UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Data Analytics Approaches in IoT Based Smart Environments

Análisis de Datos en Entornos Inteligentes Basados en el Internet de las Cosas

Dª Aurora González Vidal

2019

Department of Engineering of Information and Communications

Faculty of Computer Science

University of Murcia

Data Analytics Approaches in IoT based Smart Environments

Ph.D Thesis

Authored by:

Aurora González Vidal

Supervised by:

Dr. Antonio Fernando Skarmeta Gómez

Murcia, September 2019

Departamento de Ingeniería de la Información y las Comunicaciones

Facultad de Informática

Universidad de Murcia

Análisis de datos en entornos inteligentes basados en el Internet de las Cosas

Tesis Doctoral

Presentada por:

Aurora González Vidal

Supervisada por:

Dr. Antonio Fernando Skarmeta Gómez

Murcia, Septiembre 2019

Dedication and Acknowledgements

Quoting Ortega y Gasset “I am I and my circumstance; and, if I do not save it, I do not save myself'. For this reason, I would like to thank those who have shaped my environment and who have make this work possible.

First of all my thesis director, Antonio Skarmeta, thank you for giving me the opportunity to discover research and for enhancing the talent of those around you.

Antonio Maurandi, former mentor, now friend and colleague. With you I discovered the science to which I hope to dedicate, I owe you a lot of who I am nowadays. Our complicity has been a constant in the uncertainty.

To my supervisors and colleagues at the internships in ICS, Surrey and ISSNIP, Melbourne for accepting, helping and teaching me so much about research and about the rest of the world.

To my postdoc colleagues who were there at the starting point, Victoria and Fernando who inspired my work and to the present ones, Alfonso for those chats that take us out of our caves.

To my friends: Fran, José, Lidia, Leticia, Yoel, ... I love you deeply.

To my partner, Nicolás, who has shared with me the worst moment of this thesis, its conclusion, and stood firm. Our adventures will continue to take us far away.

To my parents for their support, the opportunities they gave me and their trust and to my sister for her sisterly love.

To all of you I am grateful.

Dedicatoria y Reconocimientos

Estoy de acuerdo con Ortega y Gasset en que “Yo soy yo y mi circunstancia, y si no la salvo a ella no me salvo yo”. Por ello me dispongo a agradecer a aquellos que han conformado mi entorno y que han dado lugar a este trabajo.

Empezando por mi director de tesis, Antonio Skarmeta, gracias por darme la oportunidad de descubrir la investigación y potenciar el talento de los que te rodean.

Antonio Maurandi, otrora mentor, ahora amigo y compañero. Contigo descubrí la ciencia a la que espero dedicarme, te debo mucho de lo que soy hoy día. Nuestra complicidad ha sido una constante en la incertidumbre.

A mis supervisores y compañeros en las estancias del ICS en Surrey y del ISSNIP en Melbourne por aceptarme, ayudarme y enseñarme tanto sobre la investigación y sobre el resto del mundo.

A mis compañeros postdoc que estaban al principio, Victoria y Fernando que inspiraron mi trabajo, y a Alfonso por esas charlas que nos sacan de nuestras cuevas.

A mis amigos: Fran, José, Lidia, Leticia, Yoel, ... os quiero mucho.

A mi pareja, Nicolás, que ha compartido conmigo el peor momento de esta tesis, su conclusión, y se ha mantenido firme. Nuestras aventuras nos seguirán llevando lejos.

A mis padres por su apoyo, las oportunidades y la confianza que me han dado y a mi hermana por su amor de hermana.

A todos vosotros os estoy agradecida.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Acronyms

AI Artificial intelligence

ANN Artificial neural network

ARIMA Autoregressive Integrated Moving Average

BRNN Bayesian Regularized Neural Network

CEP Complex Event Processing

CVRMSE Coefficient of Variation of the RMSE

DCT Discrete Cosine Transform

DR Demand Response

DTA Dense Transit Area

ENORA Elitist Pareto-based MOEA for diversity reinforcement

FS Feature Selection

GAUSS Gaussian Processes

HVAC Heating Ventilating and Air-Conditioning

ICT Information and Communications Technologies

IoT Internet of Things

LR Linear Regression

MAE Mean Absolute Error

MAPE Mean Average Prediction Error

ML Machine Learning

MLP Multilayer Perceptron

MOEA Multi Objective Evolutionary Algortihms

MPC Model Predictive Control

OCB Orion Context Broker

PCA Principal Component Analysis

RBF Radial Basis Function

RC resistor-capacitor

RF Random Forest

RMSE Root Mean Square Error

ROIs Regions of Interest

SVM Support Vector Machines

TTC Technological Transfer Centre

XGB eXtreme Gradient Boosting

CAPÍTULO 1

RESUMEN

Este capítulo presenta la motivación y la justificación del trabajo de tesis. Establece los objetivos de la investigación y los vincula a los resultados que se exponen de manera breve y conectada, dado que ciertos objetivos y resultados surgieron de necesidades que se identificaron cuando se establecieron los objetivos.

1.1 Motivación y Objetivos

El cambio climático está perturbando las economías nacionales y afectando la vida de muchas personas en todo el mundo. Sus consecuencias están costando muy caro hoy día y, si su progresión continúa, el precio a pagar será mucho mayor en el futuro.

Los fenómenos meteorológicos son cada vez más extremos, el nivel del mar está aumentando y las emisiones de gases de efecto invernadero se encuentran en los niveles más altos de la historia. Si no se toman medidas, es probable que el calentamiento global alcance los 5°C a finales de siglo[1], lo que tendrá un enorme impacto en la vida tal y como la conocemos hoy en día.

A fin de fortalecer la respuesta mundial para prevenir el cambio climático, varios países han adoptado muchas iniciativas. En 2015, se aprobó el *Programa de Desarrollo Sostenible* para 2030 y sus objetivos de desarrollo sostenible, que constituyen un llamamiento a la acción de todos los países para promover la prosperidad y proteger al mismo tiempo el planeta[2]. Dentro de las 17 metas, cuatro de ellas están directamente relacionadas con las metas de la tesis: la inclusión de energía limpia y asequible, ciudades y comunidades sostenibles, consumo y producción responsables y acción climática. En el Acuerdo de París de la COP21 (2016), los países

[1] https://www.consilium.europa.eu/en/policies/climate-change/
[2] https://www.un.org/sustainabledevelopment/climate-change/

participantes acordaron trabajar para limitar el aumento de la temperatura mundial a muy por debajo de los 2 grados centígrados [3].

Europa también está dedicando un esfuerzo considerable a reducir sustancialmente sus emisiones de gases de efecto invernadero. Para 2050, como parte de los esfuerzos requeridos por los países desarrollados, la UE se propone reducir sustancialmente sus emisiones, en un 80-95 % en comparación con los niveles de 1990 [4].

La investigación y la innovación contribuyen de manera decisiva a la lucha contra el cambio climático y a la adaptación al mismo, y las Tecnologías de la Información y las Comunicaciones (TIC) tienen el potencial de reducir un 20 % de las emisiones mundiales de CO_2 para 2030, manteniendo las emisiones a los niveles de 2015 [19]. Un informe de British Telecommunications afirma que se espera que la influencia de las TIC en la UE reduzca la huella de carbono de la UE en un 37 %, manteniendo las emisiones en los niveles de 2012.

La inteligencia artificial (IA) y sus aplicaciones particulares, como el Aprendizaje Automático (*Machine Learning* en inglés (ML) están demostrando ser muy útiles para detectar las abundantes ineficiencias de la sociedad moderna que contribuyen a la inestabilidad climática.

El trabajo de la tesis se basa en la combinación de tecnologías TIC novedosas para la recolección y gestión de los datos y su análisis a través del ML con el fin de proporcionar entornos más inteligentes que puedan hacer un uso responsable de los recursos. La aplicación de este trabajo, en un sentido amplio, contribuye a mitigar el cambio climático.

1.1.1 Internet de las Cosas (IdC) y los entornos inteligentes

Un dispositivo IdC es un objeto físico que se conecta a Internet para transferir datos. Gracias a la proliferación de los dispositivos interconectados IdC, hoy en día se están recogiendo grandes cantidades de datos. Esto permite la creación de entornos inteligentes.

Los entornos inteligentes son entornos físicos que se entrelazan invisiblemente con abundantes dispositivos IdC, es decir, sensores, actuadores, dispositivos y elementos computacionales en general, integrados sin que se aprecie en los objetos cotidianos que nos rodean, y conectados a través de una red continua. Sin embargo, no son los sensores los que hacen que un entorno sea inteligente, sino la capacidad de procesar y aprender de todos los datos que esos sensores proporcionan a través de su análisis para proporcionar servicios automáticamente.

El desarrollo y la evolución de los análisis Big Data y de las tecnologías de IdC están desempeñando un papel importante en la adopción de iniciativas de ciudades inteligentes por diversas razones. La primera razón es el crecimiento exponencial de los objetos inteligentes que pueden participar en el desarrollo de una infraestructura de IdC [20]. Cisco Internet Business Solutions Group predice que habrá 50 mil millones de dispositivos conectados para 2020 [21].

[3] `https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement`
[4] `https://ec.europa.eu/clima/citizens/eu_en`

Otras dos razones notables son el crecimiento de la población y la tendencia de urbanización que está teniendo lugar [20]. Según las Naciones Unidas, hay un total de 1.3 millones de personas que se trasladan a las ciudades cada semana, con una población urbana que crece a 6.3 mil millones, lo que representa un 68 % para el año 2050[5]. Este rápido aumento de la población urbana supone un gran estrés para las infraestructuras y el medio ambiente mundial, ya que las ciudades representan más del 70 % del consumo mundial de energía [22] y producen el 80 % de sus emisiones de gases de efecto invernadero [23].

En este sentido, las soluciones de IdC para ciudades inteligentes ayudan a promover el desarrollo económico, a mejorar la infraestructura y el medio ambiente, y a optimizar los sistemas de transporte de manera sostenible, mejorando al mismo tiempo la calidad de vida en las ciudades. Las zonas urbanas son el laboratorio perfecto para reducir las emisiones de gases de efecto invernadero, aumentar el uso de energías renovables y mejorar la eficiencia energética. En la Fig. 2.1 se muestran algunos componentes inteligentes importantes de la ciudad.

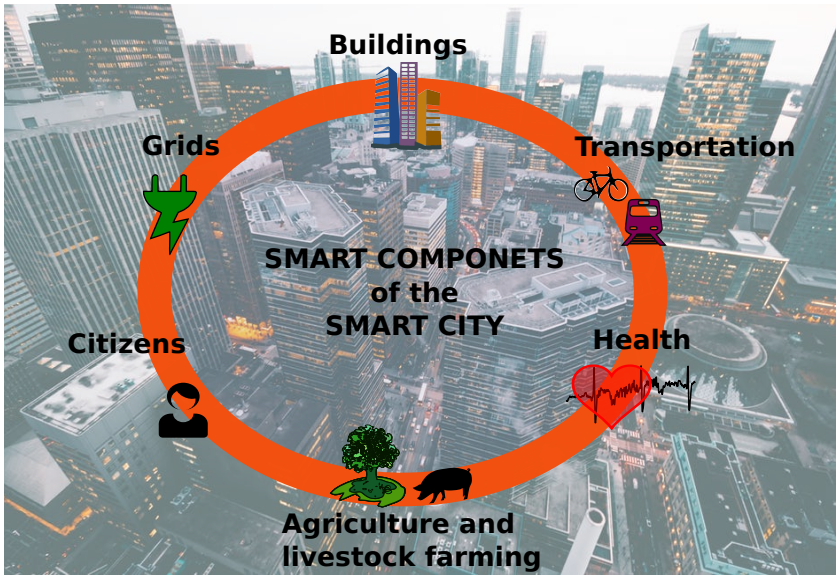

Figura 1.1: Componentes de la Ciudad Inteligente

Encontrar formas de satisfacer las necesidades energéticas de una población en crecimiento en conjunción con la creciente prosperidad económica y la escasez de recursos es un reto fundamental para lograr una sociedad sostenible. La reducción del consumo de energía y de la huella de carbono son cuestiones importantes en las ciudades inteligentes. En el desarrollo de ciudades inteligentes, la sostenibilidad se basa en la eficiencia energética y, a escala mundial, los edificios son la piedra angular de la eficiencia energética en términos de consumo de energía y emisiones de CO_2 [24].

