, results in the required energy savings throughout their operation period.

In this endeavor, efforts are currently focused on the satisfaction of energy needs for the energy efficient buildings, by assuring the operational needs with the least possible energy cost and environmental protection.

2.2 Definition of a Building Energy Management System

To achieve lower energy consumption and better energy efficiency, the role of the Building Energy Management Systems (BEMS) is significant. These systems (Fig. 1) can contribute to the continuous energy management and therefore to the achievement of the possible energy and cost savings.

The International Energy Agency [12] describes a BEMS as: *"an electrical control and monitoring system that has the ability to con-*

trol monitoring points and an operator terminal. The system can have atrributes from all facets of building control and management functions such as heating, ventilation and air conditioning (HVAC) to lighting, fire alarm system, security, maintenance and energy management".

Another definition is that *"BEMS is a control system for individual buildings or groups of buildings which uses computers and distributed microprocessors for monitoring, data storage and communication"* [10].

Other terms frequently used for these systems are Building Management System (BMS) and Energy Management System (EMS).

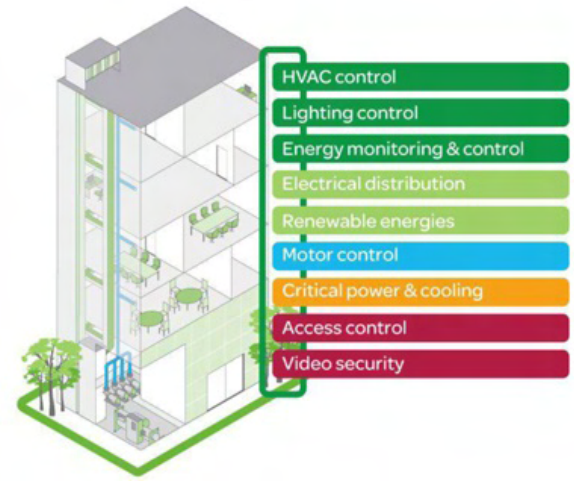


Figure 1: A common Building Energy Management System - BEMS.

In the interest of achieving being intelligent, energy efficient, green and other environmental-targeted aid, the centralized management of energy consumption is designed to improve the operation of the equipment and reduce energy usage.
In conjunction with energy management, the system can oversee and evaluate a wide assortment of other aspects of the building, whether residential or commercial.

There are many energy modules in a building, such as heating, ventilation, HVAC systems, lighting systems, elevators, office equipment, etc [5]. During operation, these devices may be controlled or disrupted by people or the environment. Hence, it is necessary to install various units of measurement and control in the building, with the system temporarily deviating from the optimal or normal operation status to be corrected back to the ideal status.

Through the BEMS real-time monitoring, the following results can be achieved:

- The level of building management is improved.
- The inefficient equipment can be found.
- Identifying abnormal energy consumption.
- Lower peak electrical demand.

2.3 Buiding Energy Management System Architecture

BEMS creates a database for energy information and processes the data to perform building energy saving by monitoring and analyzing the amount of energy and its efficiency.

The BEMS system consists of four units: monitoring system, metering system, control system and analysis system, as shown in Fig. 2.

1. **Monitoring System**
   Acquires air temperature, humidity, illumination (direct light, diffuse light), etc., indicating the indoor quality of the air and the effect of energy utilization.

2. **Metering System**
   Measures air-conditioning power, lighting and socket power, active power, special power, which reflect the amount of energy consumption.

3. **Control System**
   Optimizes the operation of equipment through the building automation system.

4. **Analysis System**
   Provides energy analysis and evaluation reports. It then provides the appropriate advice for energy-saving approaches.

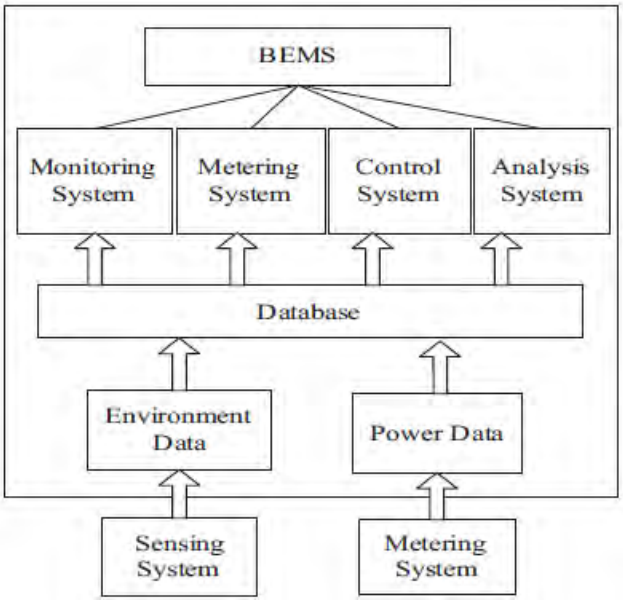


Figure 2: Performance units of a BEMS [11].

The infrastructure of the decision support model plays a leading role and is usually based on the characteristics of a typical BEMS logic [10].
The current model (Fig. 3) includes the following components:

- **Indoor sensors**

  Sensors that measure or record temperature, relative humidity, air quality, movement and brightness in the building areas.

- **Outdoor sensors**

  Sensors for external conditions such as temperature, relative humidity and illumination.

- **Controllers**

  This component category contains switches, diaphragms, valves, actuators etc.

- **Decision unit**

  A real time decision support unit is included, with the following capabilities:

  - Interaction with sensors to diagnose the condition of the building and hence the formulation of the building's energy profile.
  - Integration of intelligent systems to select the appropriate interventions, depending on the building's demands.
  - Communication with the building's controllers to implement the decision.

- **Database**

  It includes the database for the building energy features and the knowledge database, where all the basic information is recorded.

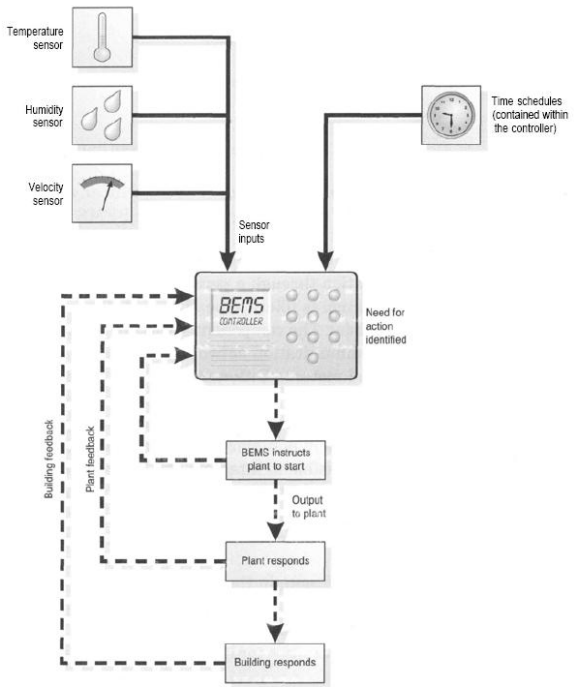


Figure 3: The main functions of a BEMS.

2.4 Analysis of Energy Efficiency

Data mining is defined as *"An interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization"*[20]. Data mining techniques are widely disseminated in research areas such as marketing, biology, engineering, medicine, and social sciences to address the issue of pattern extraction from large databases.

Data mining is also a powerful technique in providing insights into energy patterns (Fig. 4) related to the occupants' behaviour, simplifying assessments of building saving potential by improving users' energy profiles as well as driving building energy policy formulation.

Figure 4: Energy Consumption Analytics.

It must be stressed that the role of the decision support systems is extremely significant. They can contribute to the continuous energy management of the day-to-day operations of a building, in order to maintain the comfort conditions of the buildings' residents and to minimize energy consumption and cost.

Advanced control techniques based on **artificial intelligence** (neural networks, fuzzy logic, genetic algorithms, etc.) and distributed control networks offer many benefits in this direction, demonstrating a significant energy efficiency.Algorithms such as Principal Component Analysis or various Clustering techniques are able to detect outliers and abnormal activities. Thus, we are able to model and monitor the energy consumption of buildings.

3 Data Set

3.1 Building Characteristics Sheet

The data we used for this study are based on the average value of each parameter. However the implementation of the algorithms which are described below can be done with any other relative data sets.

Suppose that we have eight buildings described by their features. Let us assume that a Building Energy Management System provides us with these 9 characteristic parameters of these buildings such as the Electricity Consumption, Air Temperature, Wind Velocity, etc. (Table 1). The given values are equal to the daily average and have been calculated based on actual measurements in similar situations.

Parameters determined
Direct Light (lux, lumen/m^2)
Wind Velocity (m/s)
Wind Direction (°)
Electricity Consumption (kWh)
Air Humidity (%)
Year of Construction
Number of People
Air Temperature (°C)
Fan speed

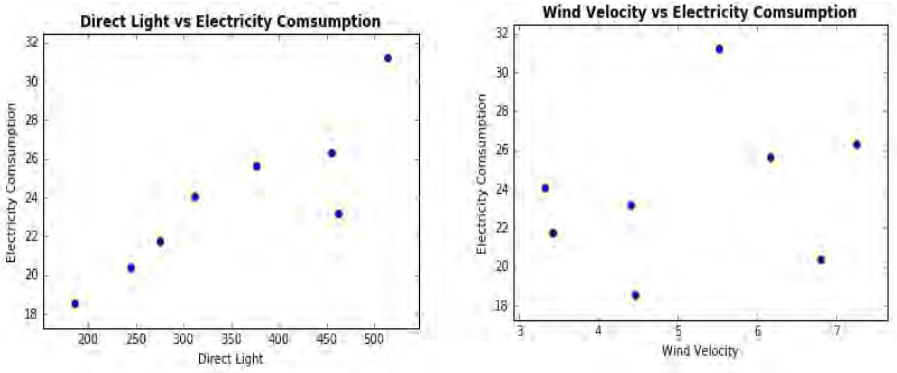
Table 1: Characteristic Parameters of each Building

Hence, a dataset is obtained which consists of 8 buildings and 9 independent variables. The actual measurements can be arranged in a table or a matrix of size 8×9. A portion of this table is shown in Fig. 5.

