Applicability of Big Data Techniques to Smart Cities Deployments

M. Victoria Moreno, Fernando Terroso-Sáenz, Aurora González-Vidal, Mercedes Valdés-Vela
Antonio F. Skarmeta, Miguel A. Zamora and Victor Chang

***Abstract*—This paper presents the main foundations of Big Data applied to Smart Cities. A general Internet of Things based architecture is proposed to be applied to different smart cities applications. We describe two scenarios of big data analysis. One of them illustrates some services implemented in the smart campus of the University of Murcia. The second one is focused on a tram service scenario where thousands of transit-card transactions should be processed. Results obtained from both scenarios show the potential of the applicability of this kind of techniques to provide profitable services of smart cities, such as the management of the energy consumption and comfort in smart buildings, and the detection of travel profiles in smart transport.**

***Index Terms*—Internet of Things; Smart City; Big Data; Predictive Models; Transit-card Mining**

I. Introduction

A Smart City emerges when the urban infrastructure is evolved through the Information and Communication Technologies (ICT) [1]. The paradigm of Internet of Things (IoT) [2] has enabled the emergence of a high number of different communication protocols, which can be used to communicate with commercial devices using different data representations. In this context, it is necessary an IoT-based platform to manage all interoperability aspects and enable the integration of optimal Artificial Intelligence (AI) techniques in order to model contextual relationships.

In urban environments there is a huge amount of different data sources. Plenty of sensors are distributed around cities, most of them installed in indoor spaces. This situation has brought new analytics mechanisms and tools that provide insight allowing us to have an effective and collaborative way to operate the machines [3]. Furthermore, there are numerous mobile data sources like smart phones, smart-cards, wearable sensors and, in the case of vehicles, on-board sensors. All these sensors provide information that makes possible to detect urban dynamic patterns. Nonetheless, most existing management systems of cities are not able to utilize fully and effectively this vast amount of data and, as a result, there is large volumes of data which is not exploited. In this direction, many AI techniques in Computer Science have been introduced to deal with the processing of huge amount of data to extract useful information (or termed by knowledge) from data [4], this trend is known as Big Data.

This paper is intended to analyze the interest of big data for smart cities. In order to face the above-mentioned aspects we propose a general architecture for smart city applications, which is modelled in four layers with different functionalities. Then, we show some applications of big data analysis in two scenarios, both dealing with sensed data coming from both static and dynamic sources. Among other objectives, the first scenario intends to create a distributed framework to share large volumes of heterogeneous information for their use in smart building applications. In this work we focus on presenting the deployments and implementations carried out in smart buildings to achieve energy efficiency. For this, different problems like indoor localization, thermal comfort characterization and energy consumption modelling have been solved through the application of big data techniques. The second example is centered on the public tram service in the City of Murcia (Spain), looking for giving insight into the great amount of data generated by the service's transit cards. In this scenario, big data techniques are applied to extract mobility patterns in public transport.

Hence, this paper faces up three aspects of nowadays smart cities which need to be solved, and for each one of them we provide some research contributions through the application of convenient big data techniques. These contributions are:

- The design and instantiation of an IoT-based architecture for applications of smart cities.
- The approach of an efficient management of energy in smart buildings.
- The extension of the data analysis for detection of urban patterns which can be used to improve public transport applied to the public tram service.

The structure of this paper is as follows: Section II enumerates the challenges that current smart cities still have to face, and proposes a general IoT-based architecture for smart-city services which is modeled in layers. Section III describes a first application of smart city where big data techniques have been applied to get energy efficiency in the buildings of a Smart Campus. Section IV presents a second smart city application that addresses the urban pattern recognitions in public transport. Section V summarizes the main benefits of applying big data techniques to the two scenarios of smart cities addressed in this paper. Finally, Section VI gives some conclusions and an outlook of future work.

M. Victoria Moreno, Fernando Terroso-Sáenz, Aurora González-Vidal, Mercedes Valdés-Vela, Antonio F. Skarmeta and Miguel A. Zamora are with the Department of Information and Communications Engineering, University of Murcia, 30100 Spain, e-mail: (mvmoreno, fterroso, auroragonzalez2, mdvaldes, skarmeta, mzamora)@um.es.

Victor Chang is with the Xi'an Jiaotong Liverpool University, China, e-mail: ic.victor.chang@gmail.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2605581, IEEE Transactions on Industrial Informatics

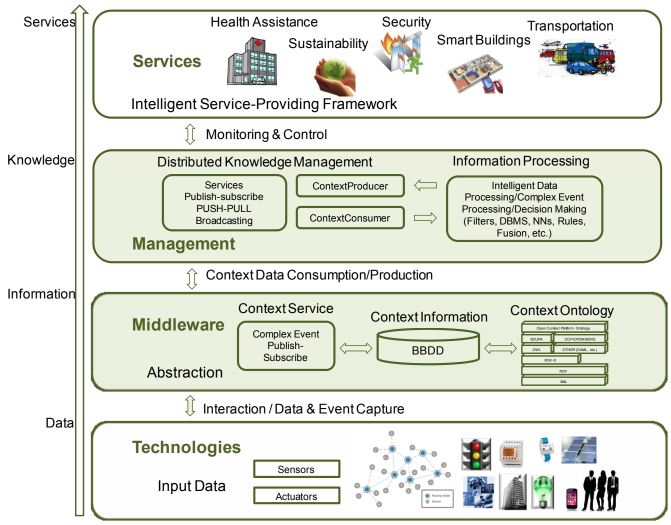


Figure 1: Layers of the base architecture for smart city services

II. IoT-based Architecture for Smart Cities

In this section we enumerate the main challenges that most current smart cities still have to face. Then, motivated by these challenges, we make a proposal of a general IoT-based architecture for smart city applications.

A. Challenges of Smart Cities

The global challenges that smart cities still have to face can be summarized in the following way:

- Sensors integration and abstraction capability. Provide means to integrate different sensor types in a common platform taking into account the different technologies, legacy systems and communication protocols with focus on IPv6 solutions.
- Individual intelligence and local reasoning. Apart from data fusion, more complex data processing can be implemented by smart objects.
- Learning and adaptation. Most of the patterns generated in smart cities are sensitive to contextual changes and are able to learn and adapt themselves to such changes as well as to human dynamicity.
- Dynamic human centric services. This work designs and implements smart mobility and smart building services that use the patterns generated to provide customized and efficient services taking into account the dynamicity of the citizens' behavior.
- User privacy and security control mechanisms. In the context of smart cities it is important to manage the way the user is able to control its data and how they are exposed to third parties and applications.

