POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Computer Engineering

Data science approaches for city-wide power consumption analysis

Supervisor:

Prof. Marco Brambilla

Author:

Sava Ristić (Student ID: 852020)

Academic Year 2017/18

I would like to thank my family and friends for all the support they gave me.

Thank you, prof. Brambilla for guiding me during my thesis work.

I could not have done it without all of you.

Abstract

Urban populations are growing rapidly, causing an immense increase in power consumption. City-level analysis of power consumption is a popular field of research with the ultimate goal to decrease consumption and greenhouse gas emissions and optimize the utilization of power.

This thesis focuses on using data science methods to perform different types of statistical analysis over power consumption in a large urban space by integrating data from various sources. Its goal is to better understand the model of power usage and to recognize which parameters are integral in predicting and optimizing consumption. It demonstrates how data science approaches can be applied to address the most difficult problems faced in the field of energy management.

Questions posed and answered in this work relate to detecting seasonality (periods with similar power consumption that can be grouped together), hourly consumption distribution, climate effect on consumption, the impact of weekends and holidays on consumption and relations between the consumption of different electricity stations with similar characteristics.

Using the city of Milan as a case study, implemented data science approaches resulted in describing the full model of electricity consumption of a large urban center. Techniques described and implemented are generic and can be reused in solving similar problems in other large cities.

Sommario

La popolazione urbana sta crescendo rapidamente, causando un immenso aumento del consumo di elettricità. L'analisi del consumo di elettricità delle città è un campo di ricerca diffuso con l'obiettivo finale di ridurre I consumi, le emissioni di gas serra e ottimizzare l'utilizzo dell'energia.

Questa tesi si concentra sull'utilizzo di metodi di *data science* per eseguire diversi tipi di analisi statistiche dei consumi in una grande città integrando i dati da varie fonti. Il suo obiettivo è comprendere meglio il modello di utilizzo dell'energia e riconoscere quali parametri potrebbero essere importanti per la previsione e l'ottimizzazione dei consumi. La tesi dimostra come si potrebbero applicare approcci *data science* per affrontare i problemi più difficili che si presentano nel campo della gestione dell'energia.

Le domande poste e risolte in questo lavoro sono legate al rilevamento della stagionalità (periodi con consumo simile che possono essere raggruppati insieme), distribuzione del consumo orario, effetti del tempo sui consumi, impatto dei fine settimana e festività e relazioni tra consumo di diverse stazioni con caratteristiche simili.

I metodi studiati sono stati validati su una città di esempio, Milano, per la quale si è fatto uso di dati reali di consumo degli cittadini. I risultati ottenuti evidenziano come i vari modelli di consumo di elettricità raggiungono buoni livelli di qualità.

Contents

Chapter 1

Introduction

1.1 Context

According to the United Nations, current population growth and rate of people moving to cities will mean that around 66 percent of global population will live in urban areas by 2050 [41]. In today's age, people overwhelmingly rely on electrical power for their everyday lives. Electric power systems are the critical infrastructure underlying all other infrastructure systems. Cities consume a tremendous amount of energy; a typical industrialized city of 1 million in the United States of America uses around 10.000 MWh per year, which is 10.000 kWh per customer [1]. Most of the electricity comes from coal and gas. With rapid urbanization and industrialization, cities have become central in addressing the problem of climate change. There is an environmental pressure to reduce overall consumption and greenhouse gas emissions. On the other side, having sufficient energy supply is vital for the community. The utility companies are responsible for maintaining and continuously improving their services. Disruptions in the power system can have severe economic and social consequences, even for short periods of time [37].

Progress in the domain of Internet of Things (IoT) and Information and Communication Technology (ICT) led to smart cities emerging as a solution for common urban problems. In particular, there is a special focus on smart grid projects that use digital communications technology over electricity supply network to detect and react quickly to changes in electrical demand. Smart electricity meters measure energy consumption and power quality at intervals of an hour or less. Readings from meters in a specific geographical area are transmitted in real-time to a central location where they are stored and analyzed.

As the cost of solar panels decreases, a considerable amount of power is being generated by smaller entities. Understanding the continuous pattern of supply and demand is even more important given storage insufficiency and constraints.

All these factors are enhanced by recent deregulations in the global energy market make analysis of energy consumption a primary research area in power systems planning and management. Monitoring and analyzing power as a part of future smart cities is important for predicting and optimizing consumption, maximizing renewable generation and finding power leaks.

Existing technology can monitor, collect, store, analyze and exploit a large amount of data involved in this process, and can do it in an easier, cheaper, and more meaningful. For this, the use of data science approaches for increasing energy efficiency is attracting even greater interest in this area.

Data science is an interdisciplinary field that unifies scientific methods, processes, algorithms, and systems from statistics, data analysis and machine learning in order to understand and analyze phenomena with data in various forms, both structured and unstructured.

Electricity forecasting presents an important challenge for amplifying future developments since inaccurate forecasting rises the operating cost of the utility company. To facilitate better planning, utility companies maintain databases with energy consumption and usage patterns of major appliances. The accuracy of the information from these databases can potentially transform demand forecasting and energy conservation from passive historical data-based activities to active data-driven operations that could identify the trends and support decision-making.

Although there is a significant number of other studies concentrated on the analysis of consumption, no other addresses a specific set of questions concerning large urban consumption posed in this work.

In order to model the system of power consumption within large cities, consumption is categorized into public lights (streetlights), households and industry. It should be taken into account that some customers also generate their own power, predominantly by utilizing solar panels.

The goal of this thesis is to understand and accurately describe the power consumption model of a large city, with a methodology that can be replicated in other urban areas.

.

1.2 Problem statement and questions

As this thesis aims to model electricity consumption with all the limitations and restrictions, questions emerge about what kinds of analyses are required. They can be divided into problems regarding household, industrial and public lights analysis of the consumption data.

These questions could be posed as:

- Can non-trivial periods of the year (that do not necessarily match astronomical or meteorological definition of seasons) be found that have the same or very similar consumption throughout the day for individual stations? Can a year be divided into intervals with similar behavior? These questions regard mainly public lights stations.

- How is the consumption distributed during the day (hourly) throughout the whole year or in specific seasons? Is the distribution Gaussian? In the case that it is not, can it be normalized? Given clusters of 'seasons', does the hourly consumption distribution change? How can results be interpreted? These questions also chiefly regard public lights stations.

- Is there a significant difference in consumption during working and non-working days? On which days of the week do people consume the most? Is it related to weekdays, working days or holidays, and in which way? These questions mainly refer to stations with mainly households and industrial consumption.

  Could information about weather conditions be used to help predict and describe consumption? How does temperature affect consumption? Does solar irradiance have a direct effect on decreased total consumption through solar power generation for the stations with a high percentage of a potential generation? Answers should be provided for household, industrial and public lights stations

- Could consumption of different stations with the same or very similar characteristics be expected to have similar behavior? Is it possible to describe consumption of one station given another station with similar metadata about their users? What are the characteristics by which they can be compared? In case it is not possible, why? With incomplete data provided for a particular station could results of methods be reused from ones with complete data and similar characteristics? These questions are of interest for all the stations: public lights, industrial and households.

1.3 Proposed solutions

Data analysis should be done by implementing different types of statistical data science methods in order to answer the previously posed questions and produce reasonable results.

Types of analysis that should be done to compare consumption are:

- per season by performing clustering over data to see if there are underlying patterns in consumption
- per hour by visualizing and using Normality tests to describe consumption distribution in general and in specific seasons
- during weekends, weekdays and holidays by using clustering and other statistical approaches to compare consumption distributions
- according to weather by checking for correlation between different meteorological conditions and consumption
- between different stations by using MSE and Student's T-test as well as approaches for correlation and covariance testing

The detailed explanation of the analysis with solutions is found in later sections of the thesis work, specifically chapters 3, 4 and 5.

1.4 Structure of the thesis

The thesis is organized as follows:

- Chapter 2, "Background for data science analysis" gives a general overview of relevant data science concepts. The chapter describes the data science pipeline and technologies used for implementing this thesis.

- Chapter 3, "Related works" presents relevant scientific research that has been done to address similar issues in fields related to energy consumption. Ultimately, it highlights the contribution of this work to existing scientific studies.

- Chapter 4, “Power consumption analysis” describes the main idea behind the thesis work. It proposes a conceptual solution to problems of power consumption in a large city while focusing specifically on the environment of this project as listed in the introduction.

- Chapter 5, “Experiments and discussion” describes the system’s implementation in analyzing the problem, validate ideas and answer the previously stated questions. It provides a report, with an interpretation of the obtained results.

- Chapter 6, “Conclusions” provides the overall summary of the thesis work with a short discussion of analysis and obtained results. Finally, it offers ideas for future works and improvements.

Chapter 2

Background for data science analysis

This chapter describes the theoretical background of the approaches and tools used in developing this thesis. The chapter is divided into two parts: the first consists of the description of scientific methods, while the second focuses on the relevant technologies and tools.

2.1 Data science pipeline

This section presents the data science workflow through explaining each step necessary in producing meaningful results from the data.

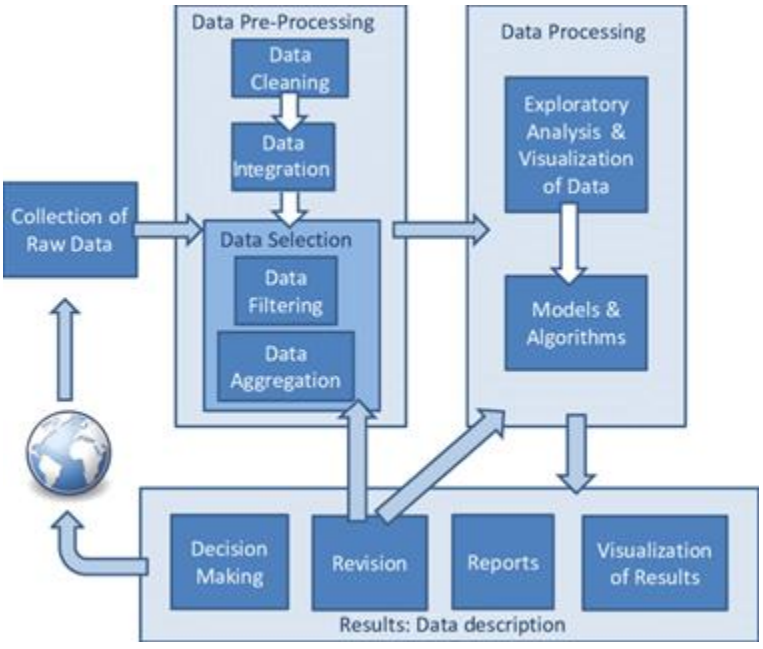

Figure 2.1 Data science processes

Figure 2.1 demonstrates data science pipeline with basic processes. Processes include data collection, data pre-processing, statistical methods for data processing and data description (specifically focusing on clustering, covariance and correlation, mean squared error, and Student's T-test).

All the methods described in this section will be explained in greater detail in chapter 4, especially the way in which they could be utilized to solve the problems posed in section 1.2.

2.1.1 Data collection

Data collection or data gathering is an established, systematic process of capturing and measuring needed information by identifying available datasets from a variety of sources, in order to answer relevant questions about the system.