	Direct_Light	Wind_Velocity	Wind_Direction	Electricity_Consumption	Air_Humidity	Year_of_Construction
Buildings						
Building_1	185.61	4.46	160	18.54	55	2012
Building_2	275.20	3.42	223	21.74	61	2008
Building_3	513.74	5.52	185	31.22	68	1975
Building_4	376.40	6.17	23	25.62	65	1998
Building_5	454.88	7.25	82	26.35	61	1987
Building_6	462.62	4.40	27	23.16	65	1992
Building_7	244.15	6.80	112	20.40	59	2009
Building_8	312.00	3.32	140	24.05	63	2001

Figure 5: A subset of the dataset.

A good starting point is to plot individual variables combined with the Electricity Concumption of each Building. Two of the variables are shown in Fig. 6.



(a) Direct Light and Electricity Consumption values for each building.

(b) Wind Velocity and Electricity Consumption values for each building.

Figure 6: Independent variables vs Electricity Consumption.

4 Principal Component Analysis

4.1 PCA as Model

Hervé Abdi and Lynne J. William presented Principal component analysis (PCA) as a "multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables" [1]. PCA provides an efficient way of capture the dominant components of an infinite-dimensional process. The main purpose of this procedure is the analysis of data to identify patterns. Finding these patterns allows us to reduce the dimensions of the dataset with minimal loss of information.

PCA is a statistical procedure that orthogonally transforms (Fig. 7) the original **n** coordinates of a data set into a new set of **n** coordinates called principal components. As a result of the transformation, the first principal component has the largest possible variance. Each succeeding component has the highest possible variance under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. Keeping only the first $\mathbf{m} < \mathbf{n}$ components reduces the data dimensionality while retaining most of the data information, i.e. the variation in the data.

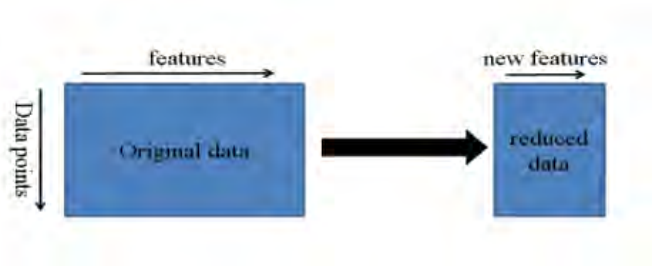


Figure 7: Data Reduction

We need to notice that the PCA transformation is sensitive to the relative scaling of the initial variables. Data column ranges need to be normalized before applying PCA. Furthermore, the new coordinates (PCs), as shown in Fig. 8, are not real system-produced variables anymore. Applying PCA to our data set loses its inter-

pretability.

One of the methods often used to graph the results obtained from the PCA is the bi-plot of PCA scores and loadings. It is the method of representation that displays the data formed by transforming the original ones into the space of the PCs (scores). The loading provides a measure of the contribution of each variable to the principal components.

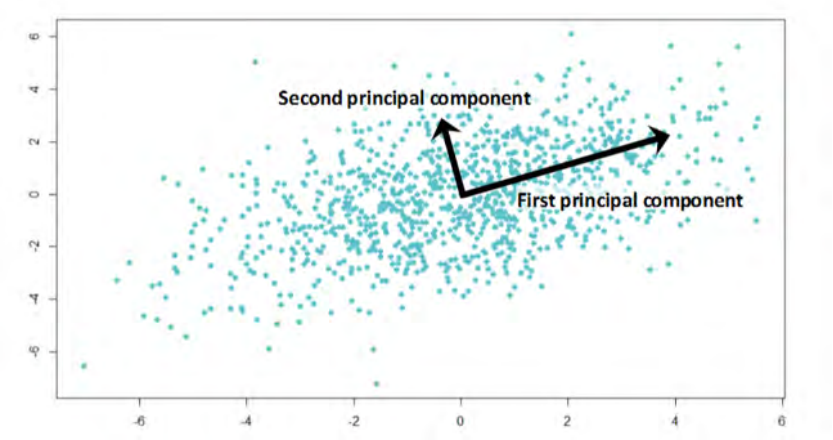


Figure 8: The fist 2 Principal Components

A Summary of the PCA Approach [17]

1. Standardize the data.
2. Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition.
3. Sort eigenvalues in descending order and choose the **k** eigenvectors that correspond to the **k-largest** eigenvalues where **k** is the number of dimensions of the new feature subspace ($k <= d$).
4. Construct the projection matrix **W** from the selected **k** eigenvectors.

5. Transform the original dataset **X** via **W** to obtain a **k-dimensional** feature subspace **Y**.

4.2 Aspects of PCA

When large multivariate datasets are analyzed, it is often desirable to reduce their dimensionality. In a project related to energy efficiency of set of buildings having available data such as Demographics, Building data, Psychographics, Room Sensor data or Building Sensor data, to separate which of them have a significant impact on the consumption each time, is of primary importance. Visualizing energy consumption patterns and understanding them could help us decide in which key factors we should pay our attention, after implementing the Principal Component Analysis.

- **Building customer segments**

  A very common approach to building and understanding customer segments [22] is through the use of clustering techniques such as Principal Component Analysis (PCA). These clustering techniques will analyze the customer data and observe if customers tend to cluster by certain features (Fig. 9), or combinations of features. Implementing PCA may be useful at identifying which groups of users are mostly presented in each room/building, finding whether there are physically vulnerable groups of users and associate their preferences for heating and cooling etc.

- **Identifying key performance indicators (KPIs)**

  A key performance indicator (KPI) is a type of performance measurement. KPIs (Fig. 10) estimate the success of a corporation or of a particular undertaking (such as projects, programs, products and other initiatives) in which it engages. Often is merely the recurring, periodic achievement of certain levels of the business objective, and sometimes attainment is realized in

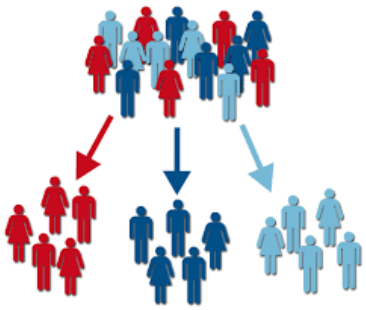


Figure 9: Different customer classes

terms of decision-making progress toward crucial goals. Thereupon, choosing the right KPIs relies on a adequate insight of what is important to the organization or the project. These assessments often lead to the identification of potential improvements, so performance indicators are routinely associated with 'performance improvement' initiatives.

Figure 10: Flow of a KPI process

The KPIs for monitoring the energy efficiency requires specific form to evaluate the realization of the goal on the consumer level. A common option in this case is the definition of consumption per entity/household. The definition of the performance indicator consists of the two components:

1. The choice of the variables involved.

2. The definition of the monitoring interval.

The variables involved into the indicator must be dependent. It is recommended that the dependence is linear. PCA is the methodology applied to detect the KPI [25]. Through this approach, it is possible to identify the relation between different variables measured. The analysis of these connections leads to detecting of key variables to build a proper KPI.

The PCA represents the variation of the data into the system, so implementing this methodology allows us to measure various data such the daily Electricity consumption or the average daily consumption per floor. We can also visualize and check Historical information related to consumption.

- **Detecting outliers**

Principal component analysis is a powerful and versatile method capable of providing an overview of complex multivariate data. PCA can also be used to detect outliers.

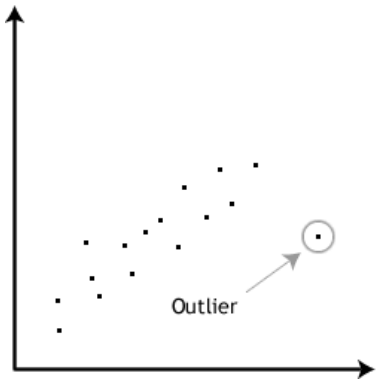


Figure 11: Observation of an outlier sample

Outliers (Fig. 11 presents a typical abnormal value) are samples that are somehow disturbing or unusual. Often, outliers

are downright wrong samples. For example, in determining the height of persons, five samples are obtained ([1.78, 1.92, 1.83, 167, 1.87]). The values are in meters but accidentally, the fourth sample has been measured in centimeters. If the sample is not either corrected or removed, the subsequent analysis is going to be detrimentally disturbed by this outlier. Outlier detection is about identifying and handling such samples.

Often outliers are mistakenly taken to mean 'wrong samples' and nothing could be more wrong. Outliers can be absolutely right, but e.g. just badly represented. In such a case, the solution is not to remove the outlier, but to supplement the data with more of the same type.

- **Detecting Moving Objects**

  PCA is not a technique commonly used in this domain. Despite this, it could be helpful sometimes. We can consider a n-frame subsequence, where each frame is associated with one dimension of the feature space, and we apply PCA [21] to map data in a lower-dimensional space where points picturing coherent motion are close to each other. Frames are then split into blocks that we project in this new space. Inertia ellipsoids of the projected blocks allow us to qualify the motion occurring within the blocks. By doing this we can identify which of the residents have kids or track new occupants (tenants).

