DATA-DRIVEN MODELS FOR BUILDING ENERGY EFFICIENCY MONITORING

Joaquim Massana i Raurich

Universitat
de Girona

PhD Thesis

Data-driven models for building energy efficiency monitoring

Joaquim Massana Raurich

2017

Universitat de Girona

DOCTORAL THESIS

Data-driven models for building energy efficiency monitoring

Joaquim Massana Raurich

2017

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:
Dr. Carles Pous

Work submitted to the University of Girona in partial fulfilment of the requirements for the degree of Doctor of Philosophy

A ma família, car sense ells jo res no sóc

To my family, because without them I am nothing

Acknowledgments

La present tesi ha estat duta a terme gràcies a la supervisió del Dr. Carles Pous, al qual m'agradaria, no només agrair el seu suport i insistència, sinó també la seva paciència infinita i afecte.

A més, m'agradaria donar les gràcies al grup eXiT i sobretot als seus membres: Dr. Meléndez, Dr. Herraiz, Dr. Colomer, Dr. Torrent, Dr. Pla i especialment al futur doctor Dr. Gay. Tampoc em vull oblidar de la resta de companys del grup amb els quals hem compartit molts moments: Llorenç, Robert i Natàlia. Sens dubte mai ha faltat una mà on agafar-se ni un somriure reconfortant.

També agraeixo a SOTIM i al Departament de Física el subministrament de les dades per a dur a terme les presents publicacions.

Dono també les gràcies a la meva família: Esteve, Concepció i Anna, per encoratjar-me a cercar el meu camí a la vida, sense renúncies. Mai m'heu fallat i sempre us he tingut a prop. Gràcies de tot cor.

I would like to thank Colin from the Limerick University for helping me to spend a beautiful, but also useful, time in Ireland.

This research project has been partially funded through BR-UdG Scholarship of the University of Girona granted to Joaquim Massana Raurich. Work developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016) and the research group eXiT (Control Engineering and Intelligent Systems) of the IIiA (Institute of Informatics and Applications), the Department of Electrical and Electronic Engineering and Automation of the University of Girona, the MESC project (Ref. DPI2013-47450-C2-1-R) funded by the Spanish MINECO (Ref. DPI2013-47450-C2-1-R) and the European Union's Horizon2020 Research and Innovation Programme under grant agreement No 680708.

Publications

The presented thesis is a compendium of the following research articles:

- Hernandez, L., Baladron, C., Aguilar, J. M., Carro, B., Sanchez-Esguevillas, A. J., Lloret, J., & Massana, J. (2014). A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings. IEEE Communications Surveys & Tutorials, 16(3), 1460-1495. Quality index: [JCR ISCS IF 2.788 Q1(1/143)]

- Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. Energy and Buildings, 92, 322-330. Quality index: [JCR CE IF 2.973 Q1(6/126)]

- Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2016). Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. Energy and Buildings, 130, 519-531. Quality index: [JCR CE IF 2.973 Q1(6/126)]

- Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2017). Identifying services for short-term load forecasting using data driven models in a Smart City platform. Sustainable Cities and Society, 28, 108-117. Quality index: [JCR CBT IF 1.044 Q2(27/61)]

The rest of publications and conferences related with this PhD thesis are the following:

Journals

- Burgas, L., Melendez, J., Colomer, J., Massana, J., & Pous, C. (2015). Multivariate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building. Energy and Buildings, 103, 338-351. Quality index: [JCR CE IF 2.973 Q1(6/126)]

Conferences

- Meléndez, J., Colomer, J., Pous, C., Burgas, L., & Massana, J. (2015). Towards a Data Driven Platform for Energy Efficiency Monitoring: Two Use Cases. In AIAI Workshops (pp. 67-82).

Acronyms

ANN Artificial Neural Network

AR Autoregressive

DG Distributed Generation

EMS Energy Management System

EV Electric Vehicle

HVAC Heating, Ventilating and Air Conditioning

MAPE Mean Absolute Percentage Error

MLP Multilayer Perceptron

MLR Multiple Linear Regression

SG Smart Grid

SM Smart Meter

STLF Short Term Load Forecasting

SVR Support Vector Regression

WSN Wireless Sensor Network

List of Figures

Contents

Abstract

Nowadays, energy is absolutely necessary all over the world. Currently, many of the scientific advances that society enjoys in the fields of health, technology or entertainment, work on the basis of energy. During the last few years, the need for energy has incremented gradually. How to obtain the energy easily, cleanly, economically and safely is one of the aims of thousands of researchers in many parts of the world.

Taking into account the advantages that it presents in transport and the needs of homes and industry, energy is transformed into electricity. Electricity is used, especially in the non-residential buildings' sector, which has never stopped growing since the arrival of electricity.

Bearing in mind the expansion of electricity, initiatives like Horizon 2020, pursue the objective of a more sustainable future: reducing the emissions of carbon and electricity consumption and increasing the use of renewable energies.

The electricity network is constantly being improved. As an answer to the shortcomings of the traditional electrical network, such as large distances to the point of consumption, low levels of flexibility, low sustainability, low quality of energy, the difficulties of storing electricity, etc., Smart Grids (SG), a natural evolution of the classical network, has appeared.

The SG, based on new technologies that have arisen in the market, provides robustness, sustainability, flexibility, quality of service and information for the user. Furthermore, it allows for the storage of energy, as in the use of electric cars, and the possibility of Distributed Generation (DG), incrementing the use of renewable energy.

One of the main components that will allow the SG to improve the traditional grid is the Energy Management System (EMS). The EMS is the element that manages all parts of the system, such as its generation, as in DG power plants, or its consumption, as in appliances. It is precisely in this framework where the electric car will allow for storing energy with the goal of reducing consumption in the hours of more demand and perform electrical load shaping; this is precisely one of the main benefits of the SG.

The current electricity system presents very high electric consumption during peak hours and very low during the valley hours. As a result, the price of energy varies: in some cases, there is an excess and, in others, a lack of energy. A flat daily consumption curve is the ideal for the electrical companies. To achieve the shaping of the consumption curve, several technologies, such as DG, energy exchange between users, energy storage using Electrical Vehicles (EV) or the control of home appliances by the electrical companies, can be used.

As has been said, the EMS is necessary to carry out the management of the power network system, and one of the main needs of the EMS is a prediction system: that is, to know in advance the electricity consumption. Besides, the utilities will also require predictions to manage the generation, maintenance and their investments.

In short, a suitable system of load prediction will allow for daily optimum management of the electrical system. It will help to perform a suitable process of decision-making, in order to plan the resources or decide about initial investments. Therefore, it is necessary to dispose of the systems of prediction of the electrical consumption that, based on the available data, forecast the consumption of the next hours, days or months, in the most accurate way possible.

It is in this field where the present research is placed since, due to the proliferation of sensor networks and more powerful computers, more precise prediction systems have been developed. This change has been undertaken at a quantitative and qualitative level, since the number of publications, with regard to the subject, has spiked and new methods, more precise and complex, have appeared. In the 1970s, most of the techniques employed were linear and, slowly, the non-linear models have been imposed, due to their advantages in dealing with complex problems.

Having said that, a complete study has been realized in [10], taking into account the need to know, in depth, the state of the art, in relation to the load forecasting topic. On the basis of acquired knowledge, the installation of sensor networks, the collection of consumption data and modelling, using Autoregressive (AR) models, were performed in the second work [13]. Once this model was defined, in the third work [12], another step was made, collecting new data, such as building occupancy, meteorology and indoor ambience, testing several paradigmatic models, such as Multiple Linear Regression (MLR), Artificial Neural Network (ANN) and Support Vector Regression (SVR), and establishing which exogenous data improves the prediction accuracy of the models. Reaching this point, and having corroborated that the use of occupancy data improves the prediction, there was the necessity of generating techniques and methodologies, in order to have the occupancy data in advance. Therefore, several attributes of artificial occupancy were designed, in order to perform long-term hourly consumption predictions. These concepts are explained in depth in the fourth work [14].

Looking more closely, in the first article [10] of the present thesis, the most relevant studies, regarding the subject of electric load prediction of the last forty years, have been reviewed. Several publications show a clear evolution from the first models, less powerful and less complex, to the current models, able to achieve high accuracy, and future trends. The models can be classified, for example, according to the area of application, the horizon of prediction or the computational cost. In the present case, the aim was to study, in depth, the Short-term Load Forecasting (STLF) in non-residential buildings; for this reason, different possibilities were analysed, in order to focus the following articles adequately.

The second article [13], presents a description of the prediction service within the Smart City architecture. The aim was to identify the needs of the services in this framework, to support the prediction data-driven models or other systems for energy efficiency monitoring. The study is based on data from the university of Girona, where a prediction service was used to explain, in a practical manner, the insertion of the service in the architecture of layers, based upon the knowledge acquired with the state of the art.

From the conclusions of the second work [13], and using newly collected data, such as weather, occupancy, calendar, indoor ambience, etc., experiments in relation to the consumption prediction models of the buildings were realized. At this point, we tested several configurations with the aim of discovering what generates better results in the prediction. Moreover, data-driven models were studied to benchmark STLF in non-residential buildings. Like the previous one, the study-case is based on data of consumption from the University of Girona. After that, a general guideline about which data, methods and parametrizations are necessary to predict the consumption in non-residential buildings in a proper manner was provided.

The third work [12] helped in identifying which methods, parameters and models generate the best results. Despite this, one of the attributes used, the occupancy of the building, is not available in advance, so it was then necessary for a system to generate this attribute artificially. Starting from the methods and data that generated better results, and on the basis of the same case study of the previous publications, an exhaustive study with regard to the possible manner of extracting information of the building's occupancy was performed. The ultimate aim was to obtain an attribute of occupancy for non-residential buildings, in order to achieve a better balance between computational cost and the accuracy obtained in the prediction.

This method, unlike other methods of the subject studied, that require weather predictions, only requires the attribute of occupancy. Besides, the method allows for hourly consumption predictions months in advance, a feature that places the method above the vast majority of hourly consumption prediction methods.

The last work [14], and closing the cycle of works of STLF in non-residential buildings that contains the present thesis, presented different indicators or attributes of occupancy, from the simplest, that uses general information, to the most complex, based on very detailed information. For each complexity level, the accuracy provided, using an SVR model and the cost of generating such indicator, was contrasted. Finally, a model that presents the best balance between the two parameters was obtained. In the same way as the previous cases, real data from the buildings of the University of Girona were used as a case study.

Thus, the thesis concludes, providing a complete state of the art of the topic STLF in non-residential buildings. Furthermore, three articles that introduce advances in the field are produced, highlighting, especially the last one, that it not only provides a new method of prediction, but a complete study of how to generate artificial indicators of occupancy.

Resum

A dia d'avui l'energia és un bé completament necessari arreu del món. Molts dels avenços científics dels que gaudim enguany en els camps de la salut, la tecnologia o l'entreteniment funcionen a base d'energia. Durant els darrers anys la necessitat d'energia ha anat incrementant-se gradualment. Com obtenir aquesta energia de manera senzilla, neta, econòmica i segura és un dels objectius de milers d'investigadors d'arreu del món.

Degut als avantatges que presenta en el transport i a les necessitats de les llars i la indústria, l'energia és transformada en energia elèctrica. Aquesta electricitat és utilitzada sobretot en el sector dels edificis no-residencials, un sector clau que no ha parat de créixer des de l'aparició de l'electricitat.

Tenint en compte la total expansió i domini de l'electricitat, iniciatives com Horitzó 2020, tenen per objectiu un futur més sostenible: reduint les emissions de carboni i el consum i incrementant l'ús de renovables.

La xarxa elèctrica està en constant millora. Partint dels defectes de la xarxa elèctrica clàssica, com són gran distància al punt de consum, poca flexibilitat, baixa sostenibilitat, baixa qualitat de l'energia, dificultats per a emmagatzemar energia, etc. apareixen les Smart Grid (SG), una evolució natural de la xarxa clàssica.

Les SG, basades en les noves tecnologies que sorgeixen al mercat, aporten solidesa, sostenibilitat, flexibilitat, qualitat de servei, informació a disposició de l'usuari, a més, permeten l'emmagatzematge d'energia a través dels cotxes elèctrics i la possibilitat de generació desagregada, incrementant l'ús de les renovables.

Un dels principals elements que permetrà a les SG millorar les xarxes clàssiques és l'Energy Management System (EMS). Els EMS, són els elements que permeten la gestió de totes les parts del sistema, tant de la generació, centrals desagregades, com del consum, electrodomèstics. És precisament dins d'aquest esquema que el cotxe elèctric ens permetrà emmagatzemar energia a fi i efecte de reduir el consum en les hores de més demanda i poder donar forma a la corba de consum, aquest és precisament un dels principals beneficis que permetran les SG.

El sistema elèctric actual presenta un consum molt alt durant les hores pic i

molt baix durant les hores vall. Això fa que el preu de l'energia varïi i que en certs moments falti energia i en d'altres en sobri. Una corba de consum diari planera seria l'ideal per a les empreses elèctriques. Per a dur a terme el modelat de la corba de consum utilitzarem diferents elements com ara la generació desagregada, l'intercanvi d'energia entre usuaris, l'acumulació d'energia amb els cotxes elèctrics o el control a distància dels electrodomèstics per part de les empreses elèctriques.

Així doncs, per a que l'EMS pugui dur a terme la gestió dels diversos elements, una de les necessitats bàsiques dels EMS serà un sistema de predicció, o sigui, saber per endavant quin consum hi haurà en un entorn determinat. A més, les empreses subministradores d'electricitat també requeriran de prediccions per a gestionar la generació, el manteniment i fins i tot les inversions a llarg termini.

En resum, un bon sistema de predicció ens permetrà una gestió diària del sistema elèctric òptima. A més, ens ajudarà a dur a terme un procés de presa de decisions adequat, per tal de planificar l'adquisició de recursos o decidir les inversions inicials. Així doncs ens calen sistemes de predicció del consum elèctric que, partint de les dades disponibles, ens subministrin el consum que hi haurà d'aquí a unes hores, uns dies o uns mesos, de la manera més aproximada possible.

És dins d'aquest camp on s'ubica la recerca que presentem. Degut a la proliferació de xarxes de sensors i computadors més potents, s'han pogut desenvolupar sistemes de predicció més precisos. Aquest canvi ha estat tant a nivell quantitatiu com qualitatiu, ja que el nombre de publicacions en relació amb el tema s'ha disparat i han aparegut nous mètodes més precisos i complexes. Cap als anys 70 la majoria de tècniques emprades eren lineals, poc a poc però els models no-lineals es van anar imposant degut als seus avantatges enfront problemes complexes.

A tall de resum, en el primer treball [10], i tenint en compte que s'havia de conèixer en profunditat l'estat de la qüestió en relació a la predicció del consum elèctric, es va fer una anàlisi completa de l'estat de l'art. Un cop fet això, i partint del coneixement adquirit, en el segon treball [13] es va dur a terme la instal·lació de les xarxes de sensors, la recollida de dades de consum i el modelatge amb models lineals d'auto-regressió (AR). En el tercer treball [12], un cop fets els models es va anar un pas més enllà recollint dades d'ocupació, de meteorologia i ambient interior, provant diferents models paradigmàtics com Multiple Linear Regression (MLR), Artificial Neural Network (ANN) i Support Vector Regression (SVR) i establint quines dades exògenes milloren la predicció dels models. Arribat a aquest punt, i havent corroborat que l'ús de dades de presència millora la predicció, es van generar tècniques per tal de disposar de les dades d'ocupació per endavant, o sigui a hores vista. D'aquesta manera es van dissenyar diferents atributs d'ocupació artificials, permetent-nos fer prediccions horàries de consum a llarg termini. Aquests conceptes s'expliquen en profunditat al quart treball [14].

Veient-ho en més profunditat, pel primer article [10] de la present tesi s'han revisat els estudis més rellevants pel que fa al tema de predicció de la demanda elèctrica dels darrers 40 anys. Les diferents publicacions mostren una clara evolució, des dels primers models, menys potents i complexes, passant pels models actuals, capaços d'assolir grans precisions, fins a les futures tendències. Dins del ventall de models que ens podem trobar, aquests es poden dividir en funció de l'àrea d'aplicació, l'horitzó de predicció o el cost computacional, per exemple. En el nostre cas l'objectiu era estudiar en profunditat els models de predicció del consum a curt termini en edificis no-residencials, per això vam analitzar les diferents possibilitats, per tal d'enfocar els següents articles de manera adequada.

En el segon article [13] es va dur a terme la descripció del servei de predicció al si de l'arquitectura Smart City. L'objectiu era identificar les necessitats dels serveis en aquest marc per a suportar els models de predicció de dades o altres sistemes de monitorització d'eficiència energètica. L'estudi se sustenta en un exemple pel cas de la universitat de Girona, on es dissenyà un servei de predicció per a explicar de manera pràctica la inserció del servei al si de l'arquitectura de capes, partint dels aprenentatges de l'estat de l'art.

A partir de les conclusions extretes del segon treball [13] es van dur a terme experiments en relació als models de predicció del consum d'edificis a partir de les noves dades recollides: dades de meteorologia, dades de presència, dades de calendari, dades d'ambient interior, etc. En aquest punt es testejaren les diferents configuracions amb l'objectiu d'esbrinar què és allò que genera millors resultats en la predicció. A més, s'estudià quins models de dades subministren millors resultats pel cas dels models de predicció del consum a curt termini en edificis no-residencials. Com en el cas anterior els casos pràctics se sustenten en les dades de consum de la Universitat de Girona. Un cop enllestit aquest treball disposàrem d'un seguit de pautes generals sobre quines dades i quins mètodes i parametritzacions cal aplicar per a predir el consum horari en edificis no-residencials de manera adequada.

El tercer treball [12] va ajudar a conèixer quins mètodes, paràmetres i models generen els millors resultats. Malgrat això, un dels atributs que es van fer servir, l'ocupació de l'edifici, no es disposava per endavant, calia doncs un sistema per a crear aquest atribut de manera artificial. Partint del mètodes i dades que van generar millors resultats, i amb el mateix cas pràctic que en les anteriors publicacions, es va dur a terme un estudi exhaustiu sobre les possibles maneres d'extreure informació de l'ocupació de l'edifici. L'objectiu final era obtenir un atribut d'ocupació pels edificis no-residencials de manera que presentés el millor equilibri entre el cost computacional de crear-lo i la precisió que genera en la predicció.

Aquest mètode, a diferència d'altres mètodes del tema estudiat, que requereixen de prediccions meteorològiques, només requereix de l'atribut d'ocupació. A més, el mètode permet fer prediccions de consum horàries mesos per endavant, cosa que el situa per davant de la gran majoria de mètodes de predicció de consum horari.

Així doncs, el darrer treball [14], i tancant amb el cicle de treballs de predicció del consum a curt termini en edificis no-residencials que engloba la present tesi, va presentar diferents indicadors o atributs d'ocupació, del més senzill, que fa servir informació general, fins al més complex, basat en informació molt detallada. I per cadascun d'aquests nivells es contraposà la precisió que subministra aquest atribut a través d'un model SVR i el cost de generar tal indicador. Finalment, es va obtenir un model que presenta el millor equilibri entre els dos paràmetres. Com en els casos anteriors es van emprar dades reals dels edificis de la Universitat de Girona com a demostració.

D'aquesta manera, la tesi es conclou subministrant un estat de l'art complet sobre la predicció de consum elèctrica, a més de tres articles que aporten avenços dins del camp, destacant sobretot el darrer que, no només aporta un mètode nou de predicció, sinó un estudi complet de com generar de manera artificial indicadors d'ocupació.

Resumen

A día de hoy la energía es un bien completamente necesario en todo el mundo. Muchos de los adelantos científicos de los que disfrutamos a día de hoy en los campos de la salud, la tecnología o el entretenimiento funcionan en base a la energía. Durante los últimos años la necesidad de energía ha ido incrementándose gradualmente. Como obtener esta energía de manera sencilla, limpia, económica y segura es uno de los objetivos de miles de investigadores de todo el mundo.

Debido a las ventajas que presenta en el transporte y a las necesidades de los hogares y la industria, la energía es transformada en energía eléctrica. Esta electricidad es utilizada sobre todo en el sector de los edificios no-residenciales, un sector clave que no ha parado de crecer desde la aparición de la electricidad.

Teniendo en cuenta la total expansión y dominio de la electricidad, iniciativas como Horizonte 2020, tienen por objetivo un futuro más sostenible: reduciendo las emisiones de carbono y el consumo e incrementando el uso de renovables. La red eléctrica está en constante mejora. Partiendo de los defectos de la red eléctrica clásica, como son gran distancia en su punto de consumo, poca flexibilidad, baja sostenibilidad, baja calidad de la energía, dificultades para almacenar energía, etc. aparecen las Smart Grids (SG), una evolución natural de la red clásica.

Las SG, basadas en las nuevas tecnologías que surgen al mercado, aportan solidez, sostenibilidad, flexibilidad, calidad de servicio, información a disposición del usuario, además, permiten el almacenamiento de energía a través de los coches eléctricos y la posibilidad de generación desagregada, incrementando el uso de las renovables.

Uno de los principales elementos que permitirá a las SG mejorar las redes clásicas es el Energy Management System (EMS). Los EMS, son los elementos que permiten la gestión de todas las partes del sistema, tanto de la generación, centrales desagregadas, como del consumo, electrodomésticos. Es precisamente dentro de este esquema que el coche eléctrico nos permitirá almacenar energía con el objetivo de reducir el consumo en las horas de más demanda y poder dar forma a la curva de consumo, este es precisamente uno de los principales beneficios que permitirán las SG.

El sistema eléctrico actual presenta un consumo muy alto durante las horas pico

y muy bajo durante las horas valle. Esto hace que el precio de la energía varíe y que en ciertos momentos falte energía y en otros sobre. Una curva de consumo diario llana seria el ideal para las empresas eléctricas. Para llevar a cabo el modelado de la curva de consumo utilizaremos diferentes elementos como por ejemplo la generación desagregada, el intercambio de energía entre usuarios, la acumulación de energía con los coches eléctricos o el control a distancia de los electrodomésticos por parte de las empresas eléctricas.

Así pues, para que el EMS pueda llevar a cabo la gestión de los diferentes elementos, una de las necesidades básicas de los EMS será un sistema de predicción, o sea, saber por anticipado qué consumo habrá en un entorno determinado. Además, las empresas suministradoras de electricidad también requerirán de predicciones para gestionar la generación, el mantenimiento e incluso las inversiones a largo plazo.

En resumen, un buen sistema de predicción nos permitirá una gestión diaria del sistema eléctrico óptima. Además, nos ayudará a llevar a cabo un proceso de toma de decisiones adecuado, para planificar la adquisición de recursos o decidir las inversiones iniciales. Así pues nos hacen falta sistemas de predicción del consumo eléctrico que, partiendo de los datos disponibles, nos suministren el consumo que habrá en unas horas, unos días o unos meses, de la manera más aproximada posible.

Es dentro de este campo donde se ubica la investigación que presentamos. Debido a la proliferación de redes de sensores y computadoras más potentes, se han podido desarrollar sistemas de predicción más precisos. Este cambio ha sido tanto a nivel cuantitativo como cualitativo, puesto que el número de publicaciones en relación con el tema se ha disparado y han aparecido nuevos métodos más precisos y complejos. Hacia los años 70 la mayoría de técnicas empleadas eran lineales, poco a poco los modelos no-lineales se fueron imponiendo debido a sus ventajas frente a problemas complejos.

En resumen, en el primer trabajo [10], y teniendo en cuenta que había que conocer en profundidad el estado de la cuestión en relación a la predicción del consumo eléctrico, se hizo un análisis completo del estado del arte. Una vez hecho esto, y partiendo del conocimiento adquirido, en el segundo trabajo [13] se llevó a cabo la instalación de las redes de sensores, la recogida de datos de consumo y el ajuste de modelos lineales de auto-regresión. En el tercer trabajo [13], una vez hechos los modelos se fue un paso más allá recogiendo datos de presencia, de meteorología y ambiente interior, probando diferentes modelos paradigmáticos como Multiple Linear Regression (MLR), Artificial Neural Network (ANN) y Support Vector Regression (SVR) y estableciendo qué datos exógenos mejoran la predicción de los modelos. Llegado a este punto, y habiendo corroborado que el uso de datos de presencia mejora la predicción, se generaron técnicas para disponer de los datos de ocupación por adelantado, o sea a horas vista. De este modo se diseñaron diferentes atributos de ocupación artificiales, permitiéndonos hacer predicciones horarias de consumo a largo plazo. Estos conceptos se explican en profundidad en el cuarto trabajo [14].

Viéndolo en más profundidad, por el primer artículo [10] de la presente tesis se han revisado los estudios más relevantes en cuanto al tema de predicción de la demanda eléctrica de los últimos 40 años. Las diferentes publicaciones muestran una clara evolución, desde los primeros modelos, menos potentes y complejos, pasando por los modelos actuales, capaces de lograr grandes precisiones, hasta las futuras tendencias. Dentro del abanico de modelos que nos podemos encontrar, estos se pueden dividir en función del área de aplicación, el horizonte de predicción o el coste computacional, por ejemplo. En nuestro caso el objetivo era estudiar en profundidad los modelos de predicción de consumo a corto plazo en edificios no-residenciales, por eso se analizaron las diferentes posibilidades, con la intención de enfocar los siguientes artículos de manera adecuada.

En el segundo artículo [13] se llevó a cabo la descripción del servicio de predicción en el sí de la arquitectura Smart City. El objetivo era identificar las necesidades de los servicios en este marco para soportar los modelos de predicción de datos u otros sistemas de monitorización de eficiencia energética. El estudio se sustenta en un ejemplo por el caso de la universidad de Girona, donde se diseñó un servicio de predicción para explicar de manera práctica la inserción del servicio en el sí de la arquitectura de capas, partiendo de los aprendizajes del estado del arte.

A partir de las conclusiones extraídas del segundo trabajo [13] se llevaron a cabo experimentos en relación a los modelos de predicción del consumo de edificios a partir de los nuevos datos recogidos: datos de meteorología, datos de ocupación, datos de calendario, datos de ambiente interior, etc. En este punto se testearon las diferentes configuraciones con el objetivo de averiguar qué es aquello que genera mejores resultados en la predicción. Además, se estudió qué modelos de datos suministran mejores resultados por el caso de los modelos de predicción del consumo a corto plazo en edificios no-residenciales. Cómo en el caso anterior los casos prácticos se sustentan en los datos de consumo de la Universitat de Girona. Una vez terminado este trabajo dispusimos de una serie de pautas generales sobre qué datos y qué métodos y parametrizaciones hay que aplicar para predecir el consumo horario en edificios no-residenciales de manera óptima.

El tercer trabajo [12] nos ayudó a conocer qué métodos, parámetros y modelos generan los mejores resultados. A pesar de esto, uno de los atributos que se usaron, la ocupación del edificio, no se disponía por adelantado, hacía falta pues un sistema para crear este atributo de manera artificial. Partiendo de los métodos y datos que generaron mejores resultados, y con el mismo caso práctico que en las anteriores publicaciones, se llevó a cabo un estudio exhaustivo sobre las posibles maneras de extraer información de la ocupación del edificio. El objetivo final era obtener un atributo de ocupación para los edificios no-residenciales de forma que presentara el

mejor equilibrio entre el coste computacional de crearlo y la precisión que genera en la predicción. Este método, a diferencia de otros métodos del tema estudiado, que requieren de predicciones meteorológicas, sólo requiere del atributo de ocupación. Además, el método permite hacer predicciones de consumo horarias meses por adelante, cosa que lo sitúa por delante de la gran mayoría de métodos de predicción de consumo horario.

El último trabajo [14], y cerrando con el ciclo de trabajos de predicción de consumo a corto plazo en edificios no-residenciales que engloba la presente tesis, presentó diferentes indicadores o atributos de ocupación, desde el más sencillo, que usa información general, hasta el más complejo, basado en información muy detallada. Y por cada uno de estos niveles se contrapuso la precisión que suministra este atributo a través de un modelo SVR y el coste de generar tal indicador. Finalmente, se obtuvo un modelo que presenta el mejor equilibrio entre los dos parámetros. Cómo en los casos anteriores se emplearon datos reales de los edificios de la Universitat de Girona como demostración.