One of the ways of data gathering is from the web Application Programming Interfaces. Application Programming Interface (API) is a set of functions, protocols, and tools that allow the creation of application software which could access the features or data of various software components.

Web API is usually an HTTP based interface that uses URLs for controls. It is a way to gather data while avoiding unnecessary visual interfaces. It can be considered as a limited shortcut for a database of a web service. APIs provide an easier and quicker access to data from a specific network.

APIs are generally coupled with a set of documentation called API specification. API specification maps resources and operations associated with an API in a format that is intuitive, independent of any programming language and both human and machine readable. It is usually a list of all possible functions and commands with their descriptions and typical requests and responses.

API call or API request is a term that refers to using URLs to call the specific function or collect resources. Each API call supplies a URL, possibly with extra parameters, and receives a response. Parameters are information appended to the end of URL used to specify which resource is needed, in which format, and to define the quantity, range, etc. Finally, the system provides a response and possibly data, most often in XML or JSON format.

Some APIs are publicly available, but most of them require user authentication in form of an API key given after registration. Another consideration refers to the rate limit on API calls. The Rate limit checks and can reject requests if too many are received at once or from the same user in order to prevent overloading the system.

2.1.2 Data pre-processing

Raw data is usually incomplete, inconsistent, and contains some errors that need modifications or transformations. Data pre-processing is used to resolve these problems by converting data into a more usable form. It is a very time-consuming task; in a typical data science and machine learning project, pre-processing step takes up around half of the time. Data pre-processing typically consists of data cleaning, data integration, data reduction and data transformation.

2.1.2.1 Data cleaning

Incorrect or inconsistent data can affect the analysis process leading to false conclusions, thereby decreasing the quality and precision of the results. Hence, one of the first and most important steps is data cleaning. Data cleaning is the process that intends to detect, correct or remove incomplete, incorrect, inaccurate, irrelevant, outdated, corrupted, incorrectly formatted, duplicated, redundant, inconsistent or missing records from a dataset. Main data cleaning tasks are:

- Fill in or removal of missing values
- Outliers detection and removal
- Correction of inconsistent data using domain knowledge

2.1.2.2 Data integration

Data integration is the step where data from multiple resources and with different formats and properties are processed and merged together in order to give the analyzer a more cohesive view. In this step, the metadata of the different datasets should be carefully examined. Finally, schema integration for matching entity is conducted.

2.1.2.3 Data reduction

In the data reduction step, the relevant data for the following analysis is identified and the dataset subsets are selected. Features that are considered irrelevant are excluded from further analysis. Data reduction consists of:

- reducing the number of attributes - Dimensionality reduction
- reducing the number of attribute values
- reducing the number of tuples

2.1.2.4 Data transformation

In the data transformation phase, the selected data are transformed or consolidated into forms necessary for analysis. Many procedures can be used depending on the data format and analysis:

- Smoothing - removing noise from the data
- Normalization - scaling values within the specific range
- Aggregation – data summarization
- Generalization - concept hierarchy climbing
- Attribute construction - new attributes construction from the given ones

2.1.3 Relevant data science methods

Data science concepts relevant to this thesis work are listed and briefly theoretically discussed in this section. Methods described are clustering, covariance and correlation, mean squared error and Student's T-test.

2.1.3.1 Clustering

Clustering is the process of grouping a set of objects into classes of similar objects. It is a form of unsupervised learning. The typical goal of data clustering is to obtain high intra-cluster similarity (making data within a cluster are similar as possible) and low inter-cluster similarity (making data from different clusters as dissimilar as possible).

Clustering can be roughly distinguished as:

- Hard clustering where each object belongs to a single cluster
- Soft clustering or fuzzy clustering where each object belongs to multiple clusters with a certain likelihood

The quality of a clustering can be evaluated both with internal and external indices. Internal indices measure the preciousness of a clustering without relying on prior knowledge. Good results of internal indices do not always provide effective results on the unseen data. External indices are alternatives to internal criterions. They are used to measure how much the cluster results match externally provided information. Both of these measures are useful for:

- choosing the correct number of clusters

- evaluating the clustering tendency of the data and understanding whether non-random structures really exist in the dataset
- evaluating the results of a cluster analysis without the need of an external information for data class labeling
- evaluating the results of a cluster analysis in the situation when data class labels are provided externally
- comparing clusters and determining which one is better

2.1.3.2 Covariance and correlation

Covariance is a measure that shows how much two random variables vary together.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)}$$

Where:

X and Y are random variables

μ_x is the mean value of X

μ_y is the mean value of Y

n is the number of items

A large covariance signifies a strong relationship between two variables. However, covariances over datasets with different scales cannot be compared. In order to get a more meaningful result, covariance can be divided the by the standard deviation to get correlation coefficient.

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where:

σ_x is the standard deviation of X

σ_y is the standard deviation of Y

Correlation represents any statistical relationship between two random variables. It shows if two quantitative variables have a linear relationship with each other and in which extent.

2.1.3.3 Mean squared error

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator does not account for information that could produce a more accurate estimate [40].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Where:

y is the true value

$\hat{y}$ is the predicted value

n is the number of items

Since the MSE squares all the differences, this measure does not have the same scale as the original measurement. To return to the original scale, MSE should be squared. This mathematical operation yields the accuracy measure called root mean square error (RMSE) [42].

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

2.1.3.4 Student's T-test

A T-test is a type of statistical test that is used to compare the means of two groups to check if groups are significantly different from each other [46]. The t-value obtained is used to show how great the difference between groups is. For each t-value calculated, there is a corresponding level of significance or p-value, which indicates how strong the evidence against the null hypothesis is. Most commonly, p-value lower than 0.05 or 0.01 is used to reject the null hypothesis.

Three main types of T-tests are:

- A 'One sample T-test' compares the mean of a group against a specified theoretical mean

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

Where:

m is the sample mean

μ is specified theoretical mean

s is sample standard deviation

n is the sample size

- A ‘Paired samples T-test’ compares means from the same group at different times

$$t = \frac{m}{s/\sqrt{n}}$$

Where:

m is the sample mean

s is sample standard deviation

n is the sample size

- An ‘Independent samples T-test’ compares the means for two unrelated groups

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

Where:

A and B are the two groups

mA is the mean of A

mB is the mean of B

nB is the size of B

nA is the size of A

S^2 is an estimator of the variance of the two groups

2.2 Relevant technologies

Technologies used in the context of the thesis work are briefly explained in this section, focusing theoretically on Apache Spark and Python background. Their practical implementation and the specific libraries used are described in section 5.2.

2.2.1 Apache Spark

Apache Spark [35] is an open-source computing framework for large-scale data processing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. Spark gives data scientists and engineers a powerful and unified processing engine built around speed, ease of use and sophisticated analytics. Spark can be used to build applications and libraries or perform ad-hoc data analysis interactively. It features a stack of components including Spark Core, SQL, Spark Streaming, and MLlib for development of machine learning pipelines, and GraphX for graph processing. All these components can be seamlessly combined in the same application. Spark can be deployed and run in different environments, e.g. as a standalone, Hadoop, Apache Mesos, Docker or on the cloud. It can read data from diverse sources including HDFS, Kafka, Apache Cassandra and MySQL and different formats like Parquet, CSV or JSON. Complete Spark ecosystem containing the most important components with possible applications, environments and data sources is shown in figure 2.2.

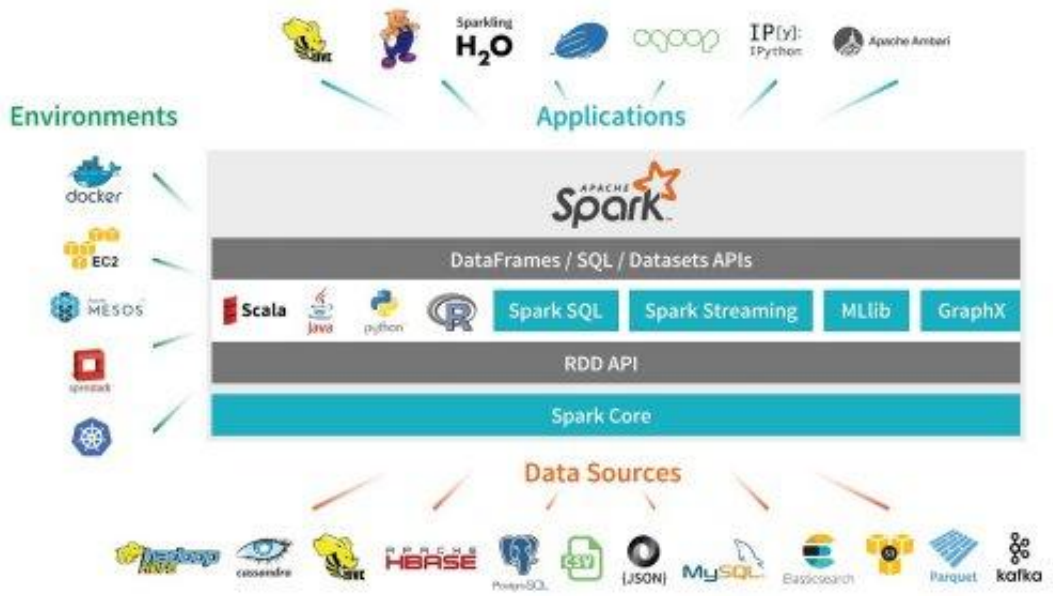

Figure 2.2 Apache Spark environment

The fundamental part of the overall project presents Spark Core. It is responsible for basic I/O functionalities, distributed task dispatching, scheduling and monitoring, networking with different storage systems, fault recovery and efficient memory management, exposed through an application programming interface for Java, Python, Scala, and R.

Apache Spark supports the notion of data abstraction as a distributed collection of objects, called Resilient Distributed Dataset (RDD). RDDs are fault tolerant and immutable collections of elements data that can be stored in memory or a disk across a cluster of machines. Whole data is divided across nodes in the cluster, in order to be operated in parallel with a low-level API that provides transformations and actions [34]. Operations over RDDs are evaluated lazily. By keeping track of operations that produced it, lost data can be automatically reconstructed in case of a failure.

In order to make large datasets processing even easier, Spark provides SQL component built on top of Spark Core. It introduces data abstraction called DataFrame, which is an immutable distributed collection of data. Dataframes provide support for structured data, similar to a table in a relational database, allowing higher-level abstraction. Spark SQL uses a domain specific language to manipulate DataFrames in Scala, Java, and Python. Furthermore, it leverages the power of declarative queries and optimized storage by providing SQL language support.

2.2.2 Python

Python [36] is a general purpose interpreted, object-oriented, interactive, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python has a fully dynamic type system and uses automatic memory management [39]. It supports object-oriented, imperative, functional and procedural programming paradigms.

Python's design philosophy emphasizes code readability, with easy to learn syntax that allows programmers to write concepts in fewer lines of code, reducing the cost of maintenance. By prioritizing readability over speed or expressiveness it emphasizes the importance of programmer effort over computer effort [39].

Python supports modules and packages, which encourages program modularity and code reuse [36]. Its standard library provides the language with a large number of additional libraries and extensions.

The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. Although there are other implementations available, the de facto standard for the language is the CPython implementation which is a bytecode compiler and interpreter [39].