El sector de los edificios también se ve muy afectado por la proliferación de contadores inteligentes y pantallas para el hogar. Esta tendencia parece ir en aumento si tenemos en cuenta

[5] https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

que la Comisión Europea ha establecido que 16 Estados miembros procederán a realizar un despliegue a gran escala de contadores inteligentes para 2020 o antes[25]. Esto, junto con los nuevos avances en materia de infraestructura de datos energéticos (o en inglés *Energy Data Infrastructure* ver [5, 6]), ha creado el entorno perfecto para la creación, entre otras tecnologías, de estrategias avanzadas de retroalimentación energética para la reducción del consumo de energía en los edificios y para la educación de los ocupantes/usuarios [26], el denominado “edificio inteligente".

Un edificio inteligente es cualquier estructura comercial, residencial o industrial en la que se han implementado procesos de automatización para controlar su funcionamiento en base a los datos recogidos por los sensores. Esto incluye tanto el ambiente interno como aparatos de aire acondicionado, iluminación, seguridad, sombreado, etc. [27] como el externo, por ejemplo el clima. Se espera que los edificios inteligentes consideren los elementos de dentro y fuera de su perímetro e interactúen con las redes eléctricas, las condiciones ambientales y los objetivos y labor de sus usuarios.

Los edificios inteligentes se consideran fundamentales para la emergencia de la ciudad inteligente. En la revista Smart Buildings Magazine, Harry G. Smeenk, vicepresidente de desarrollo de programas de la Asociación de la Industria de las Telecomunicaciones, señaló que .*el desarrollo de edificios inteligentes dará lugar a campus inteligentes, lo que fomentará comunidades inteligentes y, con el tiempo, ciudades inteligentes". En pocas palabras, los edificios inteligentes crearán una base escalable para crear la elusiva ciudad inteligente, edificio por edificio, desde cero"*[6].

En los países desarrollados, la energía consumida en los edificios representa entre el 20 y el 40 % del consumo total de energía y es superior a la de la industria y el transporte en la UE y los EE.UU. [28, 29].

Para mitigar el cambio climático, la reducción del consumo de energía junto con el uso de fuentes de energía no fósiles es crucial. Además, la reducción del consumo de energía en los edificios debe hacerse al mismo tiempo que se garantiza la comodidad de los usuarios de los edificios y se reducen los costes para luchar contra la pobreza energética. Los análisis iniciales sugieren que la conversión de edificios en edificios inteligentes gracias a la sensorización a través del IdC, junto con el análisis de datos, puede ser una opción para resolver estos problemas.

En la encuesta de 2016 de la Continental Automated Buildings Association (CABA) denominada *Intelligent Buildings and the Impact of the Internet of Things*, se identificaron los siguientes 3 retos principales a la hora de hacer edificios más inteligentes [30]:

- Mejorar las decisiones de gasto

  El hecho de que los patrones de uso de energía de los edificios a menudo no son posibles de determinar por parte de los gerentes de los edificios dificulta la identificación de las

[6] `http://www.smartbuildingsmagazine.com/features/the-smart-way-to-smart-cities-begins-with-buildings`

oportunidades adecuadas de ahorro de energía. Por lo tanto, muchas veces las medidas de ahorro de energía implementadas no mejoran la eficiencia o reducen innecesariamente la comodidad de los usuarios. Los sistemas de IdC pueden abordar este problema exponiendo datos detallados sobre el uso de la energía, permitiendo a los gestores detectar ineficiencias y crear modelos de predicción muy precisos.

- Reducir consumo energético y gasto de energía

  El control de la utilización de los equipos requiere normalmente una supervisión manual. De esta manera, es complicado reducir el consumo de energía y controlar los costes. La automatización de los electrodomésticos y otros elementos de un edificio permite un mayor control de cuánto, cuándo y cómo se consume la energía.

  Con el IdC, los gestores pueden observar y ajustar a distancia los sistemas de los edificios con sólo pulsar un botón, lo que facilita enormemente la reducción de costes. El ahorro de energía potencial puede mejorarse aún más con las tecnologías IdC.

- Mejora de la eficiencia operativa

  La mayoría de los edificios tienen sistemas separados para el aire acondicinado, iluminación, energía, calidad del aire interior, conectividad a Internet, refrigeración, etc. Esto hace que sea muy difícil optimizar las operaciones generales del edificio. El IdC ofrece la oportunidad de integrar datos de numerosas fuentes en una única plataforma analítica. De esta manera, los gerentes pueden aplicar una estrategia holística a las operaciones de construcción. La combinación de la tecnología de IdC con los edificios inteligentes puede proporcionar un sistema de mantenimiento predictivo. Cuando los parámetros del edificio se monitorean, es más fácil detectar eventos anormales. El administrador del edificio se puede informar instantáneamente para actuar en consecuencia. De esta manera, hay menos fallos en los equipos, lo que contribuye en gran medida al ahorro de costes de los edificios inteligentes.

El último reto que destacamos se refiere a los comportamientos activos y pasivos de los ocupantes con respecto a la energía. Estos comportamientos incluyen la apertura de ventanas, el uso de electrodomésticos, el uso de persianas y sistemas de protección solar, la temperatura de consigna del aire acondicionado, la elección de la iluminación, etc [31]. Para garantizar una reducción prolongada del consumo de energía, las tecnologías de ahorro de energía deben ir acompañadas del eficiente comportamiento de los ocupantes en lo que a energía respecta [32]. Como se indica en el informe de la Agencia Europea de Medio Ambiente [33], hasta un 20 % del ahorro de energía puede lograrse a través de diferentes medidas dirigidas al comportamiento de los consumidores. Para educar a los usuarios de edificios en materia de sostenibilidad, el IdC puede contribuir con la detección de tareas específicas (dependientes del contexto a tiempo real) que los usuarios pueden llevar a cabo para mejorar la eficiencia acompañadas de un razonamiento para su interiorización.

A pesar de estas claras ventajas, muchos edificios todavía no han adoptado las tecnologías IdC.

Según [34], la escasez de infraestructuras inteligentes en los edificios implica que ningún país en Europa esté completamente preparado para la revolución inteligente. Dicho de otra forma, la falta de componentes inteligentes y conectividad entre ellos en los edificios es esncial para desentrañar las posibilidades de los edificios. Considerando los siguientes componentes: aire acondicionado, enfuches, parasoles en las ventanas y automatización de los edificios, si se mejora uno solo de estos componentes de forma aislada puede dar lugar a ahorros del 5–15 %, y un sistema integrado puede conseguir un 30–50 % de ahorros en edificios existentes que de otra manera resultan ineficientes [35].

Según ENERGY STAR, más de 5 millones de edificios comerciales en USA de 4600 m^2 o menos no contienen dispositivos inteligentes para monitorear el uso de energía, la temperatura u otros factores. Se estima que estos edificios consumen hasta un 30 % más de energía de la necesaria. En todo el mundo, el número de estos edificios es mucho mayor.

1.1.2 Análisis de datos y Big Data en entornos inteligentes

La gran cantidad de datos heterogéneos que se capturan, almacenan y gestionan mediante el IdC supera las capacidades de las infraestructuras y de los motores de bases de datos tradicionales. Originalmente, las 3 Vs [36]: gran volumen, alta velocidad y gran variedad de datos se consideraron las características responsables de la aparición de las tecnologías Big Data que ayudan a resolver los problemas que superan los requisitos convencionales.

Con el tiempo, se han propuesto Vs adicionales para caracterizar al Big Data, y consideramos que las 7 Vs [37] describen mejor la complejidad del Big Data:

- Volumen: la enorme cantidad de dispositivos de IdC, incluyendo a los dispositivos portables (o *wearables* en inglés) , genera enormes cantidades de datos. Los problemas que este tamaño de los datos genera son su escalabilidad, accesibilidad y capacidad de gestión.

- Velocidad: La velocidad de transferencia de datos entre la fuente y el destino.

- Variedad: Varios tipos de datos son generados: datos estructurados o no estructurados de diferentes fuentes como imagen, vídeo, texto, sensores, etc.

- Veracidad: Los datos reales que proceden del IdC casi nunca son limpios y precisos. Es necesario encontrar mecanismos para garantizar que los datos sean fiables.

- Validez: Para pasar de explorar a procesar los datos se deben validar previamente. La validez se refiere a la exactitud de los datos con respecto al uso previsto.

- Volatilidad: la retención de datos es especialmente importante en problemas de Big Data debido a su longitud. En muchos casos es crucial determinar en qué momento los datos ya no son relevantes para el análisis actual y deben dejar de almacenarse.

- Valor: El valor representa el valor de negocio que se deriva de los datos. El interés es siempre extraer el valor máximo de los datos. El valor de los datos debe superar su coste o su propiedad y gestión, incluyendo su almacenamiento.

Aunque se recogen datos en cantidades sin precedentes, menos del 1% de estos datos se están analizando [38]. Esto se debe a las complejidades derivadas de los problemas relacionados con Big Data. Existen varios desafíos en el análisis de datos reales, tales como la alta dimensionalidad, alto volumen, ruido y los *data drifts*. Los datos proporcionados por las fuentes de IdC (dispositivos sensoriales y mecanismos de detección) son multimodales y heterogéneos.

Todas estas características dificultan la ejecución y generalización de los algoritmos, por lo que hemos identificado los siguientes retos con respecto a los datos:

- Fusión de los datos procedentes de distintos sensores

  La fusión de datos se define en [39] como la combinación de los datos de sensores procedentes de múltiples sensores para producir información más precisa, más completa y más fiable que no sería posible lograr a través de un solo sensor. En otras palabras, la fusión de datos es una técnica de procesamiento de datos que combina, mezcla, agrega e integra datos de varias fuentes.

  Se pueden desarrollar servicios innovadores mediante la fusión de datos. En ese sentido, la fusión de datos es un reto crucial que debe abordarse. En las aplicaciones de ciudades inteligentes es esencial fusionar e interpretar los datos de forma automática e inteligente [40]. La fusión de datos y el filtrado de datos se han enumerado como dos retos principales para el IdC y sus aplicaciones, como las ciudades inteligentes [41].