B. IoT-based Architecture

Several layers compound the proposed platform that was created with the goal of serving to many applications of smart cities. In Figure 1 is depicted this layered IoT-based architecture, which are detailed below.

1) Technologies Layer: In the basis part of Figure 1 it is observed that a plethora of sensors and network technologies provide the input attributes using wireless sensor networks, wired sensors, gateways, etc. which can be self-configured and remotely controlled through the Internet. Dealing with our first application that consists on the instantiation of the architecture for building management systems (BMS), in this layer it is gathered information from sensors and actuators deployed in strategic points of the building. But the aforementioned data sources in smart cities are not limited to static devices reporting measurements associated to a particular location, there are also moving ones capable to deliver measurements at different points within a geographical area. This is mainly due to the rapid development of wireless technology, mobile sensor networks and, above all, the advent of smartphones [5]. Although approaches based on mobile-phone sensing require a demanding usage of the communication, location and other attributes of the smartphone, which can bother some people due to battery draining [6], data captured by static, mobile and smart-phone sensors can be extended or enriched with the data generated by several social-media channels - like Twitter or Facebook - giving rise to a new generation of *soft sensors* from which extract relevant knowledge [7]. As a result, an alternative course of action aims at mining relevant knowledge from users on the basis of non-intrusive ways to obtain data, for example, transit cards in public transport scope.

2) Middleware Layer: The first layer provides us with a wide variety of data, so it is needed a second layer where all collected data from seamless sources are expressed in the same way, this is done in the middleware layer. The context information can be collected in an ontology defined according to the model that represents the knowledge of the specific application domain. Thus, for our energy efficiency semantic model, the devices and building concepts are borrowed by the SAREF ontology [8]. The agents representation is made using the DUL ontology [9], while the observation values of the monitored sensors are represented based on the SSN ontology [10]. However, when it comes to process the incoming data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2605581, IEEE Transactions on Industrial Informatics

in a real-time manner, it is necessary to use a lightweight representation. As a matter of fact, [11] describes sensor-data representation using a simple attribute-value schema for event-based systems.

3) Management Layer: After having extracted information from the previous layers, the management layer is in charge of determining decisions bearing in mind the target services provided in smart cities. Different big data analytic techniques can be used for the intelligent decision making processes. Algorithms like Artificial Neural Networks (ANNs) using backpropagation methods [12] and Support Vector Machines (SVMs) [13] are good to solve non-linear problems, making them very applicable to build energy prediction issues, ranging from those associated to lighting and heating, ventilation and air condition (HVAC) [14] to the prediction of the heating energy requirements [15]. For optimization problems in Building Management System (BMS) addressing energy efficiency, Genetic Algorithms (GAs) constitute a commonly applied heuristic that can be used in several optimization scenarios such as scheduling cooling operation decisions [16]. Regarding to the smart public transport application, the extraction of users behaviors from transition records have been studied by using different algorithms and techniques like maximum likelihood estimation [17], probabilistic models [18], conditional random fields [19], graphical information system (GIS)-based processing [20] or Database Management System (DBMS)-based processing [21].

4) Services Layer: Finally, the upper layer (Figure 1) shows some examples of smart city services that can be provided following the proposed architecture. Thus, this architecture can be applied to provide applications of smart cities like environmental monitoring, energy efficiency in buildings and public infrastructures [22], environmental monitoring [23], traffic information and public transport, locating citizens, manage emergencies, saving energy and other services. These actions can either involve citizens or be automatically set.

III. Smart Campus of the University of Murcia

The University of Murcia (UMU) has three main campus and several facilities deployed throughout different cities in the Region of Murcia. One of these campus is currently serving as pilot of two European Projects, the SMARTIE [24] and the ENTROPY [25] project. The goal of this use case of smart city is to provide a reference system able to manage intelligently the energy use of the most relevant contributor to the energy use at city level, i.e. buildings. The BMS implemented as part of this smart campus adapts the performance of automated devices through decisions made by the system and the interaction with occupants in order to keep comfort conditions while saving energy. We start by the most representative source of energy consumption at building level: HVAC systems.

A. System Overview

Using a BMS system, it is possible to predict users future behaviour from their recorded activities that are measured with sensors. This information allows us to provide convenient environments looking for keeping their comfort while saving energy. The first need for a building to become smart is to know location of occupants. Once solved the indoor localization problem, it is time to propose a solution to the energy efficiency of buildings associated to the thermal comfort provisioning service related to the HVAC management. For this, energy consumption models of the building need to be generated and used to implement the optimization mechanism able to maximize comfort at the same time that energy consumption is minimized. Therefore, the different problems addressed in this scenario of smart city through the application of big data techniques are:

1) Indoor localization estimation.
2) Building energy consumption prediction.
3) Comfort provisioning and energy saving through an optimization problem.

In the following subsections these problems are described with more details, as well as the techniques implemented and the results obtained.

B. Indoor Localization Estimation

As well as considering the information concerning to the identification and location of the building's occupants, it is necessary to reach the required accuracy in the location in order to provide the indoor services in a comfortable and energy efficient way. Our technological solution to cover the localization needs (i.e. those required by smart buildings to provide occupants with customized comfort services) is based on a single active RFID system and several Infra-Red (IR) transmitters. In Figure 2 we can observe the data exchange carried out among the different technological devices that compose our localization system.