Chapter 3

Related work

This chapter presents other scientific works done in related fields with taxonomy and a short description of the most relevant works. It is divided into four sections, in order to group together similar problems. The first section concerns smart cities. The second deals with smart grids with electricity monitoring. The third part focuses on general problems regarding power consumption analysis. The final part specifies previous works done analyzing climate influence on power consumption model.

3.1 Smart cities

A ‘smart city’ is a concept of interconnecting modern technologies in an urban context that provides solutions to enhance the quality and performance of urban service. It heavily relies on progress in the domain of Internet of Things (IoT) and Information and communication technology (ICT) in order to solve common urban problems.

Some of the relevant examples from the vast number of works are:

- Ibrar Yaqoob, et al. (2017) [3] discusses communication and networking technologies used in smart cities with similarities and differences among them. The paper describes future and emerging technologies, modern communication technologies, IEEE wireless technology standards, objectives, network classes, and modes of operation with the presentation of case studies on different cities.

- Andrea Zanella, et al. (2014) [4] provides a comprehensive survey of the enabling technologies, protocols, and architecture for a smart city IoT. Furthermore, the paper presents and discusses the technical solutions and best-practice guidelines adopted in the Padova Smart City project.

- Amit Kumar Sikder, et al. (2018) [12] gives an overview of the Smart Lighting System (SLS) and different IoT-enabled communication protocols, which can be used in the context of a smart city. It also describes the different usage scenarios analysis for IoT-enabled indoor and outdoor SLS and provides an analysis of the power consumption leading to a reduction of up to 33.33%.

- Arthur Souza, et al. (2017) [7] provides a data integration approach for smart cities. The main challenges in managing smart city data are related to the need of unifying information sources and the ability to share information across multiple agencies within a city and among cities in a metropolitan region, thus making it possible to obtain a complete picture of urban activity.

- Dan Puiu, et al. (2016) [8] CityPulse framework supports smart city service creation by means of a distributed system for semantic discovery, data analytics, and interpretation of large-scale (near-)real-time Internet of Things data and social media data streams. The goal is to break away from silo applications and enable cross-domain data integration.

- Takahiro Komamizu, et al. (2016) [15] develops a real-time analytical system based on StreamOLAP (an On-Line Analytical Processing system for streaming data) for smart city data analysis which enables OLAP-style analytics over streaming data and applies the analytical system for real-world smart city data.

- Mehdi Mohammadi and Ala Al-Fuqaha (2018) [13] sheds light on the challenge of underutilizing the big data generated by smart cities from a machine learning perspective. In particular, the phenomenon of wasting unlabeled data is presented. Semi-supervision is proposed for smart cities to address this challenge, with a goal of providing different levels of knowledge abstraction.

3.2 Smart grids and power systems

A ‘smart grid’ refers to an electricity supply network of transmission lines, substations, and transformers that delivers electricity from the utility to its consumers while using digital communications technology to detect and react quickly to changes in electrical demand.

Most relevant papers:

- Hassan Farhangi (2010) [6] sets the foundation for the development of the smart grid. It discusses drawbacks of conventional grids and the necessity of incorporating communication and information technologies in order to overcome them. The work describes the basic components, drivers, evolution, and standards of the smart grid.

- V.K. Sood, et al. (2009) [2] discusses some of the smart grid applications and estimates the communication infrastructure requirements of a medium data-intensive smart grid environment.

- Dilan Sahin, et al. (2011) [5] addresses critical issues on smart grid technologies, primarily in terms of information and communication technology (ICT) standards, issues and opportunities.

- M. Brenna, et al. (2012) [14] aims at moving beyond the current concept of smart grids, by providing statistic and dynamic models for the characterization of power profiles of the subsystems considered independent and integrating them into management of energy flows, a Sustainable Energy Microsystem SEM, in order to identify and classify the control variables of the power profiles.

- B. Morvaj, et al. (2011) [9] gives an overview of smart grid features and paradigms of a smart energy city in the framework set for existing cities' evolution and transformation into smart cities.

- Francesco Benzi, et al. (2011) [10] shows the way digital meters are implemented and how interfacing consumer premises can play a more active role and provide relevant benefits to the final user.

- Wil L. Kling and Johanna Myrzik (2013) [11] focuses on an optimal use of multi-energy systems in the urban environment using smart control and communication technologies and the implementation of e-mobility as the key towards highly efficient and carbon-reduced cities.

3.3 Power consumption

With electricity consumption constantly rising and the environmental pressures to reduce greenhouse gases emissions, power consumption analysis is becoming a field of interest for many research works that address similar problems.

Some of the most influential works in the field of power consumption analysis are:

- Geoffrey K.F.Tso and Kelvin K.W.Yau (2007) [16] compares regression analysis, decision trees and neural networks in predicting electrical energy consumption.

- Xinghuo Yu, et al. (2011) [17] proposes a framework for data mining, intelligent analysis and prediction of energy consumption based on electricity meter readings. A self-learning algorithm is developed to incrementally discover patterns in a data stream environment and sustain acquired knowledge for subsequent learning. It characterizes electricity consumption and thus exposes significant patterns and continuity over time.

- Sreenu Sreekumar, et al. (2015) [19] explains forecasting of the hourly electrical load using genetically tuned Support Vector Regression for the smart grid framework for the analysis and betterment of a research in the field of Electrical Power System, i.e. Short Term Load Forecasting (STLF) which is essential for utilities to control, manage and schedule generator units.

- G. Mutani, et al. (2017) [20] presents the analysis of measured energy consumption for residential buildings located in the cities of Torino and Reggio Emilia using correlation and linear regression models between temperature and solar irradiance on power consumption.

- Xiaoli Li, et al. (2010) [21] proposes an intelligent data analysis method for modeling and predicting daily electricity consumption in buildings, with the objective of enabling a building-management system to be used for forecasting and detection of abnormal energy use.

- Alias Khamis, et al. (2010) [22] focuses on the power generation and the best methods for managing excessive power generated without affecting the capacity of power in reserve. It also deals with the effect of high peak load in total power demand.

- Yuansheng Wang and Chunli Chu (2011) [23] explores the characteristics of the energy usage of Tianjin city through time series analysis and compares it with Beijing and Shanghai in order to optimize the composition of energy consumption, with the goal of reducing the CO_2 emission intensity.

- Roberto Pagani, et al. (2009) [24] aims to explore the main factors influencing energy consumption patterns and investigate which has the largest effect in Chinese developing cities with a population greater than two million people. 40 parameters related to the city energy consumption of core areas of 21 major cities were collected and analyzed using bivariate correlation and regression analysis.

3.4 Weather influence on consumption

Among the many factors influencing energy demand, a number of studies have demonstrated that weather variables have a dominating impact on energy consumption. With rapid climate changes and increased weather variability, this field is becoming the focus of scientific research.

Some works published are:

- Suchao Parkpoom and Gareth P. Harrison (2008) [25] investigates how the changing climate will affect Thailand's daily, seasonal, and long-term electricity demand. Regression models are applied to capture daily load patterns across each month of the year.

- Luis Hernández, et al. (2012) [26] presents the relationship found at an experimental level between a range of relevant weather variables and electric power demand patterns inside a smart grid, presenting a case study using an agent-based system.

- Ching Lai Hor, et al. (2005) [27] investigates the impact of weather variables on electricity demand in England and Wales by developing multiple regression models to forecast monthly electricity demand based on weather variables, gross domestic product, and population growth.

- Saša Jovanović, et al. (2012) [28] presents an analysis of the impact of weather conditions on the consumption of electricity for the city of Kragujevac. The paper tries to determine the impact of changes in the mean daily temperature on the power consumption of buildings by using different statistical methods.

- Iain Staffell and Stefan Pfenninger (2018) [29] develops an open framework for quantifying the increasing impacts of weather on electricity supply and demand using different weather models.

- Maximilian Auffhammer and Erin T. Mansur (2014) [30] specifically focuses on empirical literature in peer-reviewed economics journals that focusing on how climate, defined as a long-run average of weather, affects energy expenditures and consumption.

- Huade Guan, et al. (2017) [31] analyzes electricity supply in warm seasons in order to provide a good quantitative representation of the relationship between warm season electricity consumption and weather conditions with the necessary information for long-term electricity planning and short-term electricity management.

With respect to the works referenced above, the aim of this thesis is to use different data science approaches in order to analyze the data from the power provider integrated with the data from the weather conditions. The work focuses on extracting useful features for describing the model of a city-wide consumption in a smart grid and a smart city environment.

Chapter 4

Power consumption analysis

This chapter describes the main idea and motivation behind the project. The intent of this work is to show how data science can be used to model and solve general problems of power consumption in a large city. This chapter focuses more specifically on how to process city-wide data coming from households, industrial or public lights stations by following data science pipeline described in section 2.1. The primary goal is to provide a conceptual solution to the five problems stated in section 1.2. The aim of this chapter is to delve deeper into each problem by explaining each step in the problem and outlining how and why such steps are necessary.

The first step presents solving the problems of data collection for the power and weather conditions, followed by pre-processing of the data in order to clean it and transform it in a way that is appropriate for consumption analysis. The third step is the actual analysis performed to answer the questions related to this thesis work (seasonality, holidays...). The final step is the interpretation of the results obtained in each analysis and drawing the conclusion of the consumption model.

4.1 Data collection

The Power provider typically supplies data about its types of users, dividing them into industrial, households or public lights. Consumption and generation measurements come from secondary stations, meaning that data is sent from local stations in each neighborhood. This data represents an aggregation of all of the households, industrial buildings and public lighting measurements nearby. Local stations upload consumption to a central station, where the data is stored. Data measurements are hourly generated.

Another problem present is finding a suitable provider for the weather data. This work is focused on the data extracted from the weather API. Historical data about solar radiation and general weather conditions had to be provided with at least hourly granularity and for the period corresponding to one of the consumption data.

4.2 Data pre-processing

Data is assumed to have errors and imperfections so cleaning and detecting outliers are needed. Outliers' detection is performed using the Z-score for each point:

$$z = \frac{x - \mu}{\sigma}$$

Where:

x is a data point

μ represents mean of the whole dataset

σ is the standard deviation

Cleaned data collected from power stations and the data from the weather provider distributor should be integrated by matching the datasets by time and location. Unnecessary data should be removed and new columns of aggregated values could be added before proceeding to analyses.

4.3 Seasonality detection

Problem: Can definitive periods of the year be found that have the same or very similar consumption during the day for public light stations? How many and why?

Initial hypothesis and expected behavior: There are periods in a year when consumption is very similar and it would make sense to group them in the same class. Data can be divided into periods, more or less similar to natural seasons.

Solution: Use clustering, and by dividing data into days to represent each day as a vector and implement K-means algorithm to compare distances between vectors.

The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares [43].

Given any set of k centers Z, for each center $z \in Z$, let $V(z)$ denote its neighborhood, that is, the set of data points for which z is the nearest neighbor. Each stage of algorithm moves every center point z to the centroid of $V(z)$ and then updates $V(z)$ by re-computing the distance from each point to its nearest center. These steps are repeated until some convergence condition is met [44].

This algorithm requires the number of clusters to be specified *a priori*. It scales well with a large number of samples and has been used across a range of application areas in a variety of fields [43].

The real class labels of the data results are unknown. The evaluation methods that could be used in this case for evaluation and validation of clustering algorithm and finding the optimal number of clusters include cohesion, separation, and silhouette score.

