Smart Cities Data Classification for Electricity Consumption & Traffic Prediction

Konstantinos Christantonis[1], *Christos Tjortjis*[1], *Anastassios Manos*[2], *Despina Elizabeth Filippidou*[2], *Evangelos Christelis*[2]

[1]*International Hellenic University*, [2]*DOTSOFT SA*

Abstract. Smart cities continuously develop into highly sophisticated bionetworks, providing both smart services and ground-breaking solutions. These bionetworks consider Smart Cities as a mechanism to produce data from multiple sharing engines, creating new challenges towards the implementation of effective Smart Cities and innovative services. The purpose of this paper is to relate Data Mining techniques and Smart City projects along with a systematic literature review that distinguishes the main topics and methods applied. The survey emphasizes on various components of Smart Cities, such as data harvesting and data mining activities over city data collected. It also addresses two research questions: a) can we forecast electricity consumption and traffic load based on past data, as well as meteorological conditions? b) which attributes are more suitable for prediction and / or decision support upon energy consumption issues? Results have shown that for both cases, various models can be built based on weather data collected**.**

Key words— Smart Cities, software systems, Data Mining, Prediction, Classification

1. Introduction

Citizens living in cities with dated infrastructure may face traffic and mobility problems, thus repeatedly spending time on useless or ineffective activities, such as being stuck in traffic jams, searching for parking space, waiting in lines or traveling long distances in order to receive a service or buy goods. As a result, financial and urban lifestyle problems emerge, such as air pollution, wasteful energy consumption etc.

To overcome these challenges, there is an incremental need from the city authorities to engage with concepts such as smart economy, smart transportation, smart environment and smart infrastructures. However, to successfully design and implement such a smart infrastructure, special tools should be deployed to collect and process big datasets in order to derive useful information for citizens and local authorities. Data Mining is an appropriate and effective tool in supporting a Smart City project.

This work initially elaborates on the smart city concept and the opportunity for further exploiting data mining theory and technology, while focusing on a case study regarding data mining techniques for predicting high traffic loads at random places around a city. The study follows am approach detailed below, which focuses in analyzing available weather data collected from smart sensor devices.

The remaining of the paper is structured as follows: Section 2 provides background information on data mining and smart cities. Section 3 presents the case study on electricity consumption prediction, whilst section 4 presents the case study on traffic prediction. Section 5 discusses, compares and contracts the two case studies and section 6 concludes the paper with suggestions for future work.

2. Background

In this section we address the context of this work by presenting core concepts related to smart cities and data mining.

1.1. Smart Cities

There is a contextual shift in Smart Cities since the term originally appeared in the literature, in the late 1990's, because of the impact of new technologies and the daily human-human and human-device interaction. A highly accepted definition of the term "***Smart City***" (ISO/IEC [1]) is as follows: *"an innovative city that uses ICT and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects"*. Furthermore, a widely accepted in the literature Smart City framework suggests that urban intelligence can be derived upon analysis of the six city "dimensions" as described below and shown in Fig.1 [2].

i) **Smart economy** refers to the degree that local authorities determining economy related policies, do not only deploy ICT for operations but also for communication [3].

ii) **Smart mobility**. This dimension measures the level that citizens utilize gas-powered vehicles or an energy free means to commute. This involves a variety of aspects, including ride-sharing, car-sharing, public transportation, walking, biking, and more. The need for smart mobility became apparent due to the constant increase of traffic congestion and its related side effects, like pollution, fatalities, and idle transportation time [55].

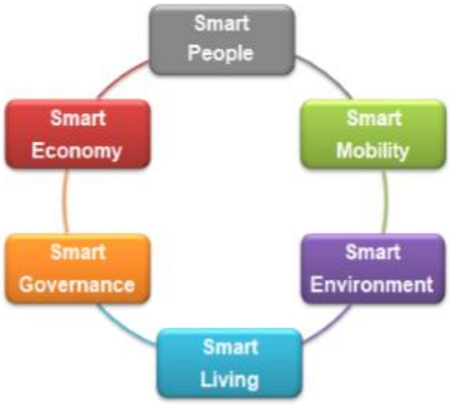


Fig.1. The six dimensions of smart cities

iii) **Smart environment**. The main concept of a Smart City refers to the proper use of technology and resources management in order to preserve a healthy environment for its habitants. When a smart city is under development, the protection of natural resources and all its infrastructures should be considered [5].

iv) **Smart people**. Digital Cities and Smart cities are not the same. They are different in a way that smart cities include people with high standards of training and skills. Smart people give high value to smart city projects that promote person-person or person-city and city-person communication. In a smart city, citizens feel active and participate in smart governance and management related activities [4].

v) **Smart living** refers to several aspects that contribute to the quality of living for a citizen, such as housing, health, safety, etc. [7].

vi) **Smart governance** refers to multiple issues, such as e-governance according to the level of active participation of a citizen in a city. It is important to smart cities because it depends on the proper setup of infrastructures that should be liable and transparent [6].

A city is considered smart if it deploys ICT solutions to deal with real life urban challenges [8]. The development of such a city originates from early urbanization times. Many estimations [9], [10] reveal that in the near future more than half of the earth population will live in urban areas. This phenomenon is more prominent in Europe, where only one out of four people will live outside urban zones. Although the 'root' is the same, the perception of 'smartness' varies from city to city depending on the existing local infrastructure and culture. It is widely accepted that smart cities are developed in order to improve citizens' quality of life and effectiveness of governance procedures.

In bibliography, one can find many different definitions and projects, which have been able to attract the interest of both scientists and ordinary citizens. One key characteristic of smart cities is that they all collect, analyse and visualise big data from various smart city applications. The origin of these data is diverse: indicative sources include data from sensors located in vehicles or in traffic lights; domestic appliances monitoring data, or even data in power line poles. The fundamental goal of a smart city project is to provide a resilient and clean city environment to citizens that promotes high standards for quality of life. The secondary goal is economic growth.

City related data are collected through sensors. These sensors connect all the objects (things) they are established and setup, exploiting the well-known Internet of Things (**IoT**) technologies. IoT is the network of all these physical devices which enables the interconnection and exchange of data among them.

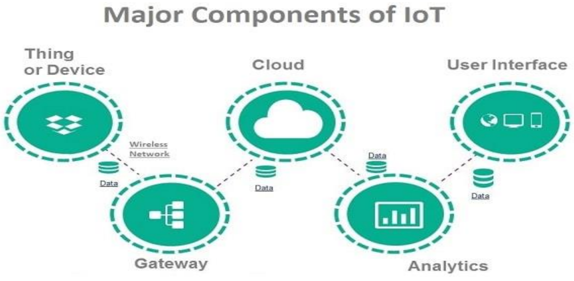


Figure 2: Major IoT components

Fig. 2 illustrates the continuous flow of data from the early stages of collection through single devices to returning meaningful conclusions to citizens [8]. The last part of the chain is the applications that citizens use.

1.1.1. Applications

Smart cities utilize multiple technologies to improve the performance of health, transportation, energy, education, and water services under the intention to deliver higher levels of comfort and quality of life to their citizens [11]. All these applications can be grouped into two categories according to the data origin: Citizen-oriented or government-oriented.

It is evident that all smart city applications are built around citizens and undoubtedly, they are the driving force behind the development of such projects. In addition, the individual data collected by governments can contribute to crowd sourcing projects, relevant with collecting citizens' opinions and beliefs upon the government or the society in general. Smart city applications are correlated. For example, safety and crime reduction are two related terms. Crime reduction can contribute to higher level of safety. Whilst crime reduction can be the main intention of a smart city project, safety is the benefit that citizens can enjoy as a result of crime reduction.

1.1.2. Concerns

It is evident that there is a global trend of people to live around cities, thus significantly increasing their population. Globally, high urban density could lead to challenges including traffic congestion, energy supply and consumption issues, high greenhouse gas emissions [12], unplanned urban development, lack of basic services, dramatic increase in waste disposal needs, as well as increases in crime and antisocial behaviour [13]. Concerns are numerous and the analysis of risks and threats that a smart city might face are numerous. Three of the most common concerns though, are summarized below:

- What happens if the system collapses?
- How sensitive are those city applications to malicious attacks?
- Have the government accommodated an essential legal 'framework' to protect personal data?

A study was conducted by a technology company in order to detect any security gaps within IoT Home Security Systems. Below are the most significant findings, as listed in [14]:

- Insufficient authorization: Most systems related to the cloud do not demand strong profile passwords (e.g. Six-alphanumeric length) resulting in a weak layer around sensitive data. Neither a fixed number of trials before profile locking existed.
- Insecure interfaces: Cloud based interfaces were vulnerable because of account enumeration, weak password policy and lack of account lockout.
- Privacy concerns: All the systems under consideration were collecting personal data such as name, address, phone number etc. from more than one user accounts. A major concern about IoT home security systems concerns accidentally publicizing personal data.
- Lack of encryption: The problem mainly appears in cloud-based connections, which are vulnerable to attacks due to the lack of encryption while transferring data, such as SSL/TLS.

In May 2018, the new General Data Protection Regulation (GDPR) became enforceable. GDPR legislation is a vast breakthrough in smart city project implementations as it regulates how city governments and administrations collect and use personal data from citizens [15]. GDPR drastically interferes with Personally Identifiable Information (PII) and regulates the issue of processing personal data, including collection, analysis, transfer, review and deletion. So far, governments and administrations were able to store personal data without time limitations.

To fully understand the way that smart cities operate, it is important to investigate the infrastructure behind such applications. Fig. 2 depicts a fundamental and basic view of a smart city project; the underlying system and applications inter-connections and data transfer are more complex, forming a complete discipline.

1.1.3. Smartest Cities

As mentioned earlier, many cities seek ways to become 'smarter'. Every year the press conducts evaluations for the smartest cities globally based on various criteria. There are cities which "traditionally" achieve near top positions and others that are "new entries" every year as far as the "smart city score level" is concerned. One of the most reliable annual assessments of smart cities was published recently. Berrone and Ricart [16] proposed a synthetic indicator named ***Cities in Motion Indexing (CIMI)***, which is a calculated based on other indicators relevant to assessing a city's quality of life. These partial indicators were governance, urban planning, technology, environment, international outreach, social cohesion, human capital, mobility and transportation, and economy. The Top-3 consists of New York, London and Paris, which traditionally belong to the top-5.

1.2. Data Mining

Data mining is a well-established field, which evolves to deal with the new types of data, including multimedia, time-series, text, spatiotemporal data and data streams [17]. It is defined as the exploration and analysis by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns [19], [56]. In simple words, data mining is the extraction of implicit, previously unknown, and potentially useful information from data [18]. The wider process is known as Knowledge Discovery in Databases (KDD), and consists of six well-defined steps [17]:

- Data cleaning
- Data integration
- Data mining
- Data transformation
- Data selection
- Pattern Evaluation

1.2.1. Challenges

KDD is a well-structured process with distinct steps in order to extract useful knowledge out of data manipulation and analysis. The process however is far from easy to follow, as there are several challenges that need to be addressed. Each project faces different technical and theoretical challenges, listed below:

- *Noisy Data.* To perform Data Mining the most essential prerequisite is Data Cleansing. This is the most time-consuming challenge. Although data are collected and stored in semi / fully automated ways, noise is inevitable. Noisy Data thus, are Data stored inaccurately due to human, sensor or transmission errors.
- *Data silos and distribution.* Data to be processed may be stored in different locations and merging these is a difficult task. A common challenge across multiple heterogeneous data sources is the absence of interoperability. Different data formats and data isolation causes these silos to raise numerous problems. Beside the lack of communication, the required infrastructure should avoid transferring data across systems, as this is a slow process.
- *Scalability and efficiency.* Training an algorithm on a typically sized dataset is turning into a standardised procedure while applying traditional algorithms into huge datasets of petabytes is turning

into a non-functional process [20]. Algorithms should be adjusted to the specific project needs and individual characteristics. For example, algorithms should reorganize computations in order to reuse intermediate results without storing them [22].