4.3 PCA Flow Diagram

A flowchart is a visual representation of the sequence of steps and decisions needed to perform a process. Each step in the sequence is noted within a diagram shape. Steps are linked by connecting lines and directional arrows. This allows anyone to view the flowchart and logically follow the process from beginning to end.

With proper design and construction, it communicates the steps in a process very effectively and efficiently. For this reason a flow diagram (Fig. 12) was constructed for better understanding of the critical steps for the PCA implementation. The key components of the diagram are also explained below.

- **Subtract the mean:**
  For PCA to work properly, we need to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. This produces a data set whose mean is zero.

- **Calculate the covariance matrix:**
  Covariance calculations are used to find relationships between dimensions in high dimensional data sets where visualization is difficult. Covariance is a measure of how much each of the dimensions varies from the mean with respect to each other. The covariance between one dimension and itself is the variance. Variance is a measure of the deviation from the mean for points in one dimension.

- **Calculate the eigenvectors and eigenvalues of the covariance matrix:**
  Calculating the eigenvectors and eigenvalues is extremely important, as they tell us useful information about our data. They provide us with information about the patterns in the data.

- **Choosing components and forming a feature vector:**
  The eigenvector with the highest eigenvalue is the principle

component of the data set. Generally, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives us the components in order of significance. To be accurate, if we initially have n dimensions in our data and thus calculate n eigenvectors and eigenvalues, and then select only the first p eigenvectors, then the final data set has only p dimensions. Feature vector s is constructed by taking the eigenvectors that we want to keep from the list of eigenvectors and forming a matrix with these eigenvectors in the columns.

- **Deriving the new data set:**
  Having selected the components (eigenvectors) that we wish to maintain in our data and formed a feature vector, we simply take the transpose of the vector and multiply it to the left of the initial data set, transposed.

$$FinalData = RowFeatureVector\ x\ RowDataAdjust$$

  where "RowFeatureVector" is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most important eigenvector at the top, and "RowDataAdjust" is the data that has been customized on average.

- **Getting the old data back:**
  Recall that the final transform is this:

$$FinalData = RowFeatureVector\ x\ RowDataAdjust$$

  which can be turned around so that, to get the original data back:

$$RowDataAdjust = RowFeatureVector^{-1} x\ FinalData$$

  However, when we take all the eigenvectors in our feature vector, it turns out that the inverse of our feature vector is actually equal to the transpose of our feature vector. This only applies because the elements of the matrix are all the unit eigenvectors of our data set.

The equation becomes:

$$RowDataAdjust = RowFeatureVector^{T} x\ FinalData$$

But, to get the actual original data back, we need to add on the mean of that original data:

$$RowOriginalAdjust=(RowFeatureVector^{T} x\ FinalData) + OriginalMean$$

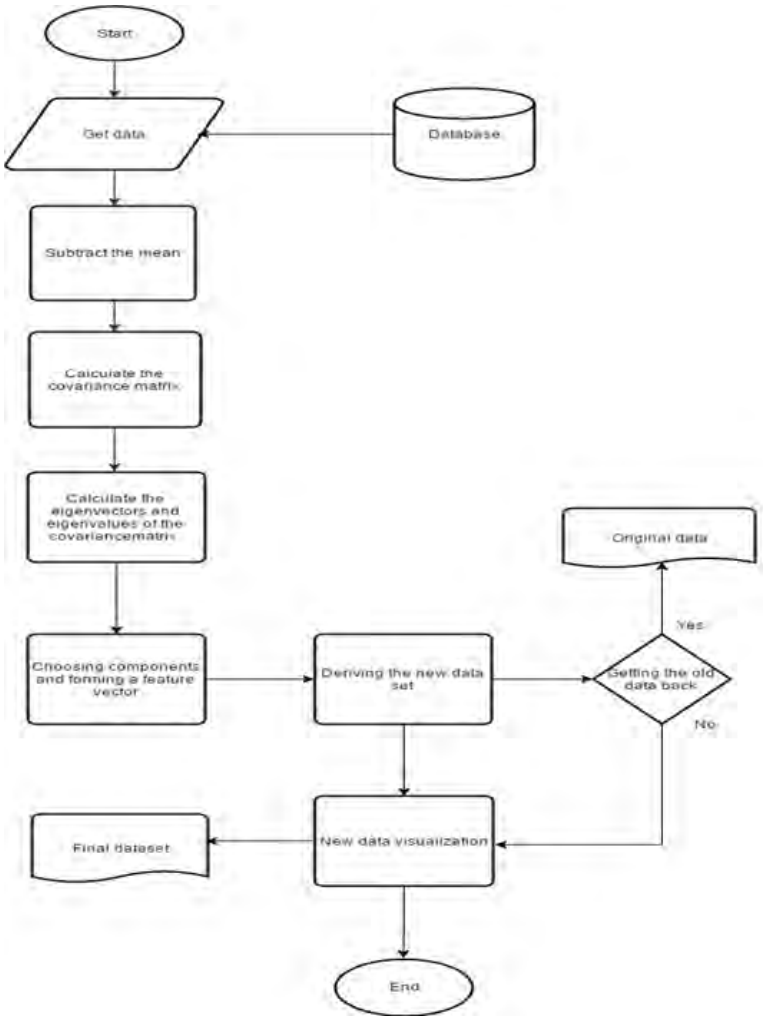


Figure 12: PCA Flow Diagram

5 Clustering Techniques

5.1 Description of Clustering Algorithms

Clustering or cluster analysis is a form of unsupervised learning, which means that the class labels of the input data are unknown. The aim of clustering is to detect groups in the data, called clusters. The input data points should be partitioned into a number of clusters, in such a way that the points belonging to the same cluster are more similar to each other than to points belonging to other clusters.

In order to accomplish this objective, the most common starting point is computing a matrix, called *dissimilarity matrix*, which contains information about the dissimilarity of the observed units. According to the nature of the observed variables (quantitative, qualitative, binary or mixed type variables), we can define and use different measures of dissimilarity.

If the observed variables are all quantitative, each unit can be identified with a point in the p-dimensional space and the dissimilarity between two objects i and j can be measured through:

1. a) Euclidean distance:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2. City-Block or Manhattan distance:

$$\sum_{i=1}^{n}|x_i - y_i|$$

Both of these two distances are specific cases, respectively for p = 2 and p = 1, of a generic distance family known as

3. Minkowski metric:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Indeed, clustering has a wide range of applications. It can be implemented in order to detect clusters among the documents suggested by search engines or to achieve market segmentation of an e-commerce store. Obviously, clustering can have also an impact on the energy sector. For instance, energy consumption patterns could be found and groups of people with the same behaviour could be formed.

More specifically, when dealing with cluster analysis there are three main issues to be resolved. Firstly, it is of utmost importance to define a score function, in order to evaluate different clustering methods and approaches. Secondly, the number of clusters is also significant as this is a crucial hyper-parameter for many clustering algorithms. Last but not least, selecting the proper clustering algorithm is also challenging. It depends significantly on the data set and on what is the aim of the clustering. Of course, there are other issues associated with choosing the right distance function and handling the different data types.

In order to implement a cluster analysis and in general a data mining task, the following steps could be followed as suggested by the Cross Industry Standard Process for Data Mining (Fig. 13), commonly known by its acronym CRISP-DM [23]. This methodology is widely used and remains the leading methodology used by industry data miners according to polls (2002, 2004, 2007, 2014) conducted by KDNuggets [16].

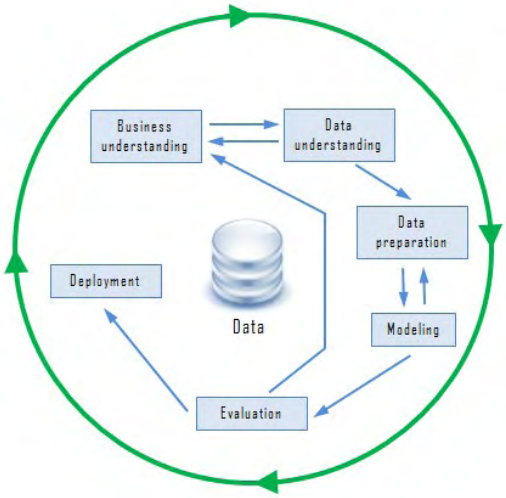


Figure 13: Cross Industry Standard Process for Data Mining.

- **Business Understanding**

  Business requirements should be understood and a specific business goal should be set before defining the clustering task.

- **Data Understanding**

  Data understanding is an integral part of the procedure. In this step, descriptive statistics could be used in order to detect trend and anomalies on the data set. Those first insights on the data through the data exploration are helpful in designing and tuning the clustering algorithms subsequently.

- **Data Preparation**

  Before selecting the clustering algorithm and modelling the problem data should be processed. This is a process, which does not have specific steps as each cluster analysis and data have different characteristics. However, it could entail stan-

dardization of the data, dimensionality reduction attribute selection and data transformation. Of course, new data could also be derived by existing data.

- **Modelling**

  In this step of the methodology various clustering algorithms are assessed and applied before getting tuned using different methods. There is a plethora of clustering methods but the most widely used and known is k-means.

  Apart from k-means algorithms hierarchical algorithms could also be used to implement a cluster analysis. Basically, there are two different types of hierarchical clustering [4]: the agglomerative, in which pairs of clusters are merged while moving up on the hierarchy and the divisive, in which all the points at the beginning belong to one cluster and then splits are performed as we are moving down to the hierarchy.