Así pues, la tesis concluye suministrando un estado del arte completo sobre la predicción de consumo eléctrico además de tres artículos que aportan avances dentro del campo, destacando sobre todo el último que, no sólo aporta un método nuevo de predicción, sino un estudio completo de como generar de manera artificial indicadores de ocupación.

Chapter 1

Introduction

In order to maintain the ecosystem and live in a sustainable world, the present collective behaviour must be changed. Nowadays, the production and transportation of electricity imply high levels of contamination and great economic expenditure. With the aim of reducing these negative effects, several initiatives have been proposed [15], [8].

As is known, the current electrical grid is divided into three levels: generation, transport and distribution. This classic design presents several shortcomings, such as dependence on large generation plants, losses in transport, high levels of contamination or centralized power generation. Control of the electricity grid is becoming more and more complicated and it is difficult to reduce the losses in transport, which is increasing, day by day. Besides, the needs of electricity users are growing in volume and features. The main concerns for the electricity grid are around infrastructure, maintenance, safety, energy efficiency and sustainability.

With the aim of solving these issues, the electricity system is changing. Generation, transport and distribution are integrating intelligence into different levels through new hardware and software being added to the electrical system. The new concept of grid, called SG [9], enables an efficient, flexible, accessible, reliable and robust behaviour with the aim of achieving proper operation with an optimal response to unexpected electrical system events. One of the main purposes of the SG is to manage the loads, in order to shape the load curve. Another main goal is to extend the use of decentralized generation. Using intelligence and the net of sensors to improve fault detection and the energy flow are also in the list of needs.

Following the same issue, the major elements in the load curve shaping scene are the EV proliferation [5], the use of DG [1] and the implantation of Smart Meters (SM) [6]. By the emergence of the EV in the electricity grid, the profile of consumption will change, becoming more flat [16]. The EV consumes electricity during the day and can be charged in the night, as when a car is in the garage, but, if it is necessary, it can return electricity to the system during the day. Therefore, the valley periods will be used to charge the EV and a fraction of this electricity will be returned during the day, in the peak periods. The DG, that has as an idea the deployment of low-power electricity generation plants close to the consumption points, provides an interesting opportunity for renewable energy sources, removing part of the shortcomings of the transport and distribution of electricity. The features of DG and renewable energy sources increase the difficulty of control; in this context, the demand for forecasting methods is on the rise. In relation to SM, the inclusion of elements that monitor and have the capacity of controlling home appliances lets the utilities shape the consumption profile.

Before continuing, some new concepts must be introduced. These new concepts have been appearing in the last few years, such as Smart City [3], microgrid [11] or smart building [18]. Due to the increase of people in cities looking for jobs, city services, such as electricity and water supply, transportation, communication, garbage management, governance and health services, are suffering degradation. In order to deal with this scenario, the management of these services must change, not only to monitor it with sensors, but also to perform corrective actions. So, the smart city is composed of a net of sensors collecting data, a process unit analysing this data and a system to implement corrective actions, with the aim of maintaining the quality of the city's services. The smart city is a synergy between software and hardware to provide city services at the highest quality levels.

A microgrid is an aggregation of multiple loads and generation technologies, with the aim of providing power and heat, in a free market context. A microgrid can work connected to the electricity grid or work in isolation, when there are perturbations in the electricity system. The size of microgrid varies between a small town and a building. Nevertheless, independent of its size, the need and the scheme are the same: generation sources, storage and electric loads.

The Smart building is a type of microgrid with its own features. The integration of intelligence in buildings, such as sensors, intelligent devices, activators, etc. implies energy saving and more control. This, in conjunction with the use of renewable microsources close to the building and the control performed with energy management systems, helps to convert the building into a more robust, efficient and flexible element.

Therefore, there is the shared need to forecast for each one of these concepts and scenarios. If the EV is used to shape the profile of consumption, there is the need to forecast. When the EMS manages the resources in a microgrid or a smart building, there is the need to forecast the demand and the production of energy. In the renewable energy plants, the production of energy, based on climatological sources (solar radiation, wind, etc.), needs to be predicted, thus the weather predictions are needed too. When a utility controls the production of energy in the power plants, there is the need to forecast. So, in the same way, there is a need to forecast the consumption of electricity in buildings, due to the utility needs, such as adjusting the generation of electricity to the consumption of users in real time. The utilities try to minimize the waste of resources covering the additional demand at the lower combustible price and the electric load forecast is the manner of adjusting the two sides of the grid.

In the building sector, the classic methods to predict electric demand used past values of consumption and calendar information, in conjunction with regression models. Nowadays, most of the works use calendar and weather data with ANN. The new proposals are in the direction of the use of weather, calendar and occupancy data with support vector regression. In each generation, the models are more complex but provide more accurate predictions. The future research line is clear; the use of user behaviour variables enhances the forecasting performance. There is another diaphanous idea; there is a big correlation between prediction sophistication and disaggregation. The more disaggregated, the more difficult to predict. So, a country is easier to predict than a city and a city is easier than a building. The aggregation of thousands of individual loads in a city flattens the load curve profile. In the case of buildings, one user using the elevator can, in a random moment, modify completely the normal consumption profile of the building. Therefore, there is the need to dispose of information to predict this kind of actuation and its effect on consumption.

In the building forecasting domain, responsible for 40% of the emissions of CO2 [4], the consumption depends on the sector. For example, the consumption profiles in malls are similar and the collected variables to predict consumption are the same. Inside the public non-residential sector there are hospitals, universities, asylums, nurseries, schools, etc. Each step in the analysis and comprehension of each sub-sector enables improvements in the prediction accuracy. Each sub-sector must understand their own features because the knowledge of it allows the implementation of the demand response system [2]. In summary, the building load curves imply non-linearities, uncertainty, abrupt changes and noise. In particular, the non-residential buildings consist of daily, weekly and seasonal patterns in their consumption profiles. In general, during the nights and holidays, consumption is extremely low and, during the day, the profile of consumption is defined by the user's behaviour. The periods, such as time of entry, time of exit, breakfast and lunch time, define the consumption. During the week, the consumption patterns are explained by the days. Similarly, the explaining factor during the year are the seasons and, in particular, the months.

When the type of building is defined, and the features are studied, it is the moment to find which is the minimal information needed to be collected to achieve the best forecasting accuracy with the lowest investment and computational cost. The use of appropriate models and attributes will deliver proper results.

On the search for attributes to correlate with consumption, not only the weather and the calendar are used; the occupancy of the building seems to be useful. There is a need to define a consistent and standard way of predicting occupancy. There are several studies, with different approaches, proposing their load prediction, based upon occupancy schemes and explaining the obtained results. The process is always the same: monitoring, modelling and predicting. There is the necessity of testing several solutions, such as calendar data, surveys, sensor data, work schedules, etc., to describe the occupancy. When the data is collected, the machine learning field presents several supervised learning methods to predict numeric attributes. The most popular and successful prediction methods are MLR [7], the ANN[19] and the SVR [17].

Chapter 2

Objectives

The main goal of the thesis is study the topic of electrical demand forecasting, in order to perform a short-term load forecasting system in non-residential buildings, on the basis of real data.

2.1 State of the art

The first work [10] has the aim of reviewing the main papers on the topic of electrical load forecasting, published during the last forty years. The works will be categorized according to different aspects. Taking into account the presented studies, there is the aim to demonstrate that there is a clear evolution in the field, showing that it is an active research area. The selected classification categories of the work are the following ones:

- According to the forecasting horizon.
- According to the aim of the forecast.
- According to the linearity of the model.
- According to the used model.
- According to the computational cost.
- According to the area of application.
- According to the used variables.
- According to the year of implementation.

- According to the used architecture.
- According to the prediction performance.

In addition, there will be an analysis of the drawbacks and advantages to understand the singularities of the field. The main objective is to understand the characteristics of the research area and enter into the topic and avoid future mistakes.

2.2 Identifying services

The objective of the second work [13] is to describe and define present efforts to embed different services in Smart City architecture. Taking into account the current framework, there is the aim to identify possible service integration difficulties.

In order to control and manage the distribution system, adjusting demand to the user's consumption, acquiring the raw materials at the best price or smoothing the consumption curve, several electric load prediction models are used by the utilities. With the aim of defining the requirements, characteristics, functionalities and possible interactions of the reference architecture is required to identify services, such as energy monitoring systems and assessment applications for urban infrastructures, based upon data-driven methods for load prediction. On the basis of the proper Smart City architecture, a service of short-term load prediction, will be explained layer by layer, trying to cover this gap.

In order to define the features of the selected smart service, a short-term load forecasting system in non-residential buildings in the University of Girona will be provided in the second work [13]. A practical explanation of the singularities of the services embedded in the described layers' architecture will be undertaken, searching for potential difficulties and management improvements; the idea can be extrapolated to other methods.

An analysis of several existent Smart City architectures will be performed, in order to select the most suitable one. There is the intention of providing an example of an implementation of how to embed services in Smart City architecture because there is a lack of them in the literature.

In the case used, the acquisition, pre-processing, modelling, and analysis procedures will be incorporated into a global service in a Smart City architecture context, in order to exploit data for energy management purposes.

2.3 Contrasting attributes

A well-known research interest is to generate short-term load forecasting models with easy parametrization, cheap implementation and reduced computational time. The current computers allow for solving some of these challenges; however, the data mining process can solve some drawbacks, such as reducing the database size, in order to diminish the computational time. In any case, there is no doubt that the main objective is to increase accuracy in the prediction.

Most of the works in this topic perform a pre-processing and then applying the models, provide the achieved performance. The first works applied linear regression [7], then the proliferation of the ANN [19] provided new improvements to the field. Nowadays, the introduction of support vector machine [17] models is giving the best forecasting performance.

Collecting data with sensors is a slow and expensive process; therefore, only crucial information must be collected. Usually, the majority of papers collect the temperature, taking into account that this is the variable most related to consumption. To a lesser extent, other weather variables, such as relative humidity, solar radiation or wind direction, are collected. The level of occupancy of the building is appearing only in some recent papers. The same happens with indoor ambience; there are only a few works in which this kind of data is used.

The main goal of the third work [12] is to provide a short-term load forecasting method for an office building at the University of Girona (Catalonia) with a Mediterranean climate with the following requirements:

- High forecasting performance.
- Low computational requirements.
- Minimum data collected.

Therefore, in order to design the best possible model, several scenarios will be proposed and the prediction accuracy will be calculated for each one. Every step of the process will be questioned with the purpose of achieving the highest forecasting performance. The main goal is to define a general framework of how to predict. The two main experiments are the following ones:

- Meteorological (temperature, relative humidity or solar radiation), indoor ambience (indoor temperature, indoor relative humidity or indoor level of light), occupancy and calendar data will be tested to discover which data is the most relevant in non-residential building load forecasting.

- Three different standard methods, such as MLR, MLP and SVR will be tested to find which of them provides better prediction performance.

2.4 Generating artificial occupancy attributes

In the topic of short-term load forecasting in non-residential buildings, regardless of the forecasting performance, all the analysed works present several drawbacks. The last work [12] of the authors indicated that SVR, using temperature and occupancy as attributes, resulted in the best load forecasting performance. So, on the basis of the last paper [12] of the authors, the present paper [14] will implement improvements. The main challenges that this paper will achieve are the following ones:

- When the model includes weather variables, such as temperature or relative humidity, in order to dispose of this data in advance, weather predictions are necessary. There is no doubt that this data contains uncertainties that will be added to the load forecasting. In addition, there is not, in every place, the option of obtaining weather predictions and these can only be obtained for a few days ahead. Therefore, there is the necessity of obtaining models without weather data.

- If occupancy data is used in the model, the data is not ready to be used in the moment of prediction. Then, techniques to artificially generate occupancy information in advance are needed.

- In most of the studies of hourly predictions, the prediction horizon is 1-hour ahead. The performance of the forecasting decays with the horizon. The further in the future is the horizon, the worst is the prediction. The aim is to perform hourly predictions, months in advance, doing hourly long-term predictions.

- The use of expert knowledge in the model is difficult and uncertain. The methodology must be defined, and a clear and objective scheme is key.

- The use of a large amount of collected data in the model implies a deployment of sensors, thus leading to a high cost.

- Some papers provide good results using past values of consumption; this makes the model dependent upon consumption measures, close to the prediction time, forcing the user to perform continuous measures.

With the aim of achieving these challenges, a real case study of short-term load forecasting for several non-residential buildings in the University of Girona will be provided. The objectives that will be accomplished in this work [14] are the following ones:

- In order to avoid the need for weather predictions, a model without temperature will be implemented.
- With the aim of disposing of the occupancy data in advance, a complete study of the different methods to generate artificial occupancy indicators will be created. On the basis of various information, such as scholar schedules, sensor data, calendar data, expert knowledge or classroom features, several occupancy indicators will be generated. Each occupancy indicator will be more sophisticated than the previous one. All the forecasting performances, achieved with each indicator, will be compared.
- A new concept, called quality factor, will be created, to provide a tool with the object of determining which is the best occupancy indicator. This indicator will consider the prediction accuracy and the workload to generate it.
- With the objective of performing consumption predictions months ahead, a model that does not depend upon non-available data in advance will be implemented.
- A simple and cheap short-term load forecasting model, without the need for continuous sensor data collection, will be provided.

Chapter 3

A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings

In this chapter, a review presents the most relevant studies on electric demand prediction over the last 40 years, and introduces the different models used as well as the future trends. Additionally, it analyses the latest studies on demand forecasting in the future environments that emerge from the usage of smart grids. This publication has been published in the following paper:

Paper published in the **IEEE Communications Surveys & Tutorials**
Volume: 16, Issue: 3, Pages: 1460-1495, Published: April 2014
DOI: 10.1109/SURV.2014.032014.00094
JCR ISCS IF: 2.788 Q1(1/143)

Luis Hernandez, Carlos Baladrón, Javier M. Aguiar, Belén Carro, Antonio J. Sanchez-Esguevillas, Jaime Lloret, and Joaquim Massana. “A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings”. *IEEE Communications Surveys & Tutorials.* Vol. 16, issue 3, p.: 1460-1495

Abstract

Recently there has been a significant proliferation in the use of forecasting techniques, mainly due to the increased availability and power of computation systems and, in particular, to the usage of personal computers. This is also true for power network systems, where energy demand forecasting has been an important field in order to allow generation planning and adaptation. Apart from the quantitative progression, there has also been a change in the type of models proposed and used. In the `70s, the usage of non-linear techniques was generally not popular among scientists and engineers. However, in the last two decades they have become very important techniques in solving complex problems which would be very difficult to tackle otherwise. With the recent emergence of smart grids, new environments have appeared capable of integrating demand, generation, and storage. These employ intelligent and adaptive elements that require more advanced techniques for accurate and precise demand and generation forecasting in order to work optimally. This review discusses the most relevant studies on electric demand prediction over the last 40 years, and presents the different models used as well as the future trends. Additionally, it analyzes the latest studies on demand forecasting in the future environments that emerge from the usage of smart grids.

Published in: IEEE Communications Surveys & Tutorials (Volume: 16, Issue: 3, Third Quarter 2014)

Date of Publication: 21 April 2014
Electronic ISSN: 1553-877X
INSPEC Accession Number: 14527953
DOI: 10.1109/SURV.2014.032014.00094
Publisher: IEEE
Sponsored by: IEEE Communications Society

Keywords
Biological system modeling,
Load modeling,
Predictive models,
Demand forecasting,
Microgrids,
Smart grids

Chapter 4

Identifying services for short-term load forecasting using data driven models in a Smart City platform

In this chapter, there is a description of the ongoing work to embed several services in a Smart City architecture with the aim of achieving a sustainable city. With this object in mind, a use case of short-term load forecasting in non-residential buildings in the University of Girona is provided, in order to practically explain the services embedded in the described general layers architecture. In the work, classic data-driven models for load forecasting in buildings are used as an example. This publication has been published in the following paper:

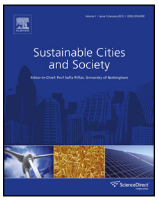
Paper published in the **Sustainable Cities and Society**
Volume: 28, Pages: 108-117, Published: January 2017
DOI: 10.1016/j.scs.2016.09.001
JCR CBT IF: 1.044, Q2(27/61)

Sustainable Cities and Society 28 (2017) 108–117

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs



Identifying services for short-term load forecasting using data driven models in a Smart City platform

Joaquim Massana*, Carles Pous, Llorenç Burgas, Joaquim Melendez, Joan Colomer

University of Girona, Campus Montilivi, P4 Building, Girona E17071, Spain

ARTICLE INFO

Article history:
Received 21 June 2016
Received in revised form 1 September 2016
Accepted 3 September 2016
Available online 7 September 2016

Keywords:
Short-term load forecasting
Data mining
Services
Building
Smart City architecture

ABSTRACT

The paper describes an ongoing work to embed several services in a Smart City architecture with the aim of achieving a sustainable city. In particular, the main goal is to identify services required in such framework to define the requirements and features of a reference architecture to support the data-driven methods for energy efficiency monitoring or load prediction. With this object in mind, a use case of short-term load forecasting in non-residential buildings in the University of Girona is provided, in order to practically explain the services embedded in the described general layers architecture. In the work, classic data-driven models for load forecasting in buildings are used as an example.

1. Introduction

The concept of Smart City appears due to the mobilization of people to the cities. This increase of people has an impact on city services such as transportation, utilities, communications, waste management, health services and much other. In order to avoid services degradation, and have an idea of the effect of such increase of people for a particular service, it is necessary to manage each service by constantly monitoring it. Therefore, it is needed to provide the system with mechanisms for collecting data. This is the first step towards getting to a Smart City. But what it really makes the city smart is to process and analyse the data and returns as response some kind of action to ensure the provision of services at satisfactory levels of quality. Hence, it is necessary to integrate these monitoring devices with the applications that perform the analysis of this data and are able to provide an action (da Silva et al., 2013).

The synergy of computational and physical components, specifically the use of cyber-physical systems (CPSs), led to the advancement of such integration. At different scale, neighbourhoods, communities or buildings can also be considered large CPS continuously operated accordingly to demand affected by the activities of users. As important is to know physical system constraints as consumer's behaviour, and interactions between both. Major Information and Communication Technology (ICT) vendors have made efforts for developing Smart City transversal platforms oriented to integrate city information and making it available to end-users. On the other hand, the utilities (water, electricity, gas, etc.) have their proprietary solutions specifically designed to operate and supervise these infrastructures and providing managing and billing services. This work falls in between these two scopes and shares the IoT (internet of Things) vision, focusing not only in making data available but also providing the required services to facilitate advanced data analysis, monitoring and assessment procedures in the domain of urban energy distribution and consumption. This paper aims to analyse a specific use case in order to identify services that are required in a platform that supports the development of energy monitoring and assessment applications for urban infrastructures.

Several general architectures for Smart Cities are proposed in the literature, but few examples of their implementation and how to embed services on them are given. According to the existing Smart City architectures, the present work proposes an implementation of a practical case, a complete short-term load forecasting system, explaining the singularities layer by layer trying to cover this gap.

The utilities are the main users of the load prediction systems, who, thanks to the load prediction, manage the maintenance and the control of the distribution systems, buying fuel at the best price or shaping the consumption curve in order to have a flat consumption curve following several strategies. So, this is a tool for the

* Corresponding author.
E-mail addresses: joaquim.massana@udg.edu (J. Massana), carles.pous@udg.edu (C. Pous), llorenc.burgas@udg.edu (L. Burgas), joaquim.melendez@udg.edu (J. Melendez), joan.colomer@udg.edu (J. Colomer).

http://dx.doi.org/10.1016/j.scs.2016.09.001
2210-6707/

utilities who manage the distribution systems, and in particular to help them to forecast the electrical consumption.

2. Context and related work

In the bibliography, taking into account the existence of different visions, several definitions of Smart City are found. In Giffinger and Pichler-Milanović (2007) the Smart City is defined as "a city well performing in a forward-looking way in economy, people, governance, mobility, environment, and living, built on the smart combination of endowments and activities of self-decisive, independent and aware citizens". Otherwise in Bowerman et al. (2000) it is said that Smart City is "a city that monitors and integrates conditions of all of its critical infrastructures, including roads, bridges, tunnels, rail/subways, airports, seaports, communications, water, power, even major buildings, can better optimize its resources, plan its preventive maintenance activities, and monitor security aspects while maximizing services to its citizens". The paper (Washburn et al., 2009) says that "the use of Smart Computing technologies to make the critical infrastructure components and services of a city which include city administration, education, healthcare, public safety, real estate, transportation, and utilities more intelligent, interconnected, and efficient".

Some papers, like Nam and Pardo (2011), coincide that Smart Cities are composed by three main dimensions. The first one is the technology dimension, where several technologies are used to monitor, control and share in the city processes. The second one is the human dimension, where creativity, relationships, education and knowledge are the base of the human infrastructure to provide social benefits to the Smart City. The third one is the institutional dimension, where the administration promotes regulations, policies and community participation to grow properly and sustainably.

On the basis of the reviewed works, the common Smart City challenges are:

- Establish a base Smart City architecture to provide a common framework for the sector.
- Dispose and extend standardized Smart City policies that lead to the growth and the proliferation of Smart City services and initiatives.
- Design a list of the essential Smart City services such as Smart water, Smart Governance, Smart buildings, etc.
- Define the basic guidelines in order to perform operations, maintenance, improvements and the scalability in the Smart Cities.

Therefore, it has sense to contribute in the field with a suitable Smart City architecture, selected for developing the services oriented to consumption prediction. It provides the basis where the smart services are going to operate. The following paragraphs summarize different works done in the field of Smart Cities covering proposed architectures and services implied and some of them particularized for short-term load forecasting (STLF). From the point of view of services, there are some papers cited.

A complete guide for design the Smart City architectures and all the functionalities from the data point of view is proposed in Wenge, Zhang, Dave, Chao, and Hao (2014). A summary of the main issues of the application systems and the difficulties and challenges in the construction of the Smart City is presented in Su, Li, and Fu (2011). A broad view of energy services and their usage, functionality and development challenges are explained in Karnouskos, Silva, and Ilic (2012). In order to improve operations and maintenance, reduce the cost of operation, provide enhanced energy management capabilities and provide scalability in the Smart City architecture a guidelines are highlighted in Al-Hader, Rodzi, Sharif, and Ahmad (2009). Several Smart City architectures and their requirements are exposed and commented in da Silva et al. (2013). The work (Morvaj, Lugaric, & Krajcar, 2011) comes up with a model for analysis of interactions with a Smart City, providing a larger scale simulation among several Smart City systems. A wide survey of technologies, protocols, and architecture for an urban internet of things in Smart Cities is shown in Zanella, Bui, Castellani, Vangelista, and Zorzi (2014).

So, there is no defined criteria about the number and the function of layers of the Smart City architecture. The work (Komninos, 2006) presents a three layers architecture: information storage layer, application layer and user interface layer. The paper (Al-Hader et al., 2009) suggests a five layers architecture: smart infrastructure, smart database, smart building manager, smart interface and integration layer. The publication (Anthopoulos & Fitsilis, 2010) proposes a five layers architecture: stakeholder layer, service layer, business layer, infrastructure layer and information layer. In Filipponi et al. (2010) the Smart City architecture is divided in two layers: knowledge processors and semantic information brokers. The paper (Lugaric, Krajcar, & Simic, 2010) proposes a Smart City architecture with three parts: the physical network, the communications infrastructure and the flow of information. The study (Al-Hader & Rodzi, 2009) divides the Smart City in two layers: monitoring layer and development layer. The work (Wenge et al., 2014) proposes a five layers architecture: data acquisition, data transmitting, data storage, support service, domain service and event application.

In relation with Smart City services, a short-term load forecasting model for non-residential building on the basis of occupancy and temperature is presented in Massana, Pous, Burgas, Melendez, and Colomer (2015). A principal component analysis is used for monitoring the electric consumption of buildings in Burgas, Melendez, Colomer, Massana, and Carles (2014). In order to organize the power production of distributed generation sources in relation with energy storage system and reduce the operational costs of microgrids a smart energy manager system is provided in Chen, Duan, Cai, Liu, and Hu (2011). In the work Lund, Andersen, Østergaard, Mathiesen, and Connolly (2012), the need to include the cogeneration power generation in electricity balancing and grid stabilization is pointed out. The benefits of a home energy control box for optimizing energy consumption from electrical vehicle charging in residential buildings are seen in Mets, Verschueren, Haerick, Develder, and De Turck (2010). In Krajačić et al. (2011) an energy system planning which incorporates renewable energy services, energy storage technologies and system regulation strategies is provided. A smart energy distribution and management system for monitoring power consumption and users' situation and controlling appliances is presented in Byun, Hong, Kang, and Park (2011). An energy information system (real data acquisition, visualization, analysis and switching) which admits the integration of several sensors is provided in Kunold, Kuller, Bauer, and Karaoglan (2011). The paper (Castro, Jara, & Skarmeta, 2013) describes a smart lighting solution which allows the integration of the communications and logic on the current street lighting infrastructure. A design and implementation of occupancy sensor platform for individual offices is presented in Agarwal et al. (2010).

Taking into account the energy signatures, in Aoun (2013) the importance of energy signatures which can help to improve the energy efficiency and monitor the consumption, is pointed out. The use of the energy signatures in order to evaluate the energy performance of chillers using several design options and operating strategies is seen in Yu and Chan (2005). In Rabl and Rialhe (1992) the addition of occupancy as a variable in energy signature model PRISM is analysed.

With regard to baseline models and measuring and verification methods, the work (Heo, Choudhary, & Augenbroe, 2012) proposes a calibration methodology of the building energy models which can deal with energy retrofit options. In Yoon, Lee, and Claridge (2003) a calibration procedure of the energy performance model on the basis of monthly data through a base load analysis approach is proposed. A statistical evaluation of the performance of various commercial building baseline models analysing the importance of the weather and the morning adjustment factors is seen in Coughlin, Piette, Goldman, and Kiliccote (2009). Measuring and verification guiding principles for the assessment of energy efficiency insisting in the need of unambiguous contractual models are highlighted in Park et al. (2011).

In the following sections, a suitable Smart City architecture, selected for developing the services oriented to consumption prediction, is detailed. It provides the basis where the smart services are going to operate. After that, a use case to explore the smart-x service in line with the proposed architecture layers is provided.

3. Vision

As it can be seen in the bibliography, there is an extensive proposal of architectures to face common challenges that arise in the Smart Cities concept. But, a reference architecture that allows the entire operation of a Smart City has not been designed yet. The subject has been treated cautiously due to the number of technologies that involves, and mainly because it has not been established a standard for integrating these technologies in order to generate a coherent, flexible, scalable, repeatable and effective system. Furthermore, some of the approaches deal with Smart Cities from a theoretical viewpoint which distances itself from the real world. The proposed architectures focus on different aspects from the point of view of technology, human-system interaction or logic (Wenge et al., 2014). Most of the proposals from the technological aspect divide the architecture in layers. There could be some slight differences, but as seen in the previous section they have some features in common.

The proposed 5-layers architecture is composed by: data acquisition layer, data transmission layer, data storage layer, preprocessing layer, services layer and application layer, as shown in Fig. 1. This architecture delivers better definition of the function of each layer and it is oriented to the services implementation.

- **Data acquisition layer** is responsible for collecting and storing external data. It can capture any kind of information including images, video, sound, and others. In particular circumstances, some preprocessing can be done here, in order to store the data filtered or more elaborated.
- **The data transmission layer** is in charge of end-to-end communications. Network technologies and protocols are taken into account at this level.
- **The data storage layer** has to be able to support large-scale complex data. Also, it has to guarantee that the data is reliable and must provide for the introduction of new data from new sensors or new available information. That is, it has to be scalable. At the same time, the layer has to provide access methods to the data.
- **Preprocessing layer**. Once the data is stored, since they come from different types of sensors or information sources, the architecture has to prevent from duplications, outliers, errors, missing values and inconsistency. These kinds of actions are carried out by the preprocessing layer.
- **Services layer**. Following with the most common layers that constitute the majority of architectures proposed for a Smart City from the technological point of view, there is the services layer. This layer makes possible the usability of the data, usually by means of modules of software that provide the data requested by the user in a transparent manner.