- *Complexity.* Data come in different formats; some are harder to manage. The more complex data types are images, raw text, audio and video.
- *Privacy and Security.* The information that may be extracted via Data Mining leads to several legal and ethical issues. The legal foundation is based on specific protocols, which establish penalization for data security and privacy Government Act [23].
- *Data visualisation*. Another technical challenge is result visualisation. Although it is of secondary importance, it affects end users. High dimensionality is usually the biggest 'curse' for both execution and visualisation.

1.2.2. Data Sources

Cities have become big engines that continuously consume and produce data. These data are produced from multiple sources in various forms. Urban Data are produced in two ways [21], [24]:

- Directly from daily activities and applications, such as social networks [25], [41].
- Collected from sensing devices, such as mobile sensors, traffic and environmental sensors [26].

In order to distinguish data sources, one needs to go back to the original sources of data production, which are typically two [1]:

- IoT: The production of data comes from sensors or actuators integrated into physical objects which are wired or wireless connected [27].
- Crowd-sensing: The production of data comes from Crowdsourcing when this is engaged with sensors. Crowdsourcing is the integration of a "crowd" in order get services or components [28].

Urban Data can also be categorized based on their ownership [29]:

- Closed Data: Contain private and personal information about their owner and are not publicly available, such as health and financial data.
- Shared Data: Data that the owner published, such as social media data and published data.
- Open Data: Data that are publicly available without the restriction of copyright.

The last category is important for Smart Cities in order to improve decision-making and data economy.

3. Electricity Consumption forecasting

The first part of our research analyzes how a system could predict total home electricity consumption. We approach the problem using two scenarios. In the first scenario, we develop a model to examine the ability of approximating total electricity consumption through the consumption of various household appliances. The second scenario attempts to investigate total consumption predictability using weather data and past activity.

3.1. Problem definition and Approach

A smart city goal is to provide citizens with information that could potentially lead to more effective decisions about the quality of life. Nowadays, besides constructing general-purpose smart projects, smart cities also focus on the development of smart homes. Smart homes transmit real time data generated by sensors. That enable us to create models. For that purpose, it is essential to secure constant tracking of the total electricity consumption per home. Unfortunately, smart metering is a newly deployed technology that still faces challenges and failures.

For that reason, this first scenario focuses on an alternative way to attain the total consumption if the central sensor fails. A very interesting survey [30] lists all the concerns and possible ways that a smart grid can fail. It is also desired to capture the most dominant sensors (appliances) around the home and examine how reliable a prediction of the total consumption could be, through a small subset of essential sensors. Although each of the homes under examination has sensors installed in different locations making results difficult to benchmark against each other, it is expected that a small subset of sensors can precisely capture adequately the original consumption. The sensors to be tested will be the subset that demonstrate the highest correlation with the total consumption.

Regarding the second scenario, the constructed model also includes usage patterns besides weather data. Since the goal most of the times is to predict the differentiation of consumption, the consuming behavior should also be considered. Authors in [31] state that "Electric demand is often considered as a function of weather variables and human social activities". In particular, most people intentionally or not, display certain patterns of consumption, depending on the conditions in their lives. For example, it is common for people to use the washing machine a certain number of times per week and rarely deviate from it. In a similar way we assume that the previous day's consumption is strongly correlated with next days. As a sub-part of this approach, we attempted to simulate grid behavior, by aggregating all the available data and examining predictability.

It is also key, to clarify that since the available data are limited and the level on information low, this study focuses on forecasting as a result of a binary classification. Reforming the problem from regression to classification was a crucial step. Ideally the outcome should be the exact consumption, but this research is conducted with a broader motivation to identify factors that affect a prediction positively or negatively, based on individual home characteristics. We investigate how effective would the model be for each of the available homes and which reasons lead to differentiation in accuracy. Therefore, the model was transformed into binary around different mean values. The two labels are 'High' and 'Low' with regards to the volume of consumption. These mean values reflect the mean consumption throughout the year, the current season or on a monthly basis. The binarization of consumption is performed around the standard mean values in order to avoid 'expensive' handcrafted data engineering. Thus, the whole process can be easily automated. Forecasting the consumption fluctuation has a totally different purpose than building a model that focuses on the detailed and precise consumption.

For further showcasing our approach to the problem, we focused on two different time intervals during the day: on and off-peak periods. That divide was made based on clues in the literature as well as the detailed examination of available consumption data during each day. Firstly, both periods last for six hours. On-peak periods were set from 15:00 to 21:00 and off-peak from 09:00 to 15:00. Obviously, off-peak period selection is complicated as there are numerous approaches that even differentiate between working days, holidays or weekends. In our case, this was a static selection throughout the year. Then, the 'supposed' consumption (target variable) for each day comes from the mean value on the period of interest. Fig. 3 is an indication of the process that was followed in the first part, split in four distinct stages.

3.2. Context

Smart* is a project that many research teams have relied on, as it offers a wide range of information. The authors in [32] used both weather and energy data in order to predict latitude and longitude of a smart meter that collects data. Another interesting approach presented the Smart Charge system, an intelligent charging system that aims to decrease the electricity bills by shifting consumption to periods when the price is lower [33]. Similarly, the SmartCap system was introduced to monitor and control electric loads while flattening electricity demands [34].

Regarding the general topic of electrical management systems there are numerous scientific efforts traced back over a decade. Machine learning and energy forecasting are appealing to many researchers who use different approaches, algorithms and domains. Some strategies seem to have different results when applied to different climate zones and continents. Different structures of Neural Networks (NNs) on forecasting were compared in [35] concluding on an instruction of form 12-16-16-1. NN's were also examined and compared with Support Vector Machines (SVMs) that resulted in the superiority of Least squares Vector Machine algorithm after an extensive review of the bibliography [36]. There are also some very interesting approaches based on customer profiles and how they could effectively be defined in order to assist providers with their strategies. In [37] customer profiles were pre-defined, based on the total consumption, thus a significant decrease on the daily expenses (electricity) was achieved. In contrast, [38-40] propose to first build customer profiles or assign some already known and then predict consumption based on these.

In general, load forecasting focuses on three different scales. As authors in [31] explain, those three have different characteristics and values for the providers; Short-Term (ST), Medium-Term (MT) and Long-Term (LT). ST focuses on reducing costs and secures constant operation of power systems and usually refers to the day ahead / week ahead forecasting. For MT the interest regards the general operation and refers to a monthly scale. Finally, LT forecasting mostly focuses on ensuring safer investments and regulations.

An important study highlights the responses in energy demand due to climate change in Massachusetts [42]. Some of the parameters refer to Heating Degree-Day, the hours of daylight and the price of electricity on a monthly scale. This study concludes that 'energy demand in Massachusetts is sensitive to temperature' while the average number of days exceeding 90°F will rise to double by 2030.

The main reason why accurate predictions cannot be guaranteed is the fact that 'energy demand has a high non-linear behavior' [43]. In addition [44] explains how a prediction's accuracy can also be negatively affected by the continuous pressure for better living standards in a disproportionate rate. The above statement is also supported by authors in [45] who analyzed the residential consumption in Brazil and concluded that the increase in electricity demand is faster than that in income.

Finally [46] explains that electric power consumption is growing rapidly and introduces a higher level of randomness, due to the increasing effect of environmental and human behavior. So, usually studies focus on ST load forecasting which is considered a more difficult task due to the noisy effect of environmental factors.

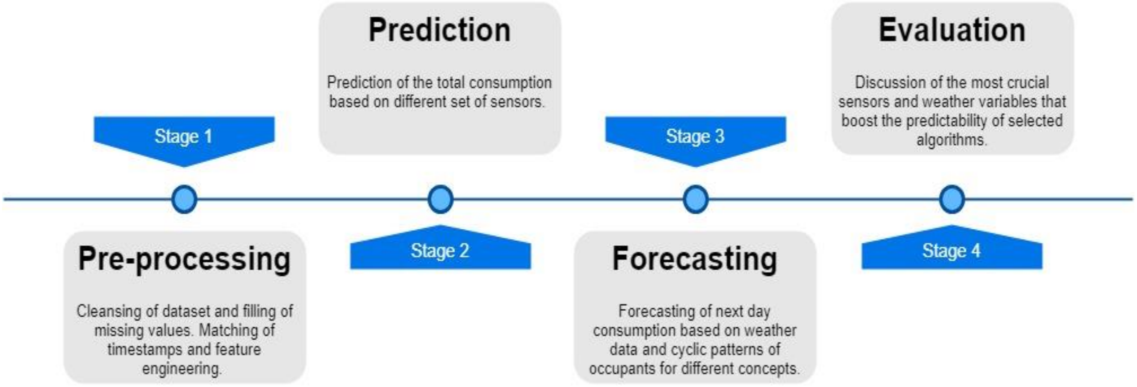

Prediction
Prediction of the total consumption based on different set of sensors.
Evaluation
Discussion of the most crucial sensors and weather variables that boost the predictability of selected algorithms.
Stage 1
Stage 3
Pre-processing
Cleansing of dataset and filling of missing values. Matching of timestamps and feature engineering.
Stage 2
Forecasting
Forecasting of next day consumption based on weather data and cyclic patterns of occupants for different concepts.
Stage 4

Figure 3: Fragmentation of work on four levels

1.3. Dataset Description and Processing

Two interrelated sets of data are used in this part from the *Smart** project. The project seeks to optimize energy consumption in homes with specific attention to modern 'smart homes'.

Initially data for seven homes labeled alphabetically (A-G) were recorded; however, their structure was not matching. After detailed processing, which mainly focused on matching timestamps and handling missing data, the usable dataset was reduced down to three homes (B, C and F) with data regarding three consecutive years and a recurring digital footprint. The Smart* project provides a description for two of them on [47]. For the rest of this work homes will be referred to by these names.

The strategy for missing values was not unified due to the multiple conditions that each data set contained. More specifically, for the “sensor selection” part of this work it was decided to exclude instances from the dataset when a sensor had failed. All these numerical values regard the recorded consumption of each installed sensor at a specific timestamp. In addition, all the redundant attributes such as Generated power, Grid load or Solar panels were removed. Each home has different meters installed; however, the number of sensors is comparable.

The dataset description claims that these homes are located in Massachusetts and therefore the weather metrics do not deviate significantly. *Home B* is a large residence across two stories with eight rooms and four full-time occupants. It is roughly 1700 square feet and it contains a central A/C as well as a gas-powered heating system. *Home C* is almost double the size of Home B, around 3500 square feet again across two stories. Unfortunately, the real number of occupants is unknown. It also generates power which not only covers some of the electricity demands, but is also possible to ‘reverse direction when the home’s generation exceeds its consumption’. Unfortunately, *Home F* does not come with a description as it is included on the dataset as an update. Information about Home F is expected to be published later in 2020. The consumption levels though are quite similar in magnitude to those of Home C.