  Due to the nature of that method, there is no need to determine the number of clusters in advance, while the result of the algorithm could be visualized with the use of a dendrogram and heatmaps. Hierarchical clustering algorithms are pretty useful for observing hierarchical structure but they are more suitable for relatively small data sets due to their time complexity. Hierarchical algorithms also usually use as a distance metric the Euclidean distance while there are different methods of agglomerative hierarchical methods.

  Finally, there are spectral clustering methods [7], which refer to algorithms that cluster points using eigenvector of matrices derived from the data. Spectral clustering methods seem easy to implement and reasonably fast. Basically, one of the main differences of the aforementioned approaches from spectral clustering is that it considers the clustering task as a graph partitioning task. Spectral clustering methods are employed in case cluster data is connected but not compact.

- **Evaluation**

  In this stage the results of the clustering should be assessed in order to determine if they meet the original purpose of the task. This step is of pivotal importance as the results will be used in order to create a list of actions and measures to be taken.

- **Deployment**

  Finally, in the deployment step the final results of the actions are presented in order to better evaluate the whole project. It could also be designed a monitoring and maintenance plan of the data mining procedure, in case it continues to take place.

5.2 K-means Clustering

K-means is a partitioning-based clustering algorithm and constitutes one of the simplest unsupervised learning algorithms. The main idea [13] is to define **k** centers, one for each cluster. These centers should be placed in an intelligent way because of different location causes different result. So, the best option is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and connect it to the nearest center. When no point is pending, the first step is completed and an early group age is done.

At this point we need to re-calculate **k** new centroids as barycenter of the clusters resulting from the previous step. After we have these **k** new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the **k** centers change their location step by step until no more changes are done, that is, centers do not move any more.

Finally the purpose of the algorithm is to minimize the within-cluster *sum of squared errors*. So, the objective function is the following where $||{x_i}^j - c_i||$ is a chosen distance measure (usually Euclidean distance) between an object and the cluster centre.

$$J = \sum_{j=1}^{n} \sum_{i=1}^{k} ||{x_i}^j - c_i||^2$$

The algorithm uses the Euclidean distance and it requires to define a priori the number of clusters. The algorithm could be described as follows:

1. Start with randomly chosen cluster centres (centroids).

2. Assign each object to the group that has the closest centroid. - *Assignment Step*

3. When all objects have been assigned, recalculate the positions of the **K** centroids. - *Update Step*

4. Repeat the Assignment and the Update step until the assignments do not change.

In this algorithm it is critical to determine the correct number of clusters and there is a wide range of methods (i.e. the elbow method, the silhouette method etc.) that could be used in order to achieve this. Of course, the most significant when specifying the number of clusters is data understanding because the proposed automatic methods do not take into account all the different parameters.

- **The elbow method**

  The idea of the elbow method is to implement k-means clustering on the data set for a range of values for the number of clusters. For each value for the number of clusters calculate the percentage of variance explained, which is defined as the ratio of the between-group variance to the total variance. Then we should plot the percentage of variance explained by the clusters against the number of clusters. If there is a point where the marginal gain will drop we will detect an angle at this point and consequently the number of clusters. Of course, it is not always easy to detect an angle and for this reason as stated before those methods are complementary and their results should not be taken for granted.

- **The silhouette method**

  The silhouette constitutes also a useful criterion for determining the proper number of clusters and it was firstly suggested by Peter J. Rousseeuw [18]. The silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. A silhouette close to 1 implies the sample is in an appropriate cluster, while a silhouette close to 1 implies the sample is in the wrong cluster.

5.3 K-means Flow Diagram

The steps described in the previous chapter can be easily represented in a flow diagram. Thus, a flow diagram (Fig. 14) was constructed for better understanding of the critical steps for the k-means implementation.

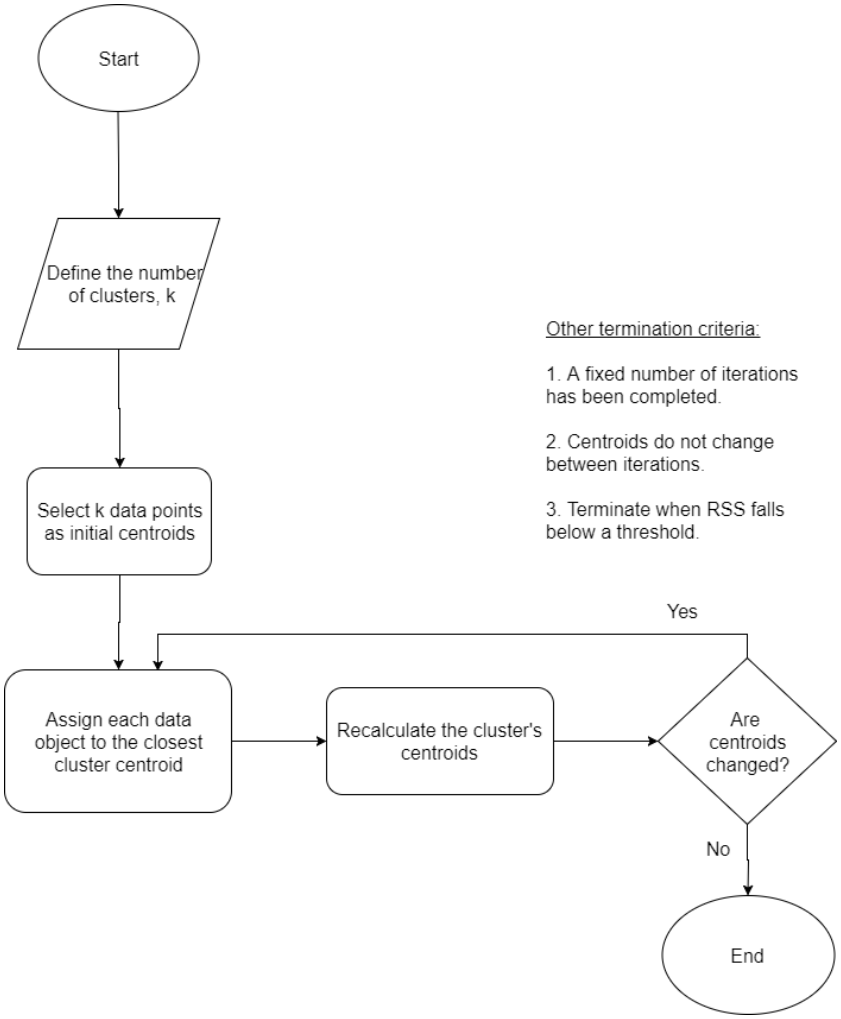


Figure 14: K-means Flow Diagram

6 Implementation of PCA

6.1 Programming Language and Tools

Python and the Jupyter Notebook were used to perform analyzes and implement the algorithms. The necessary packages to complete this study were:

- *pandas*
- *numpy*
- *matplotlib*
- *plotly*
- *scikit-learn*
- *statsmodels*

6.2 Data preprocessing

Principal component analysis is a powerful data analysis tool, capable of reducing large complex data sets containing many variables. Examination of the principal components set allows the user to spot underlying trends and patterns that might otherwise be masked in a very large volume of data.

This technique is commonly used to detect underlying correlations that exist in a (potentially very large) set of variables. The aim of this analysis is to transform a set of **n** variables, *X1, X2, X3, ..., Xn,* and to estimate the correlations. The most important of these correlations are called the Principal Components (PCs). The analysis will return vectors *Y1, Y2, Y3, ..., Yn*, each describing a different underlying variation found in the initial dataset. The vectors of *Y* are ordered by their importance. More specific, the component *Y1* is the most prevailing trend throughout the data, and accounts for more variation than *Y2*. *Y2* is a component uncorrelated with *Y1*, and will account for the second largest trend in the data. *Y3* describes the third largest component, and so forth.

PCA provides the weights required to get the new variable *Y1* better explains the variation in the entire data set in a certain sense. This new variable including the critical weights, is called the first principal component.

At this point, we need to remember the table that has been created from our data.

	Direct_Light	Wind_Velocity	Wind_Direction	Electricity_Consumption	Air_Humidity	Year_of_Construction
Buildings						
Building_1	185.61	4.46	160	18.54	55	2012
Building_2	275.20	3.42	223	21.74	61	2008
Building_3	513.74	5.52	185	31.22	68	1975
Building_4	376.40	6.17	23	25.62	65	1998
Building_5	454.88	7.25	82	26.35	61	1987
Building_6	462.62	4.40	27	23.16	65	1992
Building_7	244.15	6.80	112	20.40	59	2009
Building_8	312.00	3.32	140	24.05	63	2001

Figure 15: A subset of the dataset.

To find the first principal component of our data, it is necessary to preprocess the data. Looking at our data (Fig. 15) it is observed, that some variables such as *Direct Light* are measured in numbers that are much larger than e.g. *Wind Velocity*. For example, for Bulding 3, Direct Light is 513.74 whereas Wind Velocity is 5.52.

If this difference in scale and possibly offset is not handled, then the PCA model will only focus on variables measured in large numbers. It is desirable to model all variables, and pre-process them. This procedure is called Standardizing and will make each column have the same size so that all variables have equal opportunities to be modelled. Standardization means that from each variable, the mean value is subtracted and then the variable is divided by its standard deviation. It is crucial to notice that each variable is transformed in the same size and in the process, each variable will have negative as well as positive values because its average has been subtracted. The ***scikit-learn*** Python library supports this prepro-

cessing.