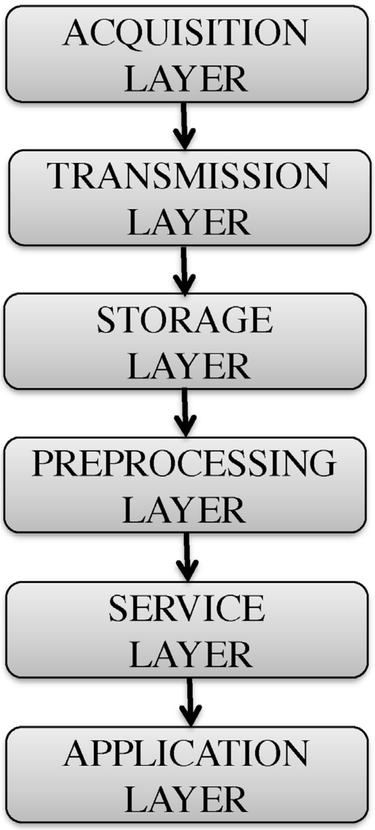


Fig. 1. Proposed layers architecture.

- **Applications layer**. The last layer is the applications layer. It is responsible for interacting directly with the user. It shows the data to the user in a comprehensible manner such as graphical form, table or other type of display, and facilitates the interaction with the platform.

  A use case focusing on forecasting electrical energy to improve its management is proposed in the next section. This example is intended to help to identify services involved, required functionalities and possible interactions. The architecture has to consider not only the infrastructure itself but also the interaction with consumers, providing performance indicators and using forecasting models. The use case shows that acquisition, preprocessing, analysis and modelling of data are required processes to provide a set of goal oriented services to systematically exploit data for energy management purposes. In this particular case, classic data-driven methods for forecasting are proposed, but the idea behind can easily be extrapolated to other methods.

4. A use case: short-term load forecasting

The aim of this use case is to present the implementation of a smart-x service following the architecture reference of the Smart Cities. The prediction of the load consumption is a need in the Smart Cities and a well-known research domain. The consumption of the non-residential buildings is determined by several factors such as previous consumption, occupancy, temperature and temperature set point. There are several ways to deal with the load forecasting depending on the horizon, available data or used model. In general, the main objective is to forecast with high accuracy and few data. The process is usually composed by data selection, data preprocessing, model selection, training process, model evaluation and results exploitation.

Table 1
Architectural features of the buildings.

Building	Floors	Year	Volume (m^3)	Frontage area (m^2)	Glass area (m^2)
PI	6	1983	26,150	3791	610
PII	6	1992	25,560	2326	1351
PIII	3	2003	11,346	1785	310
PIV	3	2003	12,000	1836	630
Faculty of Science	3	1997	34,810	4903	1233
Faculty of Law	6	1999	32,290	6420	1675
Faculty of Economics	5	1997	32,287	4770	1375

So, in the next sections, with the aim to generate an autoregressive (AR) model to predict the consumption of the buildings a complete process of short-term load forecasting using only real consumption data is explained, respecting the same architecture layers explained in Section 3.

4.1. Data acquisition

4.1.1. Sensors

There are several sources of data: electrical load data, weather data and indoor data. Different sensors, placed in distinct places are collecting data with an hourly sampling rate.

- Electric load data: electrical load data (kW) is collected using the Campus Infrastructure Monitoring System (CIMS). The CIMS is composed of several Schneider power meters installed at the university buildings. The consumption data is collected in PI, PII, PIII, PIV, Faculty of Science, Faculty of Law and Faculty of Economics buildings. There are three different types of power meters: PowerLogic ION 7350 (power, current, voltage, frequency, power factor, current and harmonic distortion), PowerLogic ION 6200 (power, current, voltage and frequency) and PowerLogic PM-810 (power, power factor and frequency). With the devices properly configured, the data is transferred by the log inserter from each device to the database every 15 min. The communication between meters and data storage is performed with the PowerLogic ION Enterprise 5.6 software.
- Weather data: data of temperature (°C) using a HMP-35AC sensor of Vaisala, relative humidity (%) using a Humicap sensor of Vaisala and solar radiation (W/m^2) using a CM11 sensor of Kipp & Zonen are collected outside the buildings by the Department of Physics.
- Indoor data: only for the case of PIV, indoor ambient and occupancy data are collected inside the building. A wireless sensor network (WSN) is collecting data of temperature (°C) using a MCP9700A sensor of Microchip, relative humidity (%) using a 808H5V5 sensor of Sencera, light level (lux) using a PDV-P9203 sensor of Optoelectronics and presence using a passive infrared sensor of Parallax. In summary, there are 6 sensors badges capturing ambient data and 2 capturing people activity.

4.1.2. Dataset

Taking into account that in this work an AR model is implemented, only consumption data is used. In the paragraphs that follow, the used data is explained with a brief introduction of its location.

The experiments are conducted using data from PI, PII, PIII and PIV, Faculty of Science, Faculty of Law, Faculty of Economics buildings located at the University of Girona, as seen in Fig. 2. The buildings have classrooms, offices and laboratories.

In Table 1, the architectural characteristics for each university building are shown.

In Table 2, the specifications of the heating system for each university building are seen.

In Table 3, the specifications of the cooling system for each university building are explained.

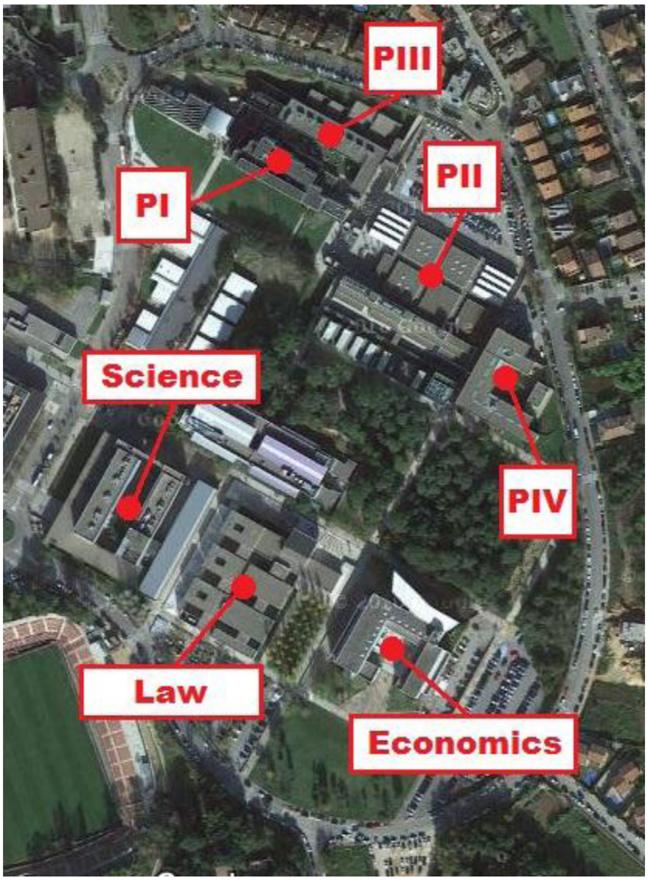


Fig. 2. Location of the buildings in the campus of the University of Girona.

The number of data instances of PI is 27,375, covering a total of 38 months, from 1st September 2011 to 15th October 2014. The total of instances of PII is 16,589, covering a total of 24 months, from 21st November 2012 to 15th October 2014. The number of instances of PIII and PIV is 16,590, covering a total of 24 months, from 23rd November 2012 to 15th October 2014. The number of instances of Faculty of Science is 27,366, covering a total of 38 months, from 1st September 2011 to 14th October 2014. The

Table 2
Heating system features of the buildings.

Building	Heating system	
	System	Boiler brand (power)
PI	Gas boiler + Fancoil	Fer (442 kW)
PII		Robur (327 kW)
PIII		Dietrich (310 kW)
PIV		Ygnis (824 kW)
Faculty of Science		Dietrich (560 kW)
Faculty of Law		Wiessman (575 kW)
Faculty of Economics		Dietrich (310 kW)

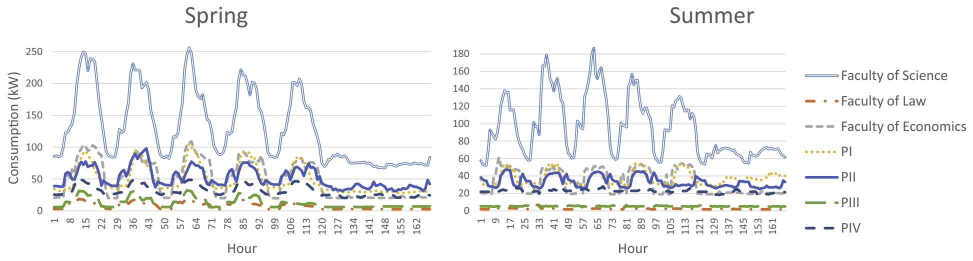


Fig. 3. Consumption vs. hour.

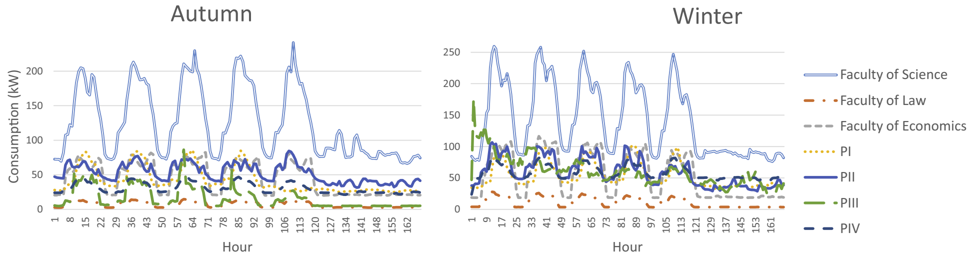


Fig. 4. Consumption vs. hour.

number of instances of Faculty of Law is 27,379, covering a total of 38 months, from 1st September 2011 to 15th October 2014. The number of instances of Faculty of Economics is 27,379, covering a total of 38 months, from 1st September 2011 to 15th October 2014.

In Fig. 3 the consumption of a week in spring and summer is observed.

In Fig. 4 the consumption of a week in autumn and winter is seen.

4.2. Data transmission

There are three data sources. The Department of Physics, serving weather data, the CIMS, disposing of consumption data and the WSN of PIV, collecting ambient and occupancy data.

The Department of Physics uses a wired sensor network to collect the data from the several instruments of the weather station. At the same time, the CIMS captures the consumption of different buildings using a wired network too. In the case of the indoor data a WSN is employed.

The WSN is composed by 8 motes of the Libelium-brand which send the measured data to a central hub, called Meshlium, trough XBee radio modules that communicate by means of the ZigBee protocol. The Meshilum data is accessed using an Ethernet connection. The Libelium technology is based on Arduino and the topology of the network is star.

In Fig. 5 the WSN of PIV building is seen.

Table 3
Cooling system features of the buildings.

Building	Cooling system	
	System	Refrigeration brand (power)
PI	Compression refrigeration system + Fancoil	Mitsubishi (160 kW)
PII		Ygnis (269 kW)
PIII		Carrier (255 kW)
PIV		Climaveneta (618 kW)
Faculty of Science		Daikin (430 kW)
Faculty of Law		Teva (1113 kW)
Faculty of Economics		Carrier (255 kW)

4.3. Data storage

The data come from 3 distinct sources: meteorological data provided by the Department of Physics, consumptions data for each buildings provided by CIMS and indoor data collected by the WSN. Each source presents distinct configurations, and even owners, that made impossible a direct actuation over the distinct databases storing the information further than data access. Data presents distinct formats for each of the sources and a homogenization step is mandatory.

The CIMS data is stored in a MSSQL database for the Schneider software ION. Department of Physics data is accessed via an SFTP server and the WSN data is stored in a MySQL database inside

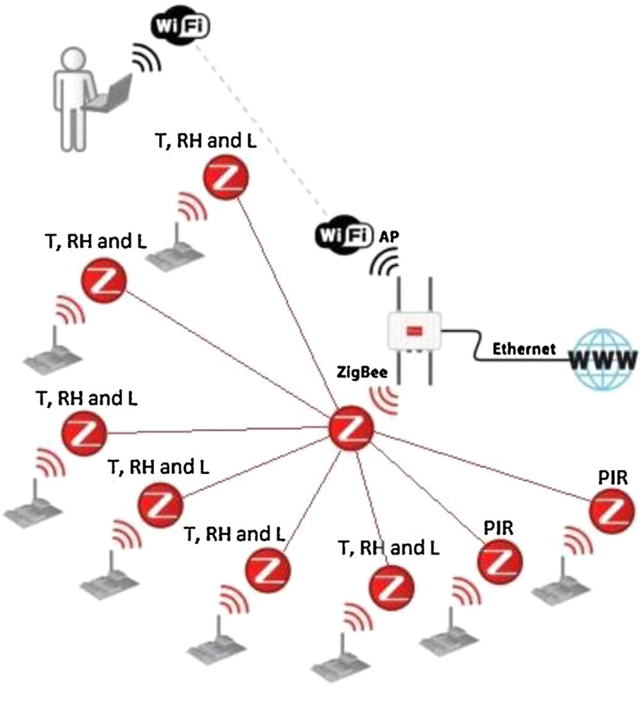


Fig. 5. Wireless sensor network of PIV building.

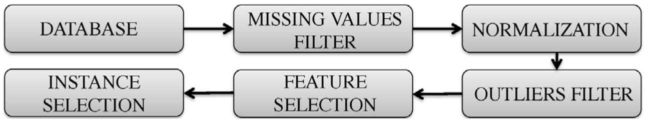


Fig. 6. Block diagram of the preprocessing.

the Meshlium. The solution implemented is an homogenization server whose tasks consist in periodically connect to the distinct data sources, check for updates and update a local MySQL database with an homogeneous format for all the data and provide a simple interface for user to select and download the desired data.

4.4. Data preprocessing

In the following sections, the steps to clean and uniform the data are explained as seen in Fig. 6.

4.4.1. Missing values

Given the mistakes in sensor readings, there is always a small amount of lost values. The percentage of missing values needs to be minimized. There are several methods used to filter the missing values such as removing or averaging them. In our case, the instances with missing values are deleted.

4.4.2. Normalization

If data has different scales and units normalization is needed. The use of the same data scale improves the forecasting. The normalization range used is from 0 to 1, as seen in Eq. (1).

$$x_{in} = \frac{x_i}{\sqrt{\sum_i (x_i^2)}} \tag{1}$$

where x_{in} is the normalized instance. x_i is the instance.

4.4.3. Outliers

The performance of the model is increased if the outliers are filtered. The more restrictive the process, the greater amount of data lost. The outliers filtering process (Ramaswamy, Rastogi, & Shim, 2000) consists in detection and substitution. In the present case, the process of detection consists in identifying n outliers based on the Euclidean distance to their k nearest neighbours. Then, according to an outlier detection process, the outliers are removed.

4.4.4. Feature selection

With the aim of removing irrelevant features, redundant and non-correlated attributes are removed. Reducing the size of the database, the computational cost is reduced. The feature selection process is composed by two blocks that perform linear correlations. The first block, in order to eliminate the useless attributes, removes the features with low correlation with the class attribute. The second block, with the aim of deleting the duplicate attributes, removes the attributes with high correlation among them.

4.4.5. Instance selection

The number of instances is reduced in order to minimize the computational cost. The selected training data is a 30% random subsample. Samples about this percentage reduce the computational time while maintain the forecasting performance levels.

4.5. Data service

There is the intention to explore the performance limits of the AR model. So, the experiments are realized using only consumption data taking into account that these data models are simpler and useful in 1-hour ahead forecasts.

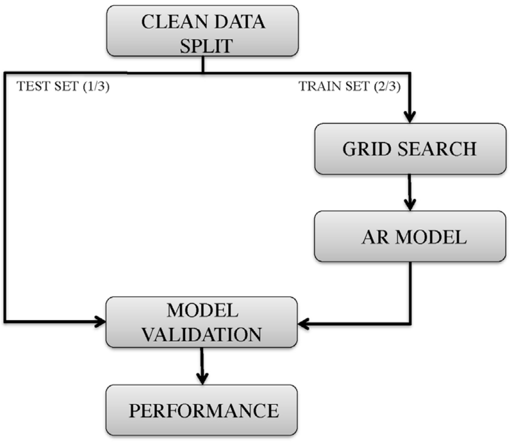


Fig. 7. Block diagram of the forecasting model.

4.5.1. Methodology

A methodology for predicting the load consumption 1-h ahead is proposed. This methodology consists of several blocks as shown in Fig. 7. The preprocessed data is split, 1/3 to test and 2/3 to train. Then, with the training data a grid search of the suitable training parameters is performed over the selected model (AR). The final step is the validation of the model using test data and the performance indicator calculation.

4.5.2. Grid search

Grid search method performs a search through the ranks of pairs of training parameters and chooses the best ones. The main tested parameters of the regression model are the following ones: ridge parameter and feature selection.

4.5.3. AR model

The AR model (Huang, 1997) specifies that the output variable depends linearly on its own previous values. Taking into account that the occupancy data is only available for PIV building and the temperature variable does not increase the accuracy of this model due to the partial disaggregation of the heating ventilating and air conditioning system, AR model is a proper model to apply in short-term load forecasting (1-h ahead). So, the consumption depends on the past values of consumption, as can be seen in Eq. (2).

$$X_t = C + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t \tag{2}$$

where X_t is the output variable. X_{t-1} are the previous values of the output. $\varphi_i, \ldots, \varphi_p$ are the parameters of the model. C is a constant. ϵ_t is white noise.

4.5.4. Validation

The validation process contrasts the model generated with training data (65%) against the test data (35%). The mean absolute percentage error (MAPE) indicator is used to validate the model due to its popularity in the forecasting field. The first period of time is used to predict the last period of time.

The MAPE performance indicator, shown in Eq. (3), does not depend on the magnitude of the unit of measurement, and is used to compare models. If the MAPE is small, the model is accurate. In the topic, a range between 1% and 20% is considered acceptable.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{m(i)} - y_{p(i)}}{y_{m(i)}}\right| \times 100 \tag{3}$$

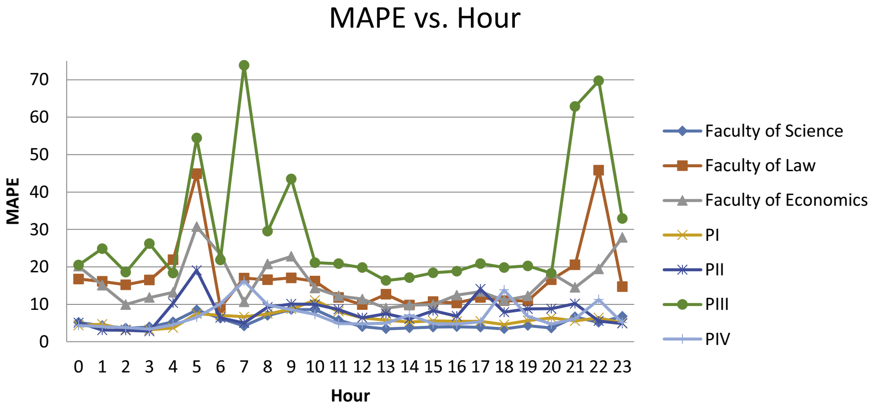


Fig. 8. MAPE vs. Hour.

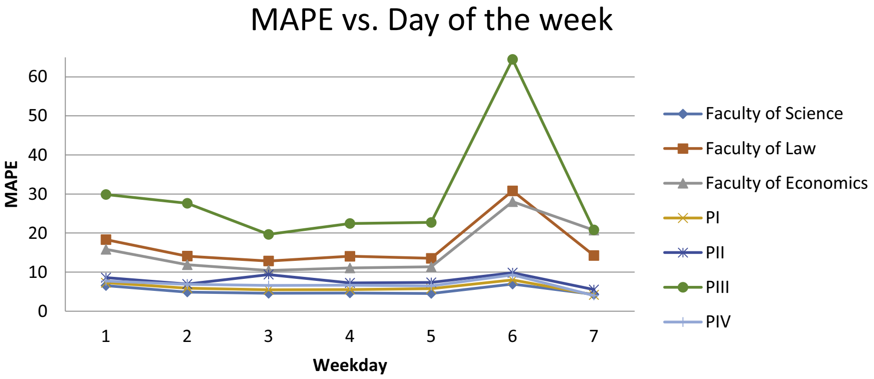


Fig. 9. MAPE vs. weekday.

where N is the number of observations. y_m is the measured output. y_p is the predicted output.

4.6. Smart application

On the basis of the outputs of the model, the smart application can offer several services such as prediction charts, energy saving information or corrective actions effectiveness. The users can access to the application interfaces to increase the expert knowledge in order to reduce the consumption or to monitor the forecasting accuracy. In the present paper analytic and graphic results are explained.

4.6.1. Analytical results

Table 4 shows the MAPE indicator for all the buildings, where CC is the correlation coefficient and the computation time is the time to perform the experiment.

The Faculty of Law and The Faculty of Economics present intermediate level of accuracy due to the variability in the consumption profile. PIII has low level of accuracy as a result of inconsistent data.

Table 4
AR model results for all buildings.

Building	Computing time (ms)	C.C	MAPE (%)
Faculty of Science	3672	0.986	5.21
Faculty of Law	4031	0.968	16.75
Faculty of Economics	3734	0.972	15.58
PI	4109	0.981	6.06
PII	2797	0.943	7.84
PIII	3235	0.876	30.11
PIV	2250	0.967	6.83

4.6.2. Graphic results

In this section several charts are presented. First three MAPE charts are presented. Then, seven prediction consumption charts are showed.

- MAPE vs. hour:

As can be seen in Fig. 8, there are three time slots which present poor quality prediction, some night hours (4:00 to 6:00) due to the cleaning and security services and at the beginning (7;00 to 10:00) and the end (20:00 to 22:00) of the school day given the variability in the human behaviour.

- MAPE vs. weekday:

As is shown in Fig. 9, Saturday presents low accuracy in the forecasting due to random activities realized in the buildings. In addition, the beginning and the end of the week, that is Mondays and Fridays, taking into account the high dispersion of human behaviours, present lower forecasting precision.

- MAPE vs. month:

As presented in Fig. 10, there is no clear conclusions about which months are better predicted but in general, months with high variability and unclear profiles such as December, June or March present worse prediction.

The following charts show the consumption prediction for each building.

- Faculty of Science, Faculty of Law and Faculty of Economics prediction:

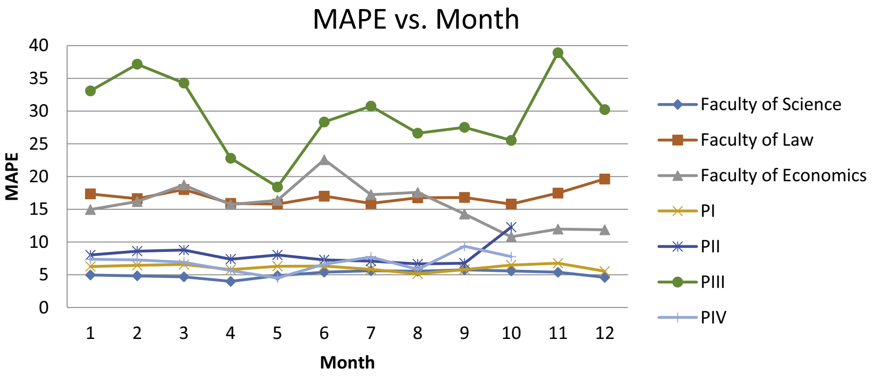


Fig. 10. MAPE vs. Month.

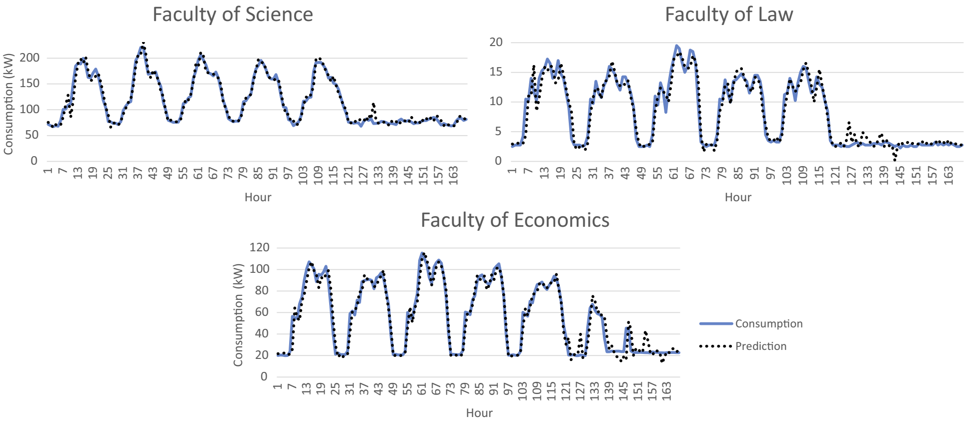


Fig. 11. Consumption vs. hour.

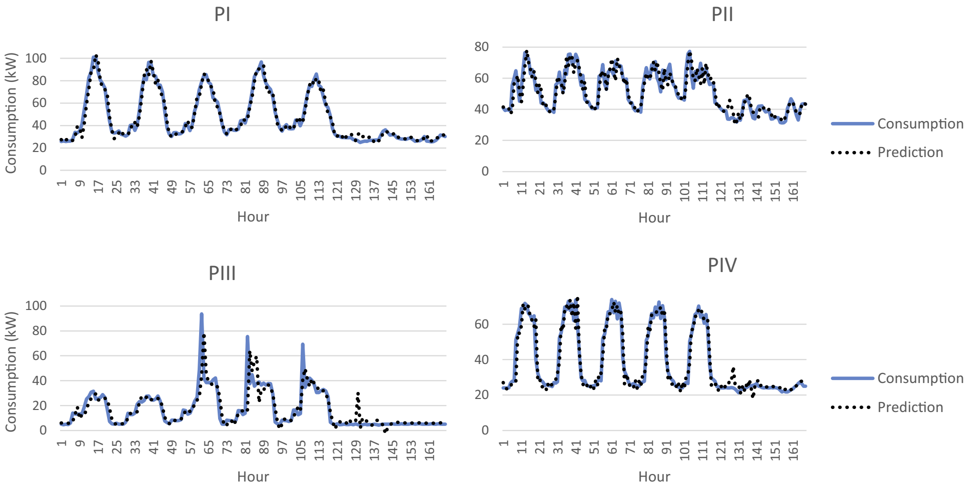


Fig. 12. Consumption vs. hour.

As is seen in Fig. 11, the worst prediction level is found in Saturdays due to the random after-school activities in some buildings. Besides, in the early hours of Mondays, due to the irregularity of some services, the prediction accuracy decays.

- PI,PII, PIII and PIV prediction:

As shows in Fig. 12, PIII presents the poorest forecasting quality due to the inconsistency of the data. As in the previous figure, Saturdays are the worst predicted days.

5. Discussion

Following the smart-x architecture layers a use case is performed:

- In the acquisition layer the robustness is key, in the present case some misbehaviours of the sensors comported data loss and outliers.
- In relation with the transmission layer, some ZigBee reception problems entailed data loss, so a previous study of the sensor distribution is needed.

- The storage layer must be safe, standard and scalable, the load consumption database presented integration difficulties.
- With reference to the preprocessing layer, outliers or missing values in the sensor measures lead to a low accuracy in the forecasting. Using a software with suitable preprocessing tools is completely necessary to obtain fine results.
- In reference to the service layer, although occupancy and weather data are variables that partially explain consumption in buildings, the AR models are simple and quick. In summary, the presented model depends only on the consumption, so weather data is not needed, making it more economic and compact. In this case, the prediction is performed 1-h ahead, that means better performance results than 24-h ahead models, where exogenous data is usually needed. From the results, it is obvious that some buildings present better forecasting accuracy than others. In order to provide fine predictions using AR models, the building must has clean and consistent data. Cyclic and well-defined consumption patterns deliver proper AR model predictions. If there are some random or undefined activities in the buildings, the autocorrelation is low. The buildings with big amount of classrooms or offices with concrete schedule are easy to predict, in the other hand, buildings with rooms with non-defined activities are hard to predict. Similarly, there are some time-slots with high variability in the human behaviour in the nights or at the beginning and the end of the school day that present difficulties to be predicted.
- Taking into account the appliance layer, there is the need to make it accessible and upgradeable. In the present case some efforts have been employed to present the results (charts, tables, etc.) through web services.