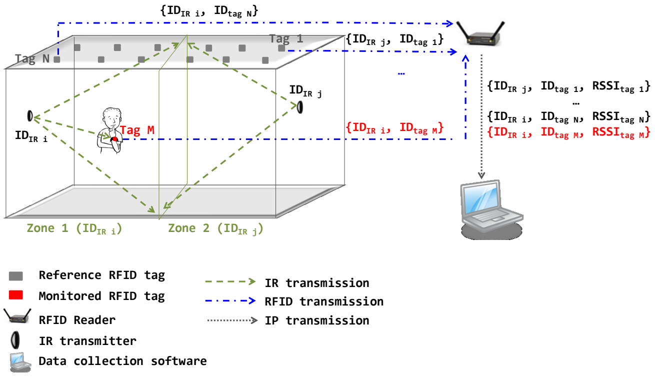

{$ID_{IR\ i}$, $ID_{tag\ N}$}
Tag N
Tag 1
{$ID_{IR\ j}$, $ID_{tag\ 1}$}
...
$ID_{IR\ j}$
$ID_{IR\ i}$
Tag M
{$ID_{IR\ i}$, $ID_{tag\ M}$}
{$ID_{IR\ j}$, $ID_{tag\ 1}$, $RSSI_{tag\ 1}$}
...
{$ID_{IR\ i}$, $ID_{tag\ N}$, $RSSI_{tag\ N}$}
{$ID_{IR\ i}$, $ID_{tag\ M}$, $RSSI_{tag\ M}$}
Zone 1 ($ID_{IR\ i}$)
Zone 2 ($ID_{IR\ j}$)
Reference RFID tag
Monitored RFID tag
RFID Reader
IR transmitter
Data collection software
IR transmission
RFID transmission
IP transmission

Figure 2: Localization scenario

The final mechanism implemented to solve the indoor localization problem is shown in Figure 3. In this figure, we can see that the first phase of the mechanism is the space division through the installation of IR devices in the walls of the building area where localization wants to be solved. Therefore, for each space division, there is an IR identifier value (ID_{ir}) associated to this region. For each one of these regions, we implement a regression method based on Radial Basis Functions (RBF) networks. The RBF estimates user positions given different RFID tags situated in the roof. Then, after the position estimation using the RBF network, a Particle Filter (PF) is applied as a monitoring technique, which takes into account previous user position data for estimating future states according to the current system model.

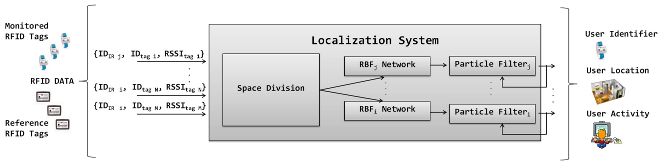


Figure 3: Data processing for location estimation

The PF used in this work is slightly different from its generic definition (which can be found in [26]). The main difference of our filter is in the correction stage. In this stage, the generic definition of the PF applies the resampling using the Sequential Importance Sampling (SIS) algorithm [26] to carry out the filtering of such particles which minimize the deviation of their predicted trajectory. In our implementation, in addition to apply the SIS algorithm to correct the particles positions, we also use in this step the information about the specific IR region at a given instant of time to benefit those particles which fall inside this area. Therefore, before applying the SSI algorithm, we filter according to the coverage area of the IR transmitter identified by the monitoring RFID tag. The main advantage of this constraint is the faster convergence of the filter, because extra information is available to carry out the correction stage of the filter.

C. Building Energy Consumption Prediction

The energy performance model of our BMS is based on the *CEN Standard EN 15251* [27]. This standard proposes the criteria of design for any BMS. It establishes and defines the main input parameters for estimating building energy requirements and evaluating the indoor environment conditions. The inputs considered to solve our problem are the data coming from the RFID cards of users, the user interaction with the building automation system through the control panels or the web access, environmental parameters coming from temperature, humidity and lighting sensors installed in outdoor and indoor spaces, the consumption energy sensed by the energy meters installed in the building, and the generated energy sensed by the energy meters installed in the solar panels deployed in our testbed. After collecting the data it is mandatory to continue with their cleaning, preprocessing, visualization and correlation calculation in order to find determining features, which can be used to generate optimal energy consumption models of buildings (management layer of the architecture presented in Section II). Over the input set, we perform the standardization and reduction of data dimensionality using Principal Components Analysis (PCA) [28], identifying the directions in which the observations of each parameter mostly vary.

Regarding the big data techniques that have been already applied successfully to generate energy consumption models of buildings in different scenarios (as such we mentioned in the management layer of the architecture presented in Section II-B3), we propose to evaluate the performance of Multilayer Perceptron (MLP), Bayesian Regularized Neural Network (BRNN) [29], SVM [30] and Gaussian Processes with RBF Kernel [31]. They were selected because of the good performance that all of them have already provided when they are applied to building modelling. All these regression techniques are implemented following a model-free approach, which is based on selecting - for a specific building - the optimal input set and technique, i.e. such input set and technique that provides the most accurate predictive results in a test dataset. In order to implement this free-model approach, we use the R [32] package named CARET [33] to train the energy consumption predictive algorithms, looking for the optimal configuration of their hyper-parameters (more information can be found in [34]). The selected metric to evaluate the models generated for each technique using test sets is the well-known RMSE (Root-Mean-Square Error), whose formulation appears in Eq. (1). This metric shows the error by means of the quantity of KWh that we deviate when predicting.

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

But in order to get a better understanding of the uncertainty of the model, we also show its coefficient of variation (CVRMSE). This coefficient is the RMSE divided by the mean of the output variable (energy consumption) for the test set (see Eq. (2)), giving us a percentage of error adjusted to the data, not just a number in general terms.

$$CVRMSE = \frac{RMSE}{\bar{y}} \tag{2}$$

D. Optimization Problem

Once the building energy consumption is modelled, we focus on the optimization of the HVAC operation trying to keep comfort conditions at the same time that energy consumption restrictions are considered. As starting point, we establish the comfort extremes considering location type, user activity and date [35]. Understanding the building thermal and energetic profiles allows us to quantify the effects of particular heating/cooling set point decisions. To derive a heating or cooling schedule, it is necessary to formulate the target outcome. In our buildings, it is possible to:

1) optimize the indoor temperature during occupation, i.e. minimize the building temperature deviations from a target temperature,
2) minimize daily energy consumption, or
3) optimize a weighted mixture of the criteria, a so-called multi-objective optimization problem.

The definition of building temperature deviation influences the results strongly: taking the minimum building temperature will result in higher set point choices and higher energy use than using, for instance, the average of indoor temperatures. Constraints on maximum acceptable deviation from target comfort levels or an energy budget can be taken into account to ensure required performance. In our optimization problem, we apply GA using the implementation provided by R (the "genalg" package [36]), to provide schedules for heating/cooling setpoints using the predictive building models (comfort and energy consumption models).