With this pre-processing of the data, PCA can be performed. The rationale behind performing PCA on a data set is the idea that hopefully much, or perhaps even most, of the variation seen can be attributed to just a few of the most important principal components. A highly correlated data set can often be described by just a handful of principal components. Equally, it is possible for the analysis to produce no useful results at all if the original variables are highly uncorrelated.

6.3 Eigendecomposition - Computing Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues are numbers and vectors associated to square matrices. Together they provide the eigendecomposition of a matrix, which analyzes the structure of this matrix. The eigenvectors and eigenvalues of a covariance matrix represent the core of a PCA.

The eigenvectors (principal components) determine the directions of the new feature space and the eigenvalues determine their size. In particular, eigenvalues explain the variance of the data along the new feature axes.

The typical approach of PCA is to perform the eigendecomposition on the covariance matrix Σ, which is a $d \times d$ matrix where each element represents the covariance between two features. The covariance between two features is calculated as follows:

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right) \left(x_{ik} - \bar{x}_k\right).$$

We can summarize the calculation of the covariance matrix via the following matrix equation:

$$\Sigma = \frac{1}{n-1} \left(\left(\mathbf{X} - \bar{\mathbf{x}}\right)^T \left(\mathbf{X} - \bar{\mathbf{x}}\right) \right)$$

where $\bar{\mathbf{x}}$ is the mean vector $\bar{\mathbf{x}} = \sum_{k=1}^{n} x_i$. The mean vector is a d-dimensional vector where each value in this vector represents the average index of a feature column in the dataset.

Python's library ***numpy*** provides us with the appropriate functions in order to calculate the covariance matrix. The covariance matrix of the standardized data is presented in Table 2.

Covariance Matrix								
1.1428	0.3351	-0.3964	0.9973	0.9605	-1.0950	0.3796	-0.0215	0.4324
0.3351	1.1428	-0.5148	0.3195	0.0186	-0.3580	0.7082	-0.0283	-0.3546
-0.3964	-0.5148	1.1428	-0.0696	-0.2682	0.1870	0.1911	0.3048	-0.0525
0.9973	0.3195	-0.0696	1.1428	0.9688	-1.0702	0.6978	0.0486	0.0608
0.9605	0.0186	-0.2682	0.9688	1.1428	-0.9043	0.4199	0.1106	0.0862
-1.0950	-0.3580	0.1870	-1.0702	-0.9043	1.1428	-0.5544	-0.1211	-0.3209
0.3796	0.7082	0.1911	0.6978	0.4199	-0.5544	1.1428	0.4188	-0.6275
-0.0215	-0.0283	0.3048	0.0486	0.1106	-0.1211	0.4188	1.1428	-0.4553
0.4324	-0.3546	-0.0525	0.0608	0.0862	-0.3209	-0.6275	-0.4553	1.1428

Table 2: The variability of the 9 parameters.

We can now perform an eigendecomposition on the covariance matrix. Numpy performs the calculations required to generate the eigenvectors. The resulting eigenvectors are shown in Table 3.

Eigenvectors								
-0.4649	-0.2123	0.0030	-0.0623	0.1472	0.0904	-0.4649	-0.1075	0.2430
-0.2110	0.2904	-0.5996	0.2457	0.3635	0.0954	-0.4023	0.0832	-0.1105
0.1289	0.1932	0.6523	0.5636	0.0457	0.0530	-0.4142	-0.1106	-0.1544
-0.4717	0.0022	0.1336	0.1665	-0.2024	-0.4938	0.0455	0.6642	0.5503
-0.4245	-0.1032	0.1746	-0.3081	-0.4388	0.5157	-0.1707	-0.0378	-0.2481
0.4743	0.0962	-0.1027	-0.0648	-0.2192	0.3917	-0.2751	0.6399	0.6822
-0.3011	0.5085	-0.0027	0.3104	-0.0303	0.4369	0.5282	0.0127	0.1518
-0.0508	0.4239	0.3726	-0.5938	0.5377	-0.0178	-0.0157	0.1831	0.1337
-0.0390	-0.6122	0.1310	0.2051	0.5217	0.3539	0.2532	0.2887	0.1843

Table 3: The 9-dimensional vector space.

By this procedure the eigenvalues are also obtained and Table 4 presents these particular values.

Eigenvalues
4.69140798e+00
2.37894062e+00
1.66148403e+00
8.31803110e-01
6.19167215e-01
8.37483937e-02
1.91629362e-02
-1.65892178e-16
-5.08833845e-17

Table 4: The resulting eigenvalues after implementing eigendecomposition.

6.4 Selecting Principal Components

The main purpose of a PCA is to reduce the size of the original feature space by projecting it into a smaller subspace, where the eigenvectors will form the axes.

In favor of deciding which eigenvector(s) can be dropped with the least loss of information for the construction of lower-dimensional subspace, the inspection of correlative eigenvalues is needed: *The eigenvectors with the lowest eigenvalues hold the least material about the dispensation of the data, thus being the ones to be dropped.*

To accomplish this, the usual procedure is to rank the eigenvalues (as we can see in Table 5) from highest to lowest and then to select the top k eigenvectors.

Eigenvalues in descending order
4.69140797635
2.37894062499
1.66148402952
0.831803110161
0.619167214858
0.0837483936743
0.0191629361635
1.658921776e-16
5.0883384456e-17

Table 5: The 9 ordered eigenvalues.

If the data is autoscaled, each variable has a variance of one. If all variables are orthogonal to each other, then every component in a PCA model would have an eigenvalue of one since the preprocessed

cross-product matrix (the correlation matrix) is identity. According to the Kaisers' rule [3], if a component has an eigenvalue larger than one, it explains variation of more than one variable.

After sorting the eigenpairs, the next question is "how many principal components are we going to choose for our new feature subspace?" At this point, it is crucial to determine the proper number of components more strongly than in the exploratory or casual use of PCA.

One of the more popular approaches is ***cross-validation***. S. Wold established cross-validation of PCA models [24] and then several slightly different approaches have been developed.
The notion of cross-validation is to count out part of the data and then determine the left-out part. If this is done prudently, the estimation of the left-out part is independent of the actual left-out part. Hence, too optimistic models due to over fitting are not possible.

6.5 Explained Variance

In statistics, ***explained variance*** measures the proportion to which a mathematical model accounts for the variation of a given data set. This practical moderation can be determined by the eigenvalues. Thus, the explained variance provides us with how much information can be traced to each of the principal components. ***Plotly*** let us visualize this measure and create an interactive chart.

Fig. 16 shows that most of the variance (45.6109% of the variance to be precise) can be explained by the first principal component alone. The second principal component still bears some information (23.1285%).

The first 4 principal components all together contain 92.9797% of the information. The rest of the principal components can safely be dropped without losing to much information. This can be verified by Fig. 17.

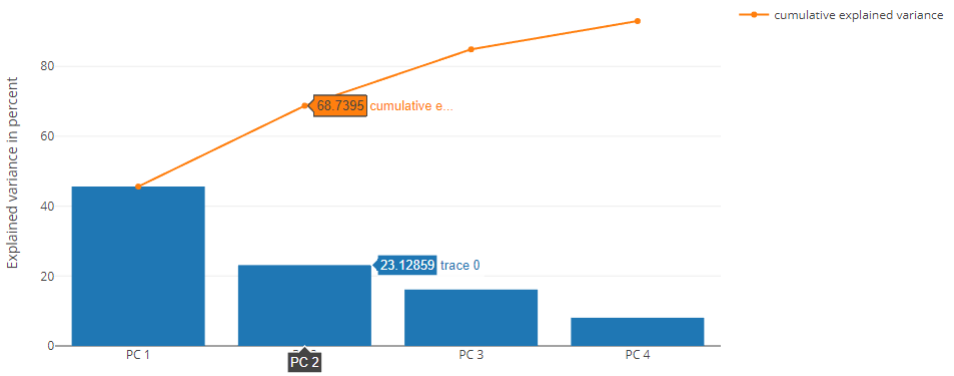


Figure 16: Explained variance (2 principal components).

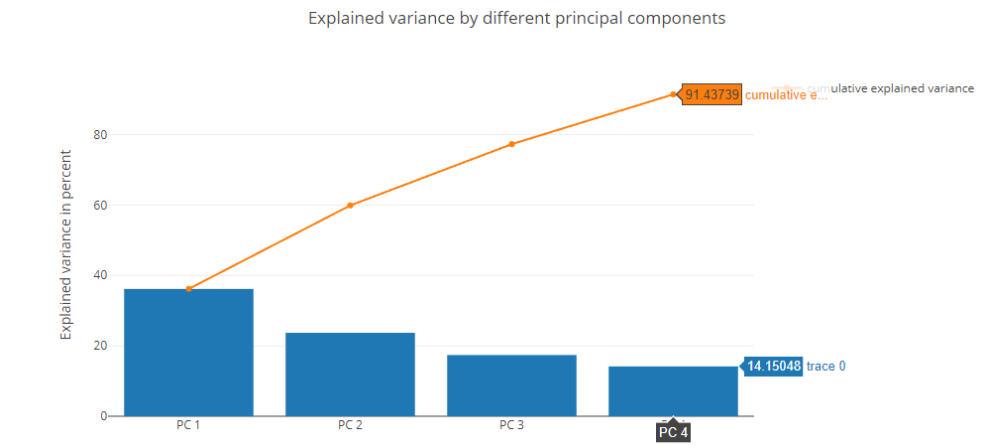


Figure 17: Explained variance (4 principal components).

For the purpose of this study, we will focus only to the first two principal components.