Regarding the first dataset, it contains consumption in kilowatts data recorded every thirty minutes from several sensors around the home. It was desired to check if there are sensors influencing the prediction model even if they only represent a small part of the total home consumption.

The second dataset consisted of weather data and processing was similar. However, the recording time in this case was hourly. Table 1 lists the weather metrics available.

Table 1: Initial weather metrics

Temperature (+Apparent)	Weather icon	Visibility
Summary	Humidity	Pressure
Wind Speed	Cloud Cover	Wind Bearing
Dew Point	Precipitation Intensity	Precipitation Probability

Out of the initial data attributes, processing and selection resulted in excluding the following metrics: visibility, pressure, summary, cloud cover, precipitation intensity and probability, as these were considered either redundant or meaningless. Briefly, *precipitation probability* and *intensity* were not included since the categorical variable *weather icon* carries similar information. Obviously, clear weather leads to zero intensity. Moreover, *visibility* was not expected to add more value to the model than complexity. C*loud cover* had numerous missing vales, while *pressure* was not considered as a metric easily understandable by residents, so as to change their living habits. On the contrary, a dummy-like variable was introduced to flag a day as weekday or not. Similarly, another variable was added to indicate national holidays. The last attribute of the model that was manually created was the time of sunset in a categorical form of five intervals (16:00 – 20:00). Obviously, that variable does not affect the consumption during off-peak periods.

1.4. Experiments

After data processing to achieve a uniform format, it was essential to appreciate consumption patterns. Fig. 4-9 present the consumption for each home averaged by month and hour. Clearly, there is a constant increase in consumption each year, yet the spread during the day is similar.

Home C does not show significant deviations throughout the years and surprisingly tends to consume more electricity during the winter. Just as for Home B, the period 15:00 – 21:00 is indeed on-peak for Home C, besides an interesting increase early in the morning. Finally, Fig. 7 clearly shows an unusual consumption behavior for the summer of 2016.

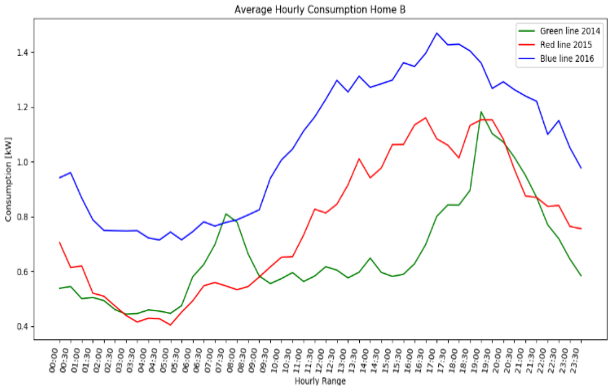


Figure 4: Home's B avg. hourly consumption

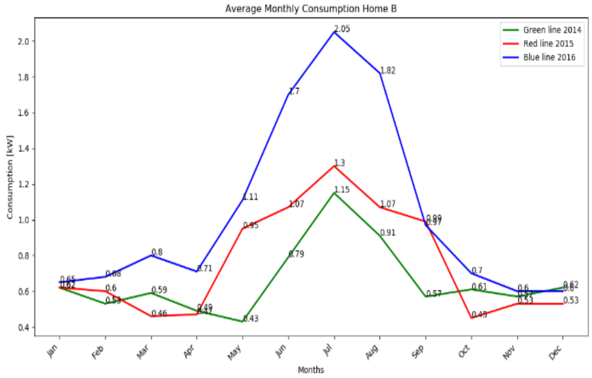


Figure 5: Home's B avg. monthly consumption

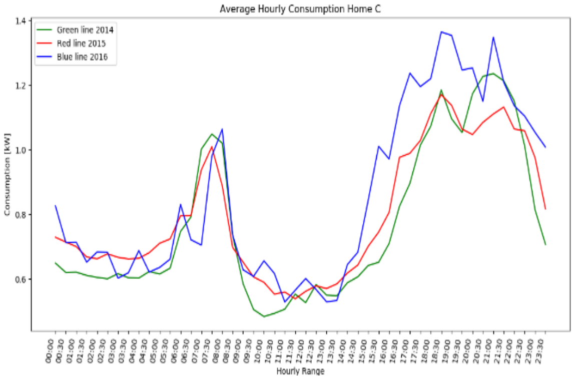


Figure 6: Home's C avg. hourly consumption

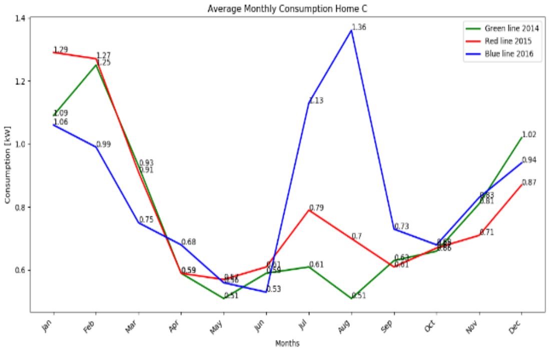


Figure 7: Home's C avg. monthly consumption

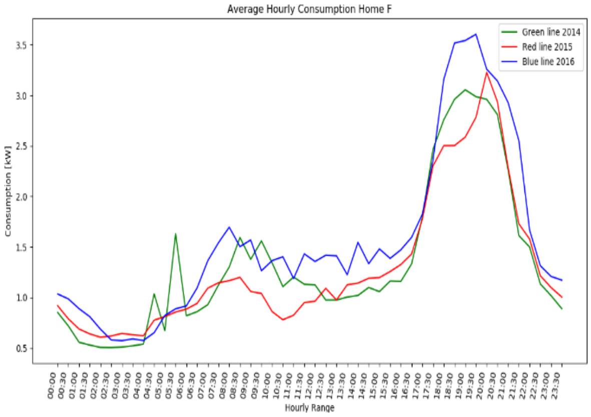


Figure 8: Home's F avg. hourly consumption

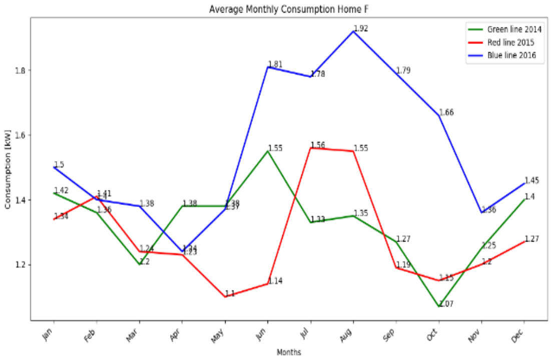


Figure 9: Home's F avg. monthly consumption

For Home F, even though there is a clear and almost identical consumption pattern throughout the day, there is not any clear pattern in consuming behavior over the months.

1.4.1. SCENARIO 1

For the regression task on this approach two different metrics and two machine learning algorithms were used. The metrics we chose were the Round Mean Squared Error (RMSE) and the Adjusted R-squared. Regarding algorithms, we used Random Forest and Gradient Boosting. We used 10-fold cross validation to avoid over-fitting. Table 2 shows the *best* results for each home. Although we show only the best results, no significant deviation in algorithmic performance was observed.

Table 2: best results for each home

	Total Results		
	Home B	**Home C**	**Home F**
RMSE	0.122	0.318	0.6023
Ad R2	0.979	0.8466	0.8584

Fig. 10-12 are heatmaps showing the correlation of each installed sensor with the total consumption for each home. The sensor names are somewhat confusing and unclear, but we kept them unique as a reference for those who plan to work with the same. Results are presented in Tables 3-11. Home B has a few more sensors than the others.

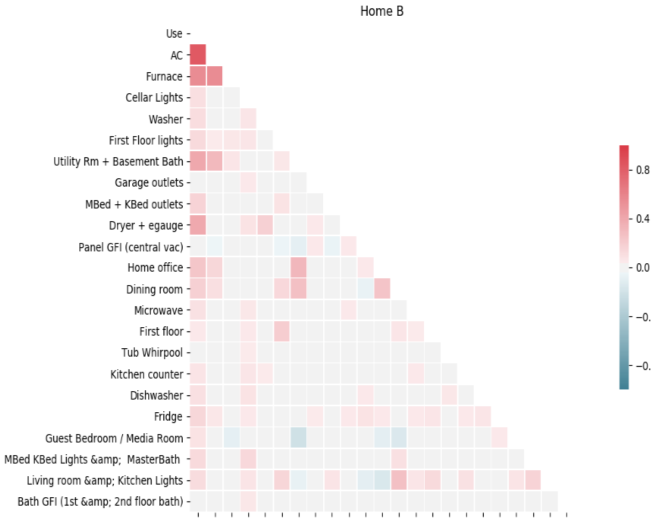


Figure 10: Correlation Heatmap – Home B

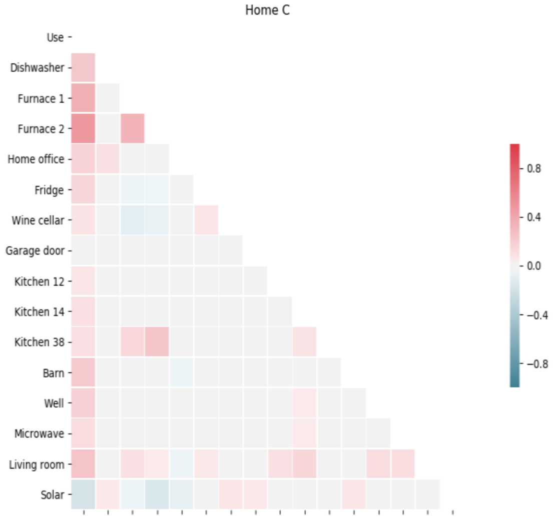


Figure 11: Correlation Heatmap – Home C

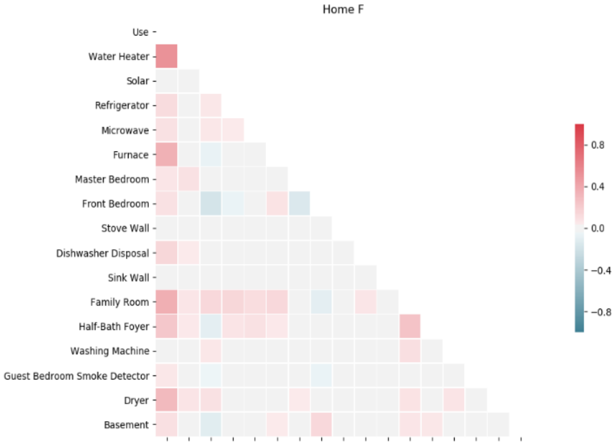


Figure 12: Correlation Heatmap – Home F

Tables 3-5 list the six most correlated sensors in absolute numbers. Tables 6-11 illustrate the predictability of algorithms when they are trained only with data regarding the top-3 or top-6 most correlated sensors.