Cohesion shows how closely related objects in a cluster are, i.e. variation within a cluster. It is computed using sum of squared errors (SSE). SSE measures the distance of each observation to its centroid and then calculates the total sum of the squared errors.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

y is the observation

$\hat{y}$ is observation's centroid

n is the number of objects

Separation can be used as a measure of how distinct or well-separated a cluster is from other clusters.

Silhouette score combines ideas of cohesion and separation by measuring how similar an object is to its own cluster compared to other clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where:

$a(i)$ is average distance between data point i and other data from the same cluster

$b(i)$ is minimum average distance of data point i from all other points in other clusters

The average $s(i)$ over all data of the entire dataset is a measure of how appropriately the data has been clustered. If there are too many or too few clusters, as may occur when a poor choice of k is used in the clustering algorithm, some of the clusters will typically display much narrower silhouettes than the rest. Thus silhouette plots and averages may be used to determine the natural number of clusters within a dataset [45].

4.4 Distribution of hourly consumption

Problem: What is the shape of hourly consumption distribution during the day for the public light stations? Could it be transformed into Gaussian distribution? Is there a difference in hourly distribution between 'seasons' obtained by 'seasonality detection' analysis?

Initial hypothesis and expected behavior: Data consumption comes from Gaussian distribution throughout a day. An increase in consumption is expected during the night and decrease during daily hours.

Solution: Visualize consumption distributions, check normality with normality tests (e.g. two-sided Kolmogorov-Smirnov); normalize distributions if needed; compare with other hours.

The Kolmogorov-Smirnov test is a non-parametric test for measuring the strength of a hypothesis that some data is drawn from a fixed distribution (one-sample test), or that two sets of data are drawn from the same distribution (two-sample test). The two-sample version of the test allows for the comparison of two (not necessarily equal-sized) datasets without any foreknowledge of the underlying distributions [47]. It is frequently used to check whether data comes from the normal distribution.

Normalization refers to sophisticated adjustments intended to bring the entire probability distributions of adjusted values into alignment.

4.5 Effect of non-working days

Problem: Is there a statistically significant difference in consumption between days? Which days have the highest or lowest levels of consumption? Do weekends affect power consumption? Do holidays have an impact on consumption and how much? These questions are more relevant to household and industrial stations since power consumption for public lights is considered to be constant.

Initial hypothesis and expected behavior: Weekends and holidays would affect consumption because more people are at home and less at work. Therefore, for stations with mainly industrial buildings, consumption should increase during weekdays.

Weekends and holidays should not have an effect on public light stations. This analysis could be used to verify this assumption by testing if consumption of public lights really follows the aforementioned pattern.

Solution: Apply K-means algorithm described in 4.3. Confusion matrix with accuracy, precision, and F-score could be used for evaluation of clustering results. Two sample Student's T-test can be used to check if consumption from both working days (from Monday to Friday) and non-working days (Saturday, Sunday, and holidays) come from the same distribution. Kolmogorov-Smirnov normality test should be used before T-test.

A Confusion matrix is a technique used to summarize the performance of a model on a set of test data for which true classes are known. Performance of the model is visualized in a table layout.

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Figure 4.1. Confusion matrix

Figure 4.1 shows the confusion matrix, where true positives (tp) represent the number of correctly predicted positive values; true negatives (tn) represent the number of correctly predicted negative values. False positives (fp) and false negatives (fn) values occur when actual class contradicts with the predicted class.

After obtaining these parameters, accuracy, precision, recall and F1 score can be calculated.

Accuracy is the most intuitive performance measure. It is a ratio of correctly predicted observation from total observations. Accuracy is a great measure but only for symmetric datasets, where values of false positive and false negatives are similar.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall or sensitivity is the ratio of correctly predicted positive observations to all observations in actual class.

$$\text{Recall} = \frac{tp}{tp + fn}$$

F measure (F1 score or F score) is the weighted harmonic mean of Precision and Recall, by taking both false positives and false negatives into account. It is not intuitively as easy to understand as accuracy, but F measure is usually more useful than accuracy, especially when false positives and false negatives are very different.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

4.6 Weather influence

.

Problem: Could weather conditions have an influence on power consumption? How does the change in temperature affect consumption? Does solar irradiance have an effect on solar power generation?

Initial hypothesis and expected behavior: Greater sun exposure of the solar panel would generate more power and decrease overall consumption. During the summer, warmer weather would lead to more power consumption due to greater use of air-conditioners. In the winter, colder temperatures would have the similar effect due to greater use of heaters.

Solution: Use correlation and covariance tests between consumption and solar irradiance. Since generation power is given only as a nominal (maximal possible) value, and without knowing if solar panels actually work, real values have to be approximated using solar irradiance. Correlation and covariance tests could be used as well for temperature and consumption.

4.7 Relation between stations

Problem: Could the consumption of two different stations be related? Can the obtained results from the analysis of one station be applied to another one with similar characteristics? What are the characteristics by which they could be compared?

Initial hypothesis and expected behavior: Data about one station could be reused in order to saves resources (time and money). There is no need for complete data for analysis if there is a relationship (correlation) between stations.

Solution: Perform MSE/RMSE, Student's T-test and correlation and covariance tests between consumption of two stations. Perform the same analysis on the station with the same nominal values in the same period of adjacent years to get the reference for comparing the results.

Chapter 5

Experiments and discussion

This chapter details the practical application of the methods and techniques described in chapter 2 on real data and provides solutions to the problems posed in section 1.2. The chapter is divided into three parts. Firstly, the experimental settings are described. Secondly, the description of the datasets and the data pre-processing in preparation for the analyses. Thirdly, the analyses report provides the results and the interpretation for the implementation of the system.

5.1 Experimental setting

This section identifies data sources for electricity and weather data and briefly describes the platform and libraries used to implement the experiments in the case of the Milan metropolitan area.

5.1.1 Data collection

The data provider for electricity consumption is 'Ricerca sul Sistema Energetico - RSE S.p.A.' [33] in collaboration with 'A2A S.p.A.' [18]. 'RSE' is an Italian company specializing in research and development in fields of electrical power, energy, and environment. Their main activity is aimed at innovation and improvement of the performance of the electricity system by focusing on economic efficiency, safety, and environmental compatibility. 'A2A' is one of the largest companies in Italy in the field of production and distribution of electricity.

Both the solar radiation data and weather data were gathered from 'Meteoblue' API, which has archived and detailed weather data available covering the past 30 years Data was collected for the city of Milan Over the course of two years (2015 and 2016).

5.1.2 Data management platform

The full analysis was implemented and run on Databricks platform.

Databricks [34] is a cloud-native, web-based platform that abstracts the complexities of Apache Spark management, resulting in a highly elastic and efficient platform to build innovative products with IPython-style notebooks. It features a Unified Analytics Platform that accelerates innovation by integrating data science, engineering, and business. Databricks provides a service to manage big data and perform advanced analytics in a simple, reliable way.

Databricks is used in this work to execute Apache Spark 2.2.0 and Python 3 code. All the data was initially uploaded to the platform. The community edition provided a cluster with 6 GB of memory.

5.1.3 Libraries used for implementation

Python libraries:

- 'numpy' and 'pandas' - for data manipulation
- 'sklearn' - for machine learning and data mining, specifically 'sklearn.cluster' for K-means clustering and 'sklearn.metrics' for silhouette score
- 'scipy' - for scientific computing, specifically 'scipy.stats' for T-Student unpaired test
- 'matplotlib.pyplot' - for visualization

Apache Spark (PySpark) libraries:

- 'Spark SQL' – domain-specific language to manipulate DataFrames
- 'ML' and 'MLlib' - distributed framework on top of Spark Core for machine learning, specifically 'ml.stat' statistics for Kolmogorov-Smirnov test and 'ml.clustering' for K-means clustering

5.2 Datasets description

The data provided from the RSE is composed of the station measurements and the details of the contract for each station for years 2015 with 365 and 2016 with 366 days.

Data measurements are generated hourly from stations located in every neighborhood around the city of Milan.

Given data was provided as comma-separated values (CSV) files.

Weather data collected from 'Meteoblue' API contained following attributes: city name, temperature, relative humidity, global solar radiation, wind speed, wind direction, rainfall and atmospheric pressure. It consists of hourly measurements from a single point in the city of Milan for years 2015 and 2016.

5.2.1 Data pre-processing

Data pre-processing was required in order to integrate, clean and transform the provided data from 'RSE' with historical data from the weather API for the city of Milan.

The largest obstacle was missing data. For some stations, more than 70% of the consumption measurements were missing for entire days. In some cases, only couple of measurements were missing during the day. The general approach was to eliminate the days which were missing all the consumption measurements. In cases of missing one or two values, filling methods were used, usually by imputing the mean of adjacent measurements. Rows of consumption data with all null or missing values were eliminated as well. In case of a missing value for the end date of station contract, the last day of measurement was used instead. Regarding the missing data of the type and description of the stations, those fields were left unchanged.

Another problem addressed in this analysis was corrupted data values. Outliers were detected by normalizing the measurement and applying a Z-score. Anomalies were either disregarded or corrected.

The original dataset contained additional information about location and events related to 'Expo' world fair 2015, as well as the different weather attributes. Redundant columns that were considered unnecessary for the scope of this work were removed.

New columns that would ease the analysis are added, some of which represent an aggregation of other columns (e.g. weekday and holiday).

Stations with most stable measurements were chosen for the analysis. Ones that had nominal consumption lower than 500 kW were not considered, since applying analysis on them would not make sense.

5.2.2 Processed datasets

After pre-processing the initial datasets, following table 5.1 of station measurement and table 5.2 of station contract were obtained. For each column, data type and descriptions are provided:

Column name	Data type	Description
ID_Station	string	Id of the station
Date_Time	timestamp	Timestamp of the measurement
Power	double	Electricity consumed (in kWh)
Weekday	int	Day of the week
Holiday	bool	Is it a holiday
Temperature	double	Temperature on the ground (in Celsius)
Relative_Humidity	double	Relative air humidity (in percentage)
Solar_Radiation	double	Amount of radiation coming from the sun (in W/m^2)
Wind_Speed	double	Wind speed (in km/h)
Rainfall	double	Amount of rain (in mm)

Table 5.1 Station measurement

Column name	Data type	Description
ID_Station	string	Id of the station
Power	double	Nominal (maximum) power designated in given period (in kW)
Main_Type	int	Station type Possible values: 1: Household if nominal fraction due to residential contracts > 0.9; 2: Industry if nominal fraction due to residential contracts < 0.05
Description	int	Whether station consumes or generates electricity. Possible values: 1: Consumption 2: Generation
Sub_Type	int	Sub-type of the station Possible values: 1: User that can generate 2: Public lighting 3: Heat pumps 4: Electric vehicle charging 5: Photovoltaic 6: Thermal 7: Hydric 8: Biomass 9: Eolic 10: Passive user
Contract_Start_Date	timestamp	Date from which the contract started
Contract_Expiry_Date	timestamp	Date of contract termination

Table 5.2 Station contract

Figure 5.1 shows example data from station 'A001' representing public lights, while figure 5.2 represents consumption data from household and industrial station ‘E001’.