In relation with the case study, some actions to take for possible energy saving improvements can be derived from the system analysis:

- Compress work schedules, reducing the hours flexibility. Specify the entering and leaving work, the mealtime and the lunchtime periods.
- Suppression of the HVAC system during weekends and holidays. Adjust the HVAC operation time downwards.
- Move the cleaning service to day hours.
- Control the HVAC in order to have temperature, relative humidity and light level inside the proper range, proposed by the authorities, taking into account the homogeneity along the building.

6. Conclusions

Urban development involves the use of intelligent services taking advantage of monitored data and providing an action to improve or maintain the quality of these city services. This paper focuses on the particular case of the electricity, identifying services that can help in increasing the energy efficiency in urban infrastructures.

The work proposes a use case of short-term load forecasting in non-residential buildings with real data in order to practically explain the services embedded in the described Smart City general layers architecture. These layers are responsible of collecting the data from the sensors, transmitting the data to the central hub, storing, cleaning and standardizing this data, applying the forecasting methodology and finally provide an application to show the results. The use case provided as a demonstration consists of predicting the consumption in 7 university buildings. The load forecasting is performed using AR models showing that the results differ according to the profile of the building and the quality of the data. When the data is complete and the consumption pattern is cyclic and clear, the results are fine. Also, the service allows to test the prediction accuracy from different points of view, such as analysing which is the best month predicted or the same for the days of the week.

As a future work, more services have to be defined to help providing more information to the users in order to improve the energy efficiency of the buildings. For example, defining an index related to the efficiency of the building can be a good contribution to the subject.

Acknowledgements

This research project has been partially funded through BR-UdG Scholarship of the University of Girona granted to Joaquim Massana Raurich. Work developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014–2016), the MESC project funded by the Spanish MINECO (Ref. DPI2013-47450-C2-1-R) and the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 680708.

References

Agarwal, Y., Balaji, B., Gupta, R., Lyles, J., Wei, M., & Weng, T. (2010). Occupancy-driven energy management for smart building automation. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building* (pp. 1–6). ACM.

Al-Hader, M., & Rodzi, A. (2009). The smart city infrastructure development & monitoring. *Theoretical and Empirical Researches in Urban Management*, *2*(11), 87.

Al-Hader, M., Rodzi, A., Sharif, A. R., & Ahmad, N. (2009). Smart city components architecture. In *International conference on computational intelligence, modelling and simulation, 2009. CSSim'09* (pp. 93–97). IEEE.

Anthopoulos, L., & Fitsilis, P. (2010). From digital to ubiquitous cities: Defining a common architecture for urban development. In *2010 Sixth international conference on intelligent environments (IE)* (pp. 301–306). IEEE.

Aoun, C. (2013). *The smart city cornerstone: Urban efficiency*. Published by Schneider Electric.

Bowerman, B., Braverman, J., Taylor, J., Todosow, H., Von, U., & Wimmersperg. (2000). The vision of a smart city. In *2nd international life extension technology workshop, Paris, Vol. 28*.

Burgas, L., Melendez, J., Colomer, J., Massana, J., & Carles, P. (2014). Principal component analysis for monitoring electrical consumption of academic buildings. *Energy Procedia*, *62*, 555–564.

Byun, J., Hong, I., Kang, B., & Park, S. (2011). A smart energy distribution and management system for renewable energy distribution and context-aware services based on user patterns and load forecasting. *IEEE Transactions on Consumer Electronics*, *57*(2), 436–444.

Castro, M., Jara, A. J., & Skarmeta, A. F. G. (2013). Smart lighting solutions for smart cities. In *2013 27th international conference on advanced information networking and applications workshops (WAINA)* (pp. 1374–1379). IEEE.

Chen, C., Duan, S., Cai, T., Liu, B., & Hu, G. (2011). Smart energy management system for optimal microgrid economic operation. *IET Renewable Power Generation*, *5*(3), 258–267.

Coughlin, K., Piette, M. A., Goldman, C., & Kiliccote, S. (2009). Statistical analysis of baseline load models for non-residential buildings. *Energy and Buildings*, *41*(4), 374–381.

da Silva, W. M., Alvaro, A., Tomas, G., Afonso, R. A., Dias, K. L., & Garcia, V. C. (2013). Smart cities software architectures: A survey. In *Proceedings of the 28th annual ACM symposium on applied computing* (pp. 1722–1727). ACM.

Filipponi, L., Vitaletti, A., Landi, G., Memeo, V., Laura, G., & Pucci, P. (2010). Smart city: An event driven architecture for monitoring public spaces with heterogeneous sensors. In *2010 Fourth international conference on sensor technologies and applications (SENSORCOMM)* (pp. 281–286). IEEE.

Giffinger, R., & Pichler-Milanović, N. (2007). *Smart cities: Ranking of European medium-sized cities*. Centre of Regional Science, Vienna University of Technology.

Heo, Y., Choudhary, R., & Augenbroe, G. A. (2012). Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings*, *47*, 550–560.

Huang, S. R. (1997). Short-term load forecasting using threshold autoregressive models. In *IEE Proceedings – Generation, transmission and distribution, Vol. 144* (pp. 477–481). IET.

Karnouskos, S., Silva, P. G. D., & Ilic, D. (2012). Energy services for the smart grid city. In *2012 6th IEEE international conference on digital ecosystems technologies (DEST)* (pp. 1–6). IEEE.

Komninos, N. (2006). The architecture of intelligent cities: Integrating human, collective and artificial intelligence to enhance knowledge and innovation. In *2nd IET international conference on intelligent environments, 2006. IE 06, Vol. 1* (pp. 13–20). IET.

Krajačić, G., Duić, N., Zmijarević, Z., Mathiesen, B. V., Vučinić, A. A., & da Graça Carvalho, M. (2011). Planning for a 100% independent energy storage for integration of renewables and CO_2 emissions reduction. *Applied Thermal Engineering*, *31*(13), 2073–2083.

Kunold, I., Kuller, M., Bauer, J., & Karaoglan, N. (2011). A system concept of an energy information system in flats using wireless technologies and smart metering devices. In *2011 IEEE 6th international conference on intelligent data acquisition and advanced computing systems (IDAACS), Vol. 2* (pp. 812–816). IEEE.

Lugaric, L., Krajcar, S., & Simic, Z. (2010). Smart city – Platform for emergent phenomena power system testbed simulator. In *2010 IEEE PES innovative smart grid technologies conference Europe (ISGT Europe)* (pp. 1–7). IEEE.

Lund, H., Andersen, A. N., Østergaard, P. A., Mathiesen, B. V., & Connolly, D. (2012). From electricity smart grids to smart energy systems – A market operation based approach and understanding. *Energy*, *42*(1), 96–102.

Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings*, *92*, 322–330.

Mets, K., Verschueren, T., Haerick, W., Develder, C., & De Turck, F. (2010). Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging. In *Network operations and management symposium workshops (NOMS Wksps), 2010 IEEE/IFIP* (pp. 293–299). IEEE.

Morvaj, B., Lugaric, L., & Krajcar, S. (2011). Demonstrating smart buildings and smart grid features in a smart energy city. In *Proceedings of the 2011 3rd international youth conference on energetics (IYCE)* (pp. 1–8). IEEE.

Nam, T., & Pardo, T. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th annual international digital government research conference: Digital government innovation in challenging Times* (pp. 282–291). ACM.

Park, K., Kim, Y., Kim, S., Kim, K., Lee, W., & Park, H. (2011). Building energy management system based on smart grid. In *2011 IEEE 33rd international telecommunications energy conference (INTELEC)* (pp. 1–4). IEEE.

Rabl, A., & Rialhe, A. (1992). Energy signature models for commercial buildings: Test with measured data and interpretation. *Energy and Buildings*, *19*(2), 143–154.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, *29*(2), 427–438.

Su, K., Li, J., & Fu, H. (2011). Smart city and the applications. In *2011 international conference on electronics, communications and control (ICECC)* (pp. 1028–1031). IEEE.

Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N., & Nelson, L. E. (2009). Helping CIOs understand smart city initiatives. *Growth*, *17*(2).

Wenge, R., Zhang, X., Dave, C., Chao, L., & Hao, S. (2014). Smart city architecture: A technology guide for implementation and design challenges. *China Communications*, *11*(3), 56–69.

Yoon, J., Lee, E., & Claridge, D. E. (2003). Calibration procedure for energy performance simulation of a commercial building. *Journal of Solar Energy Engineering*, *125*(3), 251–257.

Yu, F. W., & Chan, K. T. (2005). Energy signatures for assessing the energy performance of chillers. *Energy and Buildings*, *37*(7), 739–746.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, *1*(1), 22–32.

Chapter 5

Short-term load forecasting in a non-residential building contrasting models and attributes

In this chapter, the aim is to create a method to forecast the electric load in a non-residential building. Another goal is to analyse what kind of data, as weather, indoor ambient, calendar and building occupancy, is the most relevant in building load forecasting. This publication has been published in the following paper:

Paper published in the **Energy and Buildings**
Volume: 92, Pages: 322-330, Published: April 2015
DOI: 10.1016/j.enbuild.2015.02.007
JCR CE IF: 2.973 Q1(6/126)

Energy and Buildings 92 (2015) 322–330

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Short-term load forecasting in a non-residential building contrasting models and attributes

Joaquim Massana*, Carles Pous, Llorenç Burgas, Joaquim Melendez, Joan Colomer

University of Girona, Campus Montilivi, P4 Building, Girona E17071, Spain

ARTICLE INFO

Article history:
Received 27 October 2014
Received in revised form 2 February 2015
Accepted 3 February 2015
Available online 11 February 2015

Keywords:
Load forecasting
Multilayer perceptron
Regression
Neural networks
Support vector machines
Measured data
Mediterranean climate

ABSTRACT

The electric grid is evolving. Smart grids and demand response systems will increase the performance of the grid in terms of cost efficiency, resilience and safety. Accurate load forecasting is an important issue in the daily operation and control of a power system. A suitable short term load forecasting will enable a utility provider to plan the resources and also to take control measures to balance the supply and demand of electricity.

The aim of this paper is to create a method to forecast the electric load in a non-residential building. Another goal is to analyse what kind of data, as weather, indoor ambient, calendar and building occupancy, is the most relevant in building load forecasting. A simple method, tested with three different models, such as MLR, MLP and SVR, is proposed. The results, from a real case study in the University of Girona, show that the proposed forecast method has high accuracy and low computational cost.

1. Introduction

Electricity is the most important resource in the world economy. But, as can be seen in the Lisbon Treaty [20], electricity needs to be more economic and environmentally clean. The classic electric grid has several disadvantages: losses in transport, centralized power generation, high dependence on large generation plants, etc. The purpose of the new electrical power grid, the smart grid, is to manage loads to shape the load curve and use decentralized generation. Thanks to smart grids, the network will be robust, reliable, efficient and dynamic.

Considering these new concepts, one of the most important challenges of the utilities is to adjust power generation to a user's consumption in real time. An overestimation leads to a waste of resources, whilst an underestimation means an increase in the price to cover additional demand. Now, unlike earlier, this adjustment will be made in smaller environments like microgrids. The electric load forecast is, today, the best way to adjust the two sides of the grid.

Almost 40% of the emissions of CO_2 [2] comes from the building sector, therefore, STLF in the building field is fundamental to reduce energy consumption. The load curves of the electric power consumption of cities are different from the building ones. Building load curves present more variability, noise and non-linearity. Thus, the more disaggregated, the more difficult to predict.

Besides, there are many kinds of buildings, like residential or non-residential buildings. The non-residential buildings, as the university administration sector, has daily, weekly and seasonal patterns in their consumption profile. At nights, holidays and weekends, the consumption is extremely low. During the day there are regular patterns based on the user's activities. Within the week there are regular patterns of consumption. During the year there are regular patterns associated with the seasons. Also, at the beginning and at the end of the working day and in the lunch and breakfast times there are transition areas in the consumption. Like most buildings over 10 years old, HVAC is controlled manually with subjective criteria. In addition, the occupancy of the building and the calendar data can be significant in determining consumption.

Another important issue in the electricity forecast is to know which is the key information to be measured and collected. The goal is to obtain the maximum accuracy with the minimum data attributes and instances.

In this paper, meteorological, indoor environment, occupancy, calendar and electric consumption data have been collected in the University of Girona. Then, three different models have been tested to find which of them gives better performance. MLR, MLP and SVR models have been chosen because they are standards with

* Corresponding author. Tel.: +34 646074566.
E-mail addresses: joaquim.massana@udg.edu (J. Massana), carles.pous@udg.edu (C. Pous), joaquim.melendez@udg.edu (J. Melendez), joan.colomer@udg.edu (J. Colomer).

http://dx.doi.org/10.1016/j.enbuild.2015.02.007
0378-7788/© 2015 Elsevier B.V. All rights reserved.

Nomenclature

b	computed parameter
C	complexity parameter
H_{1-6}	6 indoor relative humidity measure points
H_e	outdoor relative humidity measure point
$k(x_i, x)$	kernel function
L_{1-6}	6 indoor luminosity measure points
L_e	solar radiation measure point
N	number of observations
n	norm of the normalized instances
p	grade of the norm
P_{1-2}	2 occupancy measure points
Q_1	first quartile
Q_3	third quartile
T_{1-6}	6 indoor temperature measure points
T_e	outdoor temperature measure point
x_{in}	normalized instance
$\lvert x_p \rvert$	vector norm
$x_1(t), \ldots, x_n(t)$	independent variables of the regression function
y_m	measured output
y_p	predicted output
$y(t)$	dependent variable of the regression function
α	interquartile range
α_i and α_i^*	Lagrangian multipliers
β	random variable of the regression function
$\beta_0, \beta_1, \ldots, \beta_n$	regression coefficients
γ	outlier factor
ϵ	epsilon insensitive loss function
σ	bandwidth of the Pearson width
ω	tailing factor of the curve fitting peak
ANFI	adaptive neural fuzzy inference
ANN	artificial neural network
AR	autoregressive
ARMA	autoregressive moving average
ARX	autoregressive exogeneous
CC	correlation coefficient
CO_2	carbon dioxide
DW	day of the week
EBP	error back propagation
GA	genetic algorithm
GS	grid search
HD	hour of day
HVAC	heating, ventilation, and air conditioning
M	month of the year
MAPE	mean absolute percentage error
MLP	multilayer perceptron
MLR	multiple linear regression
PCA	principal component analysis
PIR	passive infrared receiver
PM	polynomial model
PUK	Pearson VII universal kernel
STLF	short term load forecasting
SVM	support vector machines
SVR	support vector regression
WD	working day
WSN	wireless sensor network

suitable results. The experiment has been carried out using real data collected with a WSN.

Although, in smart building management it is necessary to forecast the load in the future, this paper focuses only on what is essential for the prediction of consumption, thus avoiding possible errors due to the prediction of the variables.

With the advance of computers, drawbacks like difficult parametrization, selection of variables and over-fitting have been solved. Furthermore, the parallel processing and the performance of the present computers will help to decrease the computational time. However, reducing the size of the database will diminish the computational cost. The features of the presented STLF method for non-residential buildings are: high accuracy, low computational requirements, low over-fitting and, minimum data collected and use as simple a model as possible.

The paper starts with related works and follows with background material. Then, the dataset is explained. This is followed by a presentation of the methodology. Next, the results are presented and the method is discussed. Finally, conclusions are shown.

2. Related works

The present state-of-the-art knowledge focuses on STLF in non-residential buildings. The analysis of the papers that follow is organized according to the following aspects: model type (MLR, ANN, etc.), used variables (load, weather, etc.), building type (malls, offices, university campus, simulated buildings, etc.) and climate type (Mediterranean, Oceanic, Continental, etc.). The works are organized in two major blocks: the works with only one tested model and the works with multiple tested models.

In the first major block, there is the case of [14], where PCA is used, through climate data, to create a new climate index *Z*. Then, this index, with MLR, is used to estimate electricity consumption. Concerning ANN, study [11] proposes to use consumption and weather data to predict load consumption in offices. This study says that variables such as temperature and solar radiation are important while the wind speed or humidity can be omitted. Ref. [7], with synthetic data of consumption and weather, states that the main virtue of ANN is simplicity. In Ref. [16], with the same data type as two previous cases but for a hotel, a type of ANN called ANFI, optimized via GA, is used. This improves ANN performance. In case [4], where the temperature and load are used to predict the consumption of a set of university buildings, a small set of similar days is selected to train a ANN based on the work activity and temperature. Both, in [25] and [18] cases, in addition to consumption and weather data, calendar related data are used to forecast consumption in offices and university buildings, respectively. The work [25] highlights the differences that can be seen in the results between real data and synthetic data, while [18] indicates that the effect of humidity and solar radiation in the prediction is smaller than the effect of the outside temperature. In [17,13], weather, consumption and a variable related to the occupancy of the building are used. Case [17], a library located in China, also presents an ANFI model optimized through GA. On the other hand [13], discusses the difficulty of obtaining data of the real occupancy of a building and which alternatives there are to collect such data in a mall in Hong Kong. Besides, Kwok et al. [13] mention that the data related to the occupancy of a building significantly improves the prediction accuracy of the MLP case presented. Related to SVR papers, the example [28] uses synthetic data of consumption, weather and occupancy and provides a new method of feature selection. Another example is [27] that with the same type of synthetic data as in the previous case, proposes a new SVR model that improves performance.

With regard to papers that contrast several methods, there are three works by the same authors. In the first paper [19], they compare mainly ANN, SVR, AR and PM for the case of a campus by using consumption, weather and calendar data. The conclusion is that the AR with daily work schedules gives the best results. In addition, they conclude that the methods should be simple and not require a difficult process of trial and error. The following work [5], compares ANN, SVR, AR and a PM for the same Campus as in the previous case, using consumption and weather data. The interpretation is that climatic variables have low influence on the variation of the load. In this case, they have used a variable sliding window to reduce the computation time. As in the two previous cases, the model that gives the best result is the AR, which is also the easiest and fastest. In the fourth study by the same authors [1], the same models as above are compared, using consumption and calendar data, but in this case selection and combination models have been applied. The conclusion is that the combination, in short increasing the complexity of the model, is only necessary if there is a model that clearly exceeds another in performance. In Ref. [10], they use consumption and weather data, for a mall, to make a comparison between the following models: ARMA, ANN and MLR. They suggest that the ANN has better accuracy but cannot be interpreted, while the MLR has lower performance but allows some interpretation.

In summary most of the papers are based on a set of instances and attributes, more or less extensive, on which they apply filters. Then one or several models are used in order to obtain the best performance. With regard to the type of model, most modern works use ANN. Related to the kind of variables, there is a widespread use of the temperature variable. Most of the studied buildings are campuses and the Oceanic climate is the most common among the analysed papers. Concerning the building occupancy, there are not many studies with real data and this also happens with the indoor ambient data.

Thus, our work forecasts the electric load of an office building at the University of Girona (Catalonia) with a Mediterranean climate. Three types of models have been contrasted: MLR, MLP and SVR. Unlike other studies, such as [17], that use opening schedules of each classroom of the library, [27,28], that use synthetic occupancy data, or [13], that calculates occupancy through the consumption of the primary air units, our proposal calculates the occupancy through real measures of the volume of entries and exits of people in the building. In addition, unlike other papers, our model adds six measurement points to reflect the atmosphere inside the building. We have installed temperature, humidity and light sensors to analyse if this kind of data can improve consumption forecasting. In contrast with other works, that use complex methods and large databases, our model has low computational requirements and a minimal database. In addition, to solve common deficiencies, the method contains weather data to react to sudden weather changes and occupancy data, to respond to unexpected human events.

3. Background

Some concepts related to the tested models are introduced before the next sections.

3.1. Models

Three paradigmatic models (MLR, MLP and SVR) are explored in order to find the best one.

3.1.1. Multiple linear regression

In the MLR case [3] the system is described as a linear equation with several independent variables and one dependent variable. The dependent variable in a given time is explained as a set of independent variables. The model is expressed with the form of Eq. (1):

$$y(t) = \beta_0 x_0(t) + \beta_1 x_1(t) + \cdots + \beta_n x_n(t) + \beta(t) \tag{1}$$

Although MLR is not the most suited method to deal with autocorrelated data (as could be ARMA or ARX), it has been chosen as a well-known and simple regression method as a basis for comparison. Also, MLR is according to the chosen methodology, unlike the time series.

3.1.2. Multilayer perceptron

The MLP is an ANN [26] formed by highly interconnected nodes and organized in layers. The structure of an MLP is: input layer, one or more hidden layers and output layer.

The input layer is where the attributes are connected. The hidden layers make non-linear mapping of the inputs. The output layer serves the output values of the MLP. The inputs pass through the different layers to the exit. Each neuron receives all the weighted inputs of the preceding layer and transforms the linear combination of it with the activity function.

The learning process is made with the aim of adjusting the weights value of each node. Among the different methods of learning, the most widely used due to its efficiency, is the EBP.

3.1.3. Support vector regression

The aim of SVM, Vapnik's idea [24] created in 1996, is to find a hyper-plane to classify data. In order to separate two classes which are not linearly separable, data is transformed using kernels functions and moved to a high dimensional feature space where data are linearly separable.

So, in the same way, SVR makes a non-linear mapping of the data with kernel functions and then proceeds to the linear regression in this new high dimensional feature space. The SVR function is detailed in Eq. (2):

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(x_i, x) + b \tag{2}$$

There are different kernel functions with different features. The PUK [23] function has the form of Eq. (3):

$$K(x_i, x_j) = \frac{1}{\left[1 + \left(2\sqrt{\left\|x_i - x_j\right\|^2}\sqrt{2^{(1/\omega)} - 1}/\sigma\right)^2\right]^{\omega}} \tag{3}$$

3.2. Parameter optimization

In order to obtain accurate models, the parameters of the models must be set carefully. There are several optimization methods, such as evolutionary algorithm, GA or particle swarm optimization. Given the number of attributes and instances of the case, we chose a simple and fast method: GS. If the volume of data was larger, we would use more sophisticated methods.

4. Dataset

The university office building has three floors, a volume of 4000 m^3 per floor. It was built in 2003. The front has an area of 1836 m^2, of which 630 m^2 are glass surface.

In relation with HVAC, the heating system (824 kW) consists in gas boilers (Ygnis) and fancoils. The cooling system (618 kW) is composed by a compression refrigeration system (Climaveneta) and fancoils. The temperature set point is adjusted manually. The setting temperature of winter is 20°C and the summer's one is 26°C. The HVAC control system detects the temperature of the offices and classrooms and modifies the fan speed to achieve the set point temperature.

We have four different sources of information: electrical load data, weather data, indoor data and calendar data. The sampling rate is hourly. The total number of instances is 7616, covering a total of 11 months, from 13th May, 2013 to 26th March, 2014.

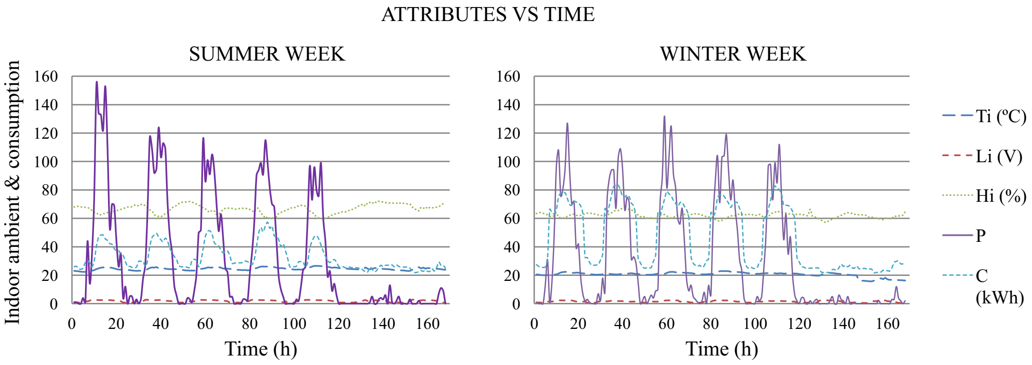


Fig. 1. Indoor ambient and consumption data for summer and winter weeks.

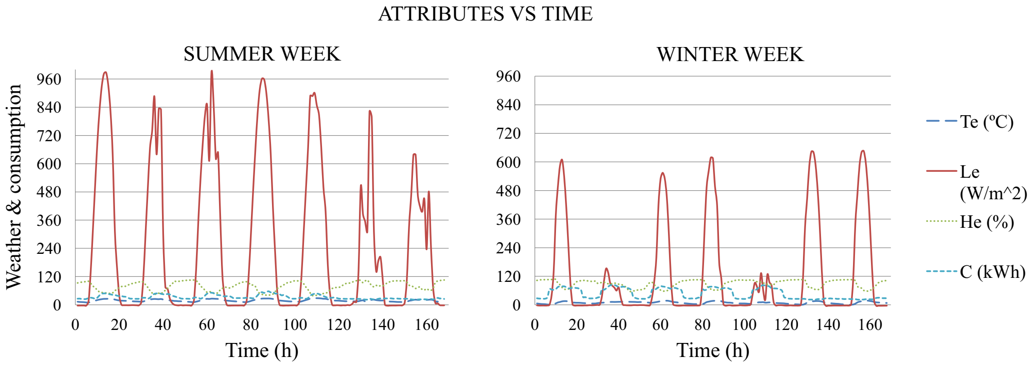


Fig. 2. Weather and consumption data for summer and winter weeks.

- Electric load data: we have collected electrical load data of the building with a power meter (PM810 Power Logics of Schneider) from the campus infrastructure monitoring system.
- Weather data: the Department of Physics has collected data of temperature (Vaisala), relative humidity (Humicap) and solar radiation (Kipp & Zonen) outside the building.
- Indoor data: in addition, with the aim of studying how the indoor ambient and the occupancy of the building influences the model, we have also installed a wireless sensor network. Among the possible solutions, we have chosen the most flexible and easiest to install with the longest battery life. The set consists of several Waspmotes (sensor and emitter) and one Meshlium (receiver and database) of the Libelium trademark. The network is configured as a star topology. Due to the characteristics of consumption, safety, cost and communication distance we have chosen ZigBee technology [12].

  In short, we have installed six electronic badges of sensors (Waspmotes) inside the building, placing them uniformly, that measure temperature (MCP9700A), relative humidity (808H5V5) and light level (PDV-P9203). There are two more electronic badges with a PIR sensors located on the two entrances of the building. We calculate the building occupancy measuring the volume of entries and exits into the building.
- Calendar data: we have also nominal attributes to predict consumption such as: hour of the day, day of the week, month and working days.

In summary, we have a total of 28 attributes, 24 of them numerical and 4 nominal. Among 24 of these attributes there are 6 indoor temperatures, 6 indoor relative humidities, 6 indoor luminosities, 1 outdoor temperature, 1 outdoor relative humidity, 1 solar radiation and 2 occupancies. The 4 calendar nominal attributes are: hour of day, working day, day of the week (contains information about the user behaviour that operates like a schedule) and month of the year (give us information related to the seasons).

For a week in summer and one in winter, Fig. 1 shows the average values of the inner measured data and Fig. 2 shows weather data.

5. Methodology

The proposed methodology is summarized in Fig. 3. First, we make a feature selection by removing insignificant attributes. In the next block, we filter missing values and outliers. At this point, we normalize and then separate training (66%) and test data (33%). The set of training data is randomly re-sampled to 30% of the data. With this 30% we perform a GS to find the parameters and the performance of the models. Then, we execute a sub-sample validation

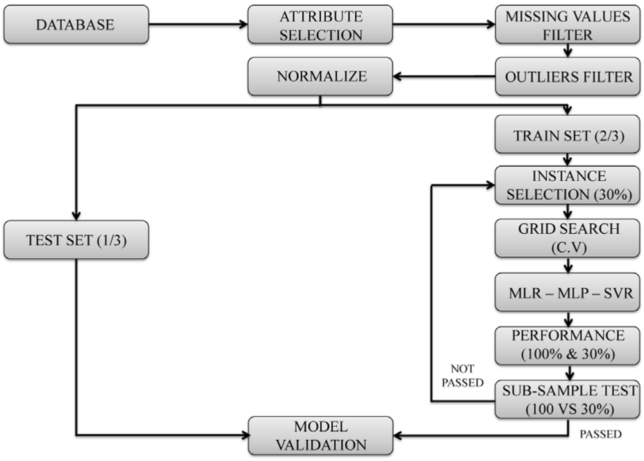


Fig. 3. Block diagram of the process.