- Identificación de patrones en la movilidad humana

  La movilidad de las personas es especialmente importante para aplicaciones como la previsión del tráfico, la planificación urbana y el modelado epidémico. Comprender los patrones de movilidad puede ayudar a tomar decisiones basadas en datos y mejorar la calidad de vida en las ciudades inteligentes. Tradicionalmente, se utilizaban técnicas no escalables para encontrar patrones macroscópicos. Hoy en día, la incorporación de la tecnología GPS en dispositivos portátiles ha permitido recoger una gran cantidad de trazas digitales de alta resolución que permiten conocer las trayectorias espacio temporales subyacentes de las personas. Al mismo tiempo, las redes sociales han incluido capacidades basadas en la localización en sus aplicaciones. Éstas abren un sinfín de posibilidades en el análisis de los patrones de movilidad humana.

- Reducción en tiempo real de información redundante

  Los algoritmos de reducción son útiles para manejar la heterogeneidad y gran volumen del Big Data reduciendo los datos a un tamaño manejable [42, 43]. Estas técnicas se aplican generalmente después de la recolección de datos [44]. Sin embargo, el almacenamiento de todos los datos complejos y de gran tamaño en bruto, redundantes, incoherentes y ruidosos que proceden de fuentes reales de IdC puede ser innecesario. La aplicación de técnicas de reducción en tiempo real puede proporcionar flujos de datos reducidos que contienen información limpia que es realmente relevante para un propósito. Por lo tanto, la aplicación de técnicas de reducción rápidas y efectivas es crucial en el desarrollo de entornos inteligentes para reducir la enorme cantidad de datos a la par que se preserva la información relevante.

- Mejora de la previsión de series temporales mediante la selección de características

  Predecir valores futuros de una serie temporal es un reto al que se han enfrentado muchos investigadores durante décadas. Como en cualquier otra tarea de modelado, el preprocesamiento es un paso esencial. En particular, la selección de características, cuyo objetivo es identificar las variables de entrada más relevantes [45]. La selección de características mejora el rendimiento de las variables predictoras al eliminar variables irrelevantes, reduce los datos para acelerar el entrenamiento y aumenta la eficiencia computacional [46] y, a menudo facilita una mejor comprensión del proceso subyacente que generó los datos.

  En lo que respecta a las series temporales, no solo debemos procesar las variables de entrada sino que hay más características que deben preprocesarse. Éstos son los *laggeados* y, en el caso de las series temporales multivariadas, el tamaño del conjunto de datos de entrada podría aumentar significativamente. La gestión de flujos de datos de series temporales multivariantes es necesaria para muchas aplicaciones de ciudades inteligentes, ya que los datos del IdC se recogen en múltiples ubicaciones distribuidas y periódicamente en intervalos de tiempo. Por lo tanto, es esencial el desarrollo de una metodología sistemática, automática y basada en datos para la evaluación de características, la construcción y la transformación de series temporales multivariadas que no requieran la aportación de expertos humanos.

  Por lo tanto, es esencial el desarrollo de una metodología sistemática, automática y basada en datos para la evaluación de características. Esta metodología debe incluir la construcción de características y la transformación de series temporales multivariantes y no requerir la aportación de expertos humanos.

- Gobernanza de los datos para el IdC

  Los datos del IdC son diferentes de los datos que las arquitecturas y plataformas típicas manejan porque son temporales, vienen en flujo y en tiempo real. Compartir y analizar la

gran cantidad de datos que generan las nuevas tecnologías en tiempo real es clave para desarrollar las aplicaciones que soportan la automatización en escenarios inteligentes. Para hacer frente a los retos inherentes a la planificación y aplicación de soluciones complejas del IdC, necesitamos gobernar nuestros datos a través de plataformas que puedan servir para los fines de todo el proceso. Dichas plataformas también deben ser capaces de gestionar la privacidad y la seguridad de los datos a lo largo de todo el ciclo de vida: recogida de datos, calidad de los datos, almacenamiento de datos, tratamiento de datos, análisis de datos y prestación de servicios.

En resumen, el objetivo de esta tesis es explorar, analizar y aplicar formas de beneficiarse del paradigma del IdC. Este trabajo se basa en la mejora y el análisis de cada paso del proceso de análisis de datos, con el fin de proporcionar mejores servicios a los ciudadanos en entornos inteligentes, es decir, ciudades y edificios inteligentes, con especial énfasis en la eficiencia energética.

Teniendo en cuenta los retos a los que se enfrentan hoy en día tanto el análisis de datos como los edificios inteligentes, establecemos los objetivos que deben alcanzarse para que este objetivo se cumpla, lo que servirá de guía para el desarrollo de la tesis.

- O1. Identificar e integrar datos para crear conjuntos de datos relativos al consumo de energía en entornos inteligentes y determinar la naturaleza de los datos en estudio (binarios, ordinales, temporales, espaciales....). Desarrollar arquitecturas para recopilar y administrar esos conjuntos de datos.

- O2. Desarrollar técnicas de reducción de datos paralelas para las series temporales y, en particular, para los flujos del IdC, preservando sus características clave en relación con las aplicaciones Big Data.

- O3. Crear metodologías y comparar modelos de predicción del consumo de energía con varios horizontes para obtener una predicción altamente precisa y extraer patrones en el uso de la energía.

- O4. Crear características y desarrollar una metodología de reducción de características para series temporales multivariadas aplicadas a la previsión del consumo de energía.

- O5. Identificar, crear y comparar modelos para encontrar patrones en el uso de sistemas de aire acondicionado que puedan ser utilizados para acciones específicas dirigidas hacia la eficiencia energética.

- O6. Identificar patrones de movilidad humana tanto a nivel macro como microscópico utilizando datos de dispositivos portables y redes sociales.

- O7. Identificar y aplicar arquitecturas analíticas de IdC a problemas reales de ciudades inteligentes que integran todas las etapas del proceso, desde la recogida de datos hasta la prestación de servicios.

- O8. Crear mecanismos de IdC para proporcionar servicios personalizados de gestión y sensibilización en materia de energía mediante el análisis de los aspectos de comportamiento relacionados con la eficiencia energética en los edificios inteligentes.

1.2 Resultados

El cuerpo de esta tesis se incluye en varios artículos y capítulos de libros publicados. Gran parte del trabajo se basa en estudios y análisis de los datos generados por los escenarios de IdC, en particular sobre cómo utilizar los datos para la predicción de la energía consumida por los edificios. Otros estudios y publicaciones derivadas de la tesis abordan aspectos específicos relacionados con la creación de infraestructuras inteligentes y otros elementos clave para la resolución de los mencionados objetivos.

El trabajo incluye la integración de 3 conjuntos de datos recogidos en relación con 2 edificios inteligentes y su limpieza, fusión y preprocesamiento, con el fin de obtener conjuntos de datos para su análisis. El primer conjunto de datos pertenece al Centro de Transferencia Tecnológica (CTT) de la Universidad de Murcia [7]. Estos datos son las observaciones ambientales externas al edificio y el consumo total de energía del edificio del 01/12/2014 al 18/02/2018 sen intervalos de 8 horas. En total, 952 observaciones y 15 variables.

El segundo conjunto de datos pertenece a la Facultad de Química de la Universidad de Murcia y está compuesto por 5088 observaciones de 50 atributos que se miden cada hora desde el 02/02/2016 hasta el 06/09/2016.

El atributo de salida es el consumo de energía medido en KWh y hemos incluido mediciones meteorológicas de 3 fuentes diferentes que rodean el edificio, predicciones con una hora de antelación proporcionadas por un servicio web y también atributos de temporada, día de la semana y días festivos.

Por último, también se ha realizado un seguimiento del uso de los sistemas de climatización en 237 aulas de la Facultad de Química. El conjunto de datos consiste en observaciones agregadas de 12 minutos sobre la temperatura ambiente, el estado de encendido/apagado y la temperatura de consigna desde el 31/01/2015 hasta el 28/02/2017.

Estos conjuntos de datos se han creado con el propósito de investigar la interacción entre las personas y los sistemas de los edificios en relación con el consumo de energía, en un intento de extraer patrones de uso y proponer formas automáticas y eficientes para evitar el derroche de energía.

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

Después de recopilar conjuntos de datos y estudiar sus características, nos dimos cuenta de la importancia de la reducción de datos y de la selección de características en entornos reales de IdC. La característica temporal de los datos procedentes de sensores (siempre vienen acompañados del momento en que se ha tomado la medición) ha sido explotada para ambos fines.

Hemos investigado métodos para la reducción de datos en entornos inteligentes, analizado sus inconvenientes y propuesto un nuevo método llamado BEATS, que cumple con los requisitos del análisis Big Data. El método propuesto se basa en la división de las series temporales (datos) en bloques que representan subconjuntos de la estructura de datos. BEATS sintetiza la información que contienen estos bloques de forma independiente, reduciendo el número de datos y conservando sus características fundamentales (perdiendo la menor cantidad de información posible). Para ello, BEATS utiliza la agregación de datos basada en matrices, la Transformada de Coseno Discreta y la caracterización de los valores propios de los datos de las series temporales. Comparamos BEATS con los algoritmos de segmentación y representación más avanzados. La mayoría de ellos asumen datos normales, no tratan los *drifts* de los datos, que son muy comunes para entornos inteligentes, y no pueden ser aplicados de manera online. BEATS está diseñado para superar estos problemas: no requiere la normalización de los datos, lo que también ayudará a preservar el valor de los datos (es decir, su magnitud), se puede aplicar de forma online mediante ventanas deslizantes y es posible calcular la distancia entre las series temporales agregadas. Para evaluar BEATS se ha utilizado en experimentos de clasificación con 6 conjuntos de datos reales. Se redujeron los datos entre un 60-70 %, mejorando significativamente el tiempo de cálculo al tiempo que se mantuvo la precisión de la clasificación. También se ha probado en técnicas de clustering donde se logró el mejor coeficiente de silueta para la mitad de los análisis, más que con ninguno de los otros métodos.

El método anterior responde a una necesidad general de los flujos de datos en el análisis de entornos inteligentes. A continuación, se ha realizado un análisis concreto del problema de la predicción del consumo de energía. Los métodos predictivos necesitan algoritmos de preprocesamiento automático que les ayuden a encontrar la mejor combinación de características para el análisis, por lo que proponemos una metodología de selección de característica multivariante que se basa en las características temporales de los datos.

La metodología se basa en *laggear* o retrasar los atributos temporales y en la configuración de una multitud de métodos diferentes de selección de características, tanto de filtro como *wrappers*, univariante y multivariante. Se han utilizado ocho métodos de selección de características para problemas de regresión y, como se esperaba, los métodos de *wrapper* han mostrado un mejor rendimiento que los métodos de filtro, y los métodos multivariantes mostraron un mejor rendimiento univariantes. Además, el Error Absoluto Medio fue mejor (EAM) que el Error Cuadrático Medio (*RMSE* en inglés) a la hora de utilizar una métrica evaluadora para los métodos de *wrapper*. Utilizando nuestra metodología, EAM se mejora en un 42.28 % y RMSE en un 36.62 % en comparación con no utilizar ninguna técnica de selección de características.

También se ha considerado la creación manual de características y su inclusión en el proceso descrito anteriormente. Se pueden crear variables derivadas de la relevancia retardada tales como: consumo energético a la misma hora y del mismo día pero de la semana anterior, consumo máximo en días laborables / fines de semana de la semana anterior, etc., para incluirlas en el proceso.