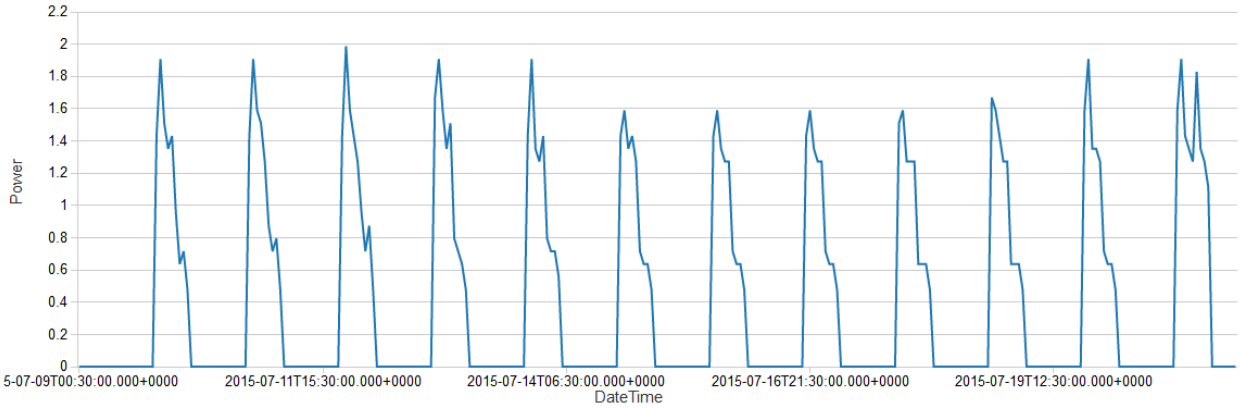

Figure 5.1 Station 'A001' sample consumption

Trivial characteristics of station 'A001': Low overall consumption, coming mainly from public lights. No obvious difference in consumption between workday and non-workday, with high consumption during the night and very little consumption during the daylight hours.

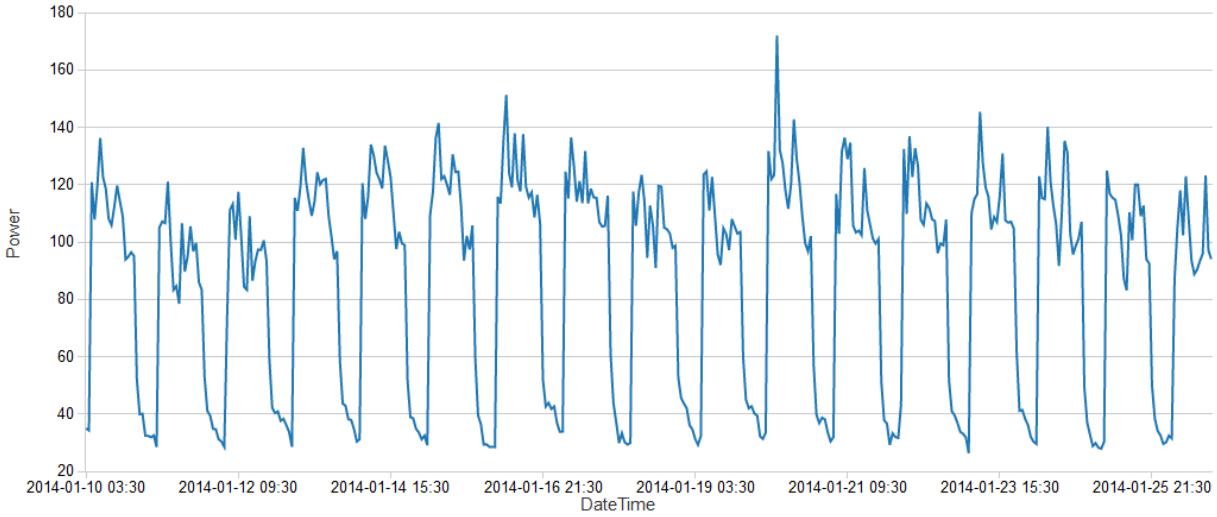

Figure 5.2 Station 'E001' sample consumption

Trivial characteristics of station ‘E001’: high consumption coming from households and industry. Uniform consumption during weekdays and weekend, with higher consumption during daylight hours.

5.3 Report of analysis and discussion

This section describes the report generated using the methods and tools previously described on the pre-processed datasets above, answering each of the questions posed in section 1.2.

The analyses are performed in batch mode on the whole set of historical data for two years of station measurements with their contract descriptions.

The implementation for each problem-solving approach is explained, followed by the results of the given method. In some cases, results are also displayed visually, using graphs and tables, for a better demonstration. The last stage is the discussion of results, in order to extract their meaning. Both quantitative and qualitative interpretations are used to describe the results in a complete way.

5.3.1 Seasonality detection

Null assumption: there are periods of a year that could be group by a similar amount of daily electricity consumption.

5.3.1.1 Implementation

Implemented using Spark and Python clustering libraries and silhouette score to evaluate the result, for public light stations. Demonstrated on the example of station ‘A001’.

5.3.1.2 Results

Applying K-means algorithm for the station ‘A001’ gave the following results:

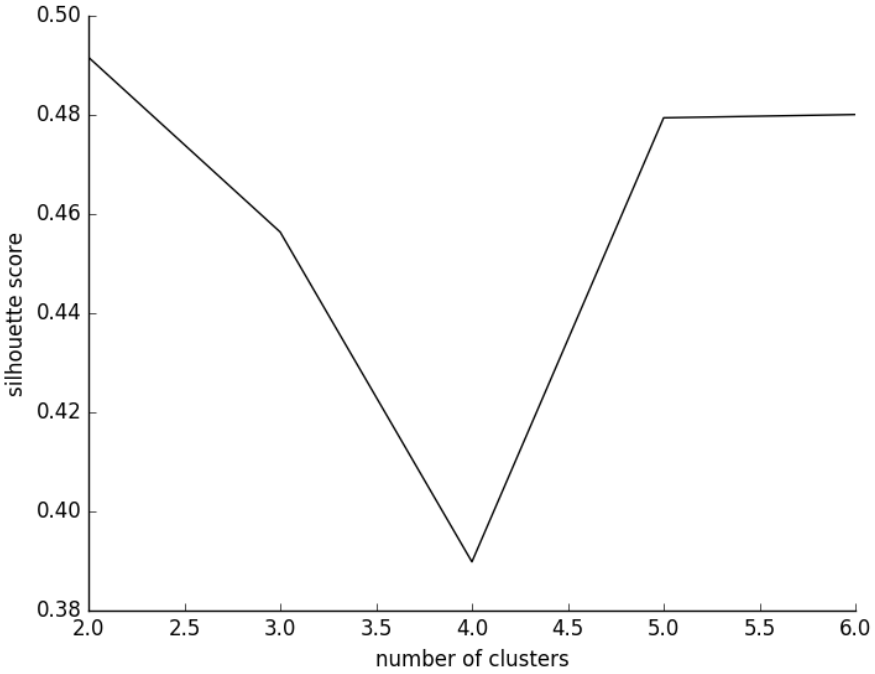

Figure 5.3 Silhouette scores for clustering of station 'A001'

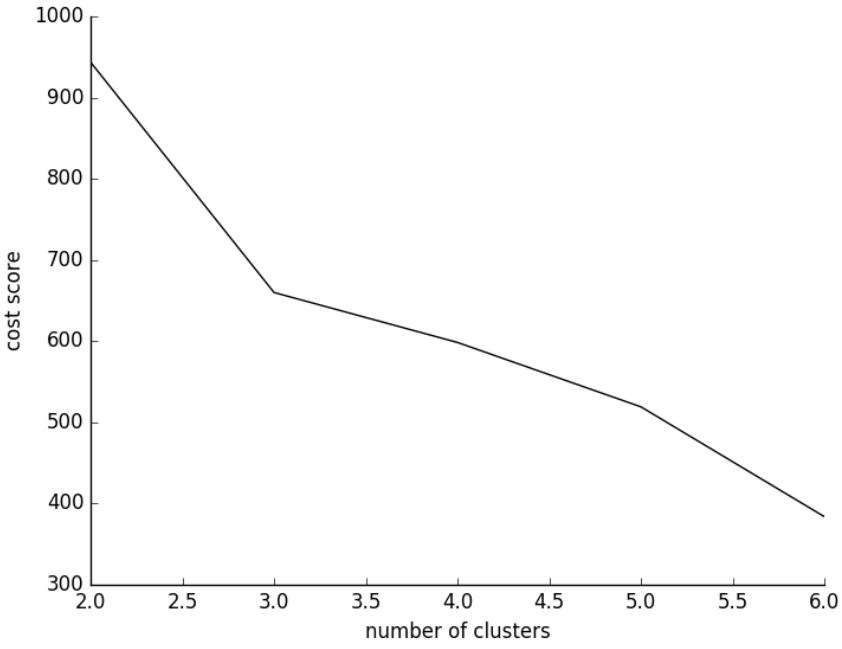

Figure 5.4 SSE scores for clustering of station 'A001'

Silhouette score from figure 5.3 and SEE score from figure 5.4 suggest that the best solution is to have two clusters. They could be labeled as 'high season' and 'low season'.

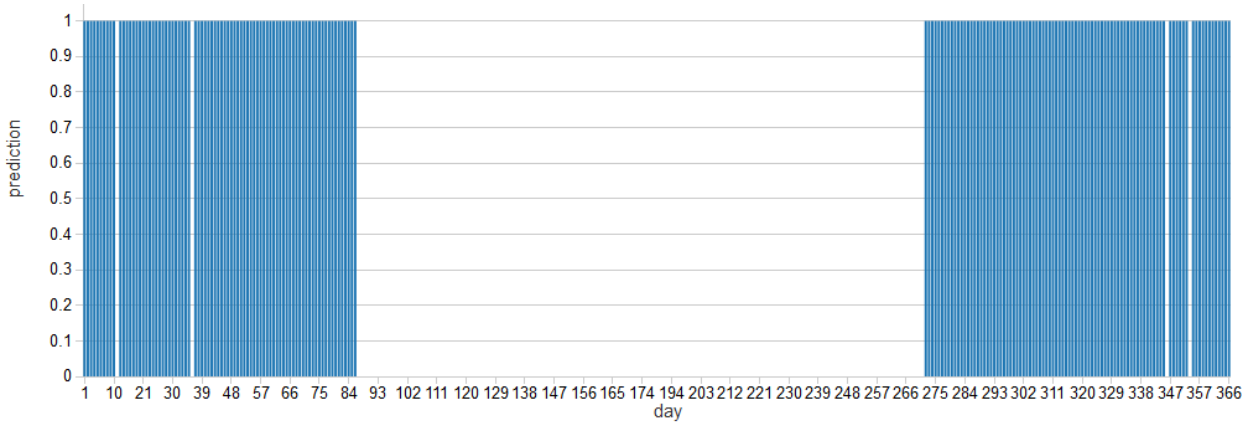

Figure 5.5 Two clusters for station 'A001'

Figure 5.5 shows the result of clustering where value 1 represents 'high season' - from 26th March until 28th September, while value 0 represents 'low season' - from 28th September to 26th March.

Average daily consumption of centroids: 'low season' 13.974, 'high season' 21.434

5.3.1.3 Discussion of results

There are periods of similar consumption that can be clustered together. The silhouette score confirms that the whole year could be clearly divided into roughly two equal parts: 'low season' and 'high season'.