Table 3: Top correlated sensors for Home B

	Correlation with total consumption
A/C	0.833791
Furnace	0.552492
Utility Room + Basement Bath	0.401344
Dryer + E-gauge	0.393643
Home Office	0.234851
Dining Room	0.187463

Table 4: Top correlated sensors for Home C

	Correlation with total consumption
Furnace 2	0.487334
Furnace 1	0.366014
Living room	0.245231
Dishwasher	0.228602
Barn	0.212722
Well	0.180083

Table 5: Top correlated sensors for Home F

	Correlation with total consumption
Water Heater	0.518500
Family Room	0.365200
Furnace	0.356082
Dryer	0.298531
Half-bath Foyer	0.232953
Dishwasher Disposal	0.143392

Table 6: Top-3 sensors – Home B

B	3 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	0.3516	0.3763
Ad R2	0.8225	0.7962

Table 7: Top-6 sensors – Home B

B	6 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	0.1483	0.1756
Ad R2	0.9690	0.9556

Table 8: Top-3 sensors – Home C

C	3 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	0.5189	0.5049
Ad R2	0.5901	0.6129

Table 9: Top-6 sensors – Home C

C	6 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	0.4068	0.4173
Ad R2	0.7527	0.7358

Table 10: Top-3 sensors – Home F

F	3 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	1.0829	1.0506
Ad R2	0.5411	0.5660

Table 11: Top 6-sensors – Home F

F	6 Top – 10-fold CV	
	Random Forest	Gradient Boosting
RMSE	0.8831	0.9618
Ad R2	0.6955	0.6362

Summarizing, it is safe to conclude that furnaces and living rooms are those that have high correlation with the total consumption in all cases. Of course, as expected the A/C was the major sensor for Home B, but unfortunately there was not any for the other two homes. In terms of predictability based on the top sensors, it was clear that there was no validity for Home F, while for Home C the results were 'marginally' acceptable. Home B, on the other hand, gave optimistic results and this was not only attributed to the A/C sensor. Home B showed smoother and more similar consumption patterns across the years.

1.4.2. SCENARIO 2

For a real-time prediction, it is necessary to record data and analyse them in a very short time. However, citizens are not only interested in learning a prediction of their instant consumption in order to adapt their consuming behaviour, but also for prior knowledge of upcoming consumption. Providers who need to adjust their production plans and form a more competitive pricing policy show similar interest. In that case, as explained in Section 3.1 the hourly information of consumption and weather metrics is transformed into one unified value that represents the average.

For example, for each day there are two values regarding consumption; one for each interval of interest that gets examined (on-off peak). These values are derived from the average values during the intervals. In addition, these daily values of consumption are transformed into binary separation around three different mean values.

Initially, we split the target variable around the general mean value for each home, in addition, we split based on the mean value of each month and season. Finally, two extra variables were added, which reflect the consumption of previous day (*Yesterday*) as well as the summation of a 7-day ahead consumption (*Past Week*). Below, Table 12 shows the number of instances that each class includes for each of the described splits.

Fully balanced datasets rarely exist in real world classification problems and that is also the case here. The instances in every case are more for class *Low*. In classification problems it is also crucial to identify the most "interesting" class in order to choose appropriate evaluation metrics. In our case, both classes are considered equally interesting, thus we evaluate accuracy.

In this scenario we introduced different algorithms, in contrast to the first, as suggested in the literature. The algorithms that were tested are Support Vector Machines (SVM), Random Forest, Stochastic Gradient Descent (SGD) and Logistic Regression. All algorithms were examined following the stages detailed below.

Stage 1 represents the initial results for each algorithm's default implementation based on the broadly known Scikit-Learn package [48].

Stage 2 reflects the results when all the data are scaled. Scaling of data is useful in cases that data represent different units and ranges. In advance we know that scaling does not affect every algorithm (Random Forest etc.).

Table 12: Balance of the two classes around the examined means

Homes -Time Interval	Consumption Per Total		Consumption Per Month		Consumption Per Season	
	Classes		Classes		Classes	
	High	Low	High	Low	High	Low
Home B ON-peak	361	735	431	665	378	718
Home B OFF-peak	372	724	400	696	393	703
Home C ON-peak	395	685	382	698	380	700
Home C OFF-peak	408	672	390	690	398	382
Home F ON-peak	553	543	553	543	547	549
Home F OFF-peak	438	658	447	649	452	644

Stage 3 returns the results after hyper-parameter tuning is performed. In this stage we aim to push the performance of the model to its limits by applying various combinations of input parameters.

It is important to clarify that for some SVM kernels the process is delayed. This happens mostly because the linear kernel is an almost identical implementation with SGD’s hinge kernel. Moreover, the polynomial kernel requires data to be scaled.

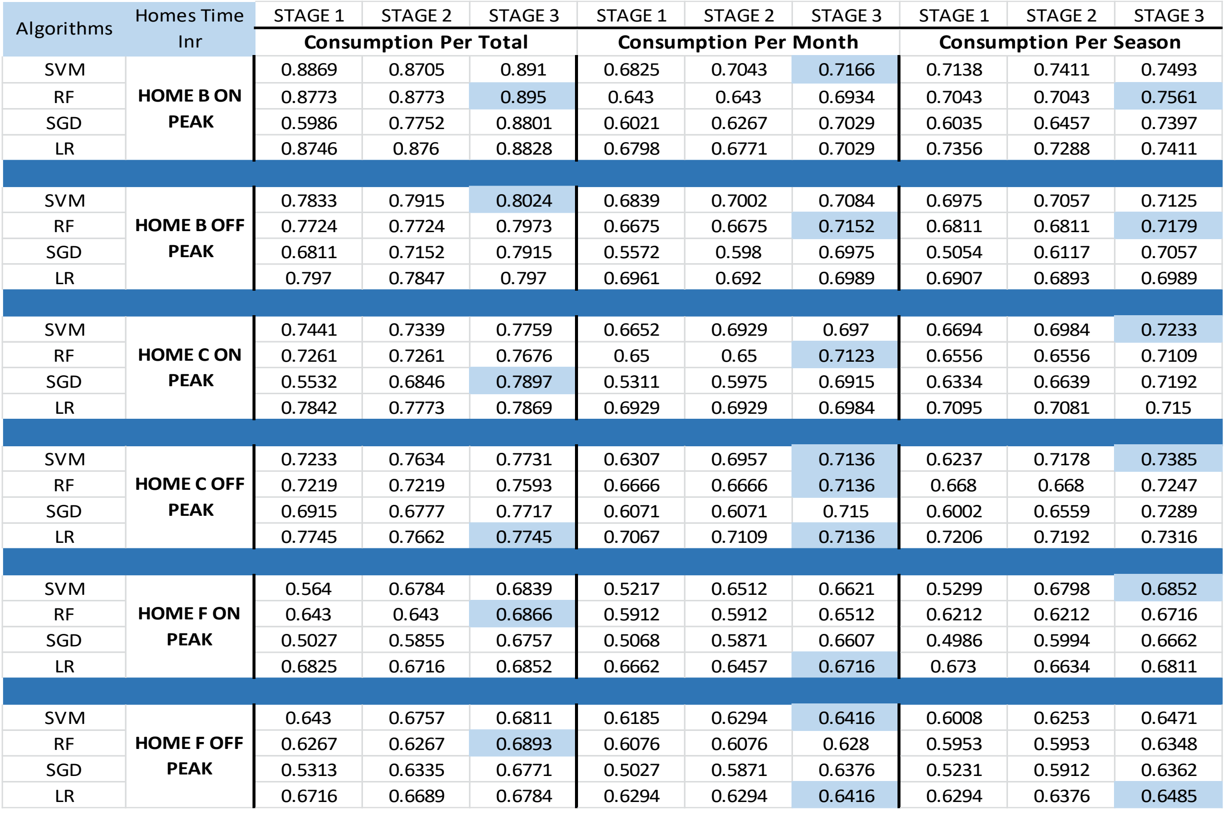
The results clearly indicate that there is no algorithm superior to all others. Logistic regression required less hyper-parameter tuning and was not highly affected by that. The most stable algorithms were Random Forest and SVM; however, the latter is slower, whilst it shows unpredictable behavior during scaling.

Regarding the two-time intervals, initially it was assumed the off-peak period would lead to better results; however, this was not supported by the final results. On the contrary, the on-peak period data returned better results. This happened mostly due to the higher fluctuations of on-peak periods which bring about more information. Table 13 shows aggregated results as they emerged from each stage for each algorithm.

1.4.3. Grid Load Simulation

As part of the second scenario we decided to examine how would a model using data aggregated from all 3-Homes would perform. An approach like that could simulate the grid and assist providers with better decision making. Since all the houses are in the same region, weather data are very similar, thus an average value for weather metrics was calculated. Regarding the electrical consumptions (target variable), values were summed both for Yesterday and Past Week variables. The same binary transformation was conducted again, only around the total mean though. The split of classes for On-peak is 474 High – 606 Low, while for Off-peak period is 433 High – 647 Low.

As seen in Tables 14-15, the performance of the model does not change drastically. However, at this point a different perspective of generalization can be examined through merging. The accuracy for both periods remains similar however, for Off-peak period a slightly higher accuracy is achieved.



Algorithms	Homes Time Inr	STAGE 1	STAGE 2	STAGE 3	STAGE 1	STAGE 2	STAGE 3	STAGE 1	STAGE 2	STAGE 3
		Consumption Per Total			Consumption Per Month			Consumption Per Season		
SVM	HOME B ON PEAK	0.8869	0.8705	0.891	0.6825	0.7043	0.7166	0.7138	0.7411	0.7493
RF		0.8773	0.8773	0.895	0.643	0.643	0.6934	0.7043	0.7043	0.7561
SGD		0.5986	0.7752	0.8801	0.6021	0.6267	0.7029	0.6035	0.6457	0.7397
LR		0.8746	0.876	0.8828	0.6798	0.6771	0.7029	0.7356	0.7288	0.7411
SVM	HOME B OFF PEAK	0.7833	0.7915	0.8024	0.6839	0.7002	0.7084	0.6975	0.7057	0.7125
RF		0.7724	0.7724	0.7973	0.6675	0.6675	0.7152	0.6811	0.6811	0.7179
SGD		0.6811	0.7152	0.7915	0.5572	0.598	0.6975	0.5054	0.6117	0.7057
LR		0.797	0.7847	0.797	0.6961	0.692	0.6989	0.6907	0.6893	0.6989
SVM	HOME C ON PEAK	0.7441	0.7339	0.7759	0.6652	0.6929	0.697	0.6694	0.6984	0.7233
RF		0.7261	0.7261	0.7676	0.65	0.65	0.7123	0.6556	0.6556	0.7109
SGD		0.5532	0.6846	0.7897	0.5311	0.5975	0.6915	0.6334	0.6639	0.7192
LR		0.7842	0.7773	0.7869	0.6929	0.6929	0.6984	0.7095	0.7081	0.715
SVM	HOME C OFF PEAK	0.7233	0.7634	0.7731	0.6307	0.6957	0.7136	0.6237	0.7178	0.7385
RF		0.7219	0.7219	0.7593	0.6666	0.6666	0.7136	0.668	0.668	0.7247
SGD		0.6915	0.6777	0.7717	0.6071	0.6071	0.715	0.6002	0.6559	0.7289
LR		0.7745	0.7662	0.7745	0.7067	0.7109	0.7136	0.7206	0.7192	0.7316
SVM	HOME F ON PEAK	0.564	0.6784	0.6839	0.5217	0.6512	0.6621	0.5299	0.6798	0.6852
RF		0.643	0.643	0.6866	0.5912	0.5912	0.6512	0.6212	0.6212	0.6716
SGD		0.5027	0.5855	0.6757	0.5068	0.5871	0.6607	0.4986	0.5994	0.6662
LR		0.6825	0.6716	0.6852	0.6662	0.6457	0.6716	0.673	0.6634	0.6811
SVM	HOME F OFF PEAK	0.643	0.6757	0.6811	0.6185	0.6294	0.6416	0.6008	0.6253	0.6471
RF		0.6267	0.6267	0.6893	0.6076	0.6076	0.628	0.5953	0.5953	0.6348
SGD		0.5313	0.6335	0.6771	0.5027	0.5871	0.6376	0.5231	0.5912	0.6362
LR		0.6716	0.6689	0.6784	0.6294	0.6294	0.6416	0.6294	0.6376	0.6485