En esta tesis se ha realizado un gran esfuerzo para encontrar formas de predecir el consumo de energía en los edificios utilizando varios métodos, horizontes y agregaciones de los datos. De los diversos trabajos que hemos desarrollado en exclusiva para la tarea de modelado del consumo, podemos resumir los siguientes:

- Evaluación del redinimento de los métodos Multilayer Perceptron (MLP), Bayesian Regularized Neural Network (BRNN), Support Vector Machines (SVM) with Radial Basis Function (RBF) Kernel, Gaussian Processes (GAUSS) with RBF Kernel, Random Forest (RF), eXtreme Gradient Boosting (XGB). Todos ellos entrenados y testados utilizando técnicas de validación del aprendizaje automático.

- Estudio del problema de la predicción del consumo de energía desde el punto de vista de las series temporales. Esto incluye la transformación de los datos y la comparación de algoritmos regresivos tradicionales y el nuevo algoritmo de código abierto Prohpet. El modelo implementado en Prophet incorpora componentes no periódicos (utilizando una curva lineal a trozos o de crecimiento logístico), un factor de tendencia que representa los cambios periódicos y los efectos de los días festivos. Éste enmarca el problema predictivo como un ejercicio de ajuste de curvas que difiere de los modelos tradicionales utilizados para las series temporales que se basan en la dependencia temporal de los datos. En este caso hemos incluido una corrección en los datos pronosticados, mejorando la precisión del modelo.

- Uso de la corrección de las predicciones meteorológicas para mejoral el RMSE, obteninedo una mejora de un 4,54% para las predicciones de las próximas 24 horas.

- Comparativa de los diferentes modelos de datos generados (que pueden considerarse del tipo *caja negra*) entre ellos y también con los modelos tradicionales de *caja gris* para la tarea de predicción del consumo diario y semanal.

- Se ha considerado una diferenciación basada en la lógica entre situaciones que dependen del tiempo para etiquetar el comportamiento con respecto al consumo. Estas son las vacaciones y fines de semana, mañanas habituales y tardes habituales. El test no paramétrico Kruskall Wallis y las comparaciones posthoc apoyan la decisión de crear 3 diferentes modelos por día.

- Evaluación no sólo del valor puntual de RMSE, sino también de si un algoritmo de aprendizaje supera estadísticamente a los demás utilizando la prueba no paramétrica de Friedman [47] con las pruebas post-hoc correspondientes para la comparación.

Tras predecir el consumo de energía, tenemos la intención de crear medidas que reduzcan el consumo esperado para obtener un uso más eficiente de la energía. El análisis de los datos de aire acondicionado es una fuente increíble de conocimiento para hacerlo. De esta manera, hemos agregado perfiles similares de variables procedentes de los aparatos (temperatura de consigna, estado de encendido/apagado y temperatura ambiente) de acuerdo a patrones de comportamiento para poder dirigir las acciones que se deben tomar cuando se detectan ajustes de temperatura anormales y el uso de aparatos. Los resultados mostraron que los usuarios pueden ser separados en dos grupos de acuerdo con su interacción con los dispositivos: uno compuesto por aquellos que interactúan con el mando de control del aparato con frecuencia y cambian la temperatura al menos una vez a la semana y otro coompuesto por aquellos que interactúan menos con los mandos.

La predicción del consumo energético de los edificios ha sido estudiada desde un punto de vista analítico, utilizando varias técnicas de preprocesamiento, horizontes y parámetros de entrada. Entendemos que existen dos escenarios principales en los que se puede predecir el consumo de energía. El primero se da cuando los modelos pueden utilizar información anterior sobre el consumo, pero solo se pueden utilizar las predicciones del resto de características o datos de entrada ya que se desea estimar el consumo a futuro, es decir, no se pueden utilizar los valores reales de las variables de entrada. El segundo se da cuando "el futuro es ahora" y queremos crear modelos de referencia para los que podamos utilizar datos de entrada reales pero sin consumo previo, ya que esto sesgaría el experimento. Dependiendo del escenario en el que nos encontremos, hemos estudiado cómo ordenar, estructurar y considerar los datos de entrada. Se ha encontrado una mejora de las predicciones al categorizar las habitaciones de acuerdo a sus patrones de uso del aire acondicionado. En ese sentido, la predicción de la movilidad humana permite a las zonas urbanas adaptar sus esfuerzos de transporte y energía a las necesidades reales de su población. Hemos desarrollado estudios preliminares basados en datos de trayectorias de dispositivos portables (*wearables* en inglés) e información geoetiquetada de redes sociales para encontrar patrones y predecir la movilidad humana.

Todos estos procedimientos analíticos que van desde el acopio y la depuración de datos hasta el análisis de los mismos y el análisis de los resultados necesitan una plataforma basada en el IdC para gestionar la interoperabilidad. La plataforma también debería permitir la integración de las técnicas óptimas de análisis de datos y aprendizaje automático para modelar relaciones contextuales y permitir la prestación de servicios. En esta tesis proponemos una arquitectura que se modela en cuatro capas: una capa de tecnologías donde se recogen los datos; una capa denominada *middleware*, donde se limpian y fusionan los datos; una capa de gestión donde se implementan las técnicas de Big Data y análisis ; y una capa de servicios donde se ofrecen

diferentes servicios que dependen del análisis previo.

Uno de los principales servicios que se han obtenido de esta tesis es la prestación de servicios personalizados de gestión y concienciación energética a los ocupantes de edificios inteligentes a través de una plataforma de IdC con el fin de aumentar la eficiencia energética. El resultado es una infraestructura que utiliza una plataforma de IdC como núcleo para administrar los datos, crear la lógica que detecta el derroche de energía, elaborar mensajes personales y cronometrados y entregar la información a través de aplicaciones móviles creadas para tal fin. Los experimentos muestran que es posible mejorar la llamada competencia de ahorro de energía, que representa el conocimiento de una persona para ahorrar energía o, en otras palabras, el potencial de un usuario para ahorrar energía utilizando las cosas que conoce. También se ha demostrado que es posible ahorrar energía a través de la retroalimentación inteligente a los usuarios del edificio.

Los resultados asociados a las contribuciones principales se presentan en la Tabla 2.1, junto al objetivo al que hacen referencia. En el capítulo 3 se explica con más detalle cómo se obtuvieron estos resultados y las principales características de las arquitecturas de IdC propuestas en esta tesis.

1.3 Organización de la Tesis

Esta tesis está organizada como un compendio de trabajos de investigación de alto impacto. Los dos primeros capítulos contienen la misma información en castellano e inglés respectivamente, y como se ha podido observar presentan tanto la motivación y justificación del trabajo como los objetivos y su vinculación con las publicaciones.

El segundo capítulo presenta la motivación y la justificación. Establece los objetivos de la investigación y los vincula a los resultados que se exponen de manera breve y conectada en el sentido de que ciertos objetivos y resultados surgieron de necesidades que se identificaron cuando se establecieron los objetivos anteriores.

El tercer capítulo es una introducción a las publicaciones donde se expone el trabajo relacionado, las brechas identificadas y los resultados, a la vez que se muestra la relación entre todas ellas. Por último, se destacan las conclusiones del trabajo.

Finalmente, el cuarto capítulo está compuesto por los 6 trabajos de investigación de alto impacto, todos ellos Q1 en el ránking de revistas científicas. Estos documentos contienen la información principal sobre los resultados presentados anteriormente. Cada uno de los trabajos de investigación va precedido de una tarjeta de presentación.

Nb	Resultado	Objetivo	Publicaciones
R1	Creación de conjuntos de datos que relacionen el tiempo atmosférico, el consumo, la ocupación y el uso de la información de los edificios. Análisis de las propiedades de éstos datos y su relación mediante análisis estadísticos.	O.1	**[4, 5, 3]**, [7, 8]
R2	Creación de un algoritmo denominado BEATS que agrega y representa datos de series temporales en bloques de vectores de valores propios (menor dimensión). BEATS se adapta a los *drifts* de los datos reales, puede combinarse con técnicas de aprendizaje automático para su posterior análisis y está pensado para una implementación paralela, siguiendo los requisitos de Big Data.	O.2	**[1]**
R3	Predicción del consumo energético para varios horizontes (horario, diario, semanal) comparando models de caja negra y de caja gris e incluyendo comparaciones estadísticas de los resultados de los métodos más precisos.	O.3	**[2, 3, 4]**, [7, 9, 10, 11, 12]
R4	Desarrollo de una metodología para la predicción de series temporales multivariables de energía basada en métodos de selección de características para la regresión de series temporales que incluye métodos univariantes, multivariantes, de filtro y *wrappers*.	O.4	**[2]**, [10]
R5	Creación de entidades de alto nivel en un edificio (grupos de usuarios / habitaciones) extrayendo perfiles de uso de los aires acondicionado usando métodos de clustering.	O.5	**[5]**, [8]
R6	Modelado de la movilidad humana basado en áreas de tránsito denso y en datos de redes sociales con *Complex Event Processing*.	O.6	[13, 14, 15, 16]
R7	Creación de una arquitectura de Big Data basada en el IdC para proveer de servicios en las ciudades inteligentes en general que se modela en 4 capas: tecnologías, fusión, gestión y servicios; integrando funcionalidades de minería de datos en la capa de gestión. La plataforma pretende ser un paso hacia la plena adaptación del paradigma del IdC en la recuperación, gestión y análisis de datos energéticos en los edificios.	O.7	**[4, 5]**, [17]
R8	Creación de una plataforma con mecanismos abiertos y extensibles para la gestión de datos de sensores. Combinando servicios relacionados con la energía y el análisis del comportamiento se construyen servicios de recomendación y se entregan a través de aplicaciones personalizadas a los ocupantes del edificio, lo que tiene un impacto directo en su comportamiento y, por lo tanto, aumenta la eficiencia energética.	O.8	**[6]**, [18, 17]

Cuadro 1.1: Resultados. Ver en negrita las publicaciones que componen la tesis. El resto son nuestras publicaciones adicionales.

CHAPTER 2

SUMMARY

This chapter introduces the motivation and justification of the thesis. It presents the research objectives and links them to the results that are briefly explained and connected.

2.1 Motivation and Goals

Climate change is already disrupting national economies and affecting lives all around the world. Its consequences are costing dearly today and, if its progression continues the cost will be much greater in the future.

Weather events are becoming more extreme, sea levels are rising and greenhouse gas emissions are now at their highest levels in history. Without action, global warming is likely to be as much as 5°C by the end of the century[1], having a huge impact on life as we know it nowadays.

To strengthen the global response to prevent climate change, countries have adopted many initiatives. In 2015, countries adopted the 2030 Agenda for Sustainable Development and its *Sustainable Development Goals*[2] which are a call for action by all countries to promote prosperity while protecting the planet. Within the 17 goals, four of them are directly related to our goals: the inclusion of affordable and clean energy, sustainable cities and communities, responsible consumption and production and climate action. In the Paris Agreement at the COP21 (2016), countries agreed to work to limit global temperature rise to well below 2 degrees Celsius[3].

Europe is also devoting considerable effort to cut its greenhouse gas emissions substantially. By 2050, as part of the efforts required by developed countries as a group, the EU aims to cut its

[1] https://www.consilium.europa.eu/en/policies/climate-change/
[2] https://www.un.org/sustainabledevelopment/climate-change/
[3] https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

emissions substantially – by 80-95 % compared to the levels in 1990 [4].

Research and innovation make a crucial contribution to fighting and adapting to climate change, and Information and Communications Technologies (ICT) have the potential to reduce 20 % of global CO_2 emissions by 2030, holding emissions at 2015 levels [19]. A report from British Telecommunications claimed that the influence of ICTs in the EU is expected to reduce the carbon footprint of EU by 37 %, holding emissions at 2012 levels.