E. Evaluation and Results

1) Scenario of Experimentation: The reference building where our BMS for energy efficiency is deployed is the Technology Transfer Centre (TTC) of the UMU[1]. Every room of this building is automated through a Home Automation Module (HAM) unit. It permits us to consider a granularity at room level to carry out the experiments.

2) Evaluation. Indoor localization mechanism: Different tracking processes are carried out in the environments considered in our tests (the TTC building), applying for this the implementation of our PF. In Figure 4 an example of some tracking processes are carried out considering transition between different spaces of the TTC. For these paths, our system was configured to acquire data every $T = 10$ s. Taking into account the target location areas involved in comfort provisioning (lighting and thermal comfort, represented in different colors), and the real and estimated location data provided by our mechanism.

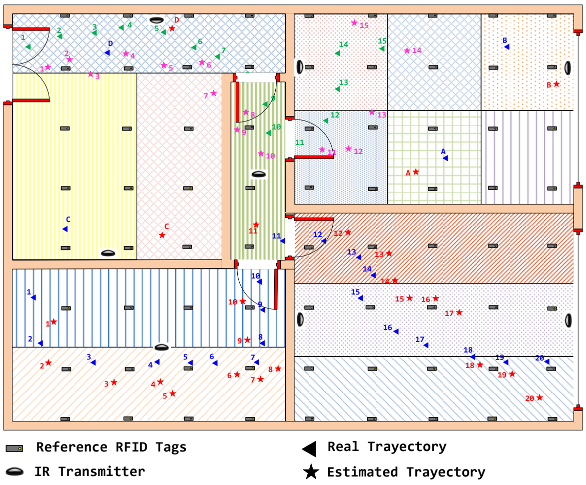
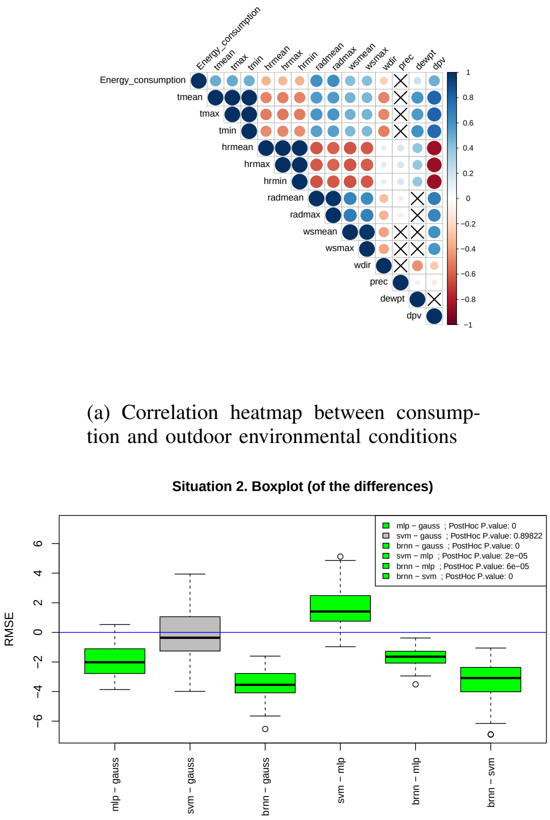


Figure 4: Tracking processes with a reference tag distribution of 1m x 1m

Thus, with a $1m$ x $1m$ distribution of reference RFID tags placed on the roof of the test room, a 65% success percentage in localization is obtained having an error lower than 1m. 98% of cases have as much 2.5m. of error. Therefore, it can be safely said that our localization system is able to track users with a sufficient level of accuracy and precision for the location requirements associated with the comfort and energy management in buildings. More details about this indoor localization system can be found in [37].

3) Evaluation. Energy consumption prediction: In Figure 5(a) it is shown the correlation heatmap between the electrical consumption of the TTC building and the outdoor environmental conditions. It is observed that energy consumption correlates significantly ($\alpha = 0.95$) and positively with temperature, radiation, wind speed variables, vapour pressure deficit and dew point; and negatively with wind direction and humidity variables. This means that we can use safely these variables as inputs of the energy consumption model of our reference building, because they have clear impact in the energy consumption. Otherwise, precipitation is so unusual that they don't have an association with the output.



(a) Correlation heatmap between consumption and outdoor environmental conditions

(b) Boxplots comparing models pairwise (situation 2)

Figure 5: Modeling results

Also, a logic differentiation between temporal situations has been considered in order to label behaviour. Situation 1: holidays and weekends; situation 2: regular mornings; and, situation 3: regular afternoons. The non-parametric Kruskall Wallis test shows that energy consumption differs significantly between situations (H(2) = 547.7, p < 0.01). Also, the post hoc pairwise comparisons corrected with Holm's method retrieve a p-value smaller than 0.01, supporting the decision of creating 3 different models [38]. Thus, for each of the three situations identified for the TTC building, we have evaluated not only the punctual value of RMSE, but also we have validated whether one learning algorithm out-performs statistically significantly the others using the non parametric Friedman test [39] with the corresponding post-hoc tests for comparison.

Let x_i^j be the i-th performance RMSE of the j-th algorithm, for this building, we have used 5-times 10-fold cross validation, so $i \in \{1, 2, ..., 50\}$ and four techniques, so $j \in \{1, 2, 3, 4\}$. For every situation, we find significant differences ($\alpha = 0.99$) between every pair of algorithms,

[1] www.um.es/otri/?opc=cttfuentealamo

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2016.2605581, IEEE Transactions on Industrial Informatics

except for SVM and Gauss RBF ($p > 0.01$), as it is shown in Figure 5(b) for the particular case of situation 2.

The three models have in common that BRNN yields a better result than the other tested techniques, based on the RMSE metric. Thus, BRNN is able to generate a model with a very low mean error of 25.17 KWh - which only represents the 7.55% of the sample (this is the most accurate result) in terms of the CVRMSE. And for the worst case, BRNN provides a mean error of 43.76 KWh - which represents the 10.29% of the sample in the reference TTC building - that is acceptable enough considering that our final aim is to save energy.

4) Evaluation. Optimization mechanism: To evaluate our GA-based optimization strategy, controlled experiments were carried out in the TTC building with different occupant's behaviours. The results show that we can accomplish energy savings between the 15% and 31%. Trying to validate the applicability of our proposal, we have also made experiments in a different scenario with limited monitoring and automation technologies, achieving energy saving of about the 23%.