The clear distinction between two seasons lead to the conclusion that public light is not automatic, but completely or partially controlled (set to start at a specific time).

5.3.2 Distribution of hourly consumption

Null assumption: hourly consumption follows a Gaussian distribution.

5.3.2.1 Implementation

Violin plots are used to visualize the distribution of hourly consumption in specific periods. Violin plots are similar to histograms and box plots in that they show an abstract representation of the probability distribution of the sample. Rather than showing counts of data points that fall into bins or order statistics, violin plots use kernel density estimation (KDE) to compute an empirical distribution of the sample [38].

Kolmogorov-Smirnov test was implemented to check if consumption follows a Gaussian distribution.

Experiments are performed for public light stations. They are demonstrated on the example of the station ‘A001’.

5.3.2.2 Results

Figures 5.6 demonstrates distributions of hourly consumption for each hour of the day by using violin plots over complete data for the station ‘A001’.

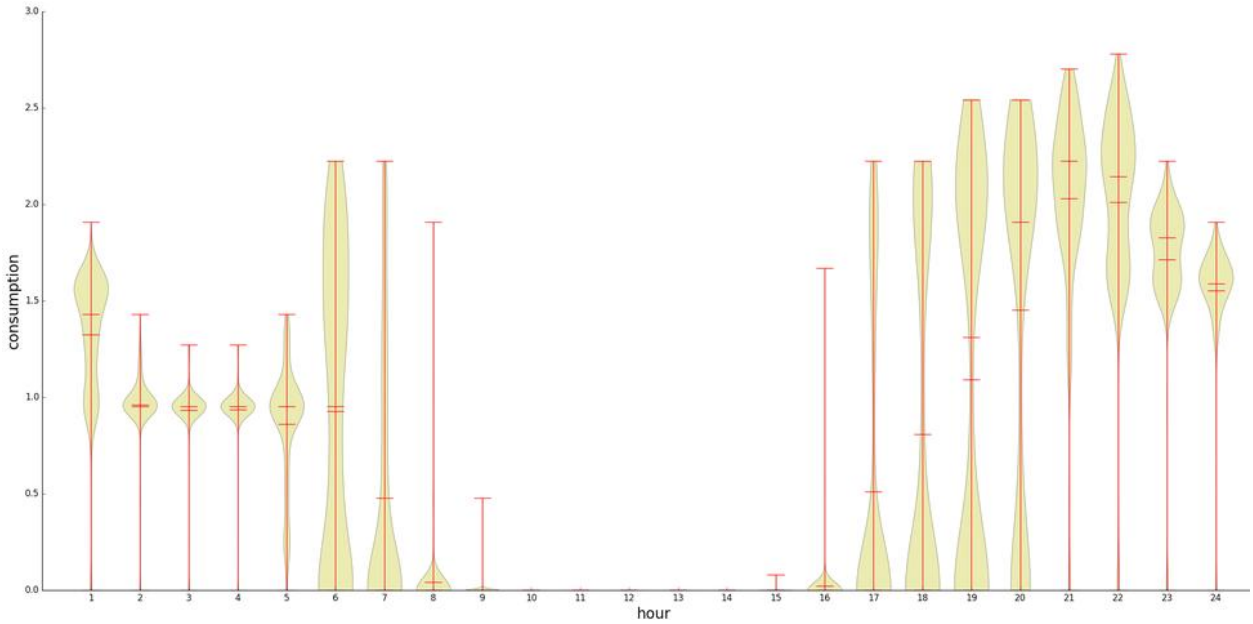

Figure 5.6 Violin plots for daily consumption per hour for station ‘A001’

Figures 5.7 and 5.8 demonstrate distributions of hourly consumption for each hour of the day by using violin plots over different parts of the year for the same station ‘A001’.

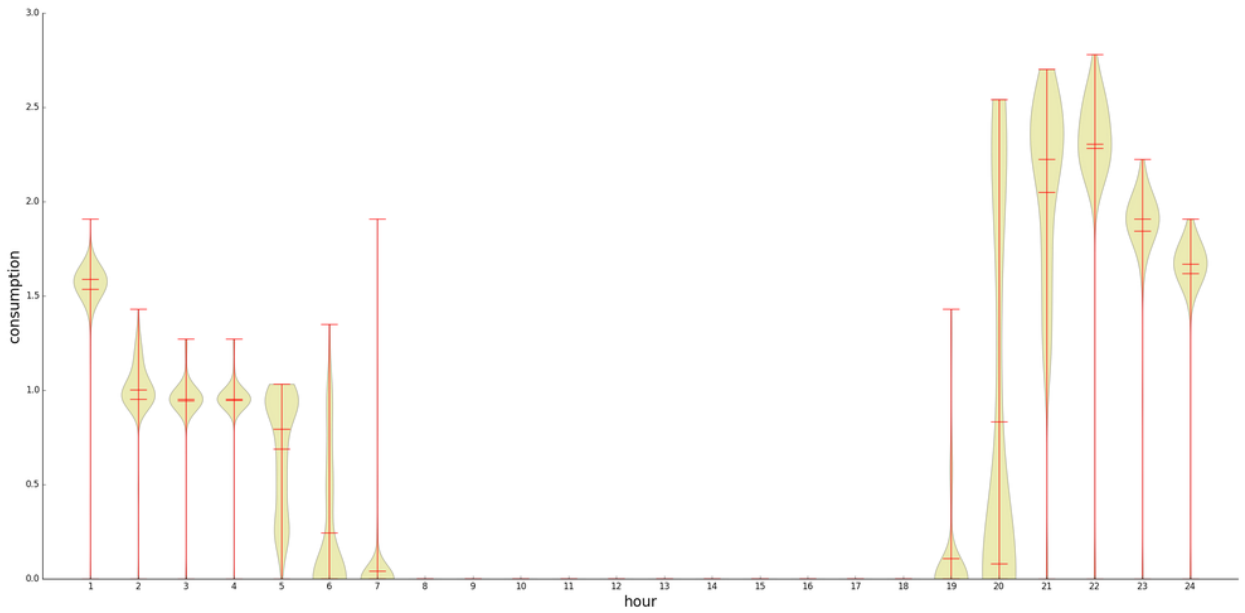

Figure 5.7 Violin plots for daily consumption in ‘high season’ per hour for station ‘A001’

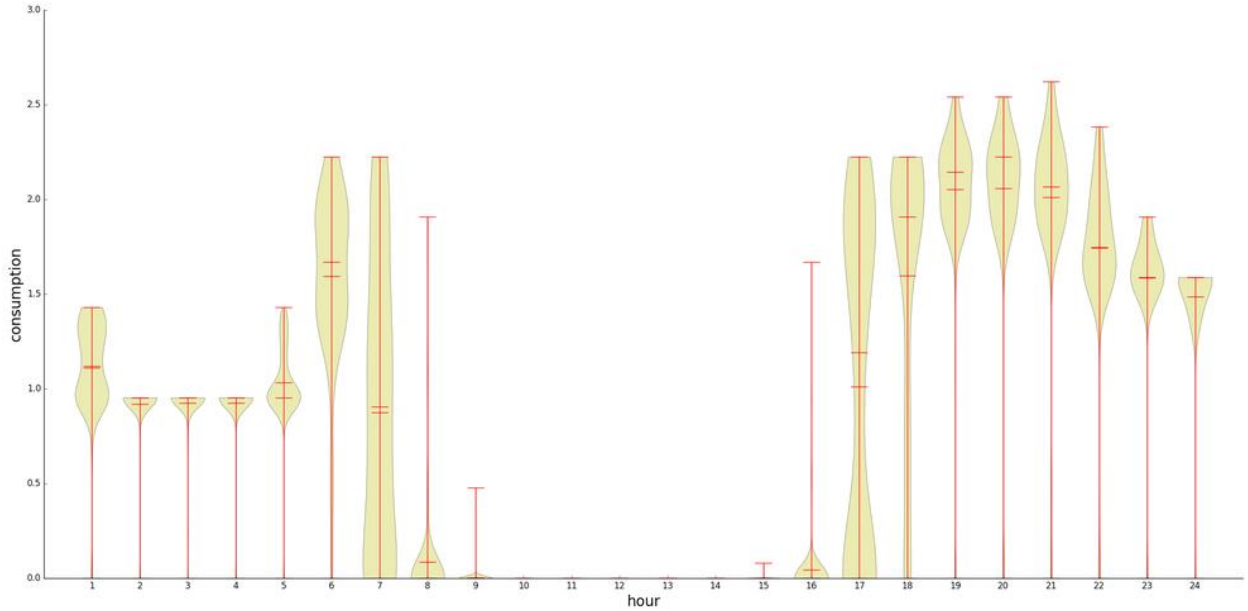

Figure 5.8 Violin plots for daily consumption in ‘low season’ per hour for station ‘A001’

Kolmogorov-Smirnov test was implemented for each hour of the initial distribution. The tests for each hour showed, with high probability that consumption did not come from Gaussian distribution.

5.3.2.3 Discussion of results

During daytime, hourly consumption amounts near zero. In transitional periods (dawns and dusks), it rises, with high variance. During nighttime hours it spikes, with very small variance. Visually, hourly consumptions do not seem to come from Gaussian distribution. In most of the cases, they are left-skewed.

Applying normality test confirmed that hourly consumptions during the day do not correspond to a Gaussian distribution.

Normalization techniques could be used to transform data into Normal distribution, but they were not applied because the data would not remain comparable.

The difference in distribution between seasons is clearly visible for the hours in the time periods 5.00-9.00 and 16.00-20.00. Distribution of consumption during other hours is almost similar.

5.3.3 Effect of non-working days

Null assumption: there is a distinction between consumption in weekdays and weekends/holidays.

5.3.3.1 Implementation

Perform 'Two samples unpaired data comparison' using a T-Student unpaired test on measurements from weekdays and weekends/holidays. The T-Student unpaired test assumes that data is normally distributed. Before starting the tests, the Kolmogorov-Smirnov test was used to verify the consumption distributions of stations.

The initial assumption was that there is no significant difference in consumption between weekdays and weekends/holidays.

Another approach used was clustering (K-means algorithm) of the data to check if the obtained clusters correspond to real weekdays and weekends/holidays. The confusion matrix is used to

display the results of predicted versus actual values. Accuracy, precision, recall, and F-score are used to evaluate the results.

Tests were performed on household and industrial stations and results are demonstrated on station ‘E002’.

5.3.3.2 Results

Average consumption for weekdays: 80.323

Average consumption for weekends: 31.76

Average consumption for holidays: 47.493

The day with maximum average consumption is Wednesday: 82.858

The day with minimum average consumption is Sunday: 30.944

Student’s T-test

Results of T-test for weekdays and weekends: t-statistic=66.326; p-value=0.0

Meaning null hypothesis can be rejected.

Results of T-test for weekdays and holidays: t-statistic=17.06; p-value=0.0

Clustering

The optimal number of clusters is 2, obtained by maximum average silhouette score of 0.585.