Artificial intelligence (AI) and its particular applications, such as Machine Learning (ML) are proving to be highly adept at spotting the many inefficiencies in modern society that contribute to climate instability.

The thesis work is based on the combination of novel ICT technologies for data collection and management and its analysis through ML in order to provide smarter environments that can make a responsible use of the resources. The application of this work contributes to mitigate climate change in a broad sense.

2.1.1 Internet of Things (IoT) and smart environments

An IoT device is a physical object that connects to the Internet to transfer data. Thanks to the proliferation of IoT devices that are interconnected, huege amounts of data are being gathered nowadays. This allows the creation of smart environments.

Smart environments are physical environments that are richly and invisibly weaved together with IoT devices; that is, sensors, actuators, gadgets, and computational elements in general, embedded seamlessly in the quotidian objects that surround us, and connected through a continuous network. However, it is not the sensors that makes an environment smart, but the ability to process and learn from all the data that they provide through its analysis in order to automatically provide services.

The development and evolution of Big Data analytics and the IoT technologies are playing a major role in the adoption of smart city initiatives for various reasons. The first reason is the exponential growth of smart objects that can participate in an IoT infrastructure [20]. Cisco Internet Business Solutions Group predicts 50 billion connected devices by 2020 [21]. Two other remarkable reasons are population growth and the urbanization trend [20]. According to the United Nations, there are a total of 1.3 million people moving into cities every week, with urban populations growing to 6.3 billion that is a 68 % by the year 2050[5]. This rapid increase in urban populations brings an intense stress on global infrastructure and environment since cities account for more than 70 % of global energy use [22] and produce 80 % of its greenhouse gas emissions [23].

In that sense, leverage IoT solutions for smart cities helps promoting economic development, upgrades infrastructure, improves environment and optimises transportation systems in a

[4] https://ec.europa.eu/clima/citizens/eu_en

[5] https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

sustainable manner while improving the quality of life in the cities. Urban areas are the perfect laboratory for cutting greenhouse gas emissions, increasing the use of renewable energy and improving energy efficiency. Some important smart city components are depicted in Fig. 2.1

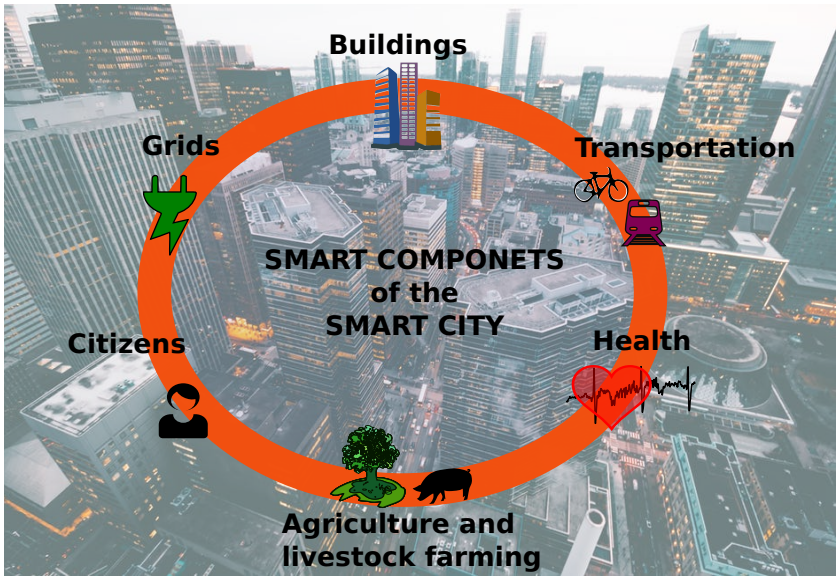

Figure 2.1: Smart City components

Finding ways to meet the energy needs of a growing population in conjunction with growing economic prosperity and resource scarcity is a fundamental challenge to achieving a sustainable society. The reduction of energy consumption and carbon footprint are important issues in smart cities. When developing smart cities, sustainability is based on energy efficiency and at a global scale, buildings are the cornerstone for energy efficiency in terms of power consumption and CO_2 emissions [24].

The building sector is also greatly affected by the proliferation of smart meters and home displays. This trend seems to be on the rise if we consider that the European Commission has established that 16 Member States will proceed with a large-scale roll-out of smart meters by 2020 or earlier [25]. This, along with new developments as regards Energy Data Infrastructure (see [5, 6]), has formed the perfect environment for the creation of, among other technologies, advanced energy feedback strategies for the reduction of energy use in buildings and for the education of building occupants/users [26], the so-called *smart building*.

A smart building is any commercial, residential or industrial structure that implements automation to control its operation based on data collected by sensors. This includes the internal environment such as Heating, Ventilation and Air Conditioning (HVAC), lighting, security, shading, etc. [27], and the external, such as the weather. Smart buildings are expected to consider elements inside and outside their perimeter and interact with electrical grids, environmental conditions, and the goals and duties of their users. Smart buildings target to improve energy efficiency, occupant comfort and environmental impact of the building as a whole.

Smart buildings are considered instrumental in bringing about the smart city. In the Smart Buildings Magazine, Harry G. Smeenk, vice president of program development at the Telecommunications Industry Association noted that *"Developing smart buildings will give rise to smart campuses, which will foster smart communities, and eventually smart cities. Simply put, smart buildings will create a scalable foundation for creating the elusive smart city, building by building, from the ground up"*[6].

The energy consumed in buildings in developed countries comprises 20-40 % of their total energy use and it is above that of industry and transport in the EU and US [28, 29].

In order to mitigate climate change, the reduction of energy use together with the use of non-fossil energy sources is crucial. Furthermore, reducing energy consumption in buildings has to be done while ensuring buildings' users comfort and lower costs in order to combat fuel poverty. Initial analyses suggest that the conversion of buildings into smart buildings thanks to the IoT sensorisation, together with data analytics might be an option by which to resolve these issues.

In the 2016 survey from Continental Automated Buildings Association (CABA) named *Intelligent Buildings and the Impact of the Internet of Things*, the following 3 main challenges when dealing with making buildings smarter were identified [30]

- Improving Spending Decisions

  The fact that buildings' energy usage patterns are often not possible to determine by building managers makes it difficult to identify the proper energy-saving opportunities. Therefore, many times the implemented energy-saving measures either do not improve efficiency or needlessly reduce users comfort. IoT systems can tackle this issue by exposing detailed energy use data, allowing managers to spot inefficiencies and creating highly accurate models for prediction.

- Reducing Energy Consumption and Expenditure

  Controlling how equipment is used normally requires manual supervision. This way, it is complicated to reduce energy consumption and control costs. Automation of a building's appliances and elements allows for greater control of how much, when and how energy is consumed.

  With the IoT, managers can remotely observe and adjust building systems with a tap of a button, making it far easier to bring costs down. The potential energy savings can be further enhanced with IoT technology.

- Improving Operational Efficiency

  Most buildings have separate systems for HVAC, lighting, power, indoor air quality, internet connectivity, refrigeration, and so forth. This makes it very difficult to optimise overall

[6] `http://www.smartbuildingsmagazine.com/features/the-smart-way-to-smart-cities-begins-with-buildings`

building operations. The IoT creates an opportunity to integrate data from numerous sources into a single analytics platform. In this way, managers can apply a holistic strategy to building operations. Pairing IoT technology with smart buildings can provide a predictive maintenance system. When the building's parameters are being monitored it is easier to detect abnormal events. The building manager can be informed instantly to act accordingly. In this way, there are fewer failures in the equipment which contributes greatly to cost savings of smart buildings.

The last challenge that we highlight relates to the active and passive behaviours of occupants with regard to energy. Those behaviours include window opening, use of appliances, solar shading and blinds, adjusting HVAC set points, lighting choice, etc. [31]. To ensure sustained reductions in energy consumption, energy-saving technologies must be accompanied by energy efficient occupant behaviour [32]. As stated in the report of the European Environment Agency [33], up to 20 % of energy savings can be achieved through different measures targeting consumer behaviour. IoT can help with the provision of specialised tasks depending on the real-time context that can lead to the education of users towards sustainability.

Despite these clear advantages, many buildings have yet to adopt IoT technology. According to [34], the lack of a smart infrastructure in buildings implies that no country in Europe is fully ready for the smart revolution. In other words, the lack of smart devices and connectivity amongst them in buildings is crucial for unlocking buildings' possibilities. Taking into account the following components: HVAC, plug load, window shading and building automation, an upgrade a single of them in an isolated way it can result in energy savings of 5–15%, and an integrated system can realise 30–50% savings in existing buildings that are otherwise inefficient [35].

2.1.2 Data analytics and Big Data in smart environments

The huge amount of heterogeneous data that is captured, stored, and managed by means of the IoT exceeds the capabilities of traditional database infrastructures and engines. Originally, the 3 Vs: [36] high volume, high velocity and high variety of data were considered the characteristics responsible for the emergence of Big Data technologies that help resolve the problems that exceed conventional requirements.

Additional V characterizations have been proposed over time, and we consider that the following 7 Vs [37] are more precise descriptors of the complexity of Big Data:

- Volume: the huge amount of IoT devices, including wearable devices, generates massive amounts of data. The concerns regarding data size are its scalability, accessibility and manageability.

- Velocity: The transfer rate of data between the source and destination.

- Variety: Several types of data: structured or unstructured data from different sources: image, video, text, sensors, etc.
- Veracity: Incoming real data from IoT are hardly ever clean and precise. It is necessary to find mechanisms to ensure that data are trustworthy.
- Validity: When the data moves from exploratory to actionable, it must be validated. Validity relates to the correctness and accuracy of data with regard to the intended usage.
- Volatility: data retention is specially important in Big Data problems due to the lengths of data. In many cases it is crucial to determine at what point data is no longer relevant to the current analysis and should not be stored.
- Value: Value represents the business value to be derived from data. The interest is always to extract maximum value from the data. Data value must exceed its cost or ownership and management, including storage.

At the same time that data are collected in unprecedented amounts, less than 1 % of these data are being analysed [38]. This is due to the complexities that problems regarding Big Data imply. There exist several challenges in the analysis of real data such as high dimensionality, high volume, noise, and data drifts. Data provided by IoT sources (sensory devices and sensing mechanisms) are multi-modal and heterogeneous.

All of the above mentioned features hinder the execution and generalization of algorithms, so we have identified the following challenges with regard to data:

- Sensor data fusion

  Data fusion is defined in [39] as the combination of data from multiple sensors to produce more accurate, more complete, and more dependable information that would not be possible to achieve through a single sensor. In other words, data fusion is a processing technique that matches, merges, aggregates, and integrates data from several sources.

  Innovative services can be created by the fusion of data. In that sense, data fusion is a crucial challenge that needs to be addressed. In smart cities applications it is essential to fuse, and interpret the data automatically and intelligently [40]. Data fusion and data filtering have been listed as two major challenges for the IoT and its applications, like smart cities are [41].

- Human mobility pattern identification

  Human mobility is especially important for applications such as traffic forecasting, urban planning, and epidemic modeling. Understanding mobility patterns can support data-driven decisions and improve quality of life in smart cities. Traditionally, non-scalable techniques were used for finding macroscopic patterns. Nowadays, the incorporation of

GPS technology in wearable devices has made it possible to collect a large amount of high-resolution digital traces that can give insight into the underlying spatiotemporal trajectories of people. At the same time, social networks have included location-based capabilities into their applications. These open up a wealth of possibilities in the analysis of human mobility patterns.