6.6 Projection onto the new Feature Space

The construction of the projection matrix that will be used to transform the data and then project them onto the new feature subspace is the indeed impressive part. It is basically the matrix of our sequence of the top k eigenvectors. Here, we are reducing the 9-dimensional feature space to a 2-dimensional feature subspace, by choosing the "top 2" eigenvectors with the highest eigenvalues to construct our 9×2-dimensional eigenvector matrix $\mathbf{W}$. Table 6 presents the projection matrix of the data.

Projection Matrix	
-0.46491564	-0.21239065
-0.21101122	0.29042947
0.12892033	0.19322503
-0.47179446	0.00220149
-0.42458333	-0.10327942
0.47437389	0.09623856
-0.30117674	0.50853169
-0.05085815	0.42397744
-0.0390087	-0.61229922

Table 6: The 9×2-dimensional eigenvector matrix $\mathbf{W}$

It is well known that the readings of a variable can be plotted. *Direct Light* is measured on 8 samples. These 8 values can be plotted in a multitude of ways.

We will use the 9×2-dimensional projection matrix $\mathbf{W}$ to transform our samples onto the new subspace via the equation $\mathbf{Y} = \mathbf{X} \times \mathbf{W}$, where $\mathbf{Y}$ is a 8×2 matrix of our transformed samples. The library ***scikit-learn***, once again, provides us with all the necessary tools to obtain the values of the coefficients for each one of the 8 samples (Buildings).

These factor scores for the first two components are given in Table 7 and the corresponding map (by using ***plotly***) is displayed in Fig. 18.

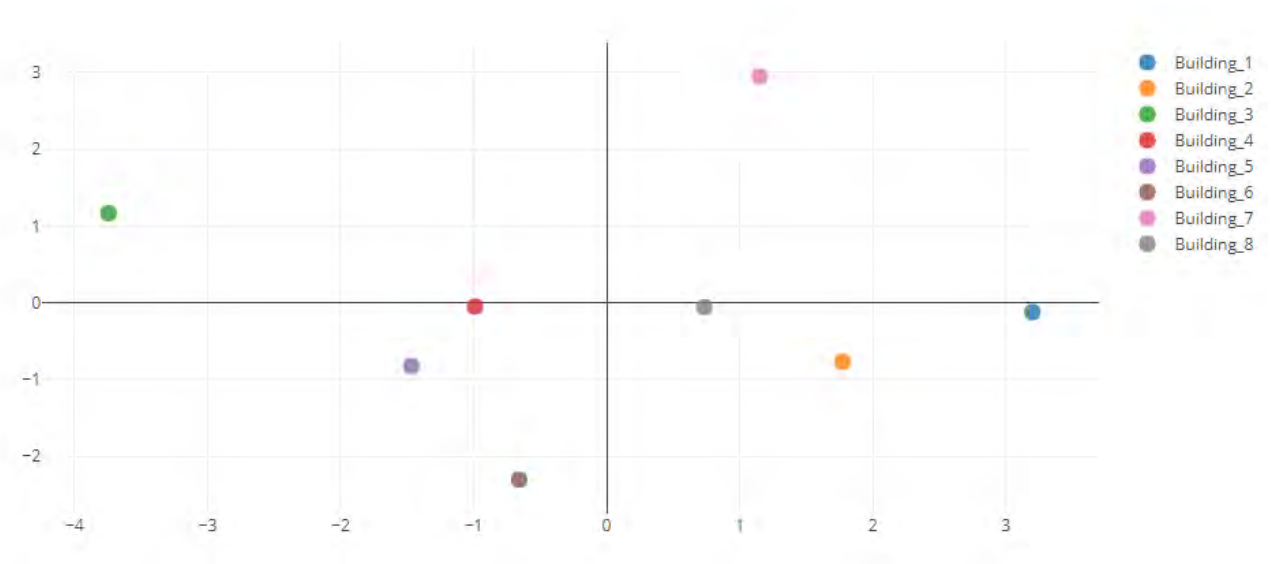


Figure 18: PCA buildings characteristics. Factor scores of the observations plotted on the first two components.

Buildings	**PC 1**	**PC 2**
Bulding 1	3.20058416	-0.1213711
Bulding 2	1.77228699	-0.769164
Bulding 3	-3.74462618	1.1669147
Bulding 4	-0.98928057	-0.04738738
Bulding 5	-1.46813611	-0.8198826
Bulding 6	-0.65747232	-2.30333569
Bulding 7	1.1516584	2.94985547
Bulding 8	0.73498562	-0.05562941

Table 7: Coefficients of each Building for the first two components.

We can see from Fig. 18 that there seem to be certain groupings in the data. For example, *Building 1* and *Building 3* seem to be almost distinctly different in this score plot.

In more detail, the first component separates *Buildings 3, 4 and 5* from *Buildings 1, 2 and 7*, while the second component separates *Buildings 3 and 7* from *Buildings 2, 5 and 6*. The examination of

the values of the contributions, shown in Table 7, complements and refines this interpretation because the contributions suggest that Component 1 essentially contrasts Building 1 with Building 3 and that Component 2 essentially contrasts Building 7 with Building 6.

This suggests that it is possible to classify a Building using these measured variables. Samples that are close are similar in terms of what the components represent which is defined by the loading vectors. Evaluation of the similarities and differences among samples in terms of the raw data is attainable. If two components explain all of the variation in the data, then a score scatter plot will reflect distances in terms of the data directly if the scores are shown on the same scale. That is, the plot must be shown as original scores where the basis is the loading vector.

6.7 Loading: Correlation of a Component and a Variable

To find the variables that account for the differences between the rooms which were observed above, we examine the loadings [8] of the variables on the first two components. The correlation between a component and a variable estimates the information they share. In the PCA framework, this correlation is called a loading. Loadings define what a principal component represents. Hence, they define what linear combination of the variables a particular component represents.

Table 8 shows the loadings of the first two components. In the variable statement, we will include these particular principal components (PC 1 and PC 2), in addition to all 9 of the original variables. We will use these correlations between the principal components and the initial variables to interpret these principal components.

With these, it is possible to explain what the scores of the model represent. For example, *Building 6* has low (negative) score for component 2. This implies that it has a lot of the opposite of the phenomenon represented in loading 2. Hence, this sample has vari-

Features	PC 1	PC 2
Direct Light	-0.46491564	-0.21239065
Wind Velocity	-0.21101122	0.29042947
Wind Direction	0.12892033	0.19322503
Electricity Consumption	-0.47179446	0.00220149
Air Humidity	-0.42458333	-0.10327942
Year of Construction	0.47437389	0.09623856
Number of People	-0.30117674	0.50853169
Air Temperature	-0.05085815	0.42397744
Fan speed	-0.0390087	-0.61229922

Table 8: Correlation of the variables with the first two components.

ation where *Wind Velocity, Number of People and Air Temperature* are low at the same time while e.g. *Fan speed* is high. Also, and this is an important point, certain variables that have low loadings close to zero, such as e.g. *Electricity Consumption and Year of Construction*, do not follow this trend. Hence, the loading tells about what the trend is and also which variables are not part of the trend.

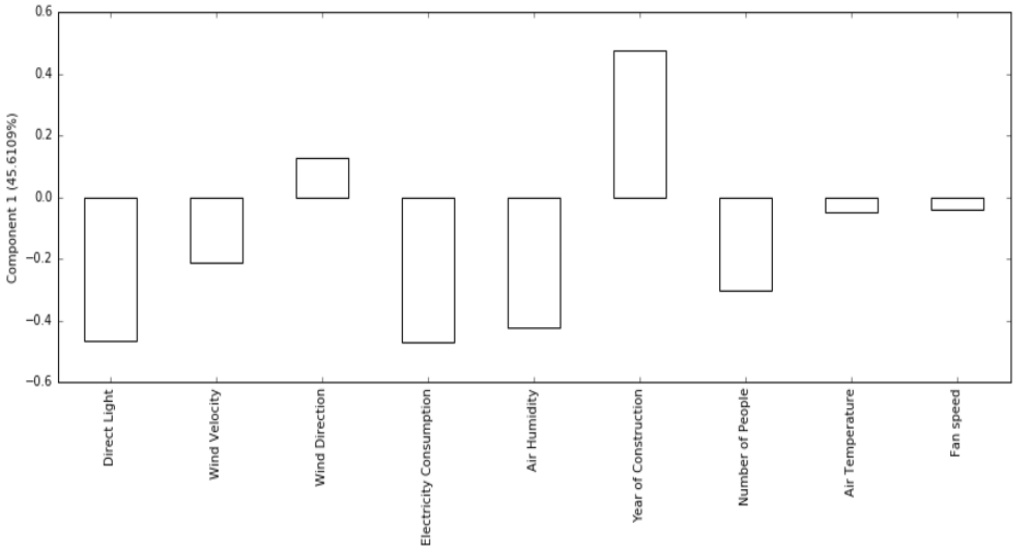


Figure 19: The loadings of the first component.

Examining the loadings of the variables on the first two components, we see that the first component contrasts the construction year with the building's lighting, the humidity of the air, the number of people in the building and the electricity consumption in this building. The second component contrasts the buildings's population, wind speed and air temperature with its lighting and the Fan Speed.

Observing more closely (Fig. 19), the first principal component is strongly correlated with five of the original variables. The first principal component increases with decreasing Direct Light, Electricity Consumption, Air Humidity and Number of People scores. This suggests that these four criteria vary together. If one decreases, then the remaining ones tend to as well. It also increases with increasing the score related to the Year of Construction.