IV. Public Tram Service of Murcia City

The second scenario is focused on the information analysis related to use of the tram service of the Region of Murcia [40]. In this case, the main goal was to perform a profiling process of the trips carried out by the users of such public service. For that aim, a fuzzy clustering algorithm is used to automatically extract tram user's profiles. Bearing in mind the architecture introduced in Section II, this system is enclosed in the management layer. The main tasks needed to reach the goal are explained in the following subsections.

A. Generation of the trip data set

According to the tram experts, information relevant to trip profiling must include data about: time (in terms of day of the week and time of the day), origin and destination stations and approximate age of the traveller. This information is being continuously recorded in different databases of the tram service. Nevertheless, certain operations of joining, transformation and preprocessing (discretization and numerization) have been performed in order to compile all this information into a set of tuples susceptible of feeding the subsequent fuzzy clustering algorithm. The two most remarkable operations are the following:

On the one hand, according to the infrastructure of the tram service, users only need to swipe the smart card when they get into the tram. Hence, the recorded data only comprises transactions at the origin of each user's trip so it can be regarded as incomplete. In order to deal with this incompleteness, a well known solution is the **trip-chaining method** which focus on recovering the origin and destination of the trips. In this case, such a method is based on the assumption that a traveller who takes the tram at an origin station, OS, ended their previous trip on that station OS. Due to the event-based nature of the card records, the Complex Event Processing (CEP) paradigm [11] was adopted to come up with a palette of event-condition-action rules to uncover the trips. While the condition part of the rules performs a match between consecutive records of the same traveller following the aforementioned trip-chaining method, the action part generates a new trip tuple (comprising the origin and destination stations) in case the condition is fulfilled.

On the other hand, as clustering techniques are based on distance calculations among data, a set of numbered (and ordered) geographical areas, each one covering some close stations are identified by the tram experts. Then, instead of having nominal values for origin and destination features these numbered areas make it easier to calculate the distance about tuples in the clustering process.

In summary, the tuples composing the data set for the subsequent clustering task are composed by the following attributes: tt_e*:{travellerAge, dayOfTheWeek, hourOfTheDay, originArea, destArea}*

B. Trip profiling

Clustering mechanisms are suitable when it comes to find out the most representative trips profiles. For that aim, the Gustafson-Kessel Clustering Method (GKCM) has been chosen since it is able to identify arbitrarily oriented ellipsoidal fuzzy clusters unlike, for instance, the Fuzzy C Means clustering Method, which impose spherical shapes to the data clusters. After the clustering task the identified prototypes (centroids) will summarize the whole data set of trips. GKCM requires to be supplied with the quantity of potential clusters (c). This is an important parameter since it determines the ability of the potential centroids to represent the real underlying structure of the data.

Therefore, several GKCM executions were performed with different values of c and the *goodness* of the different identified set of clusters was measured. One of the most used measurement is the one proposed in [41] and denoted here as r_{cs}. This magnitude quantifying both the total compactness within clusters and the total separation among them being the greater the better.

Once the number of clusters c has been decided on the basis of r_{cs}, GKCM is executed in order to find the c profiles that best represent the trip data set. Nevertheless, when exceed a time tp_{max} or a number of trips nt_{max} the algorithm is re-computed in order to detect new profiles which could rise up.

C. Evaluation and Results

The subject of evaluation is the tram service of the region of Murcia (Spain), which includes 18-km railway and 28 stations (see Figure 6). Figure 7 depicts the set of predefined geographical areas used in the experiment.

The evaluated dataset contained 378719 trips from 23400 users in November, 2013. For our experiment, the system was able to uncover 110697 trips. Expert knowledge was used to define the types of days and times of the day used in the aforementioned data pre-processing step as [*Monday-Thursday, Friday, Saturday, Sunday*] and [*0-6, 6-10, 10-12, 12-16, 16-20, 20-00*]. As a result, a generated TT_e dataset was split up into 4 different subsets based on the fact that traveller profiles depend on the day of the week (regarding, for example, differences of traffic flow between regular workdays

Algorithm 1: Cluster-based Trip profiling process.

```
Input: $TT$: dataset of raw trip tuples.
Output: $P_{TT}$: Traveller profiles extracted from $TT$.
if $t_{now} - t_{prev} > tp_{max} \vee \mid TT \mid - \mid TT_{prev} \mid > nt_{max}$ then
    $TT_e \leftarrow$ preProcessing($TT$)
    foreach $c \in \{2, .., c_{max}\}$ do
        $clust_c$ = GKCM($TT_e$, c)
        if $clust_c.r_{cs} < r_{cs}^{min}$ then
            $r_{cs}^{min} \leftarrow clust_c.r_{cs}$
            $P_{TT} \leftarrow clust_c$.centroids
    $t_{prev} \leftarrow t_{now}$
    $TT_{prev} \leftarrow TT$
    return $P_{TT}$
```

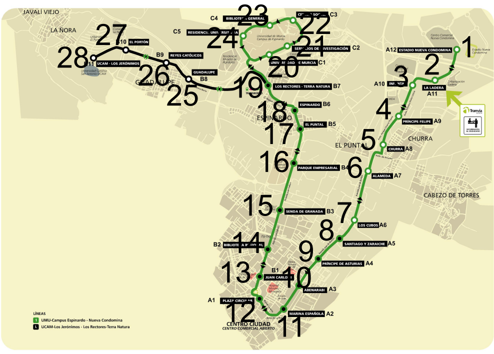


Figure 6: Line map of the tram service in Murcia.

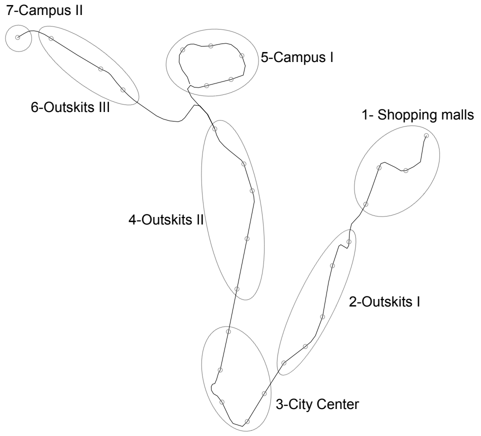


Figure 7: Geographical regions for the numerization of tuples' station fields.

and weekends). Next, the GKCM was launched with each of these subsets with different number of clusters.