Table 1
Feature selection configuration.

Starting point	Search organization	Evaluation strategy
Forward	Genetic search	Correlation-based feature selection

Table 2
Values of the parameters of the MLR.

Ridge	Attribute selection method
1.0E−8	M5

test for each model in order to ensure that it is a representative sub-sample. If we pass the sub-sample validation test we select the candidate that gives the best performance for each model (MLR, MLP and SVR). Finally, we validate it with the test data.

5.1. Attribute selection

Removing the redundant attributes, which contain irrelevant or noisy data, we achieve a faster learning process and an accurate and more compact forecasting model. The feature selection process, explained in [15], consists of a search through the space of feature sub-sets. The following aspects must be taken into account:

- The starting point is the point in the features' sub-set space from which the search starts: forward (adding attributes), backward (removing attributes) or bidirectional (starting from one point).
- The heuristic search strategies provide proper results, not the optimal one, as does the prohibitive exhaustive search. There are different search strategies: greedy hill climbing, best search, genetic algorithms, etc. Genetic algorithms, described in [6], are based on the principles of natural selection.
- The evaluation strategy, as can be seen in [9], is how feature sub-sets are evaluated. In our case the value of a feature sub-set will be considered, taking into account the maximum individual predictive capacity and minimal redundancy among them.

After testing empirically with several configurations, the values that have given best results are shown in Table 1.

5.2. Missing values

Due to errors in sensor readings or external troubles there is always a small percentage of lost samples. We must try to minimize the percentage of missing values as much as possible. The missing values can be treated in different manners such as avoiding instances that include them, making the average of close values or other more sophisticated methods based on PCA. In our case, after testing several configurations, we have found that removing instances that include missing values gives the best result. The total number of instances containing missing values is 1690 out of a total of 7604, which represents 21.16%.

5.3. Outliers

We can delete or substitute outliers by a certain value. Eliminating these values improves the performance of the model significantly. However, the more restrictive, the more amounts of data lost. It is a trade-off between the range of values considered outliers and the goodness of the model. We have used the filter of Eq. (4) to detect outliers:

$$Q_3 + \gamma \cdot \alpha < x < Q_1 - \gamma \cdot \alpha \tag{4}$$

The total number of deleted instances with outliers is 113 of a total of 7604, this represents 1.49%. After testing empirically with a range of values, $\gamma = 3$ has given the best results.

5.4. Normalization

Normalization is an important step when we have parameters of different units and scales. Usually, all parameters should have the same scale for a suitable comparison, enhancing the model performance. We have normalized using Eqs. (5) and (6):

$$x_{in} = n \frac{x_i}{|x_p|} \tag{5}$$

$$|x_p| = \left(\sum_i (x_i)^p \right)^{1/p} \tag{6}$$

After testing empirically with several ranges of parameters, $n = 1$ and $p = 2$ have given the best results

5.5. Instance selection

Because the training process of the model requires high computational cost, we have employed a sample of the complete dataset. It helps us to quickly find the optimal parameters of the model. The sample is selected randomly, therefore it is necessary to validate it. After testing empirically with several percentages of re-sampling, a sample of 30% led us to achieve a balance between reducing computational cost and maintaining the level of accuracy of the resulting models.

5.6. Grid search

GS method performs a search through the ranks of pairs of parameters and chooses the best. In the first grid search it uses only 2-fold cross validation to determine the best pair of parameters. Then, it takes the best point in the grid with the adjacent parameter pairs and makes a 10-fold cross validation. If there is no better pair the search stops here. But, if a better parameter pair is found, this pair is the next centre of the search. The process continues until there is no better pair or the pair is on the border of the grid.

The main parameters that we have analysed with the GS are the following ones: ridge parameter for MLR, hidden layers, learning rate, momentum and training time for MLP and complexity parameter, different types of kernel and the parameters for each kernel for SVR.

5.7. Models

In the next three sections we explain the questions related to the tested models.

5.7.1. Multiple linear regression

In the MLR case there is a set of parameters to specify as ridge or attribute selection method. The ridge helps to regularize the model and to avoid over-fitting problems. We need a generalized model with high accuracy with the test data. In relation to attribute selection methods, we have tested two widely used methods. The M5 method removes the attributes with the smallest standardized coefficient until no improvement is observed and the greedy selection method uses the Akaike criterion metric. The Akaike criterion handles a tradeoff between the performance and the complexity of the model.

After testing empirically with several ranges of parameters, using GS, the values that have given best results are shown in Table 2.

Table 3
Values of the parameters of the MLP.

Learning rate	Training time	Hidden layers	Momentum	Nodes
0.3	5000	1	0	15

5.7.2. *Multilayer perceptron*

In the MLP case there is a set of parameters to specify as learning rate, training time, hidden layers, momentum or number of nodes. The number of hidden layers of the network depends on the dataset. There is a hard relation between the complexity of the network and the computational cost to calculate it. The network has several nodes in the hidden layer besides the node of the output layer. All the nodes in this network are sigmoid. We chose the EBP method to train the ANN because is a widely used successful method. A suitable learning rate lets the weights converge quickly to the solution. A too large learning rate does not allow the network to converge but if it is too small it will take a great deal of time. The learning rate typically ranges from 0.2 to 0.8. The training time is the number of epochs to train the network. A long training time will enhance the performance but will increase the computational cost. Momentum helps to control the instabilities and local minimums and increases the rate of convergence.

After testing empirically with several ranges of parameters, with GS, the values that have given best results are shown in Table 3.

5.7.3. *Support vector regression*

In the SVR case there are several parameters [22] to specify such as the *C* parameter, kind of kernel and kernel parameters. To allow flexibility, the SVM handles the *C* parameter that controls the trade-off between training errors and rigid margins, creating a soft margin that allows, but penalizes, some mis-classification. With a large value of *C* the hyperplane has a small margin that will classify all the training points correctly. On the other hand, with a small value of *C* the hyperplane has a large margin that will perhaps mis-classify more points. There is no clear theory about setting *C*. Although, a reasonable proposal is to select a *C* parameter around the range of output values.

Before choosing the PUK function we have tested the most common kernel functions, such as linear, polykernel and radial basis function. PUK has a high mapping and generalization capacity and robustness. It is one of the most suitable kernels with which to face a variety of mapping problems. PUK has two adjustable parameters: σ and ω. Varying the parameter σ the kernel turns from a Gaussian shape to a Lorentzian shape. The internal parameters of the SVR model can be obtained using various algorithms. Among them, RegSMOImproved [21], due to its short implementation time, is the chosen one. The main parameter of the RegSMOImproved learning algorithm is ϵ. ϵ is related to accuracy of the approximated function. A too large ϵ will generate poor results and too small will generate over-fitting.

After testing empirically with several ranges of parameters, with GS, the values that have given the best results are shown in Table 4.

5.8. *Sub-sample validation test*

To avoid training the model with 100% of the data, given that that has a high computational cost, we have used random sub-samples. In order to determine that these samples maintain the internal distribution of the initial population we carry out a simple test with the performances of each sub-sample. Once we have the three best results for the sub-sample, we use the parameters that provide these results to calculate the models and their performances for the initial dataset. Then, we create two rankings of three positions depending on the performances of each dataset. Once done, we check that the position of the ranking for each set of parameters is the same for the sample as for the initial dataset. If so, the sample is valid and these are the selected parameters. Otherwise, we take another sub-sample and repeat the process.

Table 4
Values of the parameters of the SVR.

C	Kernel	ω	σ	ϵ
400	PUK	0.2	10	0.001

Table 5
All attributes and all instances.

Model	CC	MAPE	Computation time
MLR	0.1755	24.3%	3 s
MLP	0.2463	23.72%	1843 s
SVR	0.964	14.32%	7546 s

6. Experimental results

We have realized the experiments with Weka software [8] and using a computer with Intel Core i7-4500U processor and 8 GB of DDR3 RAM. The next sections describe the indicators used to measure the performance, the scenarios considered and the results obtained.

6.1. *Error indicator*

There are several procedures to calculate the quality of the model; MAPE and CC are two of the most common indicators in the actual literature. The MAPE, shown in Eq. (7), does not depend on the magnitude of the unit of measurement and is usually used for comparing models:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{m(i)} - y_{p(i)}}{y_{m(i)}}\right| \times 100 \tag{7}$$

The CC is a measure of the linear relation between two variables; in this case between real output and forecasted output.

6.2. *Results*

We first analyse the results of several scenarios and then we plot the results of the best scenarios.

6.2.1. *Scenarios*

In this section we analyse the results given by MLR, MLP and SVR models in different scenarios. For each scenario we provide four parameters: the performance of the model (MAPE and CC), the computation time and the attributes used. The chosen model is the one that presents the best trade-off between several indicators. The forecasting accuracy, the computation time and the number of used attributes are the most important indicators, although another interesting issue is the intelligibility of the model; that is, the information provided by the model by the naked eye. The MLR has a high level of intelligibility, while the MLP and the SVR have not.

Scenario 1. We begin with all the attributes: T_{1-6}, H_{1-6}, L_{1-6}, T_e, H_e, L_e, P_{1-2}, *M*, WD, HD and *W* and all the instances, as shown in Table 5.

In the first experiment, due to noise and redundant data, results have shown low performance indicators and large computation times. Although an SVR model achieves a high forecasting level, the unfiltered data somewhat increase the computational cost.

Scenario 2. Now we analyse, in Table 6, the effect of feature selection. In this experiment we have the following attributes: T_e, P_{1-2}, WD, HD and all the instances:

Table 6
Filtered attributes and all instances.

Model	CC	MAPE	Computation time
MLR	0.9138	4.68%	2 s
MLP	0.9992	0.97%	694 s
SVR	0.9978	0.35%	4983 s

Table 7
Filtered attributes with filtered instances.

Model	CC	MAPE	Computation time
MLR	0.9156	4.83%	1 s
MLP	0.9996	0.45%	600 s
SVR	0.9998	0.36%	4255 s

Table 8
Outdoor temperature and calendar attributes with filtered instances.

Model	CC	MAPE	Computation time
MLR	0.7335	6.26%	2 s
MLP	0.8009	8.24%	744 s
SVR	0.8716	3.43%	4628 s

Table 9
Outdoor temperature and occupancy attributes with filtered instances.

Model	CC	MAPE	Computation time
MLR	0.913	5.21%	1 s
MLP	0.9866	1%	41 s
SVR	1	0.06%	250 s

Results have indicated that the consequence of the attribute selection is an accuracy enhancement and a significant reduction of the computation time for all the models.

Scenario 3. In Table 7, with the same attributes as in the previous case, we observe the effect of instance filtering:

By treating the instances, the performance does not change much but the computation time is improved slightly. SVR and MLR are the models most positively affected by the instance filtering.

Scenario 4. From now on, we test several sets of attributes, always with filtered instances, to analyse which set produces better results and to discover how much information is contained in each set. In Table 8, we have tested the model with the following attributes: outdoor temperature and calendar indicators (T_e, M, WD, HD and W).

From the previous experiment, we can see that removing the occupancy attribute the performance is reduced significantly. The calendar attributes have less information than the occupancy attribute.

Scenario 5. In the Table 9 we have performed the experiment only with outdoor temperature and occupancy attributes (T_e and P_{1-2}) in order to confirm the previous experiment.

Taking into account the results, removing calendar attributes, MLR and MLP models have slightly worsened the performance but have improved the calculation time. However, the SVR model has an excellent forecasting accuracy and a really low computation time with only two attributes.

Table 10
Indoor ambient attributes with filtered instances.

Model	CC	MAPE	Computation time
MLR	0.6742	5.37%	1 s
MLP	0.799	3.95%	197 s
SVR	0.7291	4.9%	1218 s

Table 11
Temperature attribute with filtered instances.

Model	CC	MAPE	Computation time
MLR	0.1152	11.23%	1 s
MLP	0.0875	16.24%	22 s
SVR	0.1604	13.95%	216 s

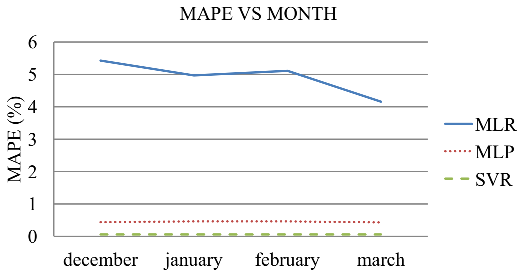


Fig. 4. MAPE for the months of the year.

Scenario 6. In Table 10 we analyse the forecasting capacity of the indoor ambient variables (T_{1-6}, H_{1-6} and L_{1-6}) when they are used alone:

Using indoor ambient variables, without calendar, occupancy or weather attributes, the performance and the computation time absolutely deteriorate.

Scenario 7. As shown in Table 11, we have tested the outdoor temperature (T_e) attribute by itself, to analyse its ability to forecast:

The temperature attribute alone, without the occupancy attribute, has given deficient accuracy indicators. So, finally, we chose the SVR model with only outdoor temperature and occupancy attributes (T_e and P_{1-2}) because it has given the best results; a trade-off between high forecast accuracy, low computation time and a low number of attributes.

6.2.2. Graphic results

In the present section we plot the average of MAPE vs months, days and hours for each model in its best scenario.

Graphic 1. In Fig. 4, we can see a MAPE vs month graphic for the three models.

The MLP and SVR models, unlike the MLR model, not only predict better but also suffer less MAPE variation in terms of the months. Heterogeneous months, as December, are hard to predict.

Graphic 2. In Fig. 5, we can observe a MAPE vs day graphic for the three kind of models.

Given that we are analysing a non-residential building and there is no consumption on the weekends, the performance on Saturday and Sunday is higher. In addition, the days with the worst forecasting level, because of their variability, are Mondays and Fridays.

Graphic 3. In Fig. 6, we can analyse a MAPE vs hour graphic for the three selected models.

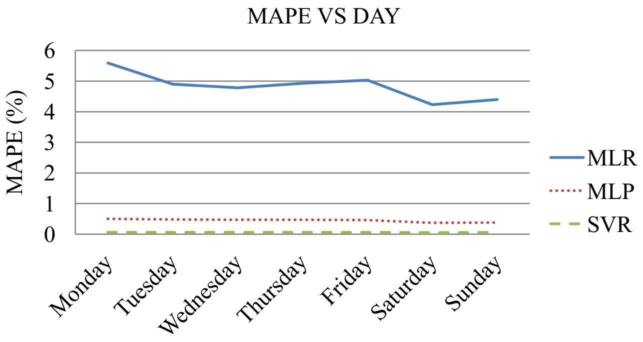


Fig. 5. MAPE for the days of the week.

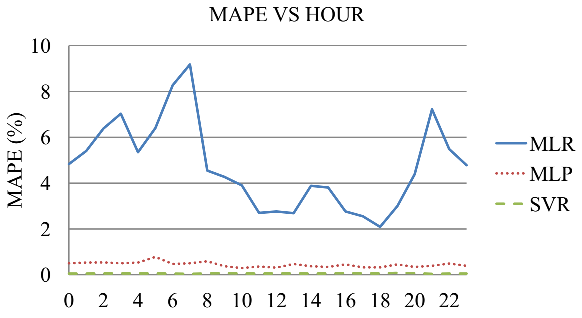


Fig. 6. MAPE for the hours of the day.

In the MLR case and, to a lesser extent, in MLP and SVR cases, we observe that night forecasting has the worst accuracy because of the hourly variability of the security and cleaning services. During the day we have three low accuracy prediction zones: the beginning and ending of the working day and lunchtime.

7. Discussion

We have contrasted three models, considering various parameters, for different sets of attributes. MLR and MLP models provide their best results using temperature, calendar and occupancy attributes. On the other hand, SVR provides its greatest results using only temperature and occupancy data. The MLR model, unlike the others, is intelligible, serving information about the relevance of every attribute. However, MLR has a lower level of goodness than MLP and SVR, which perform better against non-linear systems. The SVR model provides the highest accuracy with an acceptable computational cost but has a bad response using nominal attributes.

All the models with real occupancy data enhance the forecasting accuracy above the other with calendar data. The indoor ambient data, because of fixed HVAC operating conditions, does not provide a prediction performance improvement.

The working days of the week that have a heterogeneous profile, as does Friday, have a lower level of forecasting precision. However, weekends provide better results because they are flat and homogeneous.

Models with few attributes, thanks to the feature selection, have a lower computation time with a similar level of prediction accuracy. Filtering the instances means that the computation time is reduced and the accuracy of the model is improved.

8. Conclusions

One of the most important challenges of the utilities is to equilibrate the power generation and the consumption in the building sector, for this reason we need to predict both. So, with the aim of generating an STLF model for a non-residential building at the University of Girona we have installed a WSN to collect data. We have analysed weather, indoor ambient, calendar and real building occupancy attributes to determine which are the most relevant. Then, we have tested MLR, MLP and SVR models with different sets of attributes. So, the SVR model, with only temperature and real occupancy data, is the one that has given the best balance of accuracy and computational cost. By reducing the number of attributes, the price of the WSN installation, the size of the database and the computational cost are reduced. Therefore, the system requirements to perform the data mining process will be less demanding.

Our research leads us to discover that the ambient data, collected inside the building, do not improve the level of performance of the model because of the fixed HVAC operation condition. In the case of changes in the HVAC operation condition it would be interesting to add an internal temperature variable in the model.

The simple and economic system of measuring the occupancy, using PIR sensors on the entries of the building, provides real information of building occupancy and improves the performance of the resulting models. Furthermore, it allows maintaining an acceptable forecasting level in front of unexpected events. Other methods for detecting occupancy of buildings, like the use of cameras or the determination of CO_2 levels, are much more expensive and sophisticated.

The proposed methodology is simple, fast and, using a small number of attributes, provides a satisfactory forecasting accuracy with a low computational cost. In addition, the implemented WSN installation is economical, minimal and effortless.

In future works, several types of non-residential buildings will be tested and a method for generating simple occupancy indexes from work schedules will be proposed, having all the model data in advance.

Acknowledgements

This research has been partially supported by the Spanish Government project MESC (Ref. DPI2013-47450-C2-1-R). Also we would like to thank the Department of Physics and acknowledge the technical assistance and maintenance service of the UdG (SOTIM) which provided the weather and consumption data respectively. The authors belong to the 'Smart IT Engineering and Solutions' accredited research group (Generalitat de Catalunya, 2014 SGR 1052).

References

[1] C.E. Borges, Y.K. Penya, I. Fernández, Optimal combined short-term building load forecasting, in: Innovative Smart Grid Technologies Asia (ISGT), 2011 IEEE PES, IEEE, DeustoTech (Energy Unit) University of Deusto, Bilbao, Spain, 2011, pp. 1–7.

[2] X.G. Casals, Analysis of building energy regulation and certification in Europe: their role, limitations and differences, Energy Build. 38 (5) (2006) 381–392.

[3] N.R. Draper, H. Smith, Applied Regression Analysis, 2nd ed., John Wiley and Sons, 1981.

[4] G. Escrivá-Escrivá, C. Álvarez Bel, C. Roldán-Blay, M. Alcázar-Ortega, New artificial neural network prediction method for electrical consumption forecasting based on building end-uses, Energy Build. 43 (11) (2011) 3112–3119.

[5] I. Fernandez, C.E. Borges, Y.K. Penya, Efficient building load forecasting, in: ETFA 2011, 2011, pp. 1–8.

[6] D.E. Goldberg, J.H. Holland, Genetic algorithms and machine learning, Mach. Learn. 3 (2) (1988) 95–99.

[7] P.A. González, J.M. Zamarre no, Prediction of hourly energy consumption in buildings based on a feedback artificial neural network, Energy Build. 37 (6) (2005) 595–601.

[8] M. Hall, E. Frank, G. Holmes, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.

[9] M.A. Hall, Correlation-Based Feature Selection for Machine Learning, The University of Waikato, 1999 (Dissertation).

[10] V.A. Kamaev, M.V. Shcherbakov, N.L. Panchenko, D.P. Shcherbakova, A. Brebels, Using connectionist systems for electric energy consumption forecasting in shopping centers, Autom. Rem. Control 73 (6) (2012) 1075–1084.

[11] S. Karatasou, M. Santamouris, V. Geros, Modeling and predicting building's energy use with artificial neural networks: methods and results, Energy Build. 38 (8) (2006) 949–958.

[12] P. Kinney, Zigbee technology: wireless control that simply works, in: Communications Design Conference, 2003.

[13] S.S.K. Kwok, R.K.K. Yuen, E.W.M. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Build. Environ. 46 (8) (2011) 1681–1690.

[14] J.C. Lam, K.K.W. Wan, S.L. Wong, T.N.T. Lam, Principal component analysis and long-term building energy simulation correlation, Energy Convers. Manage. 51 (1) (2010) 135–139.

[15] P. Langley, Selection of Relevant Features in Machine Learning, Defense Technical Information Center, 1994.

[16] K. Li, H. Su, Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system, Energy Build. 42 (11) (2010) 2070–2076.

[17] K. Li, H. Su, J. Chu, Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: a comparative study, Energy Build. 43 (10) (2011) 2893–2899.

[18] A.H. Neto, F.A. Sanzovo Fiorelli, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy Build. 40 (12) (2008) 2169–2176.

[19] Y.K. Penya, C.E. Borges, I. Fernandez, Short-term load forecasting in air-conditioned non-residential buildings, in: IEEE Africon'11, 2011, pp. 1–6.

[20] J.-C. Piris, The Lisbon Treaty: A Legal and Political Analysis, Cambridge University Press, 2010.

[21] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, Improvements to the SMO algorithm for SVM regression, IEEE Trans. Neural Netw. 11 (5) (2000) 1188–1193.

[22] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.

[23] B. Üstün, W.J. Melssen, L.M.C. Buydens, Facilitating the application of support vector regression by using a universal Pearson VII function based kernel, Chemometr. Intell. Lab. Syst. 81 (1) (2006) 29–40.

[24] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 2000.

[25] J. Yang, H. Rivard, R. Zmeureanu, On-line building energy prediction using adaptive artificial neural networks, Energy Build. 37 (12) (2005) 1250–1259.

[26] B. Yegnanarayana, Artificial Neural Networks, PHI Learning Pvt. Ltd., 2009.

[27] H.-x. Zhao, F. Magoules, New parallel support vector regression for predicting building energy consumption., in: 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM), 2011, pp. 14–21.

[28] H.-x. Zhao, F. Magoulès, Feature selection for predicting building energy consumption based on statistical learning method, J. Algorithms Comput. Technol. 6 (1) (2012) 59–78.

Chapter 6

Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes

In this chapter, a simple, low-computational requirements and economical hourly consumption prediction method, based on SVR model and only the calculated occupancy indicator as attribute, is proposed. In addition, and due to the relevance of the occupancy indicator in the model, this paper provides a complete study of the methods and data sources employed in the creation of the artificial occupancy attributes. Thus, several occupancy indicators are defined, from the simplest one, using general information, to the most complex one, based on very detailed information. This work has been published in the following paper:

Paper published in the **Energy and Buildings**
Volume: 130, Pages: 519-531, Published: October 2016
DOI: 10.1016/j.enbuild.2016.08.081
JCR CE IF: 2.973 Q1(6/126)

Energy and Buildings 130 (2016) 519–531

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes

Joaquim Massana*, Carles Pous, Llorenç Burgas, Joaquim Melendez, Joan Colomer

University of Girona, Campus Montilivi, P4 Building, Girona E17071, Spain

ARTICLE INFO

Article history:
Received 9 May 2016
Received in revised form 24 July 2016
Accepted 27 August 2016
Available online 31 August 2016

Keywords:
Load forecasting
Support vector machines
Sensor data
Mediterranean climate
Occupancy indicator

ABSTRACT

An accurate short-term load forecasting system allows an optimum daily operation of the power system and a suitable process of decision-making, such as with regard to control measures, resource planning or initial investment, to be achieved. In a previous work, the authors demonstrated that an SVR model to forecast the electric load in a non-residential building using only the temperature and occupancy of the building as attributes is the one that gives the best balance of accuracy and computational cost for the cases under study. Starting from this conclusion, a simple, low-computational requirements and economical hourly consumption prediction method, based on SVR model and only the calculated occupancy indicator as attribute, is proposed. The method, unlike the others, is able to perform hourly predictions months in advance using only the occupancy indicator.

Due to the relevance of the occupancy indicator in the model, this paper provides a complete study of the methods and data sources employed in the creation of the artificial occupancy attributes. Several occupancy indicators are defined, from the simplest one, using general information, to the most complex one, based on very detailed information. Then, a load forecasting performance discrimination between the artificial occupancy attributes is realized demonstrating that using the most complex indicator increases the workload and complexity while not improving the load prediction significantly. A real case study, applying the forecasting method to several non-residential buildings in the University of Girona, serve as a demonstration.

1. Introduction

In order to build a fair and more sustainable society, new approaches and initiatives have appeared in all areas. Energy resources are limited, and there is the need to generate new technologies and legislation that allows to achieve a certain environmental balance. The Lisbon Treaty [30] and the Kyoto Protocol [10] are examples of legal initiatives that have the aim of reducing consumption and emissions. To reduce the consumption, it is necessary to improve the existing electricity grid making it more efficient and robust. The smart grid, in conjunction with decentralized power generation, could avoid many of the shortcomings of the classical electrical grid.

Thus, to increase the efficiency of the electricity grid, a balance of power generation is required such that there is no waste or lack of resources. Due to the apparition of micro-grids, there is a balance between the generation of power and the users' consumption. Given that buildings are responsible for a large part of the electricity consumption, having tools to predict their consumption is key in the adjustment process. Predicting the consumption of a city is different from predicting the consumption of a building, in that in the case of buildings there is much variability. Disaggregated environments are more difficult to predict. Thus, short-term load forecasting (STLF) methodology is used to reduce the building's consumption since it must deal with non-linearities and noise.

Recent research on energy efficiency in buildings include optimal decisions and an overall improvement in human behaviour, not just technology. The International Energy Agency's Energy in Buildings and Communities Programme (IEA-EBC) has recently completed a project related to strengthening the robust prediction of energy usage in buildings, with the goal of enabling the proper assessment of short and long-term energy measures, policies and technologies. The results of this project are collected in Annex 53 [15]. The analysis methods, developed models and results of Annex 53 were taken as the starting point for other several working areas. In particular, and due to the important effect of occupancy in energy

* Corresponding author.
E-mail addresses: joaquim.massana@udg.edu (J. Massana), carles.pous@udg.edu (C. Pous), llorenc.burgas@udg.edu (L. Burgas), joaquim.melendez@udg.edu (J. Melendez), joan.colomer@udg.edu (J. Colomer).

http://dx.doi.org/10.1016/j.enbuild.2016.08.081
0378-7788/© 2016 Elsevier B.V. All rights reserved.

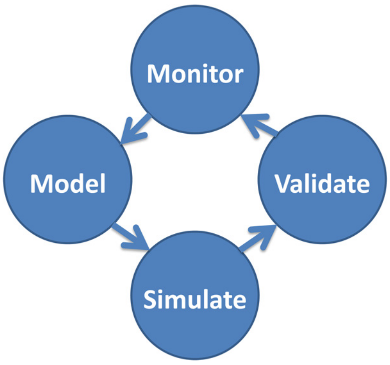


Fig. 1. Technical framework used in occupancy behaviour.

prediction, the IEA-EBC is working on Annex 66 [16]. On this annex they are trying to define and simulate occupant behaviour in a consistent and standard way. Based on these works, some new proposals have arisen, as is shown in [13]. The ontology represents energy-related occupant behaviour outlined as a DNAS (drivers, needs, actions and systems) framework, providing a systematic representation of energy-related occupant behaviour in buildings. Generally, researchers working on this topic follow a methodology that consists of monitoring, modelling and simulating, such as [13,14], as seen in Fig. 1.