This component can be viewed as a measure of whether the building is in the shade, reduced electricity consumption, understanding that there are few people in this building and the quality of the building (recall that the Year of Construction has a positive high value.) Furthermore, we see that the first principal component correlates most strongly with the Electricity Consumption and Year of Construction. In fact, we could state that based on the correlation of -0.4717 and 0.4743 accordingly, that this principal component is primarily a measure of those two Parameters. It would follow that buildings with high values on this component, would tend to be very new, in terms of foundations, heat insulation, ventilation systems etc. and extremely economical due to energy efficiency. Whereas buildings with small values would have very few of these types of innovations.

The second principal component increases with increasing Wind Velocity, Number of People and Air Temperature. Furthermore, it increases by decreasing Fan Speed. This suggests that buildings with wind velocity, huge number of occupants and very high temperatures also tend not to open the fans, or the ventilation systems do not work as they should. Finally, observing Fig. 20, it is clear that the second principal component correlates most strongly with Num-

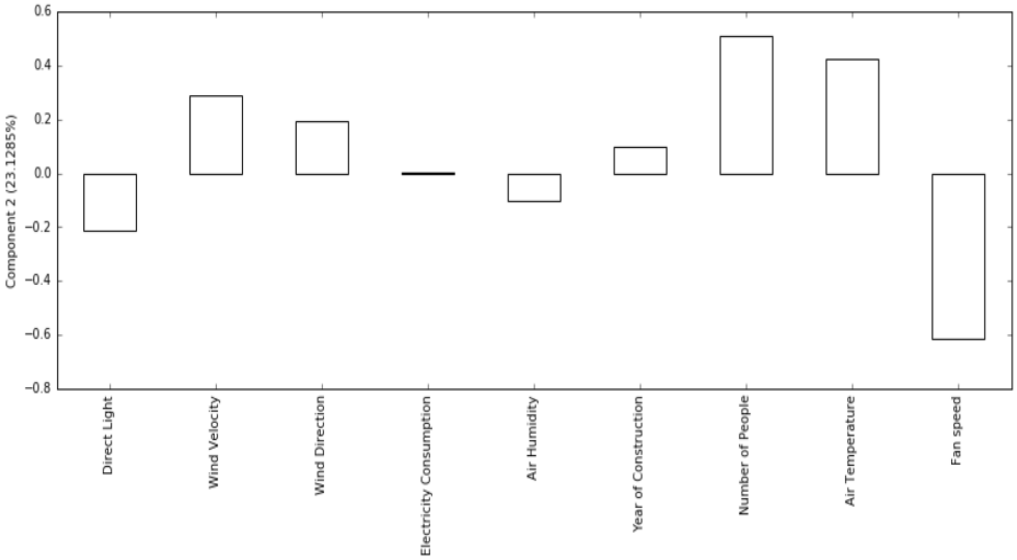


Figure 20: The loadings of the second component.

ber of People and Fan Speed. Thus, this component can be viewed as a measure of the fact that a building with many people does not function properly and effectively its ventilation system (recall that the Air Temperature has a positive high value).

7 Implementation of K-means

7.1 2 Final Clusters

As we have seen in **Chapter 7**, we have the data table in a simple structure. Using Principal Component Analysis, we plotted the samples in the first two principal components (Fig. 21).

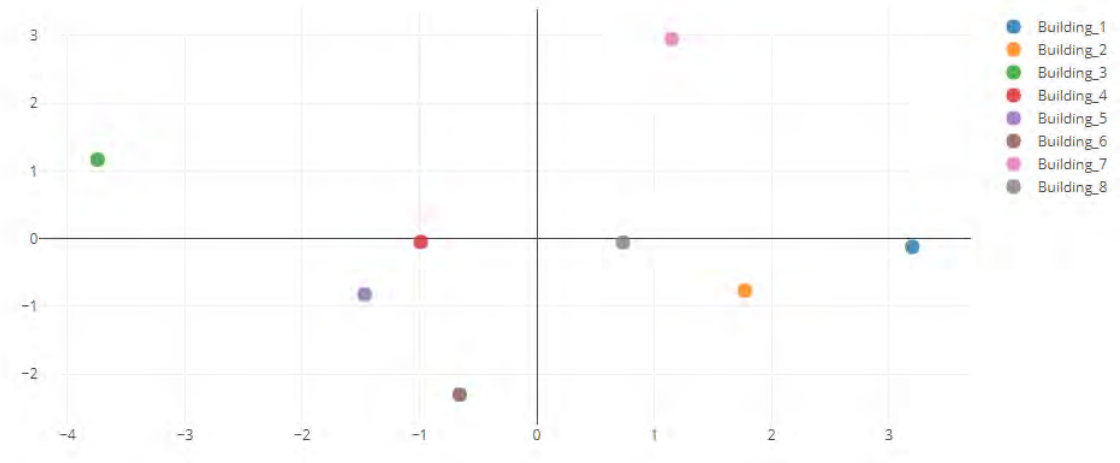


Figure 21: PCA buildings characteristics. Factor scores of the observations plotted on the first two components.

For this example, we used the Python packages ***scikit-learn*** and ***NumPy*** for computations.

Following the steps of the algorithm described in **Chapter 6** we set the k-means function to find 2 clusters. The clustering results are given in Table 9. We note that *Buildings 1, 2, 7 and 8* were classified into the *Cluster 1*, and *Buildings 3, 4, 5 and 6* into the *Cluster 0*.

Buildings	Cluster (0, 1)
Bulding 1	1
Bulding 2	1
Bulding 3	0
Bulding 4	0
Bulding 5	0
Bulding 6	0
Bulding 7	1
Bulding 8	1

Table 9: Separation of the buildings in two clusters in the 2-dim space. PCA was applied and the data were projected onto the new feature space.

Visually, we can see that the K-means algorithm splits the two groups based on their distance from the centroids. Fig. 22 shows the results. Each cluster is distinguished by a different color and mark.

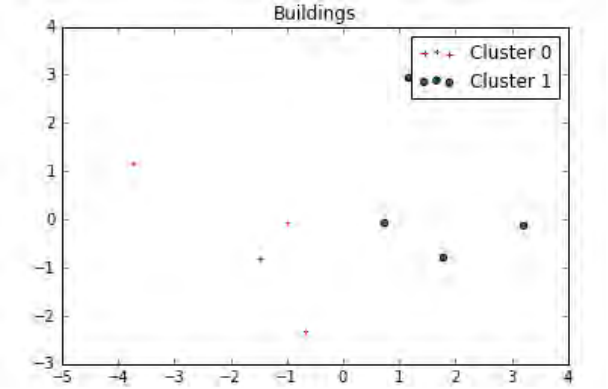


Figure 22: The two clusters in which our samples were split after implementing the K-means algorithm.

At this point, we will apply k-means to the raw data. Using the same function we aim to separate the Buildings in two clusters. However, this time, taking into account the initial 9 parameters and their values. Table 10 presents the cluster assigned to each building and we can see that the results are exactly the same as in the previous case, where we first applied the PCA on the primary data.

Buildings	Cluster (0, 1)
Bulding 1	1
Bulding 2	1
Bulding 3	0
Bulding 4	0
Bulding 5	0
Bulding 6	0
Bulding 7	1
Bulding 8	1

Table 10: Separation of the buildings in two clusters in the 9-dim space. PCA was not applied to the data.

7.2 3 Final Clusters

We would also like to classify the Buildings in three clusters. Now, we set k = 3. Hence, the k-means function will find the three clusters which separate the projection of the data in the best possible way. Table 11 shows the results of the k-mean algorithm to the data after we applied PCA. Fig. 23 allow us to observe these 3 clusters in 2-D feature space. *Buildings 1, 2, 7 and 8* form *Cluster 0*, *Buldings 4, 5 and 6* the second one (*Cluster1*) and *Building 3* is *Cluster 2*.

Buildings	Cluster (0, 1, 2)
Bulding 1	0
Bulding 2	0
Bulding 3	2
Bulding 4	1
Bulding 5	1
Bulding 6	1
Bulding 7	0
Bulding 8	0

Table 11: Separation of the buildings in three clusters in the 2-dim space. PCA was applied and the data were projected onto the new feature space.

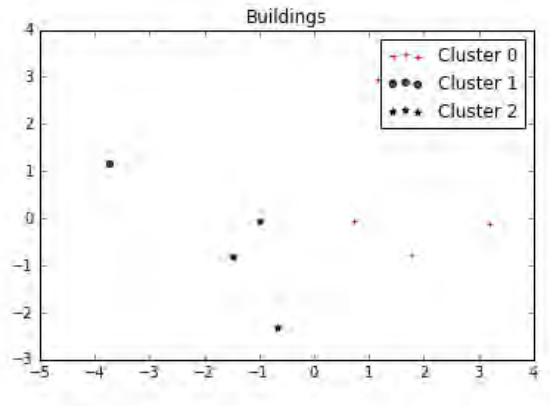


Figure 23: The three clusters in which our samples were split after implementing the K-means algorithm.

By repeating the same procedure as in subsection 8.1, we implement k-means to the raw data. The results are presented in Table 12, from which we understand that the clusters are the same as in the previous implementation of the k-means to the data obtained from the PCA.

Buildings	**Cluster (0, 1, 2)**
Bulding 1	0
Bulding 2	0
Bulding 3	2
Bulding 4	1
Bulding 5	1
Bulding 6	1
Bulding 7	0
Bulding 8	0

Table 12: Separation of the buildings in three clusters in the 9-dim space. PCA was not applied to the data.

7.3 4 Final Clusters

After we decide to separate the data into more clusters, we first apply the PCA to the data and then we apply k-means to their projection onto the 2-D feature space. We set the k-means function to find 4 clusters. Table 13 and Fig. 24 give us the details of these particular clusters. *Cluster 0* contains *Buildings 1, 2 and 8*, *Cluster*

1 comprises *Buldings 4, 5 and 6* while *Cluster 2 and 3* consist of *Buildings 7 and 3*, respectively.