In Figure 8, the cluster validation ratio r_{cs} is shown for every TT_e subset, being the lower value the better. As it can be observed, while the optimal cluster partition is reached at $c = 5$ for the Monday-Thursday subset, for the remaining subsets minima r_{cs} values are reached at higher number of clusters c. In other words, a higher number of profiles is needed to represent the weekend trips. This is reasonable given that most people postpone leisure activities to the weekend and given that there exist a quite variety of leisure activities that can be done at different hours of the day.

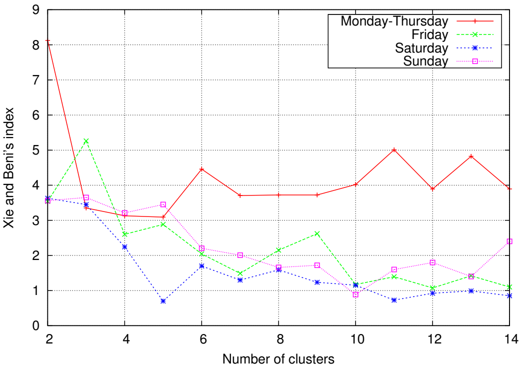


Figure 8: Cluster-validation rate for different cluster partitions.

As Table 1 shows, GKCM extracts five profiles for Monday-Thursday trips. Profiles 1 and 2 correspond to young people travelling in the morning to go towards one of the university zones from the station close the inner city. Besides, profile 5 represents a kind on traveller going back home from the university from 4 to 8 PM. Finally, profiles 3 and 4 correspond to middle-young age people (28-33 years) that take the tram around the outskirts and city center environments. These could reflect people going from residential areas.

Lastly, the heatmap shown in Figure 9 represents the membership of the Monday-Thursday trips to the defined profiles. If we interpret this plot as a time-framed sequence, a great amount of the traffic focuses on the right side of the line, which connects the city center and the university areas. Nevertheless, such load is more spread along the whole line during the evening.

V. Discussion

In this paper we propose a general IoT-based architecture which can be implemented for different applications of smart cities. This architecture is modeled in four layers, being the third one - the management layer - the layer where big data techniques are implemented to provide the different services offered then in the corresponding service layer (last layer).

The big data paradigm can be understood through the lens of 7 V's [42] (challenges). Regarding the application of different big data techniques to the specific scenarios of smart cities presented in this paper, we have overcame the challenge of *velocity* by collecting data hourly in the smart building application (consumption of energy, outdoor environmental conditions) and even in sorter intervals of time for the public transport application (many people validates their transit cards within seconds). Although we haven't tackled *volatility*, it is clearly a goal when looking for the real-time smart city because behavioural scenarios like ours change depending on many social aspects. The *veracity* of the data is guaranteed by the exhaustive pre-processing steps included in the modeling process. We have extracted *value*, making sense of the wide mentioned *variety* of data, and with the described analysis

Profile	Age	Origin Area	Dest. Area	Time of the day
P1	23.37	*City Center*	*Campus I*	0-6
P2	25.74	*City Center*	*Campus I*	6-10
P3	28.22	*Outskits II*	*City Center*	12-16
P4	32.77	*Outskits I*	*Outskits II*	6-10
P5	22.20	*Campus I*	*City Center*	16-20

Table 1: Monday-Thursday trips' profiles.

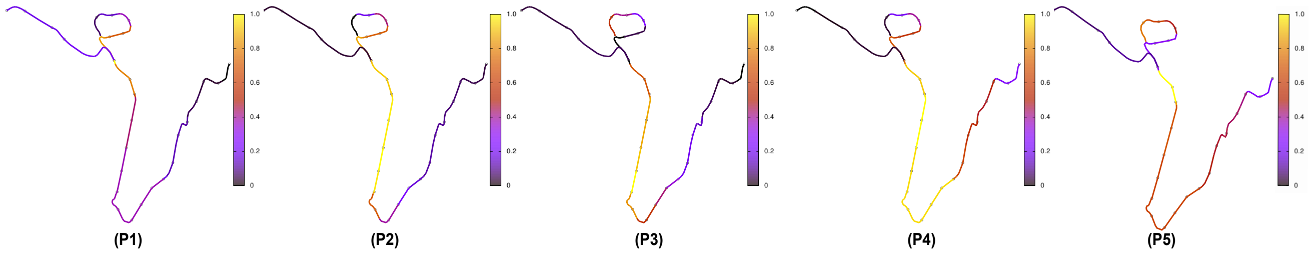


Figure 9: Tram-line heat-map of the five profiles for Monday-Thursday trips.

Smart City Application	Data	Information	Knowledge	Services
Smart Campus	IR Sensors. RFID tags. Environmental Sensors. Weather Station. Presence Sensors. Energy Consumption Meters. Weather Forecast	Data Transformation through SAREF ontology [8], DUL ontology [9] and SSN ontology [10]	Data Modelling. Predictive Regression (RBFs, SVM, ANN, RF, ARIMA). Tracking algorithm (PFs). Optimization Mechanism (GA)	Indoor localization. Building energy consumption prediction. Energy saving through the HVAC operation optimization
Public Tram Service	Mobile Sensors. Smart Cards	CEP-based filtering. Event Processing in Action [11]	Fuzzy Clustering	Infrastructure monitoring. Mobility patterns.

Table 2: Main features of the two architecture instantiations

and techniques, we have *validated* their usability for solving different problems of smart cities with high accuracy.

In both applications tackled in this paper, the huge *volume* of historical data is being stored using a NoSQL data base. At the moment, the storage system is been adapted so as to be compliant with the FI-WARE architecture[2], that intends to ease the development of novel applications based on the Future Internet. In particular, the Orion Context Broker (OCB)[3] and the COMET[4] modules are used in order to store in a NoSQL repository the historical data comprising the measurements from the different data sources.

On the whole, both instantiations of the architecture described above are summarized in Table 2. In the next subsections we summarize the main benefits obtained after applying the most suitable big data techniques to the two scenarios of smart cities addressed in this work.