These models are built after monitoring and collecting enough data about occupancy of the building. As stated in [14], this data is obtained from observational studies, occupant surveys and interviews, laboratory studies and unresolved issues in occupant monitoring, such as contextual factors. The occupancy models take into account the actions that occupant can do on the building, such as open the light, close the window, track or predict the occupant movements, and so on. It can be seen that the building must be sensorized to some extent to have this information available, a fact that is not always possible.

Although computing technology continues to develop, some forecasting models training on databases with dozens of attributes and millions of instances, may lead to high computational cost. Therefore, reducing the database is still necessary, always taking care to ensure that performance does not deteriorate. Most of the papers that propose the use of STLF methods in non-residential buildings often use weather data and, in some cases, occupancy information. Other works, such as [17,19,22], conduct comparison studies using similar models and arrive to a different conclusion, selecting other model as a best approach. The type of building and the test and training conditions can greatly affect the results. So, it is important to study different type of models in order to choose the best option in each case. According to [24], a model predicting consumption with minimal instances using support vector regression (SVR) with temperature and occupancy attributes provides excellent results for our buildings under study.

Obtaining predictions of temperature, in order to know the temperature of a particular place, is normally possible, although acquiring information of future occupancy remains difficult. In [24], occupancy information collected from passive infra-red (PIR) sensors was used. However, this information is not available in advance. The non-residential buildings usually dispose of work or scholar schedules, or other information about their occupancy. A technique designed to generate this information beforehand is needed. The goal is to obtain a model that is not dependent of any information unavailable months ago, such as previous consumption or temperature. This model can perform consumption predictions months ahead. Perhaps resulting accuracy level of the model may not be as good as the other works in this topic, but this is only a first step in this new direction.

Table 1
Occupancy related methods.

Method	Sources	Works
Calendar	Day types, months, etc.	[2,5,7,17,28,29,31,37]
Schedule	Work, student or use schedules.	[6,20,21]
Sensors	Motion, CO_2, noise, light, etc.	[8,18,23,24,26,27,33]
Expert knowledge	Surveys, interviews or inspections.	[19,25,35]

The aim of the work is to test the load forecasting accuracy using several occupancy indicators. It is not centred on occupant behaviour modelling, but estimating the occupancy is necessary, as it is one of the main factors that contributes to the accuracy of the SFTL. Concerning the occupancy estimation, we deal with buildings that are poorly sensed. That means there is not information about occupant actions, such as open/close the window, switch on/off the lights or plug a device, even if the actions are taken. There is information about scholar and working schedules, classrooms dimensions, expert knowledge, etc. Furthermore, there is only one of the buildings under study having sensors to estimate the amount of occupants inside the building by means of PIR sensors. Due to this limitation, several occupancy indexes have been defined using the available information.

In the first part of the paper, artificial occupancy indicators for the buildings are generated using different techniques and information available in advance such as academic calendars and work schedules. Then, SVR model is trained to forecast the consumption of the respective buildings, using these indicators of occupancy. Subsequently, an analysis of the relationship between the forecasting performance and the workload based on occupancy indicators, is performed. The idea is to show that there is a balance point in the artificial occupancy indicators, between forecast accuracy and workload. From a certain point on, increasing the complexity of the indicator does not improve significantly the prediction.

The paper starts with related works and follows with background material. Then, the dataset is explained. This is followed by a presentation of the methodology, where the several occupancy indicators are defined, and the test process explained. Next, the results are presented and the method is discussed. Finally, conclusions are shown.

2. Related works

There have been a large number of papers on the topic of STLF with regard to residential and non-residential buildings. The non-residential buildings are basically malls, schools, universities, hospitals and offices. Assuming that the use of information concerning the occupancy of buildings is key for improving prediction, the present state of the art focuses on the following topic: STLF in non-residential buildings based on occupancy data.

In the present state of the art, the advantages and disadvantages of the several methods associated with using the building's occupancy information in a prediction model are evaluated. The methods can be grouped into 4 blocks, as seen in Table 1.

In the first block, there are eight works that use calendar information. The first [2], is the case of a campus in Los Angeles that uses temperature and occupancy information, based on calendar data such as day of the week and holidays, with a regression tree model. In the paper [17], based on synthetic data and a non-residential building located in Athens (Greece), using meteorological data including temperature, solar flux, relative humidity and wind speed and the profile of the days of the week, the consumption is predicted using an ANN model. The work [28,29], in the campus of the University of Deusto (Spain), use weather data such as relative humidity, precipitation, temperature, wind speed and wind direction in conjunction with the use of types of day comprising Saturdays, working and non-working days using AR, ANN and SVR principally. The work [37], with regard to an office building in Hong Kong, uses weather data including temperature, solar radiation and relative humidity and also takes into account if it is a weekday or a weekend using an ANN. In [5], an ANN is trained to predict the consumption of a commercial office building in Iowa (USA) using weather data such as precipitation provability, rain indicator, outdoor dry-bulb temperature, outdoor relative humidity, wind speed and sky condition in conjunction with the use of day type indicator. The work [31] proposes a non-linear autoregressive model with exogenous inputs to forecast the load in a college campus in Texas (USA) employing weather variables including temperature, relative humidity and calendar information such as hour, day of the week or month. The paper [7] presents an ANN based on indoor and outdoor temperature and relative humidity and occupancy data including day type in a supermarket in UK.

In the second block, where schedules are used, the work [20] predicts the consumption of the university library in Zhejiang (China) using temperature data and an index of occupancy based on the opening schedule of each of the rooms of the library using a fuzzy inference system. The paper [6], a commercial building in Iowa (USA), uses SVR and ANN based on weather data such as outdoor air dry bulb temperature, outdoor air relative humidity, outdoor air flow rate, diffuse solar radiation rate, direct solar radiation rate, zone air temperature, zone air relative humidity and zone thermostat cooling set point temperature and occupancy data including schedules of building equipment, building light and HVAC operation. The work [21] proposes a model predictive control to forecast the consumption in a simulated commercial building (Energy plus) using meteorological data such as outdoor air temperature, indoor temperature and solar radiation and an equipment schedule ratio.

In the third block, there are works that employ occupancy information through the collection of sensor data. The work [8], in the case of the Research Centre in Rome (Italy), involves meteorological data such as temperature and solar radiation, and creates occupancy indicator counting the number of people who check-in using a card, and then models using an autoregressive integrated moving average, ANN and Naive Bayes. The paper [18], deals with an office building in Hong Kong involving weather data such as outdoor temperature, relative humidity, rainfall, wind speed and global solar radiation and an occupancy attribute created using the hourly total power consumption of the primary air unit, and uses an ANN to create the model. The work [24], in the University of Girona, uses temperature and occupancy data collected with PIR sensors, using MLR, ANN and SVR models. In [27], an ANN is used in conjunction with sensor data such as parking and building occupancy in the campus of the university of Lisbon (Portugal). In [33], the consumption of an office building in Sweden is forecasted based on weather data such as indoor temperature, outdoor temperature, daylight level, solar radiation and wind speed and PIR sensor data using an MLR. The case [23], the electrical load of an sports hall in Finland is predicted using autoregressive models based on meteorological data comprising indoor and outdoor air temperatures and sensor data such as CO_2 measurements. The work [26] presents an autoregressive integrated moving average model that uses outdoor temperature and sensor data such as contact closure, PIR, $\mathrm{CO_2}$ and network activity sensors to predict the consumption in an office building in Ontario (Canada).

In the fourth block, there are works that use expert knowledge such as inspections or surveys to collect information related to occupancy. The case [25], the Administration building of the University of Sao Paulo (Brazil), uses weather data and an attribute related to occupancy, generated performing expert inspections in order to describe the use and the features or the internal loads such as lighting and computers, with an ANN model. In [35], a fast-food restaurant in Cyprus, an autoregressive model based on representative indicators per energy end use of the building such as lighting, kitchen, and refrigerators is used to predict the consumption. The paper [19] proposes an ANN to predict the consumption of 19 subway stations in Hong Kong using outdoor temperature and relative humidity and expert information such as area of concourse, area of platform, shops area, plant room area, staff accommodation area and weekly amount of passengers.

All these methods provide proper results but present shortcomings. There is a demand for methods based on data available in advance which has the ability to perform hourly long-term predictions, predict with few attributes which means low computational cost and do not require continuous sensor data which is economic. In short, there is a need for simple, economic and fast systems that predict the load accurately. The main shortcomings of these methods are as follows:

- The works that employ meteorological data including temperature and solar radiation need weather forecasts which are not always available. In addition, weather forecasts can only be obtained for few days ahead and contain uncertainties. A method without weather data is needed.
- In the case of methods that use occupancy sensor data, there is no data available in advance, so there is no data to predict. Artificial occupancy data is needed.
- Most of these methods are able to forecast consumption just a few days ahead, but cannot do so months in advance. Therefore a method to predict consumption several months in advance is needed.
- The expert knowledge is not always available and contains uncertainties. A repeatable and objective method is needed.

Therefore, all the occupancy methods are employed to artificially create an index of occupancy using previously available data such as calendar, old PIR sensor data, school schedules and other information. These different artificial indicators of occupancy are then tested in order to know which gives the best consumption forecasting results. From the simplest to the more sophisticated method, an explanation of the generation process, an analysis of the load forecasting performance and a contrast of the workload of each one needs to be performed.

In short, on the basis of the existing literature, all paradigmatic data sources and different techniques are used to generate several occupancy indicators. Then, a compendium of STLF performances and the pertinent workloads for each occupancy indicator is provided. The presented methods solve the previously commented shortcomings.

3. Background

Taking into account that a large amount of instances is available, covering a broad range of weather and building conditions, three paradigmatic black box models such as an MLR [9] model, an ANN [38] model and an SVR model were tested in [24]. The results showed that the SVR model provides the most accurate prediction for this kind of data and models. In that case, a grid search algorithm was used to adjust the training parameters of the models, in this case an evolutionary algorithm is used to adjust them. This section gives a brief explanation of the SVR model and the parameter optimization method.

3.1. Support vector regression

The support vector machine [36] model consists of separate classes, that are not linearly separable, transforming them using kernel functions and moving them to a high-dimensional feature space where the data is classified through a hyperplane. On the other hand, the SVR performs a linear regression on this new high-dimensional feature space. The SVR function is seen on Eq. (1):

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)k(x_i,x)+b \qquad (1)$$

where:

α_i and α_i^* are Lagrangian multipliers.
$k(x_i, x)$ is a kernel function.
b is a computed parameter.

There are several kernel functions [3] with different features, proper for each case. However, the most common are linear, radial basis function and polynomial kernel. There is no clear rule about which is better.

3.2. Parameter optimization

There are two main reasons for using a parameter optimization method. The first one is because all the occupancy indicators in the experiments must have equal conditions. The second is because the manually search of the suitable training parameters is a slow process and the grid search is computationally expensive.

The evolutionary computation approach executes a sub-process a multiple number of times to find the optimal values for the specified parameters. The evolutionary strategies, based on the theory of Rechenberg created in 1970 [32], help us to solve an optimization problem without falling into local optimum and premature closure.

Evolutionary search [4] is based on a parental and offspring candidate solution. These solutions, called individuals, are subject to random changes and selection of best solutions iteratively. Based on the principle of biological evolution, the concepts of recombination, mutation and selection are used to solve the problem. First, a recombination selects x parents and combines their parts to create new solutions. Then, the mutation adds random changes to the preliminary solutions. Finally, n individuals are selected and constitute the parental population of the following cycle. Until the termination condition is not achieved, the process continues.

4. Dataset

In this study the experiments are performed using data from four buildings (PI, PII, PIV and Faculty of Science) located at the University of Girona. The buildings are composed mainly by classrooms, offices and laboratories. The buildings PI and PII are used in all the experiments and the buildings PIV and Faculty of Science are only used for contrasting purposes. Regarding HVAC, the heating systems consist of gas boilers and fancoils. The cooling systems are composed of compression refrigeration systems and fancoils.

Building PI, built in 1983, has 6 floors and a volume of 26,150 m^3. The frontage has an area of 3791 m^2, of which 610 m^2 are glass surface. Building PII has 6 floors, a volume of 25,560 m^3 and was built in 1992. The frontage has an area of 2326 m^2, of which 1351 m^2 are glass surface. Building PIV has 3 floors, a volume of 12,000 m^3 and was built in 2003. The frontage has an area of 1836 m^2, of which 630 m^2 is glass. The Faculty of Science building has 5 floors, a volume of 34,810 m^3 and was built in 1997. The frontage has an area of 4903 m^2, of which 1233 m^2 is glass.

The set-point temperature is manually adjusted in the summer to 26 °C, and in the winter to 20 °C. The HVAC control system detects the temperature of the offices and classrooms, and modifies the fan speed to achieve the set point temperature. The profile of these buildings in relation to the HVAC is similar, in that the four buildings have systems where most of the consumption is produced with gas boilers.

As previously stated, temperature and occupancy are the main attributes used in the non-residential buildings forecasting. The data used in this work is as follows:

- Electric load data: electrical load data of the buildings PI, PII and PIV and Faculty of Science is collected using a power meter (PM810 Power Logics of Schneider) linked to the campus infrastructure monitoring system.
- Temperature data: temperature data using a sensor (Vaisala) from the Department of Physics outside the buildings.
- Calendar data: information about working and non-working days, holidays, exams, etc.
- School schedule: the hourly schedule of each classroom.
- Working schedule: the work schedule of the teachers and employees.
- Classroom size: the number of student places for each classroom.
- Classroom devices: the list of electrical devices and their features with regard to each classroom.
- Expert knowledge information: information about the building occupancy based on interviews with experts with experience.
- Occupancy sensor data: the data of occupancy collected in PIV using PIR sensors from previous work.

Based on this information, several artificial occupancy attributes with different levels of complexity are artificially created. The main objective is to analyse which one provides the best forecast. In the search of the proper occupancy indicator, there is probably a balance point between workload and forecasting performance.

The number of data instances of PI is 27,375, covering a total of 38 months, from 1st September, 2011 to 15th October, 2014. The total of instances of PII is 16,589, covering a total of 24 months, from 21st November, 2012 to 15th October, 2014. The number of instances of PIV is 16,590, covering a total of 24 months, from 23rd November, 2012 to 15th October, 2014. The number of instances of Faculty of Science is 27,366, covering a total of 38 months, from 1st September, 2011 to 14th October, 2014.

The patterns of consumption and temperature for a summer (August 5th, 2013) and a winter (February 18th, 2013) week for the PI and PII buildings are shown in Fig. 2.

5. Methodology

This section contains the description of the artificial occupancy indicators and the forecasting method used.

5.1. Occupancy indicators

The main target is to create an artificial occupancy indicator to determine the occupancy in advance. Using some available information, there is the option of creating occupancy indicators to predict consumption some months ahead.

In this section, 43 occupancy indicators, with different levels of complexity, are created in order to find the best one. There are 7 methods used to create the indicators, ranging from low to high complexity. These 7 methods comprehend the main used techniques in the literature and some new lines are proposed, trying to cover all the possible data sources. The different indicators are tested in several experiments and finally a method is selected. The occupancy indicator is an attribute that varies from 0% to 100%. In summary, 43 short-term load forecasting models are trained and

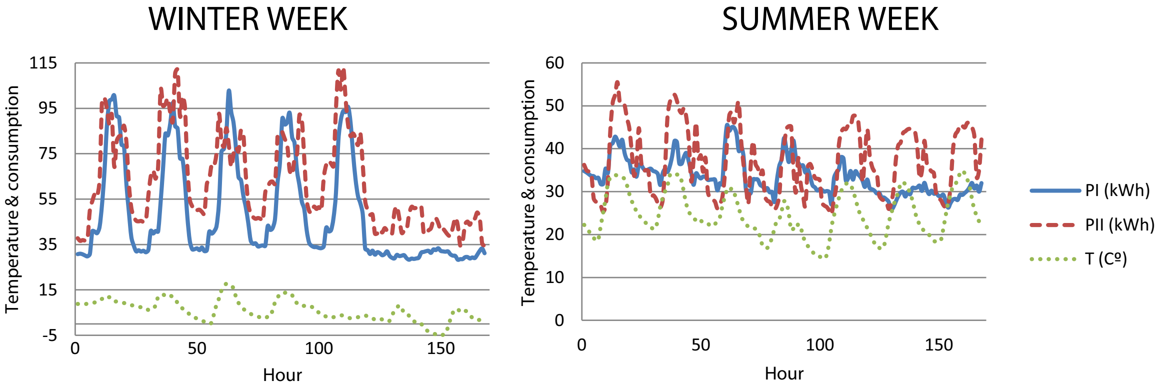


Fig. 2. Temperature and consumption data for summer and winter weeks.

tested using only one occupancy indicator set as attribute each time with the aim of finding which performs better.

The first number of the indicator is referred to as the set, while the other ones are the data sources. For example, the indicator 4.32 is generated with the method of set 4 on the basis of indicators 2.3 and 3.2. The indicators, organized in sets, are as follows:

1. Indicator set 1.
   - Binary occupancy. The simplest indicator. If the university is open, there is a 100% occupancy. If the university is closed, there is 0% occupancy.
2. Indicator set 2 (2.1–2.3). Daily profile. These 3 indicators are based on daily profiles. There are 7 different daily profiles: school day, non-school day, examination day, school-leaving examination day, August day, holiday and weekend day and, finally, Easter and Christmas holiday. Each daily profile has its own level of occupancy. Each one of the 3 indicators of this set is created using only one of the 3 data sources. The data sources to describe each daily profile are:
   - Expert knowledge. Based on the experience of the employees of the university, a level of occupancy of the building for each day type is created.
   - Sensor data. Based on the PIR sensor data collected for the previous work in the PIV building, a level of occupancy for each day type is created. The average of the level of occupancy for several days of each type of day is used.
   - Teacher scheduling. Based on the schedules of certain employees of the university, a level of occupancy for each day type is created.
3. Indicator set 3 (3.1–3.3). Hourly profile. These 3 indicators are based on hourly profiles. There are 24 different hourly profiles. Each hour has its own level of occupancy. As in the previous case, each one of the 3 indicators of this set is created using only one of the 3 data sources. The data sources to describe each hourly profile are:
   - Expert knowledge. Based on the experience of the employees of the university, a level of occupancy of the building for each hour of the day is created.
   - Sensor data. Based on the PIR sensor data collected for the previous work in the PIV building, a level of occupancy for each hour type is created. The average of the level of occupancy for several hours of each type of hour is used.
   - Teacher scheduling. Based on the schedules of certain employees of the university, a level of occupancy for each hour type is created.
4. Indicator set 4 (4.1.1–4.3.3). Aggregation function profile. These 9 occupancy indicators are created by aggregating the indicators of sets 2 and 3. The main idea is to merge the hourly information with that of the days. Up to 5 aggregation functions are tested in order to discover which provides the best results. Then, the aggregation function which provides the best performance in terms of forecasting, is selected. The aggregation functions are the following ones:

   Aggregation function A is presented in Eq. (2):

$$I_A = \frac{I_2 + I_3}{k} \tag{2}$$

   Aggregation function B is presented in Eq. (3):

$$I_B = \frac{I_2 \times I_3}{k} \tag{3}$$

   Aggregation function C is presented in Eq. (4):

$$I_C = \frac{\sqrt{I_2^2 + I_3^2}}{k} \tag{4}$$

   Aggregation function D is presented in Eq. (5):

$$I_D = \frac{(I_2 + I_3)^2}{k} \tag{5}$$

   Aggregation function E is presented in Eq. (6):

$$I_E = \frac{I_2 \times I_3}{k \times (I_2 + I_3)} \tag{6}$$

   where:
   I_2 and I_3 are the aggregated indicators of sets 2 and 3.
   k is the value to scale the output to the proper range, from 0 to 100.
5. Indicator set 5 (5.1.1–5.3.3). Summation of classes. These 9 indicators are based on the data of the previous indicator. The school, examination and school-leaving examination days instances are substituted for new values of occupancy. These new values are calculated taking into account the summation of the active classrooms for each hour. Therefore, the hour with more active classes is the hour with the maximum level of occupancy. Then, an adjustment is needed to equilibrate the instances of the previous indicator (holidays, non-school days and night hours) and the instances of the summation of classes.

   The occupancy of the building for a determined active hour is shown in Eq. (7):

$$Oh_i = \frac{\sum_{i=1}^{m} Ac_i}{Mac} \times Eaf \times 100 \tag{7}$$

   where:
   Oh_i is the level of occupancy of one building for a determined hour.

Ac_i is the number of active classrooms for a determined hour.

Mac is the maximum number of active classrooms.

Eaf is the adjustment factor. Varies from 0 to 1.

6. Indicator set 6 (6.1.1–6.3.3). Summation of weighted classes. On the basis of the previous indicator, a weighting that considers the electrical devices used in the classroom is added to the method. Each classroom is analysed and then the total electrical power of the devices is calculated. The weighting considers all types of rooms, from the laboratories which contain big electric motors, to theory classrooms which only have lights.

   The occupancy of the building for a determined active hour is shown in Eq. (8):

$$Oh_i = \frac{\sum_{i=1}^{m} Ac_i}{Mac} \times \frac{\sum_{i=1}^{m} Edp_i}{Mep} \times Eaf \times 100 \tag{8}$$

   where:

   Oh_i is the level of occupancy of one building for a determined hour.

   Ac_i is the number of active classrooms for a determined hour.

   Mac is the maximum number of active classrooms.

   Edp_i is the summation of the power of the electric devices for a determined classroom.

   Mep is the power of the classroom with more electric power.

   Eaf is the adjustment factor. Varies from 0 to 1.

7. Indicator set 7. (7.1.1–7.3.3). Summation of weighted classes with events. Using the data of indicator set 6, some variations in the occupancy are added at the beginning and at the end of certain events. The events are: the summer, the Christmas holidays, the examination period, the Easter week holidays, local festivities and university parties. In these events the occupancy is very slightly reduced.

In Fig. 3 the several occupancy indicators for a week during school term are shown. The figure shows that the complexity of the profiles increases between the occupancy indicators.

5.2. Procedure block diagram

The proposed methodology consists of several blocks as shown in Fig. 4. In the first block, the missing values are filtered. Then, the instances are normalized. In the following block, the outliers are filtered. In the next block, a feature selection is performed. Then, the data is split with 1/3 of the data to testing and 2/3 to training. In the case of PI, the training data goes from September 1st, 2011 to September 13th, 2013 and the test data goes from September 13th, 2013 to October 15th, 2014. In the case of PII, the training data goes from November 23rd, 2012 to February 23rd, 2014 and the test data goes from February 23rd, 2013 to October 15th, 2014. At that point, an instance selection (20%) is performed with the training data, and an evolutionary search of the suitable training parameters is performed over the selected model. Finally, the validation of the model is done using test data.

5.2.1. Missing values filter

Due to mistakes in sensor readings, there is always a small amount of lost values. The percentage of missing values needs to be minimized as much as possible. There are several methods used to filter the missing values such as filling or deleting. In this case, the method that provides best performance in terms of forecasting, is the deletion of the instances with missing values.

In the case of PI, the instances with missing values are 691 out of a total of 27,375, which represents 2.5%. For PII, the instances with missing values are 592 out of a total of 16,589, which represents 3.6%. In the case of PIV, the instances with missing values are 579 out of a total of 16,590, which represents 3.5%. In the case of the Science Faculty the instances with missing values are 683 out of a total of 27,366, which represents 2.5%.

5.2.2. Normalization

Normalization is needed to work with different scales and units. The use of the same data scale improves the forecasting. The normalization range used is from 0 to 1.

5.2.3. Outliers filter

By filtering the outliers the performance of the model is increased. The outliers need to be detected and can then be deleted or filled. In the filtering, the more restrictive the process, the greater the amount of data lost. In the present case, the method used to detect outliers is the local outlier factor, that consists of calculating the anomaly score according to the local outlier factor algorithm proposed by Breunig [11]. The instances with high scores are then removed.

In the case of PI, the instances with outliers are 227 out of a total of 26,684, which represents 0.85%. For PII, the instances with outliers are 119 out of a total of 15,997, which represents 0.74%. In the case of PIV, the instances with outliers are 118 out of a total of 16,011, which represents 0.74%. In the case of the Science Faculty the instances with outliers are 227 out of a total of 26,683, which represents 0.85%.

5.2.4. Feature selection

In order to remove irrelevant and duplicate data, the redundant and non-correlated attributes are removed. Reducing the size of the database, the computational cost of the training process is reduced. The feature selection consists in two blocks. In the first block the correlation with the class of each attribute is calculated and the features with low correlation are removed. In the second block the correlation between attributes is calculated and the attributes with high correlation with other attribute are removed.

That block is only for the experiments in which calendar nominal attributes are used, not for regular experiments, where only occupancy and temperature are used.

5.2.5. Instance selection

In order to reduce the computational cost of the training process, the number of instances is reduced. A random sub-sample of about 20% of the training data is selected. Some previous validations demonstrate that samples about this percentage reduce the computational time while maintain the forecasting performance levels.

5.2.6. Evolutionary search

The evolutionary search [4] is used to search the training parameters of the model. The objective of this is to deliver the same opportunities to each experiment in which all the models are trained using the same scenario. Each occupancy indicator has equal possibilities of providing the best forecasting results.

The main parameters of the evolutionary search are: maximum generations that specifies the number of generations after which the algorithm should be terminated; population size that stipulates the population size; mutation type that determines the type of the mutation operator; tournament fraction that specifies the fraction of the current population which should be used as tournament members and crossover probability that stipulates the probability of an individual being selected. The parameters of the evolutionary search method are given in Table 2.

5.2.7. Support vector machine

The main training parameters of SVR [34] are the C parameter, the type of kernel and the kernel parameters [3]. The tested

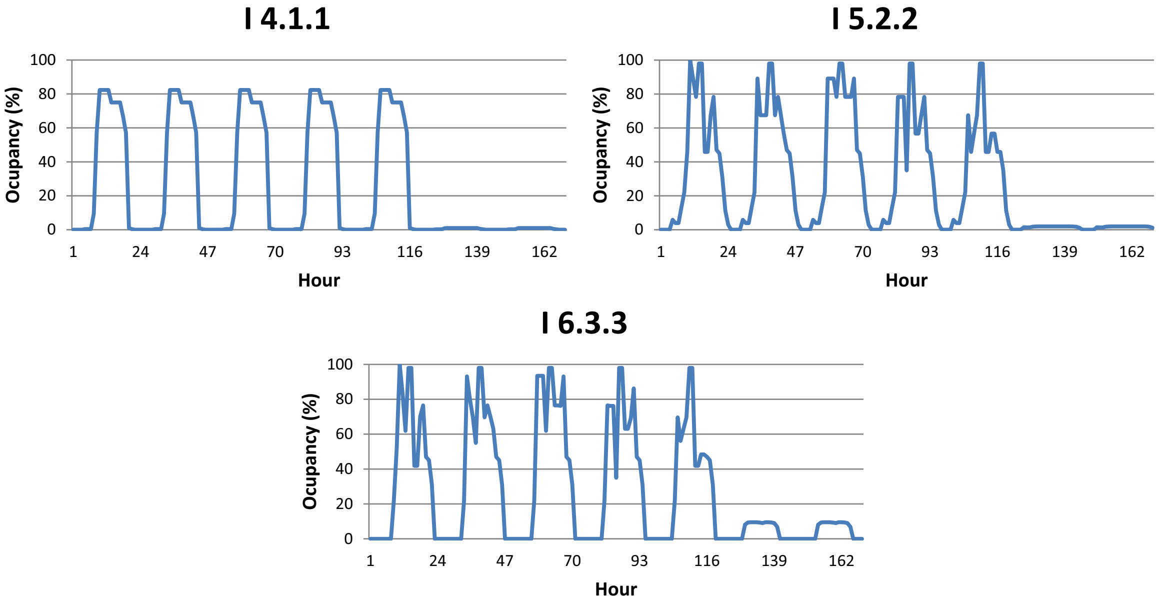


Fig. 3. Example of occupancy indicators of sets 4, 5 and 6.