Table 13: Final Results for each stage of all the algorithms for both time intervals

Table 14: On peak / general mean value

Grid	On-peak / General mean value			
	SVM	***Random Fst***	***SGD***	***Logistic Regr***
Stage 1	0.6846	0.6680	0.5892	0.7026
Stage 2	0.7358	0.6680	0.6154	0.7150
Stage 3	0.7358	0.7178	0.7247	0.7192

Table 15: Off peak / general mean value

Grid	Off-peak / General mean value			
	SVM	***Random Fst***	***SGD***	***Logistic Regr***
Stage 1	0.6376	0.6860	0.5767	0.7302
Stage 2	0.7219	0.6860	0.6528	0.7275
Stage 3	0.7275	0.7495	0.7247	0.7302

4. Traffic Prediction

Besides the approach for predicting electricity consumption detailed in the previous section, weather data can be exploited for different smart cities problems and scenarios, such as traffic prediction.

4.1. Problem Definition and Approach

Traffic prediction is a multi-dimensional problem. Research focuses on either predicting traffic loads for a given time or choosing the optimal route for a vehicle based on real-time adaptations, in order to minimize travel time. Traffic is affected by numerous factors just as electrical consumption. Accidents and social events can disrupt the normal load of vehicles in specific areas, while season and weather conditions can affect traffic in a larger scale. Based on that, we used weather data collected by sensors installed in various city spots for predicting the day-ahead volume of traffic.

The approach we followed is quite similar to that in section 3, but the scenario is reversed, meaning that we try to answer the question ‘How can you exploit sensor data that are not personalized and create meaningful conclusions for the general public?’ Deployment of smart city infrastructure requires a deep understanding of the problem of traffic. It is crucial to define your objective in advance for selecting the most appropriate locations to install sensors that either measure traffic loads or collect weather data. Busy roads do not always provide more information in comparison to less busy ones. The number of alternative routes or the location of busy buildings can affect the necessity of measuring traffic density for a specific road. Our approach aims at clarifying differences in traffic among locations, besides assessing traffic predictability based on weather data.

4.2. Context

Relating weather data with traffic is not a new subject. There are several research efforts that perform both classification and regression tasks to address that problem. In 2009, the authors of [49] examined the power of decision trees for classifying traffic load into three levels, based only on time and temperature. The results were positive and motivated our research. Moreover, in [50] the authors propose the use of volume and occupancy data. Such data, as well as speed, can be obtained by loop detectors. Loop detectors are sensors buried underneath highways and estimate traffic by collecting information related to vehicles passing past them. In general, tree-based algorithms are widely used and justified, however even more sophisticated algorithms are used such as SVM [51] and neural networks [52].

In addition, authors in [53] introduce three binary variables, other than weather conditions, for holiday, special conditions and road quality. However, in our case such features were not included for several reasons. First of all, holidays do not have the same effect on each season. For example, sunny holidays might result in lower levels of traffic in large urban centers while this is not the case for winter holidays. Special conditions, such as big social events or demonstrations are not easily modeled, since most of the times they occur suddenly. For multiclass classification it is also desired to test the predictive power of a model, not only based on accuracy but also the deviation of a wrong-labeled instances. Based on that, authors in [54] tested max entropy models on a 6-class approach and achieved nearly perfect scores on 1-level deviation class. In our case the deviations are not that high in order to justify such a detailed split of the target variable.

4.3. Dataset Description and Processing

For this task, our source of data is the newly deployed ppcity.io which is a set of platforms providing information to the citizens of Athens, Greece. The selected platform utilizes environmental, traffic and geospatial data. These data are collected by several sensors located at central points around the city. In a similar fashion to section 3, we conducted experiments relating environmental data with traffic. Initially, we selected locations where data about traffic and weather conditions were available. We focused on the seven most reliable spots, meaning spots with sensors with the most compact data flow and the widest recording ranges. Table 16 details each sensor location.

Table 16: Sensor location

Sensor 10	Kallidromiou st.: house
Sensor 13	Monastiriou st.: High School
Sensor 14	Politechneiou st.: Attica Region building
Sensor 18	Solonos st. – Sociality Office
Sensor 19	Grammou st.: High School
Sensor 20	Get Ltd. Offices
Sensor 25	Aiolou st.: Municipality building

More specifically, regarding data description, weather attributes collected were the following: Humidity, Pressure, Temperature, Wind Direction and Speed. The platform provides more weather data, but those were not included in the modeling process as they were not deemed relevant, such as ozone concentration, ultraviolet radiation (UV), nitrogen dioxide etc.

The target variable (traffic load), named Jam Factor in the platform, represents the quality of travel. It ranges between 0 and 10; 0 corresponding to a completely empty road and 10 indicating traffic in a standstill. Data were collected for almost three months including August. August is considered as the month with the lowest traffic volume in Athens since it is the period that most people go on holidays.

The way the data are processed is similar to the one selected in section 3. We define three distinct time intervals within a day and examine their differences. The first interval, named Morning, includes data collected between 07:00 and 10:00. The second, named Afternoon, includes data collected between 15:00 and 18:00 and the last one, Evening, between 19:00 and 22:00. The average value for each interval is computed. The problem is approached again by categorizing the target variable into two or three classes. Moreover, we decided to split down the model into two sub models regarding weekdays or weekends. Table 17 provides basic statistics for the target variable, as well as information for the split of classes.

TABLE 17: Statistics of traffic load for each sensor

Sensors	Intervals	Statistics					No. of classes	
		Max	Min	Median	Range	Mean	High	Low
10	07:00-10:00	2.52	0.65	1.77	1.86	1.71	45	41
	16:00-19:00	3.18	1.22	2.48	1.96	2.32	46	43
	19:00-22:00	2.69	1.27	2.09	1.41	2.04	47	41
13	07:00-10:00	2.57	0.66	1.81	1.91	1.71	49	37
	16:00-19:00	3.37	1.17	2.29	2.19	2.23	48	41
	19:00-22:00	2.37	1.26	1.87	1.11	1.84	45	43
14	07:00-10:00	2.51	0.62	1.81	1.88	1.69	46	41
	16:00-19:00	3.25	1.18	2.34	2.06	2.27	47	42
	19:00-22:00	2.55	1.24	2.01	1.31	1.95	46	42
18	07:00-10:00	2.54	0.67	1.77	1.87	1.69	45	42
	16:00-19:00	3.27	1.22	2.44	2.05	2.31	46	43
	19:00-22:00	2.61	1.27	2.05	1.34	2.01	47	41
19	07:00-10:00	3.33	0.51	1.97	2.82	1.93	45	41
	16:00-19:00	3.21	1.11	2.4	2.1	2.31	46	43
	19:00-22:00	2.77	1.24	2.11	1.52	2.04	46	42
20	07:00-10:00	2.49	0.54	1.51	1.95	1.46	45	41
	16:00-19:00	2.5	0.85	1.65	1.65	1.64	45	44
	19:00-22:00	3.17	0.97	1.66	2.2	1.68	44	44
25	07:00-10:00	2.67	0.69	1.79	1.97	1.71	46	41
	16:00-19:00	3.57	1.2	2.57	2.36	2.36	46	43
	19:00-22:00	2.67	1.27	2.11	1.39	2.05	48	40

In summary our work was conducted in three stages. Initially, we used all the available data, introducing a binary variable which indicates if the day is part of a weekend or not. At this point the target variable was split into two categories around the mean value. The second stage was to examine a sub model that included only weekday instances. However, at this point the value that we split the Jam Factor was the median, resulting in perfectly balanced classes. The final stage was to split the target variable into three categories of equal size (High, Medium, Low). The metric used for classification evaluation was accuracy, since the classes are balanced and of the same interest.

4.4 Experiments

The algorithms benchmarked for all experiments in this stage were logistic regression and random forest. Fig. 13 shows the classification results for the initial stage for each one of the described time intervals.

Fig. 14 and 15 respectively show the results obtained by excluding weekends and introducing one more class to describe the target variable.

Having observed the results for different sensors we focused on traffic volumes for each time interval. Fig. 16 verifies that August is the least busy month for large cities with a clear decline centered around the 15th of the month.