Buildings	**Cluster (0, 1, 2, 3)**
Bulding 1	0
Bulding 2	0
Bulding 3	3
Bulding 4	1
Bulding 5	1
Bulding 6	1
Bulding 7	2
Bulding 8	0

Table 13: Separation of the buildings in four clusters in the 2-dim space. PCA was applied and the data were projected onto the new feature space.

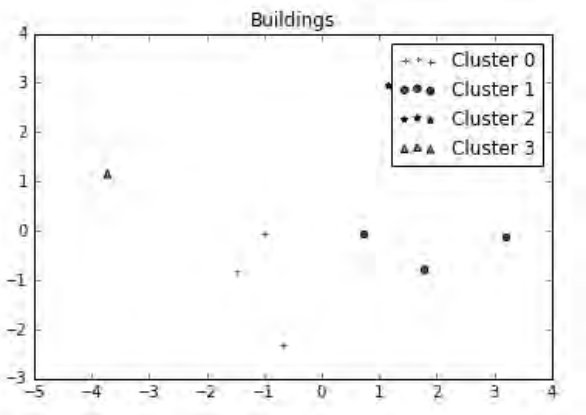


Figure 24: The four clusters in which our samples were split after implementing the K-means algorithm.

Following the same steps as in the previous subsections 8.1 and 8.2 we apply k-means to the raw data. Observing carefully the Table 14, we find that the clusters formed are different from the ones we had previously applied first PCA and then k-means. ore specifically, we see that *Clusters 0 and 2* have been altered and now consist of *Buildings 1 and 7*, and *Buildings 2 and 8*, respectively.

Buildings	Cluster (0, 1, 2, 3)
Bulding 1	0
Bulding 2	2
Bulding 3	3
Bulding 4	1
Bulding 5	1
Bulding 6	1
Bulding 7	0
Bulding 8	2

Table 14: Separation of the buildings in four clusters in the 9-dim space. PCA was not applied to the data.

This inaccuracy is due to the lack of information of the first two components. After applying k-means to the raw data we understand that Building's 1 characteristics are more similar to those of Building 7. Table gave us the impression that the characteristics of Building 1 are closer to Building 2. If the two components had explained 100% - all information - Table and Table would be exactly the same.

8 Conclusions

PCA and K-means clustering methodologies have been introduced for modelling and monitoring energy consumption in buildings. PCA is the basic methodology while K-means allows to compare and confirm the findings of PCA.

Visualization of the correlation matrix allows representing relationships among variables and helps selecting variables to be used for monitoring. Projecting the data onto a new 2-dim feature space we were able to identify the different groups of buildings formed. Abnormal electricity consumption values can be detected (in case we have a bigger dataset) and which independent variables that are mainly responsible for these high values may occur with the aid of the loadings of each component. Furthermore, relations between the buildings can be found by analyzing the correlations between the principal components and the original variables.

Implementing the K-means algorithm we were able to confirm the groups of data which were formed by applying PCA. What we have observed is that when we divided our samples in two and three clusters the results were exactly the same. When we split the data in four clusters, though, K-means presented discrepancies in the way it separated the data before and after the PCA was applied. This is explained if we remember the variation related to the two first principal components. 68.7395% of the variation is explained by these two components. If the two components had explained 100%, all information would be contained in these two components, but for this particular model, almost one third of the variation is still retained in other components, so when we applied K-means after PCA the components were not fully indicative of variations in the data, and big part of the information was lost.

It is true that PCA and K-means clustering are considered to have very different goals and initially do not seem to be related. Nevertheless, as Chris Ding and Xiaofeng He explained [6], there is a deep relationship between them. PCA represents the n data vectors as linear combinations of a small number of eigenvectors. Through this procedure it minimizes the mean-squared reconstruction error.

In contrast, K-means represents the n data vectors as linear combinations of a small number of cluster centroid vectors. This is also carried out to minimize the mean-squared reconstruction error. So the agreement between K-means and PCA is quite good, but it is not exact.

The method can extend towards energy forecasting applications through developing a regression model [14] that directly predicts the value of Electricity Consumption, using all of the potential predictor variables we obtain from a BEMS.

We fit a linear model (Fig. 25) for one of the predictor variables, Air Temperature, to the Electricity Consumption using the ***statsmodels*** library. We can use this linear model for any of the features.

Additional object of future study could be the development of regression models using the the coefficients of the loadings for each one of the predictor variables. In this case, we would be able to investigate and determine how precise is our prediction by using a certain number of principal components.

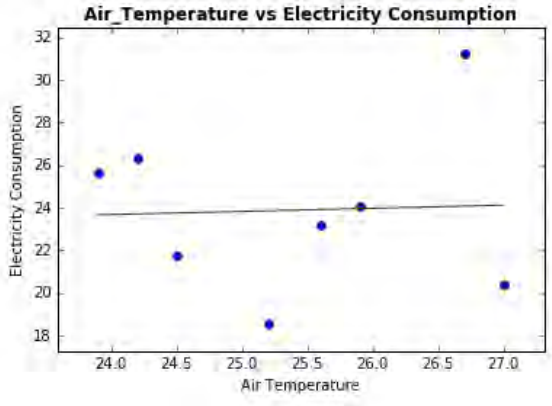


Figure 25: Linear regression between Electricity Consumption and Air Temperature.

References

[1] Hervé Abdi and Lynne J Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[2] AS Ahmad et al. "A review on applications of ANN and SVM for building electrical energy consumption forecasting". In: *Renewable and Sustainable Energy Reviews* 33 (2014), pp. 102–109.

[3] Darren T Andrews and Peter D Wentzell. "Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer". In: *Analytica Chimica Acta* 350.3 (1997), pp. 341–352.

[4] David M Blei. "Hierarchical clustering". In: *Lecture Slides, February* (2008).

[5] Daniel Coakley, Paul Raftery, and Padraig Molloy. "Calibration of whole building energy simulation models: detailed case study of a naturally ventilated building using hourly measured data". In: *First building simulation and optimization conference, Loughborough, UK*. 2012, pp. 10–11.

[6] Chris Ding and Xiaofeng He. "K-means clustering via principal component analysis". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 29.

[7] Denis Hamad and Philippe Biela. "Introduction to spectral clustering". In: *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*. IEEE. 2008, pp. 1–6.

[8] Ian T Jolliffe. "Principal Component Analysis and Factor Analysis". In: *Principal component analysis*. Springer, 1986, pp. 115–128.

[9] D Kolokotsa et al. "Predictive control techniques for energy and indoor environmental quality management in buildings". In: *Building and Environment* 44.9 (2009), pp. 1850–1863.

[10] Geoff J Levermore. *Building Energy Management Systems: Applications to low-energy HVAC and natural ventilation control*. Taylor & Francis, 2000.

[11] Xudong Ma et al. "Supervisory and Energy Management System of large public buildings". In: *Mechatronics and Automation (ICMA), 2010 International Conference on*. IEEE. 2010, pp. 928–933.

[12] McIntyre D Mansson L G. *Technical Synthesis Report: A Summary of Annexes 16 17 Building Energy Management Systems. Energy Conservation in Buildings and Community Systems*. URL: `http://www.ecbcs.org/annexes/annex17.htm`.

[13] Matteo Matteucci. "A Tutorial on Clustering Algorithms, K-Means Clustering". In: *Dip Di Elettron Infomazione E Bioingegneria, Politec Di Milano* (2007).

[14] Peter McCullagh. “Regression models for ordinal data”. In: *Journal of the royal statistical society. Series B (Methodological)* (1980), pp. 109–142.

[15] Luis Pérez-Lombard, José Ortiz, and Christine Pout. “A review on buildings energy consumption information”. In: *Energy and buildings* 40.3 (2008), pp. 394–398.

[16] Gregory Piatetsky. “CRISP-DM, still the top methodology for analytics, data mining, or data science projects”. In: *KDD News* (2014).

[17] S Raschka. *Implementing a Principal Component Analysis (PCA) in Python step by step*. 2014.

[18] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[19] S Schimschar et al. “Towards nearly zero-energy buildings—Definition of common principles under the EPBD”. In: *Ecofys, Politecnico di Milano, University of Wuppertal (Unpublished) for European Commission* (2013).

[20] Sai Sumathi and SN Sivanandam. *Introduction to data mining and its applications*. Vol. 29. Springer, 2006.

[21] Nicolas Verbeke and Nicole Vincent. “A PCA-based technique to detect moving objects”. In: *Image Analysis* (2007), pp. 641–650.

[22] Jaap E Wieringa and Peter C Verhoef. “Understanding customer switching behavior in a liberalizing service market: an exploratory study”. In: *Journal of Service Research* 10.2 (2007), pp. 174–186.

[23] Rüdiger Wirth and Jochen Hipp. “CRISP-DM: Towards a standard process model for data mining”. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. 2000, pp. 29–39.

[24] Svante Wold. “Cross-validatory estimation of the number of components in factor and principal components models”. In: *Technometrics* 20.4 (1978), pp. 397–405.

[25] Shen Yin, Xiangping Zhu, and Okyay Kaynak. “Improved PLS focused on key-performance-indicator-related fault diagnosis”. In: *IEEE Transactions on Industrial Electronics* 62.3 (2015), pp. 1651–1658.

[26] Zhun Yu et al. “A systematic procedure to study the influence of occupant behavior on building energy consumption”. In: *Energy and Buildings* 43.6 (2011), pp. 1409–1417.