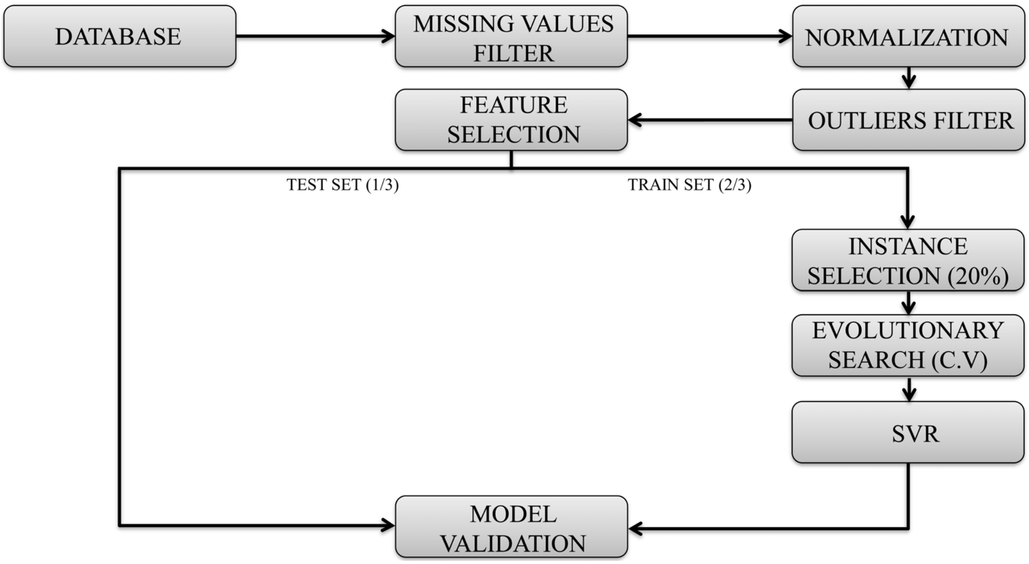


Fig. 4. Block diagram of the process.

Table 2
Parameters of the evolutionary search.

Parameter	Value
Max generations	35
Population size	5
Mutation type	Gaussian
Tournament fraction	0.25
Crossover probability	0.9

kernels and their parameters, are linear (C), Polynomial (C and degree) and Radial Basis function (C and gamma). The C parameter is the complexity constant and adjusts the misclassification tolerance. If C is too large there is an over-fitting, but if it is too small there is an over-generalization. The polynomial kernel is defined by $k(x, y) = (x \times y + 1)^d$ where d is the degree of polynomial. The radial kernel is defined by $k(x, y) = exp(-g||x - y||^2)$ where g is gamma.

The optimization of the training parameters of SVR for each experiment, is performed using the evolutionary search method. A range for each parameter is defined before undertaking each experiment. Then, when the training process is finished, the proper parameters are found.

5.2.8. *Validation*

In the validation process, the model generated with training data (65%), is used to calculate the class attribute of the test data (35%). This data is then validated with a MAPE (mean absolute percentage error) indicator in front of the real values. The MAPE performance indicator is chosen due to its popularity in the forecasting field. The data is chronologically selected, so that the first period of time is used to predict the last period of time.

6. Experimental results

The experiments have been realized using Rapid Miner [12] and a computer with an Intel Core i7-4500U processor and 8 GB of DDR3 RAM. In the next section the indicator used to measure the performance is described. Then, several scenarios and the results obtained are described.

6.1. Error indicator

Among the different methods used to calculate the quality of the model, mean absolute percentage error (MAPE) is the most common indicator found in the forecasting literature. The MAPE performance indicator, shown in Eq. (9), does not depend on the magnitude of the unit of measurement, and is used to compare models. The smaller the MAPE, the more accurate is the model.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{m(i)} - y_{p(i)}}{y_{m(i)}}\right| \times 100 \qquad (9)$$

where:

N is the number of observations.
y_m is the measured output.
y_p is the predicted output.

6.2. Quality factor

The quality factor is a parameter calculated performing the weighted average of the MAPE and the workload, and then scaling it between 0 and 100, as seen in the Eq. (10). The workload is created calculating the hours invested in the generation of each indicator set and then translating it in a 0 to 100 range.

$$Qualityfactor = \frac{MAPE + \times 0.1 * WL}{2 \times MQF} \times 100 \qquad (10)$$

where:

MAPE is the mean absolute percentage error in a 0–100 range.
WL is the workload.
MM is the maximum value of MAPE.
MQF is the maximum value of quality factor.

6.3. Analytical results

In this section the results of several scenarios are analysed, and then the results of the best model are plotted. In each scenario a comparison is performed with the aim of resolving doubts and reaching a conclusion. The main purpose of the experiments is to discover which method generates the best occupancy indicator, a trade-off between workload and forecasting accuracy. The performance of the model (MAPE) is the main output in each experiment.

In the first scenario, the average performance of each indicator set is calculated. In the second scenario, the influence of the temperature attribute is analysed. In the third scenario, the prediction accuracy of each data driven-model is studied. In the fourth scenario, the performance of each data source is examined. In the fifth scenario, the several aggregation functions are assessed. In the sixth scenario, the performances of the SVR kernel functions are contrasted. In the seventh scenario, the proposed model is compared to a model based on several calendar nominal attributes. In the eighth scenario, the presented model is compared to an autoregressive model. In the ninth scenario, the performance of the model for several buildings is presented. In the tenth scenario, the forecasting accuracy for each sensor data treatment is evaluated. In the eleventh scenario, the workload for each indicator set is provided.

The several experiments, unless otherwise indicated, have been performed with PI and PII data, using occupancy and temperature attributes, with an SVR model, linear kernel, aggregation function E, 65% of the training data, 35% of the test data and 20% of the training sub-sample.

Table 3
Performance for indicator set.

Indicator set	PI Avg. MAPE (%)	PII Avg. MAPE (%)
Set 1	49.81	48.34
Set 2	24.34	22.90
Set 3	30.08	23.91
Set 4	18.11	18.05
Set 5	17.05	18.01
Set 6	17.03	17.96
Set 7	16.99	17.93

Table 4
Performance for temperature attribute.

Indicator set	PI		PII	
	Without Temp.	Temp.	Without Temp.	Temp.
	Avg. MAPE (%)		Avg. MAPE (%)	
Set 1	49.16	49.81	52.40	48.34
Set 2	23.21	24.34	27.01	22.90
Set 3	30.52	30.08	25.27	23.91
Set 4	17.13	18.11	18.67	18.05
Set 5	16.29	17.05	18.65	18.01
Set 6	16.28	17.03	18.53	17.96
Set 7	16.25	16.99	18.51	17.93

6.3.1. Scenario 1. Performance according to the indicator set

The first experiment is performed with temperature and each of the artificial occupancy attributes. For each building, a total of 43 of training and test processes are realized, one for each occupancy indicator. Then, the MAPE average is calculated for each occupancy indicator set. The main idea is to discover which is the best occupancy indicator set, that is the method that provides the best predictive accuracy, as seen in Table 3.

Table 3 indicates that most sophisticated indicator set, set 7, presents the best results. However, the improvement between sets 4, 5, 6 and 7 is very slight. So, there is a relationship between the complexity of the indicator and the MAPE, but it is not linear.

6.3.2. Scenario 2. Performance according to the temperature attribute

The objective of the second experiment is to determine if the temperature attribute improves the forecasting quality. The second experiment is implemented first with the temperature and then without the temperature attribute. The same steps as in the first experiment are undertaken, and the outcome is shown in Table 4.

As seen in Table 4, the outcomes do not show an improvement based on the use of the temperature attribute. The forecasting performance variation is minimal and in opposite directions.

6.3.3. Scenario 3. Performance according to the model

The third experiment has the aim of evaluating which is the most appropriate model. An MLR [9] and an ANN [38] models are compared with the SVR model. The adjustable training parameters are as follows: for the MLR the ridge factor and the feature selection method, and for the MLP the learning rate, the momentum and the training cycles. The experiment is conducted for indicator set 4, as seen in Table 5.

According to Table 5, the SVR model provides the most suitable results, but there are no clear evidences as to whether the MLR or the MLP model is better. As seen in [24], the SVR model is the correct model for dealing with load consumption forecasting in non-residential buildings.

6.3.4. Scenario 4. Performance according to the data source

Experiment 4 consists of assessing the several data sources in order to know which is the most suitable: expert knowledge, sensor

Table 5
Performance for model (set 4).

Model	Avg. MAPE (%)	
	PI	PII
MLR	23.69	21.13
MLP	21.31	34.96
SVR	18.11	18.05

Table 6
Performance for data source (set 4).

Data source	Avg. MAPE (%)	
	PI	PII
Expert knowledge	17.67	17.91
Sensor data	17.63	18.35
Teacher schedule	19.55	19.17

Table 7
Performance for aggregation function (indicator 4.2.2).

Aggregation function	Avg. MAPE (%)	
	PI	PII
A	27.22	27.14
B	14.60	18.01
C	28.01	29.47
D	19.36	20.62
E	14.42	16.99

data or teacher schedule. The results for indicator set 4 are depicted in Table 6.

The conclusion is that expert knowledge and sensor data sources enhance the forecasting accuracy of the teachers' schedule data source. The use of the sensor data source is preferable because is an impartial and repeatable method compared with the expert knowledge data source. The fact that the sensor data were collected in PIV indicates that there is room for improvement. Therefore, the differences in the results between the three data sources are slight, as seen in Table 6.

6.3.5. Scenario 5. Performance according to the aggregation function

Experiment 5 analyses the effect of the aggregation function as described in 5.1. In the generation of the occupancy indicator 4, a process of aggregation between indicators 2 and 3, is carried out. To find out how to perform the aggregation process properly, five aggregation functions are tested. The results for indicator 4.2.2 are shown in Table 7.

As shown in Table 7, the aggregation function is absolutely crucial. The aggregation functions B and E far exceed the results of the rest. The multiplicative aggregation functions improve the forecasting performance of the additive ones. The aggregation function E slightly exceeds performance method B.

6.3.6. Scenario 6. Performance according to the kernel

In experiment 6, the performance of each SVR kernel is tested. Linear, radial basis function (RBF) and polynomial kernels are analysed with the aim of comparing their forecasting accuracy. The parameters for the RBF kernel are C and gamma, for the polynomial kernel are C and the polynomial degree and for the linear kernel is C. The outcomes for indicator set 4 are listed in Table 8.

The experiments presented in Table 8, show that the linear kernel is the most efficient of all three kernels. In addition, the linear kernel involves a lower computational cost than the RBF and the polynomial.

Table 8
Performance for kernel type (set 4).

Kernel	Avg. MAPE (%)	
	PI	PII
Linear	18.11	18.05
Polynomial	25.47	22.55
RBF	20.59	19.45

Table 9
Performance for attribute type (indicator 4.2.2).

Attribute type	Avg. MAPE (%)	
	PI	PII
Calendar nominal attributes	26.67	21.63
Indicator 4.2.2	14.42	16.99

Table 10
Performance for model (indicator 4.2.2).

Model	Avg. MAPE (%)	
	PI	PII
ARMA-X	26.84	19.87
SVR (indicator 4.2.2)	14.42	16.99

6.3.7. Scenario 7. Performance for nominal attributes

The principal purpose of the experiment 7 is to prove that the utilization of one single occupancy attribute in the load forecasting model is more appropriate than the use of several calendar nominal attributes, such as year, month, week day, holiday, type of day and hour of the day.

In the experiment, all the nominal attributes are converted into numeric class to deal with the SVR. A comparison is performed between the nominal attributes model and the 4.2.2 occupancy indicator model. The results are shown in Table 9.

As seen in Table 9, the presented model provides better prediction results than the model with the calendar nominal attributes. The presented model outperforms the models with a large set of attributes if the data used in the occupancy attribute creation is processed correctly. Significant differences can also be seen in terms of computational time, in favour of one single attribute model.

6.3.8. Scenario 8. Performance for auto-regression model

In the experiment 8 the main issue is to show that the presented model increases the forecasting accuracy compared with the auto-regressive models [1]. So, a 24-hour ahead ARMA model with exogenous variables including temperature, is contrasted with the SVR model with the 4.2.2 occupancy indicator. The results are presented in Table 10.

This experiment contrasts with an ARMA-X model where the past values of consumption and temperature are used as attributes, with the presented model. Usually, the auto-regressive models provide suitable results in this field. However, the presented results show that the ARMA model does not improve the occupancy indicator SVR model, as seen in Table 10. Moreover, due to the amount of attributes used in the ARMA-X, the difference in computational time, in favour of the single occupancy attribute, is remarkable.

6.3.9. Scenario 9. Performance according to the building

Experiment 9 is done to test the method in other buildings of the university. PI and PII buildings are compared to PIV and the Faculty of Science buildings. The results of the comparison for indicator set 4 are presented in Table 11.

The results of PIV and the Faculty of Science buildings are similar to the results of the buildings used in the experiments. The PI, PII

Table 11
Performance for building (set 4).

Building	Avg. MAPE (%)
PI	18.11
PII	18.05
PIV	16.35
Science	18.75

Table 12
Performance for sensor data treatment (indicator 4.2.2).

Sensor data treatment	Avg. MAPE (%)	
	PI	PII
Hour per day	14.76	17.19
Aggregation function (indicator 4.2.2)	14.42	16.99

Table 13
Workload for each indicator set.

Indicator set	Workload units
1	10
2	20
3	20
4	25
5	80
6	90
7	100

and Faculty of Science buildings have the same profile in terms of offices, laboratories and classrooms. However, the profile of PIV is different as it consists mainly of offices. Due to the fact that the sensor data collection was realized in PIV, the prediction accuracy for PIV is higher.

6.3.10. Scenario 10. Performance according to the sensor data treatment

The experiment 10 analyses which is the most suitable method for processing the sensor data. The performance comparison is between the presented model, the 4.2.2 indicator, and an occupancy indicator generated by calculating for each of the 7 day profiles the 24 hourly occupancy levels, based on the average of the available sensor data, for an each hour of each day profile.

The results show that the presented method is slightly better than the hour per day method, as Table 12 shows.

6.3.11. Scenario 11. Workload according to the indicator

Experiment 11 clarifies the workload product of the creation of each occupancy indicator.

The aim of the experiment 11 is to show the great difference in the generation of indicators 1, 2, 3 and 4, that involve a small amount of work, and indicators 5, 6 and 7, the production of which requires more labour hours, as shown in Table 13.

6.4. Graphic results

In the present section some of the previous experiments are plotted. The following figures show the output of the model based on the indicators, with the best ratio between accuracy in terms of forecasting and the workload needed to produce it, appearing with regard to indicator 4.2.2.

6.4.1. Chart 1. MAPE vs. hour

In Fig. 5, a chart of MAPE vs. hour for both buildings is shown.

Fig. 5 shows that the prediction in the class hours presents good results. There are 4 hourly zones where the prediction is of poor quality. These time-slots are at the beginning and the end of the school day, at lunch time and during some night time hours.

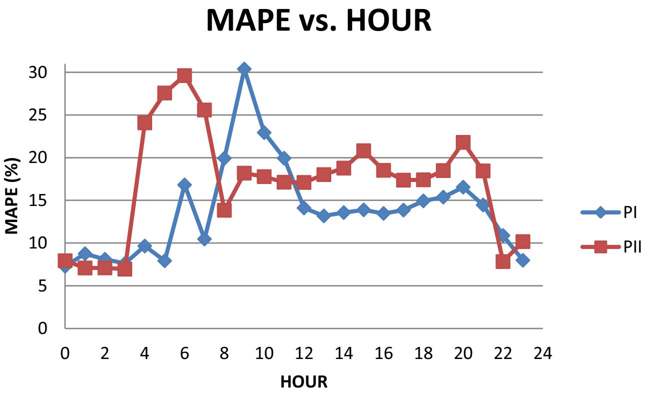


Fig. 5. MAPE vs. hour.

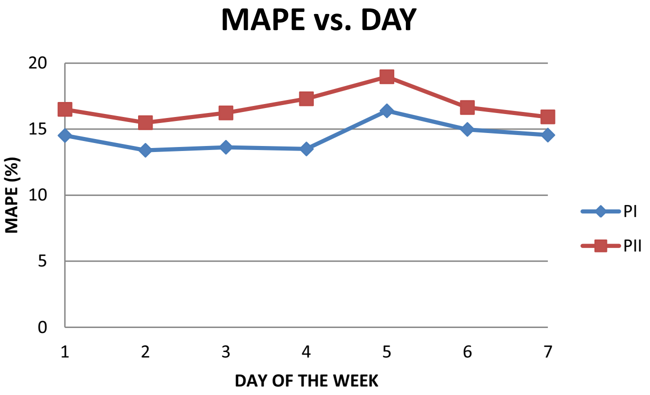


Fig. 6. MAPE vs. day of the week.

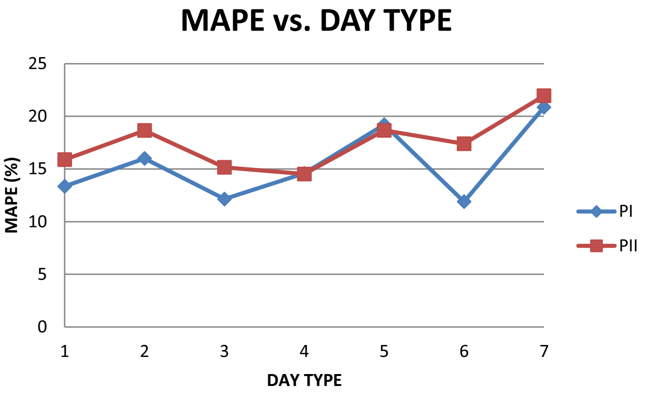


Fig. 7. MAPE vs. day type.

6.4.2. Chart 2. MAPE vs. day of the week

In Fig. 6, a MAPE vs. day plot for both buildings is presented, where the numbers 1–7 represent Monday to Sunday respectively.

As seen in Fig. 6, the forecasting of the midweek days is suitable. The prediction performance decays principally at the beginning and at the end of the working week and on Saturdays.

6.4.3. Chart 3. MAPE vs. day type

In Fig. 7, a MAPE vs. day type chart for both buildings is plotted.

Among the several profiles of days: school (1), exam (3), school-leaving examination (4), holiday and weekend (6) days are well predicted by the model. At a lower level of prediction performance there are: non-school (2), Easter week and Christmas (7) and August days (5).

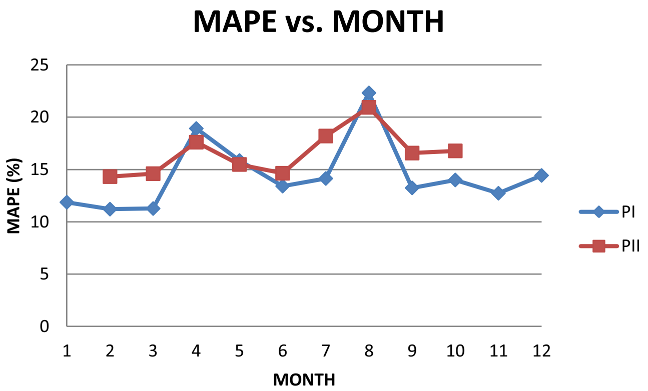


Fig. 8. MAPE vs. month.

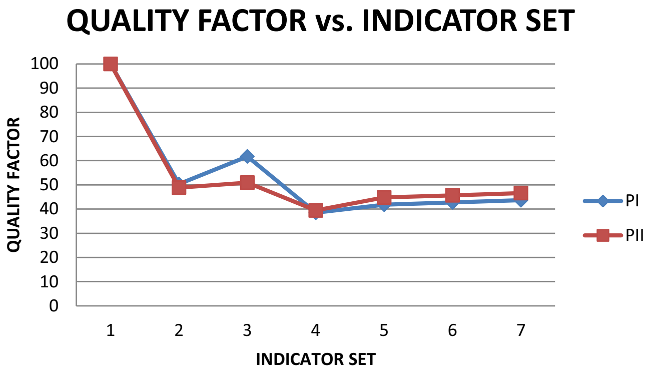


Fig. 9. Quality factor vs. occupancy indicator set.

6.4.4. *Chart 4. MAPE vs. month*

In Fig. 8, a MAPE vs. month chart for both buildings is shown.

Overall, both models offer a poor level of prediction in April and August. Equally, the following months are better in terms of prediction: January, February, March, June, September, October, November and December.

6.4.5. *Chart 5. Quality factor vs. indicator set*

In Fig. 9, a plot of the quality factor vs. the occupancy indicator set is presented.

The smaller the quality factor, the better it is. Fig. 9 shows that indicator set 4 presents a balance between prediction accuracy and workload.

7. Discussion

Taking into account the experiments, there is a non-linear relationship between occupancy indicator complexity and forecasting accuracy. The computational cost is not the main issue in this work because most of the experiments use few attributes. In this work, the workload to generate the artificial attributes is a major concern. The more sophisticated indicators (5, 6 and 7) predict better than the simple ones (1, 2 and 3) but require a large amount of workload. There is a balance point situated on indicator set 4, where forecasting precision and workload are suitable, as the quality factor indicates in Fig. 9. In addition, the occupancy indicators created using expert knowledge and sensor data sources provide a superior prediction than the teacher schedule ones, although the method based on collected sensor data is more expensive, delivers more impartiality and repeatability. In relation to the aggregation functions, it is shown that multiplicative aggregation functions such as aggregation function E, are much better than the additive ones. In addition, the experiments show that the temperature attribute, in this work, is not necessary, so it does not improve the load forecasting. This is due to the partial disaggregation of the HVAC system from the electric consumption, since it is composed by gas boilers and fancoils, and a portion of the energy consumption is not electricity. Furthermore, the proposed sensor data treatment, based on the aggregation functions, enhances the hour per day treatment.

As in [24], SVR model outperforms the other tested models (MLR, ARMA-X and MLP), however the computational cost slightly increases due the low number of attributes, though it is entirely acceptable. Moreover, among the several tested SVR kernels, the linear kernel not only provides more accurate predictions, but also involves a short computational training time. Furthermore, the utilization of a single attribute of occupation in comparison with several calendar nominal attributes such as hour, day of the week, day type and month has resulted in a more compact and precise model. Additionally, the method has proved to work satisfactorily with other university buildings such as the PIV and Faculty of Science ones, as shown in Table 11.

Analysing the charts, can be seen that the worst consumption prediction periods are in the non-well defined human conduct intervals, and in the high variability intervals. In Fig. 5, referring to MAPE vs. hour chart, the load forecast is less efficient in the nocturnal hours, at the beginning and the end of the school-day and at lunch time. In the nocturnal hours, this is due to the uncertainties generated by the cleaning and security services. At lunch time and at the beginning and the end of the school-day it is due to the variability in the individual behaviours of the users.

Comparing Fig. 6, which refers to the MAPE vs. day chart, the prediction performance decreases principally with regard to the beginning and the end of the working week and Saturdays. Mondays and Fridays contain a large variability, especially Fridays, because there are no classes in the afternoon, but some teacher's offices are occupied. Saturdays are complicated with regard to prediction due to random activities in the university installations, which is not the case on Sundays.

In relation to Fig. 7, referring to MAPE vs. day type chart, the accuracy of the model is reduced in non-school, Easter week, Christmas and August days. The profile of days that are adequately predicted are uniform days. For example, in school, examination and school-leaving examination days, the university is open and there are students. In the same way, in holidays and weekend days the university is closed and there are neither students nor teachers. On the other hand, non-school, Easter week and Christmas days are not accurately predicted. Given the dispersion of human behaviours, there are no students in the building, but some employees tend to work during these periods. In relation to August, the HVAC system is not running, so the consumption pattern is slightly different, and there are some employees who work with non-defined schedules in some laboratories.

As shown in Fig. 8, referring to the MAPE vs. month chart, the prediction presents the lowest forecasting levels during April and August. The month of April mainly contains Easter week, therefore is hard to predict, as explained previously. The month of August has been explained previously. In general, months with classes, where consumption patterns are mainly defined by the students' behaviour, the low-dispersion human behaviour periods, are the months that present the highest accuracy level with regard to prediction.

Among the improvements for future work, there is no doubt that enhancing the descriptive level of human behaviour in terms of the worst-defined time periods would improve forecasting accuracy. Furthermore, the chosen model that appears in the charts is based on the sensor data (indicator 4.2.2). For this reason the largest deviations in terms of prediction are located in specific hourly or daily periods. If another data source has been used, the forecasting divergences would be located in other intervals. So, the data sources could be analysed to know which are better for each time-slot, and then apply them selectively or mixed in order to achieve optimal performance. Finally, a revision and an improvement in terms of the occupancy levels of special days including Easter week, Christmas and non-school days, is necessary. Perhaps, the model could be improved by incrementing the number of captured special days in the sensors' database. Besides, although the results are not poor, the sensor data source uses general data from the PIV building. Performing a short data collection procedure in the other buildings to obtain specific data could improve the accuracy of the predictions. It is important to note that by improving the adjustment between indicator sets 3 and 4, some additional prediction accuracy could be obtained.

8. Conclusions

One of the most prevailing needs in terms of utilities is to adjust electricity generation to consumption. For this reason, consumption forecasting is a well understood domain. Also, 40% of electricity consumption is in the building sector. In a previous paper [24], the authors presented an STLF model for non-residential buildings for the University of Girona. The main results obtained showed that using occupancy and temperature as attributes, and as a model the SVR model provides the best load forecasting. However, that model used continuous occupancy sensor data, unavailable in advance. In fact, the main purpose of the paper was to determine the appropriate attributes and models. Now, a fully operational STLF model for the non-residential buildings of the University of Girona is presented.

This paper aims to dispose of the occupancy data in advance. Therefore, several artificial occupancy attributes from different data sources have been created. Then, to find which is the best artificial occupancy indicator, several methods and data sources including sensor data, expert knowledge, class schedules and school calendar, are tested and analysed through the SVR model. Furthermore, this information is compared in terms of the workload resulting from the creation process associated with each occupancy attribute, searching for the most balanced occupancy indicator between performance and workload. Finally, some experiments are conducted to compare the proposed model to other classic models and attributes.

Although the prediction accuracy is lower with respect to previous work [24], the main objective of the presented work is to generate a model based only on artificial attributes, tracing a new path towards artificial occupancy attributes generation methods. The results show that the model which has the best ratio between forecasting precision and workload is an SVR model with a linear kernel trained only with one occupancy attribute generated from the aggregation of the hourly and daily profiles, based on sensor data. So, the SVR model provides the best results in comparison with other data-driven models (ARMA-X, MLR or ANN). Moreover, taking into account the partial disaggregation of the HVAC system, the model does not depend on temperature, converting it in a more compact and simple model and reducing the computational cost. Unlike the other models, this new model can perform hourly consumption predictions months in advance, using only occupancy data. In addition, the proposed method could interpolate the new consumption levels if new classrooms would be constructed, which differs from other works.

In summary, an STLF method for non-residential buildings is provided. This simple and compact model predicts the hourly consumption, months in advance, and is based only on occupancy. Other methods are based on auto-regression or on the need for previously unavailable exogenous variables, and thus require weather forecasts or consumption data to perform the prediction, making a long-term hourly forecast impossible. Moreover, this paper explains the methods for the generation of these occupancy indicators. Every occupancy attribute is assessed in order to determine which method and data source provide the best results in terms of prediction. In future work, departing from the presented methods, some indicator adjustments and revisions of the data sources will be performed in order to improve the forecasting precision of the method.

Acknowledgements

This research project has been partially funded through BR-UdG Scholarship of the University of Girona granted to Joaquim Massana Raurich. Work developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016) and the MESC project funded by the Spanish MINECO (Ref. DPI2013-47450-C2-1-R).