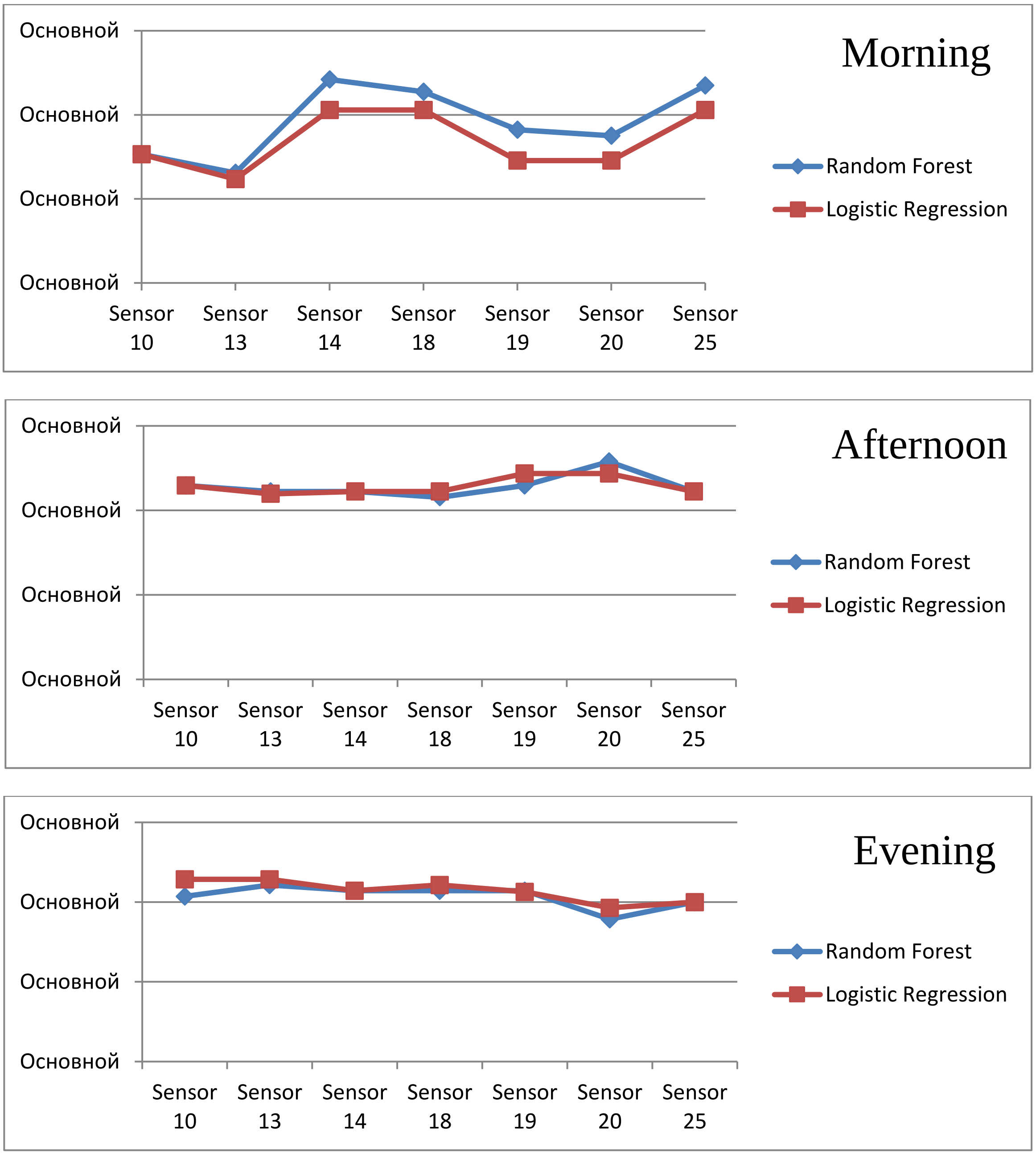


Figure 13: Accuracy for the whole set of sensors

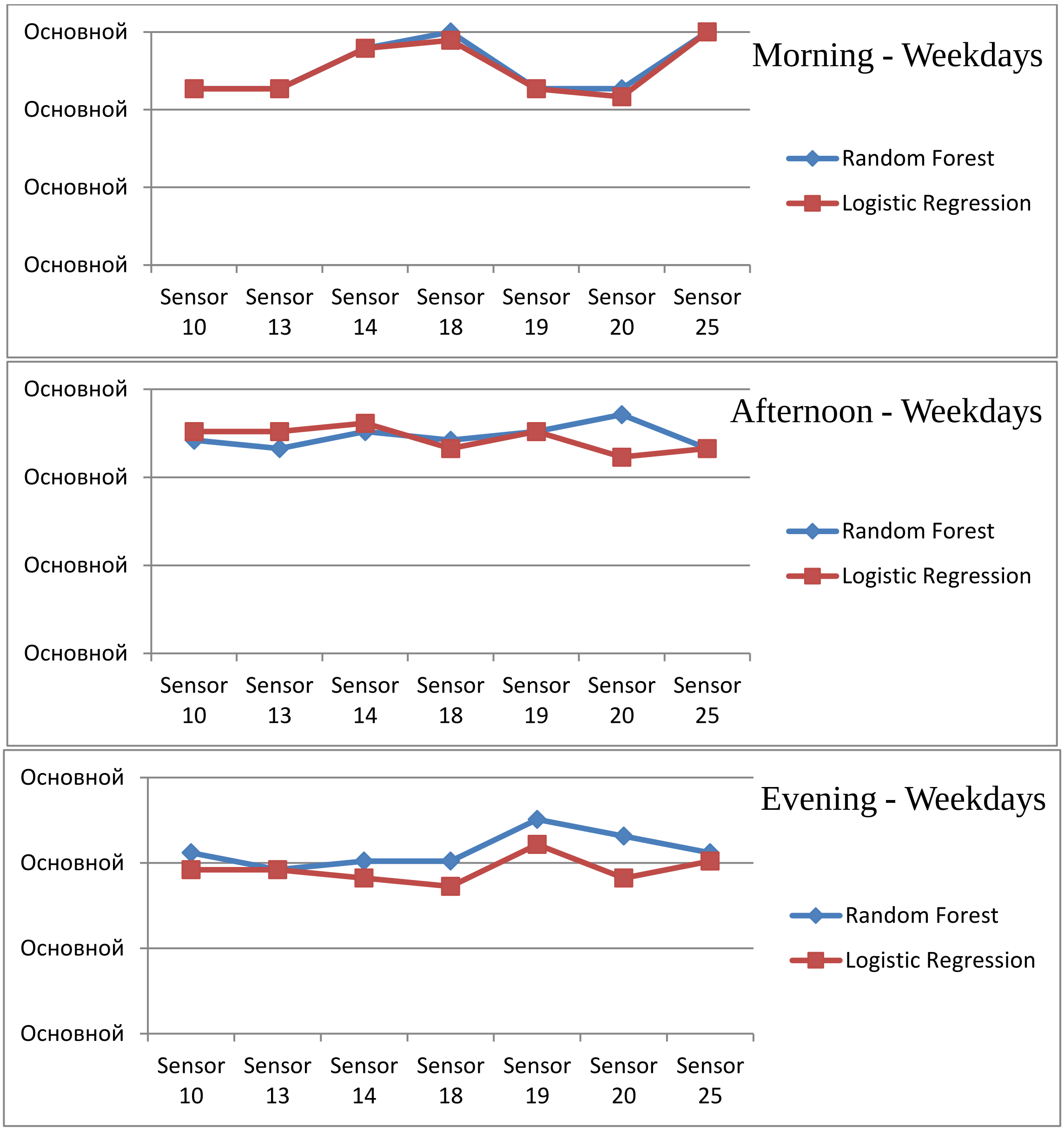


Figure 14: Accuracy for weekday data

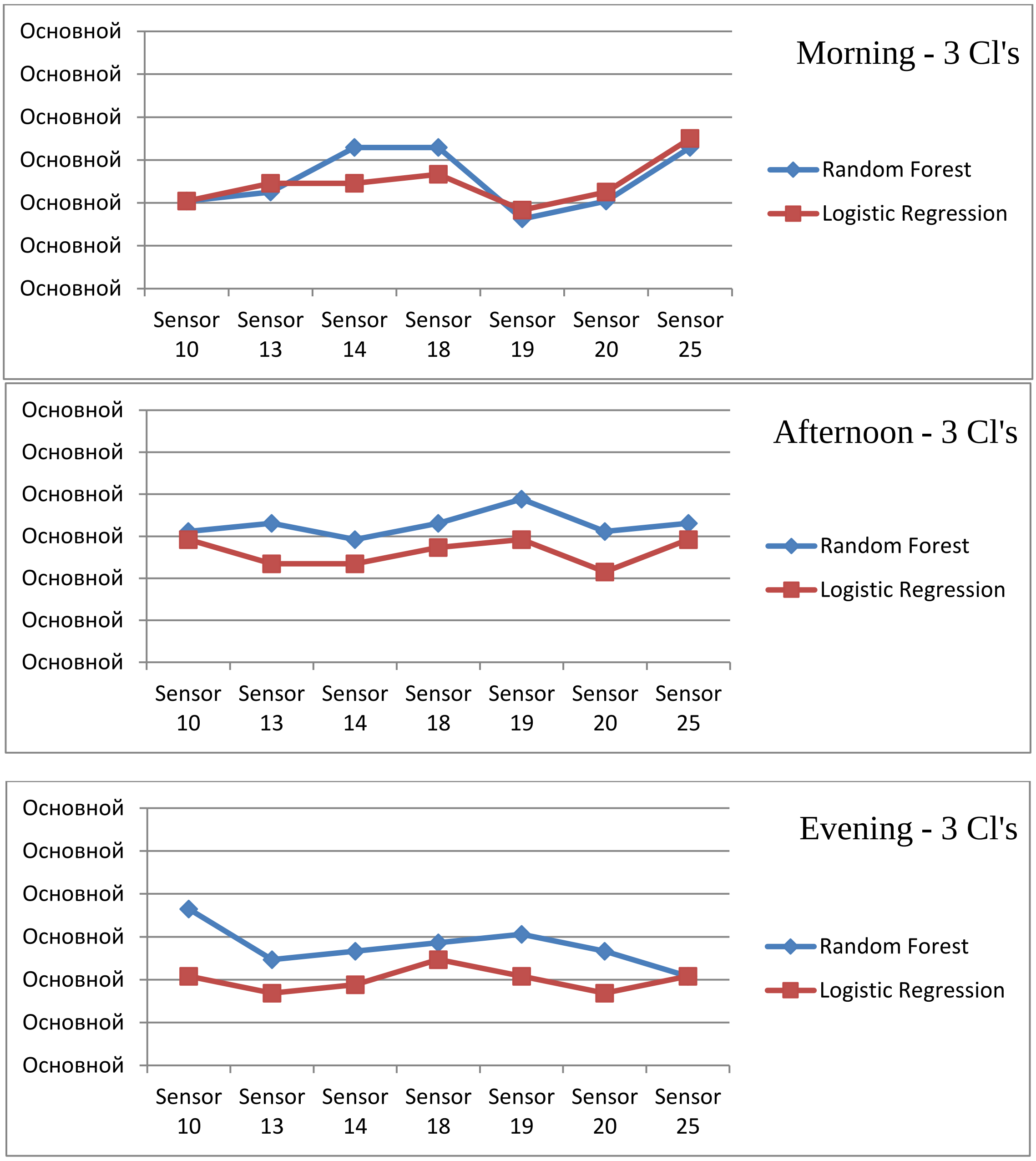


Figure 15: Accuracy for the three-class target variable

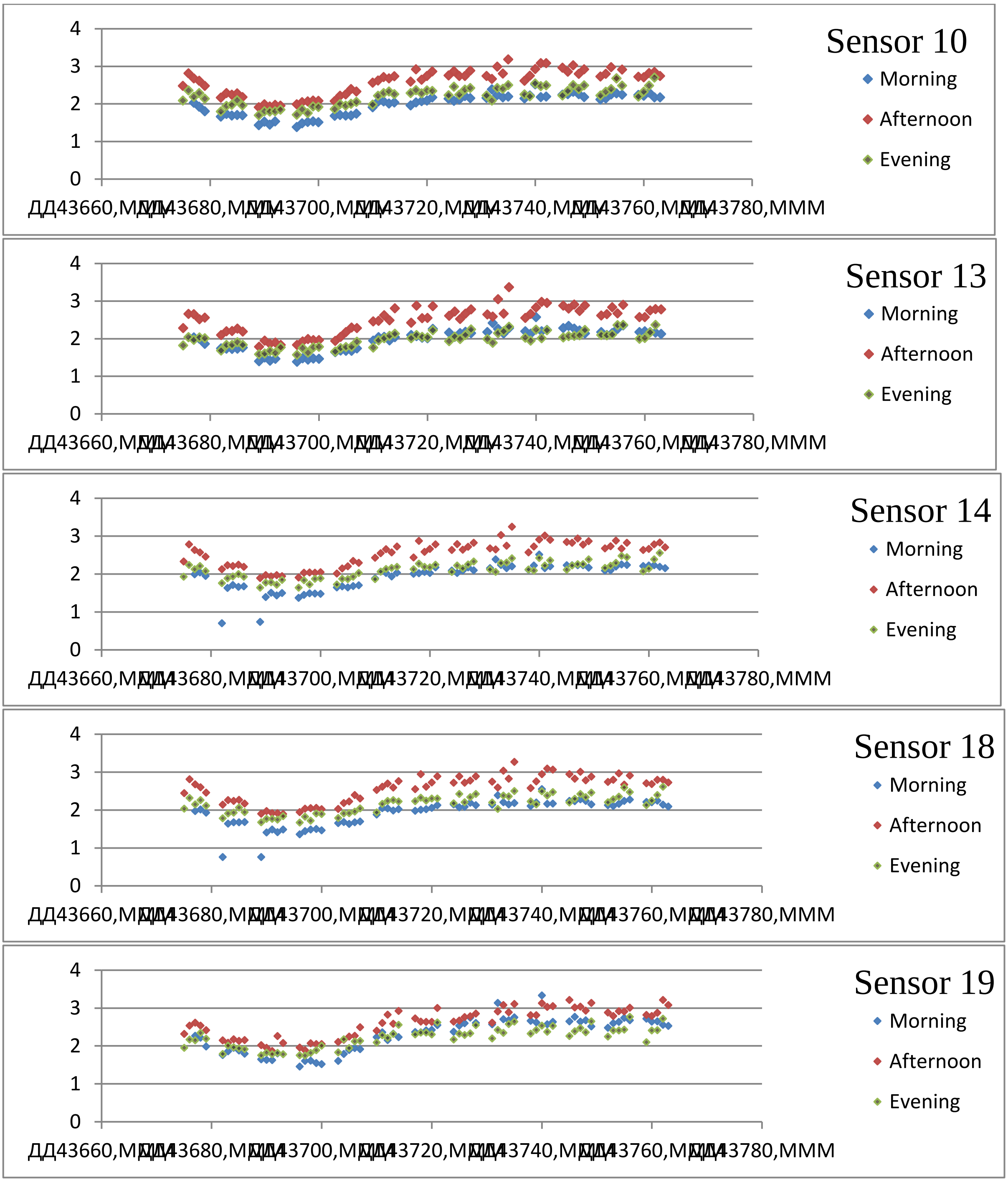


Figure 16 (Part 1): traffic volumes per sensor for each time interval

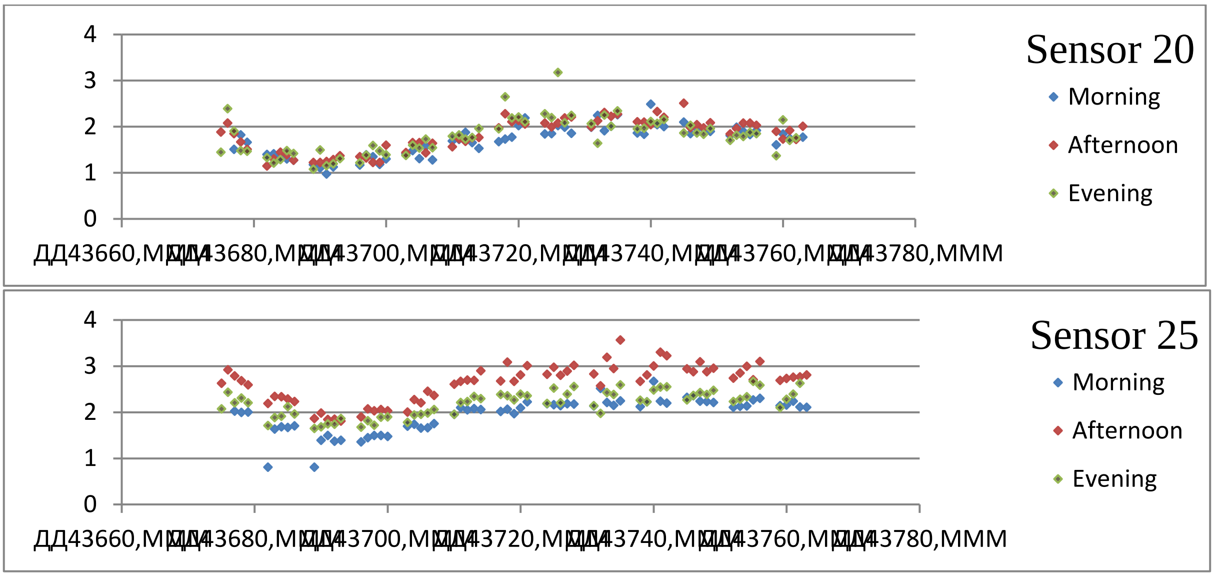


Figure 16 (Part 2): traffic volumes per sensor for each time interval

We can also observe that the highest volume of traffic on a weekday occurs in the afternoon, for every sensor except from Sensor 20. For the other two periods of the day, it is not clear which one dominates over the other. It appears that traffic is heavier in the evenings in August, while from the begging of September both mornings and evenings showcase equal traffic volumes.

5. Evaluation

We evaluate this work following the KDD process including all the essential steps. We start with similarities across both case studies. Firstly, data acquisition revealed the importance of acquiring enough data. Data collection should be steady and clearly defined in advance, supported by domain expertise wherever possible. In the first case study mostly, we had to discard a lot of data due to sensor recording failures.

Regarding the step of pre-processing, on both case studies, initially we used all the available features with questionable results. Better results were achieved after discarding features which introduced noise. Especially for weather data exploitation, it is crucial to fully understand the correlation between variables, otherwise redundant information might be captured.

The evaluation of machine learning algorithms was conducted uniformly. Even though algorithms showcased similar performance, Random Forest was the most stable. In addition, the fact that it does not require data scaling and it can easily handle missing values, resulted in its superiority. The remaining of this section discusses the findings for each case study individually.

The first case study focused on individual households, unlike many studies which attempt to predict electricity consumption of the grid or large blocks of apartments. The first scenario for this case study demonstrated that one can calculate the total home consumption with marginal error using readings from a well-defined set of sensors, measuring the consumption of certain appliances. Thus, it is recommended to install smart meters in key appliances, besides the main electricity meter for smart homes, thus facilitating data availability and analysis. In contrast with commercial buildings, where electricity consumption follows a pattern (for example 08:00 to 17:00 with a decline around 13:00 during the lunch break), occupant behavior affects individual households. Thus, at the single household scale, a thorough prediction model is hard to be established.

Accuracy was over 75% for two out of the three houses. Much higher accuracy was not expected in this analysis, as important factors, such as occupancy or activity inputs, were not available and could not be incorporated into the model. We could have drawn more robust conclusions if the data from all seven homes were available and met the criteria set. Furthermore, the decomposition of this time-series problem into explanatory input variables gives more space for creativity and understanding of the problem itself.

For the second case study, the general problem in terms of real-time adjustment of routes in order to avoid traffic congestion is important for many cities. However, the day-ahead prediction of the volume assumes that there will not be any unexpected incident. Starting from this point, the factor of environmental conditions is crucial, since many citizens and visitors decide in advance the way they want travel around the city in the upcoming day. The results of the approach in section 4 are encouraging and justify what was stated above. All three stages achieve accuracy above the expected levels. For instance, at the first stage, afternoon and evening periods result in accuracy higher than 0.8 for almost all sensors. The morning period is a bit unstable, with significantly lower accuracy especially for sensors 10, 13, 19 and 20. However, sensors 13 and 19 are installed in school areas and that might have a negative impact, since environmental conditions are not the determining factor affecting the volume of traffic in August. Surprisingly, excluding weekends from the initial model results in much higher accuracy in the morning. However, the same sensors return worse results. For the remaining two periods that exclusion had different impact. The afternoon was affected positively, while the evening slightly negatively.