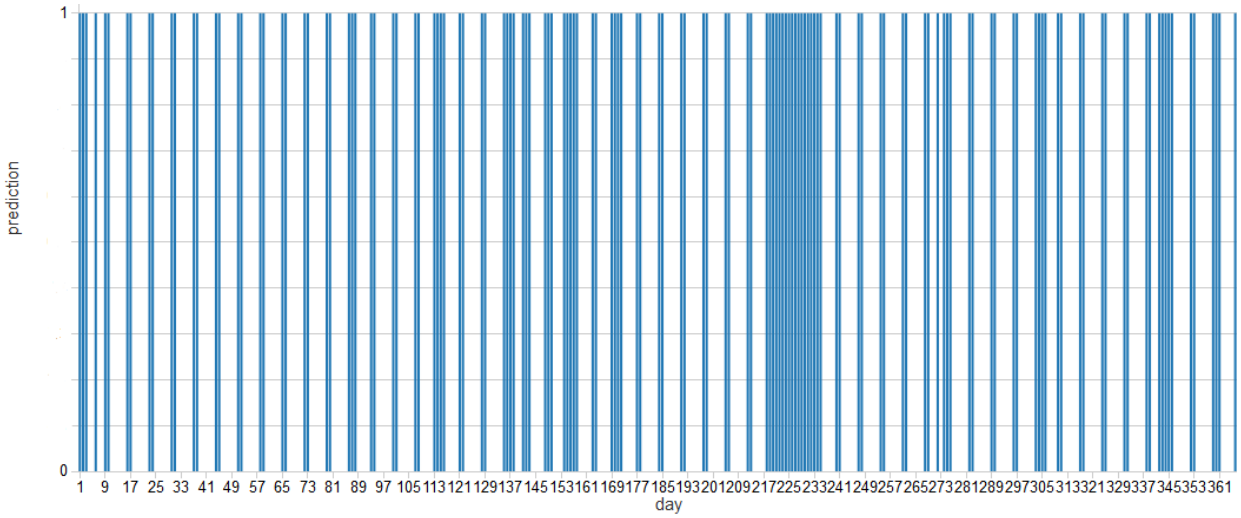

Figure 5.9 Two clusters for station 'E002'

Figure 5.9 demonstrates clusters of 'working' and 'non-working' days. The cluster of 'non-working' days includes a period in the month of August, which is considered holiday month in Italy. It also includes connecting of holidays with weekends which is visible by periods of three to five connected consecutive days.

	Predicted weekday	Predicted non-weekday	Sum
Actual weekday	tp=224	fn=25	249
Actual non-weekday	fp=3	tn=114	117
Sum	227	139	366

Table 5.3 Confusion matrix for clustering of station 'E002'

Table 5.3 shows confusion matrix for two clusters with calculated values for true positive, false positive, false negative and true negative.

After defining confusion matrix, following parameters are calculated:

- Precision=98,687%
- Recall=89,96%
- Accuracy=92,35%
- F1-score=94,12%

5.3.3.3 Discussion of results

Working days have significantly higher consumption. There is a notable difference in consumption between workdays and weekends. Wednesday has the maximum average consumption while Sunday has the minimum. There is a significant effect of holidays on consumption as well, with both clustering and T-test confirming this conclusion.

5.3.4 Weather influence

Null assumption: there is an influence in consumption that is caused by weather conditions; Temperature increases consumption in summer and decreases it in winter. Solar radiation decreases total consumption by increasing power generation.

5.3.4.1 Implementation

Correlation and covariance tests are done on household, industrial and public light consumption stations to test the influence of temperature. Using previously observed pattern of how temperature affects consumption, the next task was to expend the analysis by testing solar irradiance impact on power consumption and generation. The results are demonstrated on station ‘E002’.

5.3.4.2 Results

Temperature influence

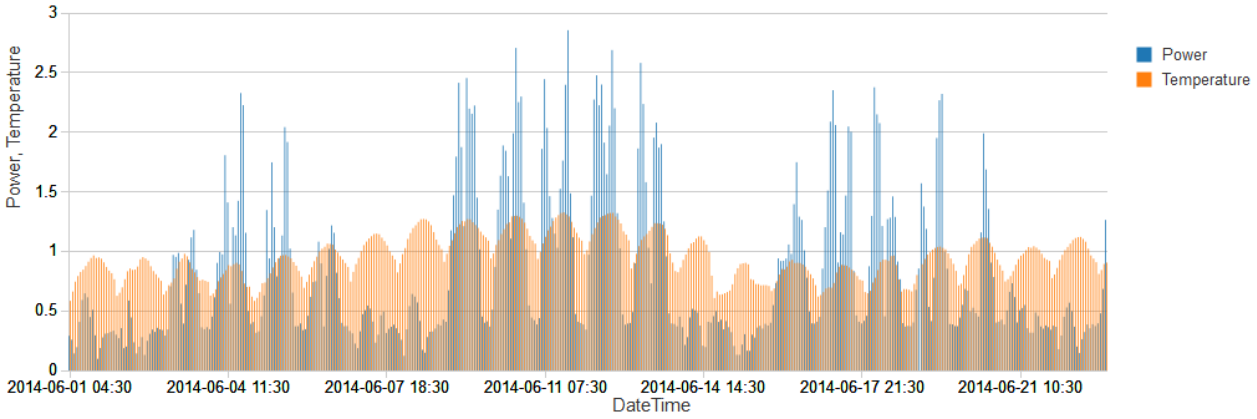

Figure 5.10 Temperature and power consumption in the summer for station 'E002"

In order to recognize possible effects of temperature on power consumption, temperature and consumption are demonstrated together in figure 5.10. Measurements values were standardized in order to get a better visual understanding.

The relation between temperature and consumption in the summer:

Covariance: 97.223

Correlation: 0.506

The relation between temperature and consumption in the winter:

Covariance: 603.914

Correlation: -0.096

Solar radiation influence

Correlation between solar radiation and temperature in different seasons:

Summer: 0.474

Winter: 0.567

The relation between solar radiation and consumption in the summer:

Covariance: 7869.237

Correlation: 0.408

The relation between solar radiation and consumption in the winter:

Covariance: 603.913

Correlation: 0.0414

5.3.4.3 Discussion of results

Tests clearly confirm that temperature has a high, direct correlation to power consumption in the summer. In the winter, there is a slight negative correlation.

There are no clear indications of the effect of solar radiation on power generation since the influence of solar radiation on the hypothesized decreased power consumption could not be detected. In fact, the opposite effect is recognized, especially in the summer. The reason for this is likely due to the high correlation between temperature and solar irradiance. As stated before,

temperature and power consumption are directly correlated. This effect could be masking the potential effect of solar radiation on consumption. The main problem was that the data for solar generation is provided only as a nominal value.

5.3.5 Relation between stations

Null assumption: there is no significant difference in consumption between two stations.

5.3.5.1 Implementation

Stations with the most similar characteristics were found by comparing nominal consumption from station contracts from the same period of the year. MSE and T-test were applied to find stations. Correlation and covariance tests were applied as well to confirm the results.

5.3.5.2 Results

The ‘closest’ stations found were ‘E003’ and ‘E001’. They had nominal consumption difference of 15 percent with both of the stations having a power range between 526.8 and 561.74 in the period between '2016-11-24' and '2017-07-31’. Figure 5.11 visualizes consumption of these two stations in the given period where orange line presents power consumption for station ‘E003’ and blue line power consumption for station ‘E001’.

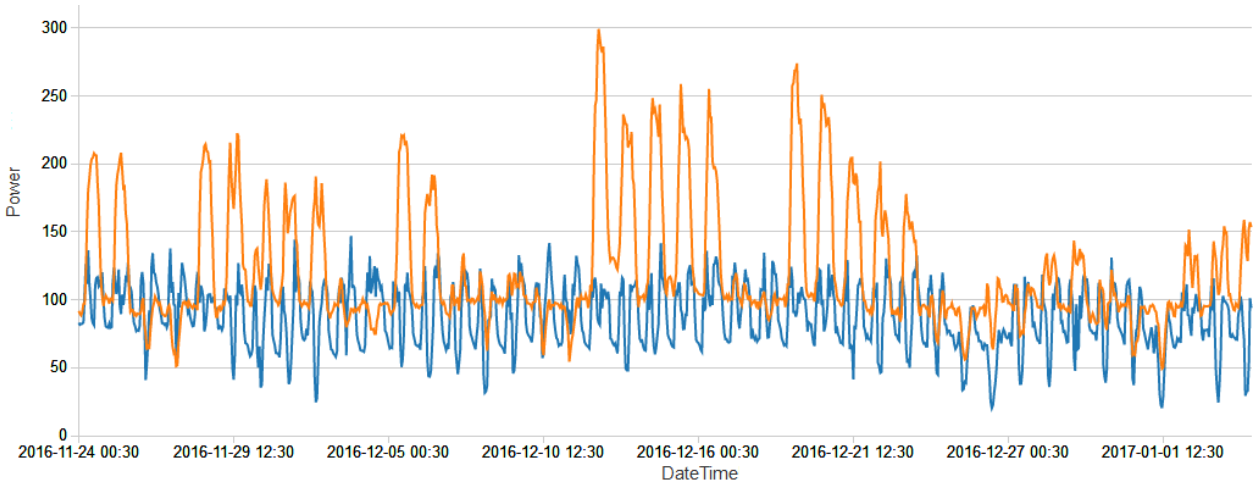

Figure 5.11 Hourly consumption for stations 'E003' and 'E001'

Following results are obtained:

Covariance: 81.174

Correlation: 0.056

MSE: 5155.401

RMSE: 71.8011196753571

T-test for two stations: t-statistic=-32.157; p-value=0

Meaning null hypothesis can be rejected.

5.3.5.3 Discussion of results

According to the results, the analysis could not find significant similarity in measured consumption for the station with similar nominal consumption.

Research limitations include insufficient data from the data stations, both in terms of the time periods selected, and the number of stations themselves. Therefore, there were not enough stations with similar nominal consumption for patterns to be found.

One recommendation that arose was to collect more data to rerun the experiment and to enrich metadata from station contracts.

Chapter 6

Conclusions

6.1 Short summary of the work

The main focus of this work is to model and analyze power consumption in a large urban space. Data was collected, integrated, cleaned and chosen before being analyzed. Different statistical methods were used (clustering, correlation, MSE, T-test, etc.) to perform analysis over two years of power consumption measurements and contractual data for each station integrated with weather data for the city of Milan. The goal of the analysis was to answer questions regarding hourly distribution, the difference in daily consumption between working and non-working days, seasonal trends, and the relationship between stations with similar characteristics.

For each of the stated problems, a solution was provided, both theoretically and practically, and through multiple approaches. In order to get a better understanding of the results, they are presented both visually and numerically and interpreted in quantitative and qualitative ways.

Having a clear understanding of the underlying model provides the possibility for optimization and making reasonable predictions regarding power supply and demand.

The ultimate goal is to use these results in order to save electricity, maximize power generation and decrease the emission of greenhouse gases.

6.2 Critical discussion of the results

6.2.1 Aggregation of the results

Overall analysis of the problems for the power consumption data gave the following results:

- There are periods of data with similar consumption (‘low’ and ‘high’ seasons).
- Distribution of the hourly consumption is not Gaussian; it is skewed. However, overall consumption distribution is Gaussian.
- The difference in consumption of electricity between working days and non-working days is significant.
- The weather conditions did influence power consumption and power generation, although in a limited scope.
- Similarities in measured consumption could not be found between stations of similar characteristics.