- Real-time redundant information reduction

  Reduction algorithms are useful to manage the heterogeneity and the big volume of Big Data by reducing data into a convenient size [42, 43]. These techniques are usually applied after data collection [44]. However, storing all the complex and large raw, redundant, inconsistent, and noisy data that come from real IoT sources might be unecesasary. Applying reduction techniques in real time can provide reduced data streams containing clean information that is really relevant for a purpose. Therefore, the application of fast and effective reduction techniques is crucial in the development of smart environments to reduce the massive amount of data while relevant information is preserved.

- Improvement of time series forecasting using Feature Selection (FS)

  Forecasting future values of a time series is a challenge that many researchers have faced for decades. As in any other modelling task, preprocessing is an essential step. In particular, FS which aims to identify the most relevant input variables [45]. FS consists of eliminating inputs that are irrelevant for the task in order to enhance predictive algorithms' performance. In that sense, FS achieves data reduction, serving for accelerating training and increasing computational efficiency [46]. Furthermore, FS can provide a better understanding of the process that generated the data.

  Regarding time series, there are extra candidate features to be preprocessed. Those are the lagged values and, in the case of multivariate time series, the size of the input dataset might increase significantly. Handling multivariate time series data streams is necessary for many smart city applications since IoT data is collected from multiple, distributed locations and periodically over a time range.

  Therefore, it is essenctial the development of a systematic, automatic, data driven methodology for feature evaluation. Such methodology should include feature construction and transformation of multivariate time series and not require input from human experts.

- Data governance for IoT

  IoT data is different from the data that typical application architectures and platforms handle because it is temporal, on streams and real-time. Sharing and analysing the vast amount of data being generated by new technologies in real time is key in order to develop the applications that support automation in smart scenarios. To address the challenges inherent in planning and implementing complex IoT solutions, we need to govern our data

through platforms that can serve for the purposes of the whole process. Those platfomrs should also be capable of managing the privacy and security of the data across the entire lifecycle: data collection, data quality, data storage, data processing, data analysis and service provision.

In short, the aim of this thesis is to explore, analyse and implement ways to benefit from the IoT paradigm. This work is based on the improvement and analysis of every step in the data analysis process, leading to provide better services to citizens in smart environments, namely smart cities and smart buildings, with a special emphasis on energy efficiency.

Considering the challenges that both data analytics and smart buildings face nowadays, we set out the objectives that must be attained for this aim to be fulfilled, which will serve as a guide for the development of the thesis.

- O1. To identify and integrate data in order to create datasets relative to energy consumption in smart environments and to determine the nature of the data under study (binary, ordinal, temporal, spatial...). Develop architectures to collect and manage those datasets.

- O2. To develop parallel data reduction techniques for time series and, in particular, for IoT streams preserving their key characteristics regarding Big Data applications.

- O3. To create methodologies and compare models for energy consumption forecasting with several horizons in order to obtain highly accurate forecasting and to extract energy usage patterns.

- O4. To create features and develop a feature reduction methodology for multivariate time series applied to energy consumption forecasting.

- O5. To identify, create and compare models for finding patterns in using HVAC systems which can be used for target actions towards energy efficiency.

- O6. Identify human mobility patterns in both macro and microscopic levels using data from wearable devices and social network.

- O7. To identify and apply IoT analytic architectures to real smart city problems that integrate all steps of the process, from data collecting to service provision.

- O8. To create IoT mechanisms in order to provide personalized energy management and awareness services by analysing behavioural aspects related to energy efficiency in smart buildings.

2.2 Results

The body of this thesis is included in several published articles and book chapters. Much of the work is based on studies and analysis of the data generated by IoT scenarios, particularly on how to use the data for the prediction of the energy consumed by buildings. Other studies and publications stemming from the thesis tackle specific aspects related to the creation of smart infrastructures and other key elements for resolving the aforementioned objectives.

The work includes the integration of 3 datasets collected from 2 smart buildings and their cleaning, fusion and preprocessing in order to obtain datasets to be analysed. The first dataset belongs to the Technological Transfer Centre (TTC) at the University of Murcia[7]. These data are the environmental outdoor observations and the total energy consumption of the building from 2014-12-01 to 2016-02-18 in intervals of 8 hours. In total, 952 observations and 15 variables.

The second dataset belongs to the Faculty of Chemistry at the University of Murcia and it is composed of 5088 observations of 50 attributes that are measured hourly from 2016-02-02 until 2016-09-06. The output attribute is the *energy consumption* measured in KWh and we have included meteorological measurements from 3 different sources that surround the buidling, hour ahead predictions provided by a web service and also season, day of the week and holiday attributes.

Finally, we have also monitored the use of HVAC systems in 237 rooms of the Faculty of Chemistry. The dataset consist of 12-minutes aggregated observations regarding room temperature, on/off status and set point from 2015-10-31 until 2017-02-28.

These datasets were created with the purpose of investigating phenomena related to the interaction between people and buildings' systems regarding energy consumption in an attempt to extract usage patterns and propose automatic and efficient ways to avoid wasting energy.

After collecting datasets and studying the characteristics of the data, we realised the importance of both data reduction and FS in real IoT environments. Data coming from real sensors presents a temporal characteristic that has been exploited for both purposes.

We have investigated methods for data reduction in smart environments, analysed their drawbacks and proposed a novel method called BEATS, that complies with requirements of Big Data analysis. The proposed method is based on splitting time series data into blocks which represent subsets of the whole data structure. BEATS synthesizes the information that the blocks contain independently, by reducing the data points while still preserving their fundamental characteristics (loosing as little information as possible). For such purpose, BEATS uses matrix-based data aggregation, Discrete Cosine Transform (DCT) and eigenvalues characterization of the time series data. We compare BEATS with the state-of-the art segmentation and representation algorithms. Most of them assume normal data, do not handle data *drifts* —which are very common for smart environments— and they cannot be applied in a online manner. BEATS is designed to overcome those issues: it does not require normalization of the data, which will

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

also help to preserve the value of the data points (i.e. magnitude of the data), can be applied in an online way using sliding windows and it is possible to compute the distance between the aggregated time series. For BEATS evaluation, it was used in 6 public real datasets. Data were reduced between 60-70 % of and computation time was significatively improved while keeping accuracy in classification. It was also tested for clustering where it achieved the best silhouette coefficient in half of the analysis, more than any of the other methods.

The prior method responds to a general necessity for data streams in smart environments analysis. Following that, a more specific analysis was carried out in the particular problem of energy consumption prediction. Predictive methods need automatic preprocessing algorithms that can help them find the best combination of features for the analysis, so we propose a multivariate FS methodology that is based on the temporal characteristics of the data. The methodology is based on lagging the temporal attributes and configuring a collection of different methods for FS, both filter and wrapper, univariate and multivariate. We applied eight different FS methods for regression and, as expected, wrapper FS methods showed better performance than filter FS methods, and multivariate FS methods showed better performance than univariate FS methods. Also, Mean Absolute Error (MAE) was better than the root mean squared error (RMSE) as metric performance in evaluators for wrapper FS methods. Using our methodology, MAE is improved by 42.28 % and RMSE by 36.62 % compared to not using any FS technique. The manual creation of features and its inclusion in the process described above has also been considered. Variables derived from the lagged-relevancy such as: energy load of the same hour and the same day but previous week, maximal load of working days / weekends during previous week and so forth can be created in order to include them in the process.

A great effort has been made in this thesis in order to find ways to predict energy consumption in buildings using several methods, horizons and aggregations of the data. From the several works that we have develop in the exclusive modelling task, we can summarise the following:

- Evaluation of the performance of Multilayer Perceptron (MLP), Bayesian Regularized Neural Network (BRNN), Support Vector Machines (SVM) with Radial Basis Function (RBF) Kernel, Gaussian Processes (GAUSS) with RBF Kernel, Random Forest (RF), eXtreme Gradient Boosting (XGB). All of them trained and tested using ML validation techniques.

- Study of the energy consumption forecasting problem from a time series point of view. This includes the transformation of the data and the comparison of traditional regressive algorithms and the novel open source library Prohpet. The implemented model on Prophet incorporates non-periodic components (using piecewise linear or logistic growth curve trend), a trend factor that represents periodic changes and holidays effects. It frames the forecasting problem as a curve-fitting exercise which differs from the traditional models used for time series that account for the temporal dependence structure in the data. In this case we have included a correction on forecasted data, improving the accuracy of the model.

- Use of weather forecasting correction for improving RMSE, obtaining a 4,54 % improvement on average for 24 hours hourly predictions.
- Comparison of the several generated data driven models (that can be considered black box) amongst them and also with traditional grey box models for the task of daily and weekly consumption prediction.
- A logic differentiation between temporal situations was considered in order to label behaviour. Those are holidays and weekends, regular mornings and regular afternoons. The non-parametric Kruskall Wallis and the posthoc pairwise comparisons support the decision of creating 3 different models per day.
- Evaluation of not only the punctual value of RMSE, but also of whether one learning algorithm out-performs statistically significantly the others using the non parametric Friedman [47] test with the corresponding post-hoc tests for comparison.

After predicting energy consumption, we intend to create measures that will reduce the expected consumption going towards a more efficient use of energy. The analysis of HVAC data is an incredible source of knowledge in order to do so. In that way, we have aggregated similar profiles of HVAC variables (setting point, onf/off status and room temperature) into behaviour patterns in order to be able to direct the actions that need to be taken when detecting abnormal temperature settings and usage of HVACs. Results showed that users can be separated in two groups according with their interaction with the devices: one composed by those who often interact with the controllers and change the temperature at least once a week and another one comppose by those who interact less with the controllers.

The energy consumption prediction in buildings has been studied from an analytical point of view, using several preprocessing techniques, horizons and input parameters. We understand that there are two main scenarios where energy consumption prediction takes place. The first one is when models can use available past information regarding consumption but they can just use predictions for future inputs since we want to estimate future consumption. The second one happens when “the future is now" and we want to implement baseline models for which we can use the real inputs but no prior consumption because it would bias the experiment. Depending on the scenario we are at, we have studied how to tidy, structure and consider input data. An enhancement of predictions is encountered when categorising rooms according to their HVAC usage patterns. In that direction, the prediction of human mobility allows urban areas to adapt their transport and energy efforts to the real needs of their population. We have developed preliminary studies based on trajectory data from wearable devices and geotagged information from social network in order to find patterns and predict human mobility.

All those analytic procedures that go from data collection and cleaning to the analysis of the data and analysis of results need an IoT-based platform in order to manage interoperability

aspects. The platform should also enable the integration of the optimal data analysis and machine learning techniques in order to model contextual relationships and allow service provision. In this thesis we propose an architecture that is modelled in four layers: a technologies layer where data is collected; a middleware layer where data is cleaned and fused; a management layer where Big Data techniques and analysis are implemented; and a service layer where different services that depend on the previous analysis are offered.

One of the main services that were obtained from this thesis was the provision of personalised energy management and energy awareness services to smart buildings occupants through an IoT-platform in order to increase energy efficiency. The result is a framework that uses an IoT platform as the core to administer the data, create the logic that detects energy waste, elaborates messages that are personal and timed, and deliver the information via created-for-purpose mobile apps. The experiments show that it is possible to improve the so-called energy saving competence, which represents the knowledge of a person to save energy or, in other words, the potential of a user to save energy by using things they know. It has also been proven that it is possible to save energy via intelligent feedback to building users.