Reforming the target variable into three classes of equal size was promising. First, the accuracy was steadily higher than 0.6 for all sensors. At this point, for the first time it was obvious that one algorithm was performing consistently better than the others. Random Forest was for every sensor equal or better than Logistic Regression for the afternoon and evening periods. However, it is worth noting that parameter tuning for Random Forest was much slower than for Logistic Regression which is widely considered to be among the fastest classifiers.

6. Conclusions and Future Work

Both case studies produced some clear and informative results. The main conclusion is that the systematic recording and manipulation of weather data can be supportive for decision making.

With regards to predicting electricity consumption, we conclude the following:

- The accuracy of the model increases when consumption follows standard patterns throughout the year. For instance, Home F demonstrates a smooth hourly average consumption, but that is not true for the average monthly consumption.
- Even not strongly supported by data, there is an intuition that the bigger a home the harder it was for the model to predict consumption. For instance, Home C is twice as big as Home B.
- It is safe to rely on a set of 3 up to 6 sensors in order to predict total consumption. In both cases predicted results slightly deviated from the actual values.
- The most important part of the analysis is feature selection, as the selection of algorithms or the evaluation techniques (i.e. Cross-validation, Manual split) did not affect the result that much.
- Unexpectedly, Home F and Home C show no difference in predictions for On or Off-peak periods. For Home B there is a significant decrease for the Off-peak period. Possibly because none of the periods is actually of low consumption for both homes.
- Past consumption data had a slight positive impact on the model.

Regarding the traffic load prediction, we have more concrete conclusions since we collected information from more sensors. The list below summarizes the conclusions.

- In the afternoon roads are busier and for most of the sensors even evenings have higher volume of traffic in comparison to mornings.
- The exclusion of weekends from the model resulted in slightly better accuracy.
- Morning data do not return stable results for all the sensors. As discussed, this may be because some sensors are located in schools.
- Transforming the target variable into three classes resulted in admittedly good results, far better than the baseline model.
- Regarding the algorithms, only in the case of three classes it appears that one algorithm, Random Forest, stands out for all sensors.

6.1 Threats to Validity

The biggest threat on such approaches, is that those day-ahead models rely on weather data which also are predicted. Thus, it is crucial that we have accurate predictions of weather conditions. In addition, the size of the homes for first case study is much bigger than a typical house or a regular apartment. Since the electricity consumption is also affected from the size, in smaller residences the results may have been different. Another threat could be sufficient deseasonilising; the factor of time could be possibly analyzed into more explanatory variables.

6.2 Future Research Directions

Research like the one presented here requires in-depth analysis and utilization of different approaches. For that reason, we attempted to cover as many aspects of the research problem of prediction as possible. On both data repositories it is expected more data to be released in 2020. Of course, having more data leads to better approaches and more concrete results.

Having more data could also lead to deployment of even more sophisticated algorithms. Artificial neural networks have already been implemented for similar works with good results. Regarding feature selection, on the first case study it is desired to introduce the concept of the *week of the month*, which is increasingly used by similar efforts. More specifically, it is claimed that people tend to consume higher amounts of electricity during a specific week of the month.

For the second case study, besides testing more sensors in Athens, it would be beneficial to analyze data from Thessaloniki as well and to compare and contrast results. Lastly, it would be interesting to compare the results for both case studies on different time intervals. As described in sections 3 and 4, our selection of periods during the day was static.

The work was implemented within the project ppCity - T1EΔK-02901, co-funded by the European Regional Development Fund, the European Comission and the Greek National Structural Funds.

References

[1] L.G. Anthopoulos. 2017. Understanding Smart Cities: A Tool for Smart Government or an Industrial Trick? Springer. DOI: 10.1007/978–3–319–57015–0

[2] R. Giffinger and H. Gudrun. 2010. Smart Cities Ranking: An Effective Instrument for the Positioning of Cities. ACE 4, 12 (February 2010), p.p.7–25, URI: http://hdl.handle.net/2099/8550

[3] Komninos N, (ed.) the architecture of intelligent cities: Integrating human, collective and artificial intelligence to enhance knowledge and innovation. Intelligent Environments, 2006 IE 06 2nd IET Int'l Conf. on; 2006: IET.