The results obtained in each analysis can be used to confirm the initial, intuitive hypothesis and formalize the model and expected behavior of power consumption in the city of Milan.

6.2.2 Threats to validity

- Although results provide meaningful explanations, it should be taken into consideration that the analysis was not been done for the whole dataset, but only on the most stable stations, a subset of the data. Methods were implemented on multiple stations, but not all of them produced meaningful results on the same level.
- Another limitation presented was an insufficient amount of data. Although two years of historical data was collected, many stations had incorrect or inaccurate measurements, or most had null or zero values. This made a lot of data unreliable or unusable. Measurements were aggregated and published from central stations located in every neighborhood, but not from the houses directly. There is no clear distinction between electricity coming from industry or from households. Measurements for power generation were not available, only nominal (maximum possible) power generation was provided for the given period.
- Correlation does not imply causation – correlation analysis results should be taken with caution. Even if there is a correlation between two things, it does not mean they cause each other.

6.2.3 Ways to improve results

Much can be done to improve the results and make them more applicable for other urban areas. The addition of many or all of these suggestions would result in more conclusive analyses

- Using domain expert knowledge to clean the data and performing the same analysis. Expert knowledge can be used to interpret data, and verify the results as well.
- Dividing industrial consumption into sectors with more detailed metadata about each sector. Enriching contract data with social-economic and demographic factors.
- Weather data could be more detailed and specific. Location of weather measurements could correspond to a location of a consumption device, in order to take into consideration the effects of microclimate.
- Instead of providing aggregated consumption for the neighborhood, devices that measure consumption directly should be installed in every consumer's building. Direct measurements of power generation would lead to more accurate data available for analysis.
- Increased frequency of data uploaded and accuracy of the measurement devices would make this power consumption model more reliable.

6.3 Future work

The approaches proposed in the thesis work are meant as a starting point for a more extensive analysis of consumption. This research can be expanded in different aspects.

- Integration with social networks, such as Twitter or Foursquare, for checking the activity in the region. Detecting special events could describe power consumption model more comprehensively.
- Real-time data predictive analysis can be implemented to create an early warning system, notifying the power provider in advance and giving them time to make informed decisions and take appropriate actions.
- Consumption of the stations could be visualized to users. Gamification techniques could be used to encourage participation in energy-saving competitions.

This kind of analysis could be applied with little modification for other urban areas, and furthermore, whole regions. With proper adaptation, it could be conducted for water or gas consumption as well.

Bibliography

[1] U.S. Energy Information Administration–EIA https://www.eia.gov

[2] “Developing a communication infrastructure for the Smart Grid” V.K. Sood, D. Fischer, J.M. Eklund, T. Brown; Published in: IEEE Electrical Power & Energy Conference (EPEC), 2009

[3] “Enabling Communication Technologies for Smart Cities” Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Yasir Mehmood, Abdullah Gani, Salimah Mokhtar, Sghaier Guizani; Published in: IEEE Communications Magazine (Volume: 55, Issue: 1, January 2017)

[4] “Internet of Things for Smart Cities” Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, Michele Zorzi; Published in: IEEE Internet of Things Journal (Volume: 1, Issue: 1, Feb. 2014)

[5] “Smart Grid Technologies: Communication Technologies and Standards” Vehbi C. Gungor, Dilan Sahin, Taskin Kocak, Salih Ergut, Concettina Buccella, Carlo Cecati, Gerhard P. Hancke; Published in: IEEE Transactions on Industrial Informatics (Volume: 7, Issue: 4, Nov. 2011)

[6] “The path of the smart grid” Hassan Farhangi; Published in: IEEE Power and Energy Magazine (Volume: 8, Issue: 1, January-February 2010)

[7] “A data integration approach for smart cities: The case of natal” Arthur Souza, Jorge Pereira, Juliana Oliveira, Claudio Trindade, Everton Cavalcante, Nelio Cacho, Thais Batista; Published in: International Smart Cities Conference (ISC2), 2017

[8] “CityPulse: Large Scale Data Analytics Framework for Smart Cities” Dan Puiu, Payam Barnaghi, Ralf Tönjes, Daniel Kümper, Muhammad Intizar Ali, Alessandra Mileo; Published in: IEEE Access (Volume: 4), 2016

[9] “Demonstrating smart buildings and smart grid features in a smart energy city” B. Morvaj, L. Lugaric, S. Krajcar; Published in: Energetics (IYCE), Proceedings of the 3rd International Youth Conference on Energetics, 2011

[10] “Electricity Smart Meters Interfacing the Households” Francesco Benzi, Norma Anglani, Ezio Bassi, Lucia Frosini; Published in: IEEE Transactions on Industrial Electronics, 2011

[11] “Energy efficiency in smart cities” Wil L. Kling, Johanna Myrzik; Published in: IEEE Power and Energy Society General Meeting (PES), 2013

[12] “IoT-enabled smart lighting systems for smart cities” Amit Kumar Sikder, Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, Kemal Akkaya, Mauro Conti; Published in: IEEE 8th Annual Computing and Communication Workshop and Conference, 2018

[13] “Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges” Mehdi Mohammadi, Ala Al-Fuqaha; Published in: IEEE Communications Magazine (Volume: 56, Issue: 2, Feb. 2018)

[14] “Challenges in energy systems for the smart cities of the future” M. Brenna, M.C. Falvo, F. Foiadelli, L. Martirano, F. Massaro, D. Poli, A. Vaccaro; Published in: Energy Conference and Exhibition (ENERGYCON), IEEE International Energy Conference and Exhibition, 2012

[15] “Towards Real-Time Analysis of Smart City Data: A Case Study on City Facility Utilizations” Takahiro Komamizu, Toshiyuki Amagasa, Salman Ahmed Shaikh, Hiroaki Shiokawa, Hiroyuki Kitagawa; Published in: High Performance Computing and Communications, 2016

[16] “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks” Geoffrey K.F.Tso, Kelvin K.W.Yau; Published in: Energy (Volume 32, Issue 9, September 2007, Pages 1761-1768)

[17] “A Data Mining Framework for Electricity Consumption Analysis From Meter Data” Daswin De Silva, Xinghuo Yu, Damminda Alahakoon, Grahame Holmes; Published in: IEEE Transactions on Industrial Informatics (Volume: 7, Issue: 3, Aug. 2011)

[18] A2A S.p.A., https://casa.a2aenergia.eu/

[19] “One day forth forecasting of hourly electrical load using genetically tuned Support Vector Regression for smart grid framework” Sreenu Sreekumar, Jatin Verma, Sujil A, Rajesh Kumar; Published in: Recent Advances in Engineering & Computational Sciences (RAECS), 2015

[20] “Modeling hourly profile of space heating energy consumption for residential buildings” G. Mutani, F. Giaccardi, M. Martino, M. Pastorelli; Published in: Telecommunications Energy Conference (INTELEC), 2017

[21] “Classification of Energy Consumption in Buildings with Outlier Detection” Xiaoli Li, Chris P. Bowers, Thorsten Schnier; Published in: IEEE Transactions on Industrial Electronics (Volume: 57, Issue: 11, Nov. 2010)

[22] “Energy & electricity consumption analysis of Malaysian power demand” Alias Khamis, Annas Alamshah, Azhar Ahmad, Azhan Ab Rahman, Mohd Hendra Hairi; Published in: 4th International Power Engineering and Optimization Conference (PEOCO), 2010

[23] “Time Series Analysis of Energy Consumption for Tianjin City” Yuansheng Wang, Chunli Chu; Published in: International Conference on Management and Service Science (MASS), 2011

[24] “Analysis of the Factors Influencing the Energy Consumption Pattern of Two Million Level Cities in China” Roberto Pagani, Wei Yu, Lei Huang; Published in: International Conference on Management and Service Science, 2009

[25] “Analyzing the Impact of Climate Change on Future Electricity Demand in Thailand” Suchao Parkpoom, Gareth P. Harrison; Published in: IEEE Transactions on Power Systems (Volume: 23, Issue: 3, Aug. 2008)

[26] “A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework” Luis Hernández, Carlos Baladrón, Javier M. Aguiar, Lorena Calavia, Belén Carro, Antonio Sánchez-Esguevillas, Diane J. Cook, David Chinarro, and Jorge Gómez; Published in: Sensors 2012 12(9), pages 11571-11591

[27] "Analyzing the impact of weather variables on monthly electricity demand” Ching-Lai Hor, S.J. Watson, S. Majithia; Published in: IEEE Transactions on Power Systems (Volume: 20, Issue: 4, Nov. 2005)

[28] “Weather conditions impact on electricity consumption” Saša Jovanović, Zorica Djordjević, Milorad Bojić, Slobodan Savić, Biljana Stepanović; Published in: Conference of mechanical engineering technologies and applications, 2012

[29] “The increasing impact of weather on electricity supply and demand” Iain Staffell, Stefan Pfenninger; Published in: Energy (Volume 145, 15 February 2018, pages 65-78)

[30] “Measuring Climatic Impacts on Energy Consumption: A Review of the Empirical Literature” Maximilian Auffhammer and Erin T. Mansur; Published in: Energy Economics (Volume 46, November 2014, pages 522-530)

[31] ”Incorporating residual temperature and specific humidity in predicting weather-dependent warm-season electricity consumption” Huade Guan, Simon Beecham, Hanqiu Xu and Greg Ingleton; Published in: 20 IOP Publishing Ltd Environmental Research Letters (Volume 12, Issue 2, February 2017)

[33] Ricerca sul Sistema Energetico - RSE S.p.A., http://www2.rse-web.it/home.page

[34] Databricks: https://docs.databricks.com/

[35] Apache Spark: https://spark.apache.org/documentation.html

[36] Python: https://docs.python.org/

[37] “Local Power: Tapping Distributed Energy in 21st-Century Cities”, Scientific American, June, 2010

[38] Matplotlib: https://matplotlib.org/examples/statistics/violinplot_demo.html

[39] Python Programming: http://www.computer-books.us/python.php

[40] ‘Theory of Point Estimation” (2nd ed.), Lehmann, E. L., Casella, George; Published in: New York: Springer. ISBN 0-387-98502-6. MR 1639875, 1998

[41] United Nations World Urbanization Prospects: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html

[42] “The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance”, Bruno A. Walther and Joslin L. Moore; Published in: Ecography 28, pages 815-829, 2005

[43] "Developing and Implementing the Data Mining Algorithms in RAVEN", Sen Ramazan Sonat, Maljovec Daniel Patrick, Alfonsi Andrea, Rabiti Cristian; Idaho National Lab. (INL), Idaho Falls, ID (United States), 2015

[44] “An efficient k-means clustering algorithm: analysis and implementation”, T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu; Published in: IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 24, Issue: 7, Jul 2002)

[45] “Measuring the Distance of Moving Objects from Big Trajectory Data”, Khaing Phyo Wai, Nwe Nwe; Published in: International Journal of Networked and Distributed Computing, (Vol. 5, Issue: 2, April 2017, pages 113–122)

[46] “T test as a parametric statistic”, Tae Kyun Kim; Published in: Korean Jornal of Anesthesiology 2015 Dec; 68(6): pages 540–546

[47] “Data streaming algorithms for the Kolmogorov-Smirnov test”; Ashwin Lall; Published in: IEEE International Conference on Big Data, 2015

List of Figures

List of Tables