The results associated with the main contributions are presented in Table 2.1, alongside the objective referred to. In Chapter 3 we explain in more detail how these results were obtained, and the principal characteristics of the IoT architectures proposed in this thesis.

2.3 Organisation of the Thesis

This thesis is organised as a compendium of high impact research papers. The first two chapters contain the same information in Spanish and English respectively and as it could be seen, they introduce both the motivation and rationale of the work and the objectives and their linkage with the publications.

The second chapter introduces the motivation and rationale. It sets the research objectives and links them to the results that are exposed in a brief and connected manner in a sense that certain objectives and results arose from necessities that were identified when other objectives were set.

The third chapter is an introduction to the research publications where the related work, the identified gaps and the results are exposed while showing the linkage between all of them. At the end it highlights the conclusions of the work.

Finally, the fourth chapter is composed by the 6 high impact research papers -all of them are ranked as Q1. Those papers contain the main information regarding the results presented above. Each of the research paper is preceded by with its presentation card.

Nb	Result	Objective	Publications
R1	Creation of datasets relating weather, consumption, occupation and usage of buildings information. Analysis of data properties and their relationship using statistical analysis.	O1	**[4, 5, 3]**, [7, 8]
R2	Creation of an algorithm named BEATS that aggregates and represents time series data in blocks of lower dimensional vectors of eigenvalues. BEATS adapts to drifts in real data, can be combined with machine learning techniques for further analysis and it is thought for a parallel implementation, following Big Data requirements.	O2	**[1]**
R3	Energy consumption forecasting for several horizons (hourly, daily, weekly) comparing black-box and grey-box models including statistical comparison of results between highly accurate methods.	O3	**[2, 3, 4]**, [7, 9, 10, 11, 12]
R4	Development of a methodology for energy multivariate time series forecasting based on FS methods for time series regression that includes univariate, multivariate, filter and wrapper methods.	O4	**[2]**, [10]
R5	Creation of higher level entities in a building (groups of users / rooms) by extracting profiles on HVAC data using clustering methods.	O5	**[5]**, [8]
R6	Human Mobility Modelling Based on Dense Transit Areas and Social Media with Complex Event Processing	O6	[13, 14, 15, 16]
R7	Creation of an IoT-based Big Data architecture for smart city services in general that is modelled in 4 layers: technologies, fusion, management and services integrating data mining functionalities as built-in features the management layer. The platform intends to be a stage towards the full adaptation of the IoT paradigm in the retrieval, management and analysis of energy data in buildings.	O7	**[4, 5]**, [17]
R8	Creation of a platform with open and extensible mechanisms for sensor data management. Combining energy and behavioural analytics and recommendation services actions are built and delivered through personalized applications to the building occupants, having a direct impact on their behaviour and, thus, increasing energy efficiency.	O8	**[6]**, [18, 17]

Table 2.1: Results. In bold the publications composing the thesis. The others are our additional publications

CHAPTER 3

THESIS CONTRIBUTIONS

This chapter is an introduction to the research publications where the related work, the identified gaps and the results are exposed while showing the linkage between all of them. At the end the conclusions of the work are highlighted.

3.1 Related Work

In this section, we make a thorough search and description of the attempts to solve the previously identified challenges.

First, we expose the concerns regarding the needs of building management with real continuous data instead of audits present in literature. We also review some machine learning approaches in which energy consumption prediction has been involved. After that, we expose the related work for the data processing problems that we faced: time series representation and FS for energy efficiency prediction. Finally, the state-of-the-art IoT architectures for smart cities, energy management and behavioural analytics for energy efficiency are presented.

At the end of each subsection, the identified gaps have been included. We have developed a series of methodologies and techniques for filling those gaps.

3.1.1 Why energy consumption prediction is useful and how has it been carried out according to literature

The application of data analytics for researching how to improve buildings' operation is widely spread in the literature. This includes, amongst others, model-based predictive controls for energy consumption and heating, Demand Response (DR), occupancy detection, and forecasting and automated fault detection. However, building managers typically rely on in-house or external

energy audits carried out yearly at best to determine suitable energy conservation measures [48]. An energy audit entails a revision of the energy efficiency of existing equipment, the operating conditions at facilities, and data collection and analysis with the aim to optimise energy usage and identify energy efficiency measures. Such activities are invasive and labour-intensive and neglect inter-building diversity and metrics that are not related to energy such as occupant satisfaction. Furthermore, several studies show that building performance decreases over time after the implementation of the findings of an energy audit [49, 50]. Therefore, the need for non-invasive energy audits over the complete life of a building motivates the data analytics research in buildings since it renders the potential for metrics that define fairer load profiles in buildings [48]. For example, [51] utilises a combination of clustering and deep learning methods, in conjunction with a weighted aggregation mechanism in order to improve load forecasting accuracy over a short period of time.

Artificial neural networks (ANNs) are able to learn the key information patterns within a multidimensional domain. These have been applied in the field of solar energy, for modeling and design of a solar steam generating plant [52], for the estimation of heating-loads of buildings [53], etc. They have also been used in HVAC systems, solar radiation, modeling and control of power-generation systems, load-forecasting and refrigeration [54]. BRNN are a type of ANN and have been used in the prediction of a series of building energy loads from an environmental input set [55]. Also, RF model has been applied in order to predict energy consumption in residential buildings [56].

Likewise, SVM have been used to predict both the total short-term electricity load and the short-term loads of individual building service systems (air conditioning, lighting, power, and other equipment) in buildings that have electricity sub-metering systems installed [57].

Another common technique for non-linear regression proposed in the literature to be applied are GAUSS with RBF Kernel [58]. It has already been used to forecast electrical load [59] or to estimate the number of occupants in a room according to data related to the room status: motion detection, CO_2 reading, sound level, ambient light and door state sensing [60].

In the reviewed literature, several time-series modelling techniques have been used for different load forecasting problems. For example, Kawashima [61] explored AutoRegressive Integrated Moving Average (ARIMA), Exponentially Weighted Moving Average (EWMA) and Linear Regression (LR) [62], to forecast cooling loads one-day in advance.

In those cases in which limited amounts of data are available and the information concerning the building architecture is partially known, grey models are suitable alternatives for the prediction of energy consumption [63]. Grey-box models use simplified physical descriptions to simulate the behaviour of a building's energy systems, and with them identify important parameters and characteristics using statistical analysis [64]. According to this nomenclature, the previously mentioned ML models are known as black-box models.

Since it was shown that resistor-capacitor (RC) networks can accurately represent the thermo-

dynamics of buildings [65], grey-box models have been used to represent the thermodynamics of buildings. Nowadays, programs such as EnergyPlus, include thermal networks in their codes [64]. This has motivated research into ideal model topologies and methodologies for these models so as to ensure that they accurately represent the responses of buildings [66]. Also, other works are now focusing on how these methods can be used for the characterization of the thermal envelope [67].

One of the objectives of load forecasting in buildings is to make short term near real-time predictions of energy demands. These forecasts can be used in planning and allocating resources to meet the demand. However, there are more research categories on data analytics-driven for buildings in which energy consumption predictions are used. Those are: (a) baseline creation, (b) Model Predictive Control (MPC) , (c) DR, and (d) occupant-centric controls.

- Baseline creation

  Baselines can be used not only for performance monitoring but also for confirming and measuring energy savings derived from the implementation of energy savings actions [6].

- MPC is composed of at least two elements: (a) a forecasting algorithm and (b) an optimization algorithm to determine the optimal control sequence. Many forms of MPC have been applied for the control of HVAC equipment [68]. MPC offers the possibility of anticipating the energy needs of a building taking into account the usual requirements (e.g., comfort ranges) and the possible events that alter consumption (thanks to the prediction model), being able to optimise the building's thermal behaviour on the basis of the defined control goals. This facilitates the use of energy storage capabilities for optimising the use of renewable energy generated on-site.

- Demand Response (DR) also uses predictive models. DR energy requests, based on the current grid status, are generated from the utility or system operators and sent to the building [69]. DR can be used by building owners and operators for responding to real-time pricing [70] and also a utility company can use it to send emergency signals [71]. The objective of DR is to achieve energy cost savings through load shifting and peak load reduction strategies in response to almost real-time variations in the utility rates.

- Occupant centric control: Occupant's detection implies certain automation for lightning, however, thermal systems are not immediate and they need a certain occupancy prediction in order to improve a building's intelligence. Consumption baselines give insight into people's whereabouts, being an important element of occupant behaviour modelling. Occupancy schedules can be derived from the building's electricity consumption [72] and, at the same time, variations on consumption sometimes mean anomalous events.

In many of the reviewed studies for energy consumption forecasting, we encountered the following gaps: (i) the temporal characteristics of energy consumption and/or inputs have not

been considered, (ii) ML models have not been compared to other models that include physical parameters of the buildings and (iii) the inclusion of external predictors is not realistic or adapted to the predictive model context.

3.1.2 Time series representation

There are several approaches to represent a numeric time-dependent variable (i.e. a time series). Using basic statistics would not represent all the information that the time series contains. A classical example that supports this claim is the Anscombe's Quartet, [73] that shows how four very different datasets have identical simple statistical properties: mean, variance, correlation and regression coefficients.

In order to reduce the number of data points in a series and create a representation, segmentation methods can be used as a pre-processing step in data analytics.

Given a time series T containing n data points, segmentation is defined as the construction of a model $\bar{T}$, from l piecewise segments ($l < n$) such that $\bar{T}$ closely approximates T [74].

The segmentation algorithms that aim to identify the observation where the probability distribution of a time series changes are called change-point detection algorithms. Sliding windows, bottom-up, and top-down methods are popular change-point detection based approaches. For sliding windows, each segment is grown until it exceeds an error threshold. In the bottom-up methods, the segments of data are merged until some stopping criteria is met and top-down methods partition the time series recursively until a stopping criteria is met [75].

Another way of classifying the algorithmic methods for segmentation is considering them as online and offline solutions [76]. While offline segmentation is used when the entire time series is previously given, the online segmentation deals with points that arrive at each time interval. In offline mode, the algorithm first learns how to perform a particular task and then it is used to do it automatically. After the learning phase is completed, the system cannot improve or change (unless we consider incremental learning or retraining). On the other hand, online algorithms can adapt to possible changes in the environment. Those changes are known as “drifts". Whereas top-down and bottom-up methods can only be used offline, sliding windows are applicable to both circumstances.

After segmentation, the representation of the time series based on the reduction can be regarded as an initial step that reduces the load and improves the performance of tasks such as classification and clustering. The use of such algorithms can be generally regarded in two ways:

- Representation methods: Extracting features from the whole time series or its segments and applying ML algorithms in order to classify them or compute the distance between the time series representation for clustering.

- Instanced based methods (similarities): Computing the distance matrix between the whole series and using it for clustering or classification applying a k-nearest neighbour approach

[77] by finding the most similar (in distance) time series in the training set.

We review the work made using both approaches since the ultimate goal of our time series representation is to make the time series data more compact for further processing.

- Whole series similarities: Similarity measures are used to quantify the distance between two raw time series. The list of approaches is vast and the comparison between well-known methods has lead to the conclusion that the benchmark for classification is dynamic time warping (DTW) since other techniques proposed before 2007 were found not significantly better [78].