References

[1] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. Stat. Math. 21 (1) (1969) 243–247.
[2] S. Aman, Y. Simmhan, V.K. Prasanna, Improving energy use forecast for campus micro-grids using indirect indicators, in: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), IEEE, 2011, pp. 389–397.
[3] S. Amari, S. Wu, Improving support vector machine classifiers by modifying kernel functions, Neural Netw. 12 (6) (1999) 783–789.
[4] T. Bäck, Evolutionary algorithms in theory and practice, in: Evolutionary Algorithms in Theory and Practice, Oxford University Press, 1996.
[5] Y. Chae, R. Horesh, Y. Hwang, Y. Lee, Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings, Energy Build. 111 (2016) 184–194.
[6] C. Cui, T. Wu, M. Hu, J.D. Weir, X. Li, Short-term building energy model recommendation system: a meta-learning approach, Appl. Energy 172 (2016) 251–263.
[7] D. Datta, S.A. Tassou, D. Marriott, Application of neural networks for the prediction of the energy consumption in a supermarket, in: Proceedings of the International Conference CLIMA, 2000, pp. 98–107.
[8] M. De Felice, X. Yao, Neural networks ensembles for short-term load forecasting, in: 2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG), IEEE, 2011, pp. 1–8.
[9] N.R. Draper, H. Smith, E. Pownell, Applied regression analysis, in: Applied Regression Analysis, vol. 3, Wiley New York, 1966.
[10] M. Grubb, C. Vrolijk, D. Brack, The Kyoto protocol. A guide and assessment, in: The Kyoto Protocol. A Guide and Assessment, 1999.
[11] Z. He, X. Xu, S. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9) (2003) 1641–1650.
[12] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.
[13] T. Hong, S. D'Oca, W.J.N. Turner, S.C. Taylor-Lange, An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework, Build. Environ. 92 (2015) 764–777.
[14] T. Hong, H. Sun, Y. Chen, S.C. Taylor-Lange, D. Yan, An occupant behavior modeling tool for co-simulation, Energy Build. 117 (2016) 272–281.
[15] IEA-EBC, Annex 53, Total energy use in buildings: analysis & evaluation, methods, Technical report, International Energy Agency's Energy in Buildings and Communities Programme, 2013.
[16] IEA-EBC, Annex 66, Definition and simulation of occupant behavior in buildings, Technical report, International Energy Agency's Energy in Buildings and Communities Programme, 2015.
[17] S. Karatasou, M. Santamouris, V. Geros, Modeling and predicting building's energy use with artificial neural networks: methods and results, Energy Build. 38 (8) (2006) 949–958.
[18] S. Kwok, E. Lee, A study of the importance of occupancy to building cooling load in prediction by intelligent approach, Energy Convers. Manage. 52 (7) (2011) 2555–2564.
[19] P.C.M. Leung, E.W.M. Lee, Estimation of electrical power consumption in subway station design by intelligent approach, Appl. Energy 101 (2013) 634–643.
[20] K. Li, H. Su, J. Chu, Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: a comparative study, Energy Build. 43 (10) (2011) 2893–2899.
[21] X. Li, J. Wen, Building energy consumption on-line forecasting using physics based system identification, Energy Build. 82 (2014) 1–12.
[22] X. Li, J. Wen, E.W. Bai, Developing a whole building cooling energy forecasting model for on-line operation optimization using proactive system identification, Appl. Energy 164 (2016) 69–88.

[23] X. Lü, T. Lu, C.J. Kibert, M. Viljanen, Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach, Appl. Energy 144 (2015) 261–275.
[24] J. Massana, C. Pous, L. Burgas, J. Melendez, J. Colomer, Short-term load forecasting in a non-residential building contrasting models and attributes, Energy Build. 92 (April) (2015) 322–330.
[25] A. Neto, F. Fiorelli, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy Build. 40 (12) (2008) 2169–2176.
[26] G.R. Newsham, B.J. Birt, Building-level occupancy data to improve ARIMA-based electricity use forecasts, in: Proceedings of the 2nd ACM workshop on Embedded Sensing Systems for Energy-efficiency in Building, ACM, 2010, pp. 13–18.
[27] J.A. Oliveira-Lima, R. Morais, J.F. Martins, A. Florea, C. Lima, Load forecast on intelligent buildings based on temporary occupancy monitoring, Energy Build. (2016).
[28] Y.K. Penya, C.E. Borges, D. Agote, I. Fernández, Short-term load forecasting in air-conditioned non-residential buildings, in: 2011 IEEE International Symposium on Industrial Electronics (ISIE), IEEE, 2011, pp. 1359–1364.
[29] Y.K. Penya, C.E. Borges, I. Fernández, Short-term load forecasting in non-residential buildings, in: AFRICON, 2011, IEEE, 2011, pp. 1–6.
[30] J.C. Piris, The Lisbon Treaty: A Legal and Political Analysis, Cambridge University Press, 2010.
[31] K. Powell, A. Sripr asad, W. Cole, T. Edgar, Heating, cooling, and electrical load forecasting for a large-scale district energy system, Energy 74 (2014) 877–885.
[32] I. Rechenberg, Evolution Strategy: Optimization of Technical Systems by Means of Biological Evolution, Frommann-Holzboog, Stuttgart, 104, 1973.
[33] C. Sandels, J. Widén, L. Nordström, E. Andersson, Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data, Energy Build. 108 (2015) 279–290.
[34] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[35] E. Spiliotis, A. Raptis, Z.N. Legaki, V. Assimakopoulos, Forecasting electrical consumption of commercial buildings using energy performance indicators, Int. J. Decis. Support Syst. 1 (2) (2015) 164–182.
[36] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.
[37] S.L. Wong, K.K.W. Wan, T.N.T. Lam, Artificial neural networks for energy analysis of office buildings with daylighting, Appl. Energy 87 (2) (2010) 551–557.
[38] B. Yegnanarayana, Artificial Neural Networks, PHI Learning Pvt. Ltd., 2009.

Chapter 7

Main results and discussion

Taking into account the first publication [10], the survey, some interesting information has been extracted:

- The use of complex models implies a disconnection from the problem knowledge base and these kinds of models do not provide extra information about the correlations.
- The increase of the frequency of the data update allows to enhance the performance of the model.
- Most of the papers published in the last 5 years are based on non-linear models.
- Only a few works are addressed to perform forecasting models in disaggregated environments such as buildings or small towns, concretely the last ones.
- Almost 60% of the works are focussed on short-term load forecasting.
- Almost 50% of the works use only load values; 32% use load and weather.
- There is an average of two years for the data collected in the non-linear models.
- There is a clear correlation between average error and the volume of data collected. The more data collected means a better prediction.
- There is a clear relationship between the aggregation level and the prediction performance. The more disaggregated, the more difficult it is to predict.

Concerning the second publication [13], the main factors to consider at the time of implementing a service according to the smart city architecture are the following ones:

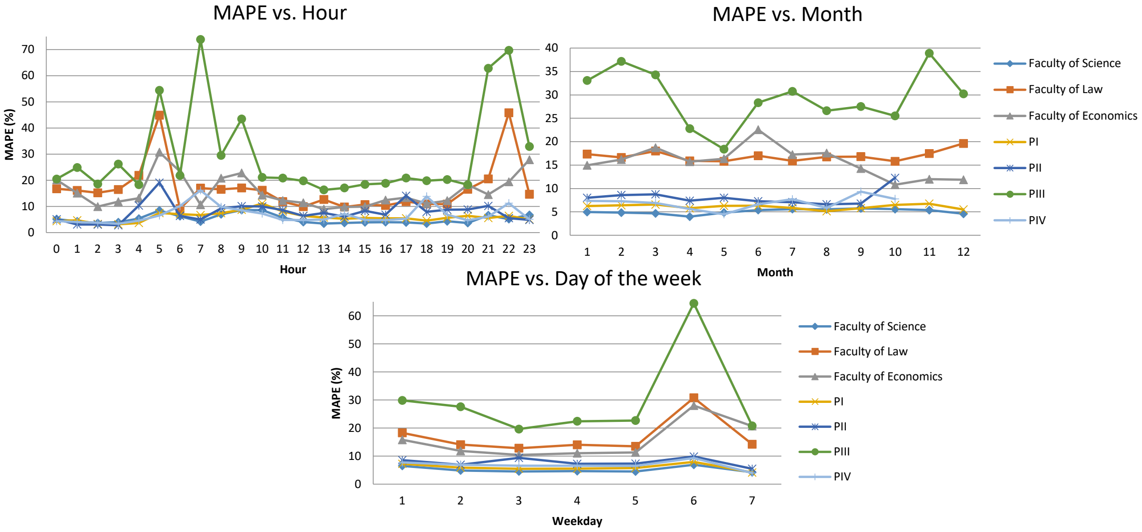


Figure 7.1: The MAPE indicator according to several conditions for the AR model.

- The process of acquisition and the transmission of data must be robust to avoid data loss and outliers. In the same way, the safety, scalability, anonymity and redundancy of the storage system are key. To dispose of a versatile preprocessing layer is absolutely necessary, in order to filter the raw data and achieve proper results in the data mining process.

- The service layer receives clean data, applies the algorithms and returns the desired data to the appliance layer. The time, consistency and robustness of the process are vital.

- The appliance layer must be accessible, visually pleasing and provide all the required information in a comprehensive manner such as tables, charts, etc.

In relation to this case study, some appreciations must be performed:

- The AR model is simple and quick and depends only upon the load data reducing the economic and the computational cost.

- AR models perform better in the 1-hour ahead scenarios than in the 24-hour ahead ones.

- Taking into account the PIII building data, it is obvious that the particularities of the building and the quality of the data allow for providing accurate forecasting results, as seen in Figure 7.1.

- AR models deliver proper results in front of cyclic and well-defined consumption data patterns.

- Some time-slots, due to some variabilities, are more difficult to predict as seen in Figure 7.1.

As a result of the study, some advices, such as to compress the work schedule, rearrange the HVAC operation schedule, move some services or reduce the HVAC power are introduced.

Regarding the third work [12], and in order to obtain the best forecasting performance, the main paradigmatic black box models have been compared, taking into account various parameters, such as the set of attributes or training parameters.

While SVR provides its greatest accuracy in the forecasting, using only temperature and occupancy, other methods, such as MLR and MLP, also need calendar information. The MLR model, focused on linear systems, performs worse than MLP and SVR but provides information about the significance of each attribute. As seen in Figure 7.2, the highest performance is achieved by the SVR model, produced with a medium computational cost.

With regard to all the models:

- Real occupancy data increases the prediction accuracy in comparison with calendar data.

- Indoor ambience data, due to the HVAC operating conditions, does not improve the performance of the models.

- Instance and feature selection reduces the computational time, whilst maintaining the performance of the model.

- The days with heterogeneous profiles, such as Friday, present lower prediction accuracy; otherwise, weekends, due to a homogeneous profile, provide a higher forecasting performance, as shown in Figure 7.2.

With respect to the fourth work [14] and on the basis of the experiments and the experience, there are some evident conclusions:

- There is a non-linear relationship between forecasting performance and the occupancy indicator complexity.

- The computational cost issues are solved using few attributes.

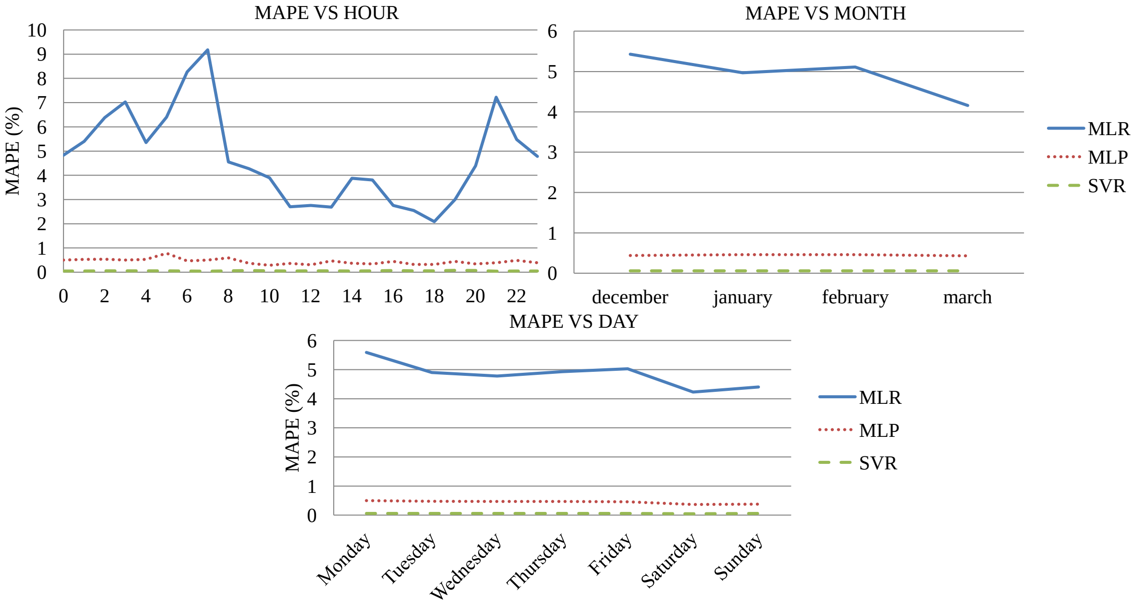


Figure 7.2: The MAPE indicator according to several conditions for the SVR model.

- The workload invested in generating the occupancy attributes is the main point of the paper.
- The model used worked for several buildings.

After testing all the occupancy attributes in several scenarios and conditions, these are the main considerations:

- The more complex indicators perform better than the simple ones, but require more workload.
- The quality factor indicator explains the relationship between prediction accuracy and workload. The quality factor indicator suggests that the balance point is found in the indicator set 4, as seen in Figure 7.3.
- Some data sources, such as sensor data or expert knowledge, perform better than other ones. Although sensor data source is more expensive, it is impartial and repeatable.
- The attribute of temperature does not improve the performance of the prediction, due to the singularities of the HVAC system.
- The SVR model outperforms the other models tested, such as neural networks, regressions and auto-regressions.

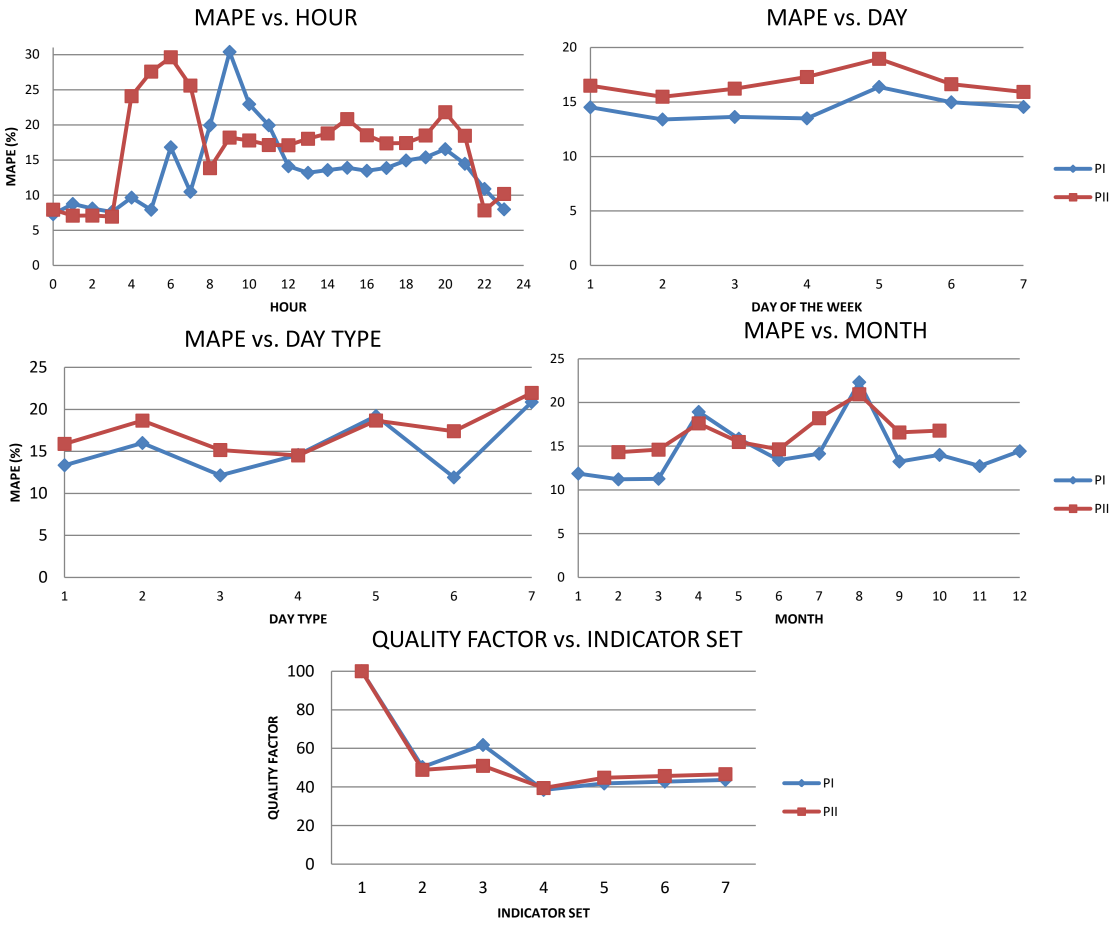


Figure 7.3: Quality factor and the MAPE indicator according to several conditions for the SVR model.

- The linear kernel provides better accuracy and speed in the prediction.

Looking at the charts, it seems obvious that the non-defined intervals, mainly due to human behaviour, contain the worst prediction ranges. So, on the basis of the best model, the main notes are the following ones:

- In the Mean Absolute Percentage Error (MAPE) vs. hour chart, in Figure 7.3, the main unpredictable periods are focussed on the beginning and the end of the school-day and at lunch time, due to the variability in services and in human behaviour.

- Referring to the MAPE vs. day chart, in Figure 7.3, the major deficiencies in the prediction are located at the beginning and the end of the working week, as a result of the variability in human conduct and random activities.

- With regard to MAPE vs. day type chart, as seen in Figure 7.3, the main prediction difficulties are placed in the heterogeneous days, such as non-school, Easter week, Christmas and August days, mainly due to the dispersion of human behaviour and HVAC conditions. On the other hand, when the university is working in regular mode, the prediction is better.

- From the last chart, the MAPE vs. month one, seen in Figure 7.3, the months of April and August are hard to forecast because of Easter week and HVAC issues.

Chapter 8

Conclusions

From the first paper [10], the survey, the main conclusions are the following:

- The introduction of the EMS in several levels of the Smart City services increases the need for energy predictions. Smart buildings, disaggregated environments or microgrids are the main users of minute-by-minute, hourly or daily forecasts.
- With the deployment of Wireless Sensor Networks (WSN), the forecasting models are improving in performance, due to the continuous actualization of the data. In addition, the proliferation of new types of sensor data provides explanations about economic and social changes.
- The EMS is able to perform energy balancing and resource planning by means of suitable predictions.
- Usually, load and weather predictions are the main information that EMS needs, in order to manage and operate the loads in a proper schedule, with the aim of balancing grid consumption, ensuring stability and reliability.

The main challenges in relation to Smart City services are to design a framework and to define common elements. At a low level, to enhance the quality, increase the efficiency and the robustness of the grid are the shared objectives of the utilities. The main conclusions at the end of the second work [13] are the following ones:

- The paper identified intelligent city services in the smart city context. The work explained and described the singularities and difficulties of the integration of these city services, resulting from the monitoring, the management or the operation of the electricity grid.

- The paper is based upon a particular case; a STLF system for seven non-residential buildings located at the university of Girona.
- According to the smart city layers' architecture, the paper depicted the features of each layer of the embedded service.
- The function of the layers consists of: collection of data, transmission of data, storage of data, cleaning and standardization of data, forecasting process and, finally, to display the results to the user.
- The selected model is an autoregressive model and was chosen due to its simplicity and notoriety, taking into account that only the class attribute is needed in this model.
- In the case of the AR models, the integrity and quality of the data, and the presence of well-defined and cyclic patterns in consumption, are the key factors to obtain the higher accuracy in the prediction.
- Several tests have been performed, taking into account diverse scenarios and analysing the forecasting process for months or the days of the week.

As has been said, the purpose of STLF models is to feed the EMS, with the aim of managing, operating and balancing the electricity grid. There is a lot of scientific literature around this issue and the main interest of the utilities is to perform the process in a more accurate, economical, quick and simple manner. On completion, the third work [12], the main conclusions are the following:

- The goal of the paper was to perform an STLF model for a non-residential building at the University of Girona. For this reason, data of weather, indoor ambience, calendar and real building occupancy have been collected, in order to explore which are the most meaningful.
- Several models, such as MLR, MLP and SVR, and diverse groups of attributes, have been tested with the aim of finding the optimal combination. At the end of the tests, the combination that has provided the highest forecasting performance is the SVR model with occupancy and temperature data.
- The main advantages of a small number of attributes in the database are a reduction of the computational cost and the price of the WSN.
- Another finding is that temperature, relative humidity and light level data collected inside the building are not useful in increasing the accuracy of the model. This is the result of the HVAC operation conditions.

- The main advantages of collecting information about the occupancy of the building using PIR sensors are: a reduced price, in face of other occupancy detector methods, such as cameras, CO2 sensors or batch readers; simplicity in the installation and an acceptable level of accuracy.

The authors presented an STLF model for non-residential buildings for the University of Girona in the last paper [13], using temperature and occupancy information to perform the prediction. Although the accuracy achieved in the prediction was extremely high, the model is based on unavailable data in advance, due that to the fact that real occupancy data is not available, at the moment of prediction. In order to complete the last work [13] and dispose of a completely functional STLF model for the non-residential buildings of the University of Girona, the fourth paper [14] proposed a solution generating occupancy attributes in advance. These are the main conclusions from [14]:

- The quality factor, a compromise between workload and accuracy, provided a general idea of the most balanced occupancy indicator.

- The multiple experiments, in order to compare the proposed model in relation to the other models, using in each case different attributes, have been descriptive and useful in determining the best one.

- Finally, the paper not only provides a model to obtain the highest accuracy but also to create a model based only upon artificial attributes, presenting a guide of the artificial occupancy attributes' generation methods.

- Observing the results, it was clear that the combination that performed better was the SVR model, using a linear kernel and one occupancy attribute, created using sensor data combined from the hourly and daily profiles. The model delivered better results than any other data-driven model, such as AR, ANN or MLR.

- One of the major advantages of the present proposal is that the resulting model does not depend on weather attributes; that makes the model simpler and less demanding, computationally.

- Another important virtue is that the presented model, based only upon occupancy, can predict the consumption, hour by hour, months in advance.

- It is also interesting to note that one of the presented models allows us to perform electricity consumption predictions, taking into account modifications in the university, such as variations in the number of classrooms.

- In short, in the present work, an STLF model for non-residential buildings, based upon occupancy data, has been developed, which can perform hourly predictions, months in advance. In order to perform long-term hourly forecasts, the model is not dependent upon exogenous variables that need to be predicted or measured, such as weather or consumption.
- Besides this, a complete analysis of the occupancy indicators and their relative accuracy is provided in conjunction with the guidelines to generate them. The major objective of the paper was to determine which methods, techniques and information produces the best STLF models.

On the basis of the last work [14], these are the proposed improvements for future works:

- Create a method to adjust the models in relation to the occupancy indexes.
- Review and test new data sources, in order to increase the accuracy of the predictions.
- Increase the level of description or definition of the occupancy attributes during the daily time-slots, where the prediction accuracy is low. Select the best occupancy data source for each time-slot, mixing them if it is necessary, in order to increase the performance of the model.
- Increase the collected sensor data of special days, such as Easter week, Christmas and non-school days, with the aim of increasing the performance during these days.
- Obtain occupancy sensor data of other buildings, realizing short campaigns, with the objective of enhancing the forecasting performance in these buildings.
- In addition, in a general manner, these services can be complemented with other functionalities, such as energy efficiency building rating or building benchmarking. An energy saving recommendation system will be helpful in this area too.

Chapter 9

Erratum

On page 76, the Eq. 7 should be:

$$Oh_j = \frac{\sum_{i=1}^{m} Ac_{ij}}{Mac} \times Eaf \times 100 \qquad \forall j = 1..24 \qquad \forall i = 1..m \tag{9.1}$$

Where:
Ac_{ij} is the number of active classrooms for a determined classroom i and hour j.
Eaf is the adjustment factor. Varies from 0 to 1.
Mac is the maximum number of active classrooms.
m is the total number of classrooms.
Oh_j is the level of occupancy of one building for a determined hour j in %.

On page 77, the Eq. 8 should be:

$$Oh_j = \frac{\sum_{i=1}^{m} Ac_{ij}}{Mac} \times \frac{\sum_{i=1}^{m} Edp_{ij}}{Mep} \times Eaf \times 100 \qquad \forall j = 1..24 \qquad \forall i = 1..m \tag{9.2}$$

Where:
Ac_{ij} is the number of active classrooms for a determined classroom i and hour j.
Eaf is the adjustment factor. Varies from 0 to 1.
Edp_{ij} is the summation of the power of the electric devices for a determined classroom i and hour j.
Mac is the maximum number of active classrooms.
Mep is the power of the classroom with more electric power.
m is the total number of classrooms.
Oh_j is the level of occupancy of one building for a determined hour j in %.

On page 79, the Eq. 10 should be:

$$QF = \frac{(\frac{MAPE}{MM} \times 100 + 0.1 \times WL)}{(100 + 0.1 \times WLm)} \times 100 \tag{9.3}$$

Where:
$MAPE$ is the mean absolute percentage error in a 0 to 100 range.
MM is the maximum value of MAPE.
QF is the quality factor in %.
WL is the workload.
WLm is the minimum value of workload.

Bibliography

[1] T. Ackermann, G. Andersson, and L. Söder. Distributed generation: a definition. *Electric power systems research*, 57(3):195–204, 2001.

[2] M. Albadi and E. El-Saadany. Demand response in electricity markets: An overview. In *Power Engineering Society General Meeting, 2007. IEEE*, pages 1–5. IEEE, 2007.

[3] B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Von Wimmersperg. The vision of a smart city. In *2nd International Life Extension Technology Workshop, Paris*, volume 28, 2000.

[4] X. García Casals. Analysis of building energy regulation and certification in Europe: Their role, limitations and differences. *Energy and Buildings*, 38(5):381–392, may 2006.

[5] S. Deilami, A. Masoum, P. Moses, and M. Masoum. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *IEEE Transactions on Smart Grid*, 2(3):456–467, 2011.

[6] S. Depuru, Li. Wang, and V. Devabhaktuni. Smart meters for power grid: Challenges, issues, advantages and status. *Renewable and sustainable energy reviews*, 15(6):2736–2742, 2011.

[7] N. Draper and H. Smith. *Applied regression analysis*. John Wiley & Sons, 2014.

[8] European Commission (EC). Europe 2020: A Strategy for smart, sustainable and inclusive growth. *Working paper {COM (2010) 2020}*, 2010.

[9] H. Farhangi. The path of the smart grid. *IEEE power and energy magazine*, 8(1), 2010.

[10] L. Hernandez, C. Baladron, J. M. Aguiar, B. Carro, A. J. Sanchez-Esguevillas, J. Lloret, and J. Massana. A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings. *IEEE Communications Surveys & Tutorials*, 16(3):1460–1495, jan 2014.

[11] R. Lasseter and P. Paigi. Microgrid: A conceptual solution. In *Power Electronics Specialists Conference, 2004. PESC 04. 2004 IEEE 35th Annual*, volume 6, pages 4285–4290. IEEE, 2004.

[12] J. Massana, C. Pous, Ll. Burgas, J. Melendez, and J. Colomer. Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings*, 92:322–330, 2015.

[13] J. Massana, C. Pous, Ll. Burgas, J. Melendez, and J. Colomer. Identifying services for short-term load forecasting using data driven models in a Smart City platform. *Sustainable Cities and Society*, 28:108–117, 2017.

[14] J. Massana, C. Pous, Ll. Burgas, J. Melendez, and Jo. Colomer. Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. *Energy and Buildings*, 130:519–531, 2016.

[15] J. C. Piris. *The Lisbon Treaty: A Legal and Political Analysis.* Cambridge University Press, 2010.

[16] S Rahman and G B Shrestha. An investigation into the impact of electric vehicle load on the electric utility distribution system. *IEEE Transactions on Power Delivery*, 8(2):591–597, 1993.

[17] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[18] T. Weng and Y. Agarwal. From buildings to smart buildings—sensing and actuation to improve energy efficiency. *IEEE Design & Test of Computers*, 29(4):36–44, 2012.

[19] B. Yegnanarayana. *Artificial neural networks.* PHI Learning Pvt. Ltd., 2009.