[4] Chourabi H, Nam T, Walker S, Gil-Garcia JR, Mellouli S, Nahon K, et al., editors. Understanding smart cities: An integrative framework. System Science (HICSS), 2012 45th Hawaii Int'l Conf. on; 2012: IEEE

[5] Caragliu A, Del Bo C, Nijkamp P. Smart cities in Europe. Journal of urban technology. 2011;18(2):65-82.

[6] Albino V, Berardi U, Dangelico RM. Smart cities: Definitions, dimensions, performance, and initiatives. Journal of Urban Technology. 2015;22(1):3-21.

[7] Palvia SCJ, Sharma SS, editors. E-government and e-governance: definitions/domain framework and status around the world. Int'l Conf. on e-governance; 2007.

[8] Khan Z, Kiani SL (2012) A cloud-based architecture for citizen services in smart cities. In: ITAAC Workshop 2012. IEEE Fifth International Conference on Utility and Cloud Computing (UCC), Chicago, IL, USA. pp 315–320. IEEE

[9] A. Alkandari, M. Alnasheet, I. Alshekhly, (2012) Smart Cities: Survey

[10] S. Dirks, M. Keeling, J. Dencik. (2009) How smart is your city? Helping cities measure progress

[11] Al Nuaimi et al. (2015) Journal of Internet Services and Applications 6:25 DOI 10.1186/s13174-015-0041-5

[12] United Nations Environment Programme. (2015) The Global Initiative for Resource Efficient Cities

[13] E. L. Glaeser and B. Sacerdote "Why Is There More Crime in Cities?" Journal of Political Economy, Vol. 107, no. 6, part 2 (December 1999) 225

[14] HP study finds alarming vulnerabilities with IoT home security systems (2015) http://www8.hp.com/us/en/hp-news/press-release.html?id=1909050

[15] How the EU‟s GDPR will be a game changer for cities using open data (2018) https://eu.smartcitiescouncil.com/article/how-eus-gdpr-will-be-game-changer-cities-using-open-data

[16] P. Berrone, J. E. Ricart (2018) IESE Cities in motion index

[17] J. Han, M. Kamber, J. Pei (2012) Data Mining Concepts and Techniques 3rd Edition

[18] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Ma-chine Learning Tools and Techniques 4th edition

[19] Ghafari, S.M.; Tjortjis, C. 'A Survey on Association Rules Mining Using Heuristics', WIREs Data Mining and Knowledge Discovery, Vol. 9, no. 4, July/August 2019, (Wiley).

[20] Tzirakis P. and Tjortjis C., 'T3C: Improving a Decision Tree Classification Algorithm's Interval Splits on Continuous Attributes', Advances in Data Analysis and Classification, Vol. 11, No. 2, pp. 353-370, 2017, (Springer).

[21] P. Sotres, J. R. Santana, L. Sanchez, J. Lanza, L. Munoz (2017) Practical lessons from the deployment and management of a smart city IoT infrastructure: The SmartSantander Testbed case

[22] H. Blockeel, M. Sebag (2004) Scalability and Efficiency in multi-relational data mining

[23] D. K. Singh, V. Swaroop (2013) Data Security and Privacy in Data Mining: Research Is-sues and Preparation

[24] V. Moustaka, A. Vakali, L.G. Anthopoulos. 2018. A systematic review for smart city data analytics. ACM Computing Surveys. DOI: https://doi.org/10.1145/3239566

[25] D. Rousidis, P. Koukaras, C. Tjortjis, "Social Media Prediction A Literature Review", Multimedia Tools and Applications, Springer, 2019, DOI: 10.1007/s11042-019-08291-9.

[26] K. Christantonis, C. Tjortjis, "Data Mining for Smart Cities: Predicting Electricity Consumption by Classification", 10th IEEE Int'l Conf. on Information, Intelligence, Systems and Applications (IISA 2019), pp. 67-73, 2019.

[27] Mi. Chui, M. Löffler and R. Roberts. 2010. The Internet of Things. McKinsey Quarterly (March 2010). Retrieved June, 2017 from: http://www.mckinsey.com/industries/high–tech/our–insights/the–internet–of–things

[28] E. Estellés–Arolas and F. González–Ladrón–de–Guevara. 2012. Towards an integrated crowdsourcing definition. JIS, 1–14. DOI: http://dx.doi.org/0.1177/016555150000000

[29] E. Broad. 2015. Closed, shared, open data: what's in a name? (September 2015). Retrieved June 2017 from: https://theodi.org/blog/closed–shared–open–data–whats–in–a–name

[30] Y. Mo, T. Hyun Jim Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, B. Sinopoli (2011) Cyber-Physical security of a Smart Grid Infrastructure

[31] C. Xia, J. Wang, K. McMenemy (2010) Short, medium and long term load forecasting model and virtual load forecaster based of radial basis function neural networks

[32] D. Chen, D. Irwin (2017) Weatherman: Exposing Weather-based Privacy Threats in Big Energy Data

[33] A. Mishra, D. Irwin, P. Shenoy, J. Kurose, T. Zhu (2012) SmartCharge: Cutting the Electricity Bill in Smart Homes with Energy Storage

[34] S. Barker, A. Mishra, D. Irwin, P. Shenoy, J. Albrecht (2012) SmartCap: Flattening Peak Electricity Demand in Smart Homes

[35] A. Azadech, S.F Ghaderi, S. Sohrabkhami (2007) Forecasting electrical consumption by integration of Neural Network, time-series and ANOVA

[36] A. Ahmad, M. Hassan, M. Abdullah, H. Rahman, F. Hussin, H Abdullah, R. Saidur (2014) A review on applications of ANN and SVM for building electrical energy consumption forecasting

[37] A. Sauhats, R. Varfolomejeva, O. Linkevics, R. Petrecenko, M. Kunickis, M. Balodis (2015) Analysis and prediction of electricity consumption using smart meter data

[38] X. Li, C.P. Bowers, T. Schnier (2010) Classification of Energy Consumption in Buildings With Outlier Detection

[39] M. Espinoza, C. Joye, R. Belmans, B.D Moor (2005) Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series

[40] Q. Wang (2009) Grey Prediction Model and Multivariate Statistical Techniques Forecasting Electrical Energy Consumption in Wenzhou, China

[41] P. Koukaras, C. Tjortjis, D. Rousidis, "Social Media Types: Introducing a Data Driven Taxonomy", Computing, Springer, Vol. 102, no. 1, pp. 295-340, 2020.

[42] A. D. Amato, M. Ruth, P. Kirshen, J. Horwitz (2005) Regional energy demand responses to climate change: Methodology and Application to the commonwealth of Massachusetts

[43] R. Torkzadeh ET. Al. (2014) Medium Term Load Forecasting in Distribution Systems Based on Multi Linear Regression & Principal Component Analysis: A Novel Approach

[44] V. Bianco, O. Manca, S. Nardini (2009) Electricity consumption forecasting in Italy using linear regression models

[45] G. Jannuzzi, L. Schipper (1991) The structure of electricity demand in the Brazilian house-hold sector

[46] E. Almeshaiei, H. Soltan (2011) A methodology for Electric Power Load Forecasting

[47] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy (2012) Smart*: An open data set and tools for enabling research in sustainable homes

[48] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

[49] Applying Data Mining in Prediction and Classification of Urban Traffic. S. K. Nejad, F. Seifi, H. Ahmadi and N. Seifi, 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, 2009, pp.674-678. doi: 10.1109/CSIE.2009.906

[50] Dynamic Traffic Prediction Based on Traffic Flow Mining. Y. Wang, Y. Chen, M. Qin and Y. Zhu, 2006 6th World Congress on Intelligent Control and Automation, Dalian, 2006, pp.6078-6081. doi: 10.1109/WCICA.2006.1714248

[51] Short-Term traffic condition prediction of urban road network based on improved SVM. H. Yan and D. Yu, 2017 Int’l Smart Cities Conf. (ISC2), Wuxi, 2017, pp. 1-2 doi: 10.1109/ISC2.2017.8090856

[52] Short-Term Traffic Flow Prediction Considering Spatio-Temporal Correlation: A Hybrid Model Combing Type-2 Fuzzy C-Means and Artificial Neural Network. J. Tang, L. Li, Z. Hu and F. Liu, in IEEE Access, vol. 7, pp. 101009-101018, 2019 doi: 10.1109/ACCESS.2019.2931920

[53] Prediction of Road Traffic Congestion Based on Random Forest Y. Liu and H. Wu, 2017 10th Int’l Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2017, pp. 361-364, doi: 10.1109/ISCID.2017.216

[54] Road Traffic State Prediction with a Maximum Entropy Method. H. Dong, L. Jia, X. Sun, C. Li, Y. Qin and M. Guo, 2009 Fifth Int’l Joint Conf. on INC, IMS and IDC, Seoul, 2009, pp. 628-630 doi: 10.1109/NCM.2009.411

[55] https://www.geotab.com/blog/what-is-smart-mobility/

[56] Tjortjis C., “Mining Association Rules from Code (MARC) to Support Legacy Software Management”, Software Quality Journal (SQJ), pp. 1-13, 2020, DOI: 10.1007/s11219-019-09480-3.

Information About the Authors

Konstantinos Christantonis BSc, MSc, is a Research Assistant at the International Hellenic University. He works on several projects involving intelligent systems. The main areas of his interest are data mining, machine learning and natural language processing. Lately, he is involved in projects regarding Smart Cities. He is a production and management engineer with an MSc in Data Science.

Christos Tjortjis is an Assoc. Prof. in Knowledge Discovery & Software Eng. systems at the Int’l Hellenic University, School of Science & Technology. He is Director for the MSc in Data Science, the MSc in ICT systems and the MSc in Smart Cities & Communities. He holds a Deng from Patras, Computer Eng. & Informatics, a BSc from Democritus Law School, Greece, an MPhil in Computation from UMIST, and a PhD in Informatics from Univ. of Manchester (UoM), U.K. He was Lecturer at UMIST, Computation, and the Schools of Informatics and Computer Science, UoM. He published over 60 papers in int’l journals and conferences. He received over 830 citations (h-index 15). He leads the Data Mining and Analytics Research Group.

Mr. **Anastasios Manos** is the Public Sector Director of DOTSOFT. He is responsible for public sector strategy and project implementation and coordination in many ICT integration projects in Greece and Cyprus. He started his career from Web Application Development, and he worked as an assistant project manager and later as a project manager in SW development in numerous successful projects throughout the years. He holds a Bachelor in Informatics from the Aristotle University of Thessaloniki, an MBA from University of Macedonia and an MSc in Computer Science from the same university.

Dr. **Despina Elisabeth Filippidou** is a Senior ICT Programme Management Officer. She works as an independent consultant and entrepreneur specialising in managing complex ICT R&D projects Her late aspirations involve IoT (Internet of Things), AR and VR (Augmented and Virtual reality), Gamification as well as Artificial Intelligence. She is a Information Systems and Informatics Engineer with a Master’s M.Sc. and Phd.D in Requirements Analysis Engineering. Her experience spans for more than 25 years, having been appointed to high-end positions for both private and public bodies and agencies.

Evangelos Christelis holds a BSc in Computer Science (AUTH) and Master in Knowledge, Data and Software Technologies (AUTH). He has significant experience in projects related to the development and customization of software applications and the development of mobile applications for android and iOS.

The paper was received on 10/01/2020