A. Benefits of Big Data Applications in Smart Buildings for Energy Efficiency

Here we summarize the main findings extracted from all the experiments and analysis carried out during the application of big data techniques to the smart campus of the UMU.

1) **The resolution of the indoor localization problem.** Applying regression techniques based on RBFs and a tracking algorithm applying PFs to data coming from RFID and IR sensors installed in buildings, it was possible to solve the indoor localization problem with a mean accuracy of 1.5 m. Then, indoor localization data can be used to provide customized services in buildings.
2) **The resolution of the building energy consumption estimation.** Applying PCA and BRNN techniques to data related to outdoor environmental conditions and energy consumption of buildings, it was possible to generate energy consumption predictive models of buildings with a very low mean error of 43.76 KWh - which only represents the 10.29 % of the sample - in the worst case. Then, energy consumption predictions can be used to design the optimal strategies to save energy in buildings.
3) **The resolution of the optimization problem related to the maximization of thermal comfort and minimization of energy consumption in buildings.** Applying optimization methods based on GAs to optimize the energy consumption of buildings meanwhile comfort conditions are satisfied, and after including user localization data and user comfort preference prediction, it was possible to get energy savings in heating of about 23% compared with the energy consumption in a previous month without any energy BMS.

[2] https://www.fiware.org [Available Feb. 2016]

[3] http://catalogue.fiware.org/enablers/publishsubscribe-context-broker-orion-context-broker [Available Feb. 2016]

[4] https://github.com/telefonicaid/fiware-sth-comet. [Available Feb. 2016]

B. Benefits of Big Data Applications in Urban Pattern Recognition to Improve Public Tram Service

After applying Big Data techniques to the urban pattern extraction in the public tram service, all the results from the experiments allowed the service staff to draw up quite interesting conclusions. These are summarized below:

1) **Regarding the resolution of the trip extraction**. The formal discovery of the stations' load in terms of trips' origin and destination would allow the service provider and the city council to better plan the whole public transport service of the city. This way, the more important stations might be considered as "hub" points where commuters can easily transfer from tram to another kinds of transport. Moreover, such an information could be also useful so as to forecast future infrastructure needs in each part of the tram line (e.g. location and number of places of new parking lots for bicycles close to tram stations).
2) **Concerning the resolution of the urban profiles generation**. Experiments pointed out the importance of undergraduates as tram users. Hence, most of the traffic load concentrated in the line between the city center and the campuses. This was really helpful in order to design promotional campaigns for these type of travellers. Moreover, results also confirmed that the line segment towards the shopping-mall areas was underused. Thus, campaigns to promote the use of the tram to go shopping was also considered.

VI. Conclusions and Future Work

This paper displays the benefits of applying big data techniques over data originated by IoT-based devices deployed in smart cities. A general architecture modelled in four layers is proposed to be applied in smart city applications considering big data issues. As part of this overview, a differentiation between static and mobile data sources is made, proposing for each one of them suitable techniques to extract relevant knowledge from their data. Then, we describe two big data applications for smart city services. Specifically, the services of energy efficiency and comfort management in the buildings of a smart campus, and the public transport service of a city. In the first scenario of smart city we have demonstrated that, after applying appropriate big data techniques to problems like indoor localization, energy consumption modeling and optimization, we are able to provide mean energy savings of 23% per month, while indoor comfort is ensured. Regarding to the urban pattern recognition carried out using data related to the public tram service of the city of Murcia, experiments were performed to confirm that the proposed patterns ended up being of great interest for the service provider in order to better understand how travellers make use of the transportation system. This was fairly useful in order to come up with better planning protocols and more tempting promotional campaigns.

The ongoing work is focused on the inclusion of people behaviour during the operational loop of this kind of systems for smart cities. Thus, for the case of smart building applications, users will be encouraged to participate in an active way through their engagement to save energy. On the other hand, in the case of the public tram service, data coming from crowdsensing initiatives will be integrated to improve the estimation of the urban mobility patterns.

Acknowledgments

This work has been funded by the Spanish Seneca Foundation by means of the PD program (grant 19782/PD/15), the MINECO TIN2014-52099-R project (grant BES-2015-071956) and ERDF funds, and by the European Commission through the H2020-ENTROPY-649849 and FP7-SMARTIE-609062 EU Projects.

References

[1] N. Komninos, "Intelligent cities: variable geometries of spatial intelligence," *Intelligent Buildings International*, vol. 3, no. 3, pp. 172–188, 2011.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] L. Da Xu, W. He, and S. Li, "Internet of things in industries: a survey," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 4, pp. 2233–2243, 2014.

[4] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big data analytics: Computational intelligence techniques and application areas," *Int. J. Inf. Manage*, pp. 10–15, 2016.

[5] Z. Yan and D. Chakraborty, "Semantics in mobile sensing," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 4, no. 1, pp. 1–143, 2014.

[6] A. Carroll and G. Heiser, "An analysis of power consumption in a smartphone." in *USENIX annual technical conference*, 2010, pp. 1–14.

[7] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[8] L. Daniele, F. den Hartog, and J. Roes, "Created in close interaction with the industry: The smart appliances reference (saref) ontology," in *Formal Ontologies Meet Industry*. Springer, 2015, pp. 100–112.

[9] K. Janowicz and M. Compton, "The stimulus-sensor-observation ontology design pattern and its integration into the semantic sensor network ontology," in *Proceedings of the 3rd International Conference on Semantic Sensor Networks-Volume 668*. CEUR-WS. org, 2010, pp. 64–78.

[10] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog *et al.*, "The ssn ontology of the w3c semantic sensor network incubator group," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 17, pp. 25–32, 2012.

[11] O. Etzion and P. Niblett, *Event Processing in Action*, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2010.

[12] T. Maniak, C. Jayne, R. Iqbal, and F. Doctor, "Automated intelligent system for sound signalling device quality assurance," *Information Sciences*, vol. 294, pp. 600–611, 2015.

[13] A. H. Neto and F. A. S. Fiorelli, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," *Energy and Buildings*, vol. 40, no. 12, pp. 2169–2176, 2008.

[14] H.-x. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586–3592, 2012.

[15] B. B. Ekici and U. T. Aksoy, "Prediction of building energy consumption by using artificial neural networks," *Advances in Engineering Software*, vol. 40, no. 5, pp. 356–362, 2009.

[16] F. Ascione, N. Bianco, C. De Stasio, G. M. Mauro, and G. P. Vanoli, "Simulation-based model predictive control by the multi-objective optimization of building energy performance and thermal comfort," *Energy and Buildings*, vol. 111, pp. 131–144, 2016.

[17] W. Wang, J. P. Attanucci, and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, p. 7, 2011.

[18] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[19] N. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun, "Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach," in *2013 IEEE 13th International Conference on Data Mining (ICDM)*, Dec. 2013, pp. 877–886.

[20] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[21] J. P. Attanucci and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, p. 131, 2011.

[22] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *Industrial Informatics, IEEE Transactions on*, vol. 7, no. 3, pp. 381–388, 2011.

[23] S. Fang, L. Da Xu, Y. Zhu, J. Ahati, H. Pei, J. Yan, and Z. Liu, "An integrated system for regional environmental monitoring and management based on internet of things," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 2, pp. 1596–1605, 2014.

[24] EU Smartie Consortium. (2013-2016) EU Smartie Project. [Online]. Available: http://www.smartie-project.eu

[25] EU Entropy Consortium. (2015-2018) EU Entropy Project. [Online]. Available: http://entropy-project.eu/

[26] A. Haug, "A tutorial on Bayesian estimation and tracking techniques applicable to nonlinear and non-Gaussian processes," *MITRE Corporation, McLean*, 2005.

[27] E. Standard *et al.*, "Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics," *EN Standard*, vol. 15251, 2007.

[28] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[29] L. Hawarah, S. Ploix, and M. Jacomino, "User behavior prediction in energy consumption in housing using bayesian networks," in *Artificial Intelligence and Soft Computing*. Springer, 2010, pp. 372–379.

[30] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using support vector machine to predict next day electricity load of public buildings with sub-metering devices," *Procedia Engineering*, vol. 121, pp. 1016–1022, 2015.

[31] M. Alamaniotis, D. Bargiotas, and L. H. Tsoukalas, "Towards smart energy systems: application of kernel machine regression for medium term electricity load forecasting," *SpringerPlus*, vol. 5, no. 1, pp. 1–15, 2016.

[32] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[33] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[34] A. González-Vidal, V. Moreno-Cano, F. Terroso-Sáenz, and A. F. Skarmeta, "Towards energy efficiency smart buildings models based on intelligent data analytics," *Procedia Computer Science*, vol. 83, pp. 994–999, 2016.

[35] J. A. Orosa, "A new modelling methodology to control hvac systems," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4505–4513, 2011.

[36] E. Willighagen, "Genalg: R based genetic algorithm," *R package version 0.1*, vol. 1, 2005.

[37] M. V. Moreno, M. Zamora-Izquierdo, J. Santa, and A. F. Skarmeta, "An indoor localization system based on artificial neural networks and particle filters applied to intelligent buildings," *Neurocomputing*, vol. 122, pp. 116–125, 2013.

[38] J. M. Andy Field and Z. F. Niblett, *Discovering Statistics Using R*, 1st ed. Sage Publications Ltd, 2012.

[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[40] [Online]. Available: www.tranviademurcia.es

[41] S. Miyamoto, H. Ichihashi, and K. Honda, "Algorithms for fuzzy clustering," *Methods in c-Means Clustering with Applications. Kacprzyk J, editor Berlin: Springer-Verlag*, 2008.

[42] M. Ali-ud-din Khan, M. F. Uddin, and N. Gupta, "Seven v's of big data understanding big data to extract value," in *American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the*. IEEE, 2014, pp. 1–5.

Dr. M. Victoria Moreno received the B.S. (Hons.) and the M.S. degrees in Telecommunications Engineering in 2006 and 2009, respectively, both of them from the School of Telecommunication Engineering of Cartagena, Spain; and the Ph.D degree in Computer Science in 2014 from the University of Murcia, Spain. Currently, she is a post-doctoral researcher of the Seneca Foundation. Research interests include data analysis and modelling, and energy efficiency in smart environments.

Dr. Fernando Terroso-Sáenz graduated from University of Murcia with a degree in Computer Science in 2006. He also received the master degree in Computer Science at the same university in 2010. Then, he finished his PhD at the Dept. of Communications and Information Engineering in 2013. Since 2009, he have been working as a researcher in this group. His research interests include Complex Event Processing (CEP), Ubiquitous Computing and Intelligent Transportation Systems (ITSs).

Aurora González-Vidal graduated in Mathematics at the University of Murcia in 2014. In 2015 she got a fellowship to work in the Statistical Division of the Research Support Services, where she specialized in Statistics and Data Analyisis. During 2015, she started her PhD studies in Computer Science, focusing her research in Data Analytics for Energy Efficiency.

Dr. Mercedes Valdés-Vela received the Computer Engineering degree (1998) and the Ph.D. degree (Hons.) in Computer Science (2003) at the University of Murcia (Spain). In 2000, she started to work as research staff in the Dept. of Information Engineering and Communications. She is currently a full time assistant professor of the same Department. Her main research areas are Soft Computing, Complex Event Processing and Ambient Intelligence.

Dr. Antonio F. Skarmeta received the M.S. degree in Computer Science from the University of Granada, Spain, and the B.S. (Hons.) and the PhD degrees in Computer Science from the University of Murcia. Currently, he is a Full Professor in Dept. of Information and Communications Engineering at the same university. He is involved in numerous projects, both European and National. Research interests include mobile communications, artificial intelligence and home automation.

Dr. Miguel A. Zamora received the M.S. degree in Automation and Electronics and the Ph.D. degree in Computer Science in 1997 and 2003, respectively, both of them from University of Murcia, Spain. Currently, he is a Senior Professor in the Dept. of Information and Communication Engineering of the same university. His research interests include consumer electronics, home and building automation and sensor fusion.

Dr. Victor Chang is an Associate Professor (Reader) at Xi'an Jiaotong Liverpool University, China since June 2016. Within four years, he completed PhD (CS, Southampton) and PGCert (Higher Education, Fellow) part-time. He helps organizations in achieving good Cloud design, deployment and services. He won a European Award on Cloud Migration in 2011, best papers in 2012 and 2015, and numerous awards since 2012. He is one of the most active practitioners and researchers in Cloud Computing, Big Data and IoT in UK.