- Intervals: For a series of length m, there are $m(m-1)/2$ possible contiguous intervals.

  Piecewise Linear Representation (PLR) [79] methods are based on the approximation of each segment in the form of straight lines and include the perceptually important points (PIP), Piecewise Aggregate Approximation (PAA) [80], and the turning point (TP) method [81].

  The state-of-the-art models Time Series Forest (TSF)[82] and Learned pattern similarity (LPS)[83] generate many different random intervals and classifiers on each of them, ensembling the resulting predictions.

- Symbolic Aggregate approXimation (SAX) [84] Among all the techniques that have been used to reduce the number of points of a time series data, SAX has attracted the attention of the researchers in the field. SAX allows a time series of length n to be reduced to a string of length l ($l < n$). The algorithm has two parameters: window length w and alphabet size α, and it involves three main steps [85]:

  1. Normalization: standardizes the data in order to have a zero mean and a standard deviation of one;
  2. Piecewise Aggregation Approximation (PAA): divides the original data into the desired number of windows and calculates the average of data falling into each window; and
  3. Symbolization: discretizes the aggregated data using an alphabet set with the size represented as an integer parameter α, where $\alpha > 2$.

- Shapelets are subsequences of time series that identify with the class that the time series belongs to.

- Ensembles. COTE algorithm [86] uses a collective of ensembles of classifiers on different data transformations.

  The ensembling approach in COTE is unusual because it adopts a heterogeneous ensemble rather than resampling schemes with weak learners. COTE contains classifiers constructed

in the time, frequency, change (autocorrelations), and shapelet transformation domains (35 in total) combined in alternative ensemble structures. Each classifier is assigned a weight based on the cross validation training accuracy, and new data are classified with a weighted vote.

In the reviewed literature for time series representation we can see that there are some facts that remain unsolved:

- Outliers and noise: when data are coming from sensors and physical devices usually contains noise and outliers that affect the identification of the correct parameters of the distribution.
- Data follows different distribution: some scenarios in which data does not follow a normal distribution, as assumed by some methods in the literature [84], are: radioactive decay (exponential distribution), number of cars passing through a point in a period of time (Poisson distribution), queuing models (gamma distribution), batting averages of baseball players (beta distribution).
- Fast data: two of the V's from the 7V's Big Data challenges [37] are *velocity* and *variety*. Traditionally in data mining, already collected data are processed in an offline manner using historical data. However, in IoT applications, we need to consider short-term snapshots of the data which are collected very quickly. The data are represented as streams and it can change (locally or globally) over time. Thus, we need adaptive methods that catch up with the changes and update the models online during their operation.

Taking the above-mentioned cases into account, we seek an algorithm that does not require normalization of the data. The latter will also help to preserve the value of the data points (i.e. magnitude of the data). The lack of sensitivity to magnitude in the algorithms that make assumptions about the normalized distribution and use Z-normalization makes them less efficient in analysing correlations and relatedness measures. Another requirement is the application of the algorithm in an online way and using sliding windows. Nonetheless, we have to be able to compute the distance between the aggregated time series.

3.1.3 Feature selection

An FS method is a search strategy where the performance of candidate subsets is measured with a given evaluator. A stopping criterion establishes when the FS process must finish. FS methods are typically categorized into wrapper, filter and embedded, univariate and multivariate methods. Wrapper methods [87] use a predetermined learning algorithm to determine the quality of selected features according to an evaluation metric [88]. Filter methods apply statistical measures to evaluate the set of attributes [89, 90]. Embedded methods achieve model fitting and

FS simultaneously [91]. Multivariate methods evaluate features in batches. Univariate methods evaluate each feature independently. Figure 3.1 illustrates graphically the FS flow.

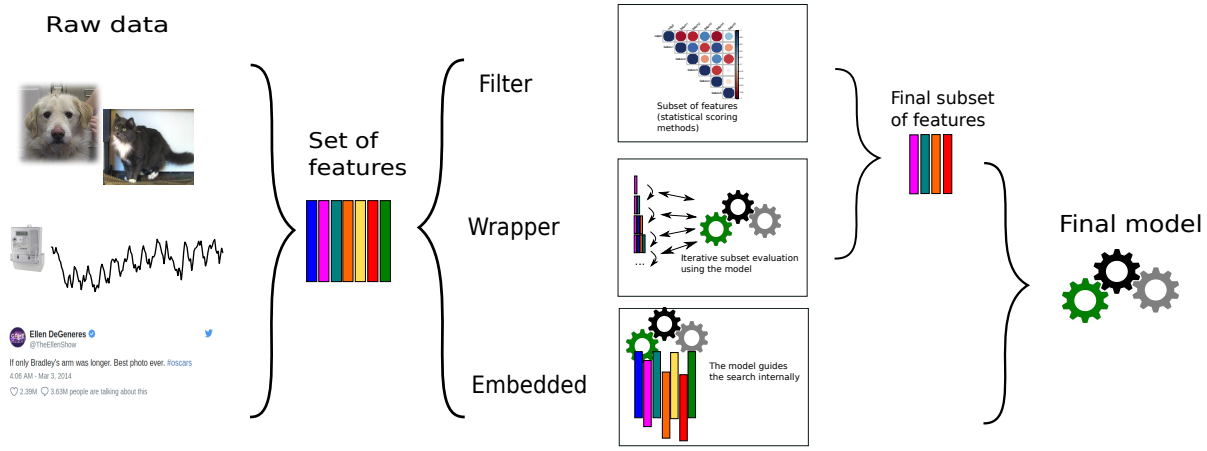

Figure 3.1: The FS flow

We have done an extensive search in order to find academic works that have carried out FS. Together with the works that address FS for energy consumption time series, we have also considered important to review FS for energy consumption when not treated as time series, and FS for time series problems in general, i.e. other approaches not specifically related to energy consumption.

The first paper that studied how the selection of subsets of features associated with building energy behaviours influences a ML model performance for energy consumption prediction used some filter methods for FS and support vector regression for forecasting [92]. A bit later, in the thesis [93], Fast Correlation-Based Filter (FCBF) is used for FS in load prediction error problems in four building areas. A meteorological dataset from several locations and also, the geographical factor are exploited by selecting variables from different locations. The baseline comparisons are done with e-SVR. According to this work, how the relationships between features change with distance motivates a greedy FS method for the electrical load forecasting. In the works [7, 4], correlation and principal components analysis (PCA) are used for FS and transformation.

FS for time series prediction has been carried out using neural networks [94]. By combining contemporaneous and lagged realisations of the independent variables and lagged dependent variables more general models of dynamic regression, autoregressive (AR) transfer functions and intervention models are constructed. It has also been done using the Granger causality discovery [95] to identify important features with effective sliding window sizes, considering the influence of lagged observations of features on the target time series.

The optimal time-windows and time lags for each variable based on feature pre-processing and sparse learning in order to configure the input dataset were searched in [96].

The forecasting of solar radiation time series in enhanced by using a train set of bootstrapped SVM in order to perform FS was done in [97]. They assure that this method is more robust than a

regular FS approach because using the later, small changes on the train set may produce a huge difference on the selected attributes. For solar radiation prediction, [98] masks the inputs as a FS step . They create their own features by defining night, sunrise, day and sunset according to the moment that their instruments perceive those. This provides certain improvements in forecast accuracy. A data-driven multi-model wind prediction methodology using a two-layer ensemble ML technique is developed in [99]. A deep FS framework is employed where four different approaches are used in order to get the input vector: PCA, Granger Causality Test, Autocorrelation and Partial Autocorrelation Analysis, and Recursive Feature Elimination. Another ensembles way of selecting features is presented in [100] and it is used for predicting the number of incoming calls for an emergency call centre in a time series manner. They use five algorithms (ReliefF, PCA, Freq. Discretization, Information Gain and K-means) that are different in nature and combine the rankings computed grouping similar approaches and computing new weights as the mean of the individual weights. After that, all variables that are ranked among the top five positions in at least three of the groups compound the selected features. In the thesis work [101] they present three case studies in which FS is a step in the model creation. They used the following methods: sequential forward/ backward selection (SFS, SBS), sequential forward/ backward floating selection (SFFS, SBFS), the n best features selection (nBest) and the best individual features.

The main data characteristics of energy time series have been specifically analysed in [102]. To explore such data from different perspectives they consider two main categories: nature (nonstationarity, nonlinearity and complexity characteristics) and pattern (cyclicity, mutability or saltation, and randomicity or noise pattern). After that, FS for electricity load forecasting was done in a time series manner using correlation and instance based methods [103]. In [104] it is presented a survey on data mining techniques for time series forecasting of electricity. The survey focuses on the characteristics of the models and their configuration. Wrapper methods, ANNs, mutual information, autocorrelation and ranking based methods are mentioned as FS techniques used in the prediction of energy consumption. Finally, the work [9] uses temperature time series together with the day of the week in order to estimate energy consumption.

To conclude, in several studies for energy consumption forecasting FS is not discussed and when it is, it does not account for the temporal characteristics of the data and it is not carried out systematically looking for the best method. Regarding the papers that focus on FS for time series prediction [94, 95], we highlight that the focus of [94] is narrowed to neural networks which is not the best for every situation since usually, neural networks are more computationally expensive and require much more data than traditional algorithms. Also, the *No Free Lunch* theorem [105] suggests that there can be no single method which perform bests on all datasets. [95] is focused on the Granger causality as FS so none of them provides a systematic comparison between the possibilities available in the FS field. There is no paper that carries out a systematic combination of univariate, multivariate, filter and wrapper methods and also checks their performance using

several predictive algorithms.

There are two main objectives that have not been considered at the same time, that is, minimise the forecasting error and also the number of variables to be used. This is an important gap that is fulfilled in this work.

3.1.4 HVAC usage patterns

Buildings occupants undertake adaptive actions in indoor environments either to change or adapt to it [106]. Within buildings, it has been seen that occupants have a substantial impact on the energy consumption [107], and it is for that reason that several studies have been carried out to understand the behaviour of building occupants [108, 109]. Also to try to reduce energy use via the change in occupant's habits [110]. Some adaptive actions that occupants can carry out for adjusting their thermal conditions are opening/closing windows and doors [106], adjusting thermostat set points [111], too hot/too cold complaint calls [112], and repositioning their window shades [113]. Research on studying these adaptive occupant behaviours has been focussing on understanding their impact on the energy performance of a building [114]. It has been shown by the literature that this change in habits can result in some cases to 20% savings [115, 116]. Given that the 50% of EU's final energy consumption is used for heating and cooling, of which 80% is used in buildings [117] HVAC are a crucial subject of study if we aim to reduce energy consumption.

The majority of studies that involve HVAC information focus on identifying anomalous behaviours in order to detect faulty equipment [118, 119]. However, the thermostat control is the main behaviour to regulate the thermal comfort, that the most energy-consuming aspect of a building. A survey of 1134 homes in England found a wide variation in thermostat settings which, in the interest of energy efficiency and sustainable development, could form the foundation of a "social norm" campaign aimed at reducing temperatures and energy use in "overheated" homes [120]. Also, the heating set point was related to outside air temperature, relative humidity and wind speed using data from 13 Danish dwellings [121].

Little research has been found in the extraction of thermostat behaviour. Two of the most relevant studies [120, 121] only consider set point for heating, leaving unstudied the cooling part of the thermostatic control. Also, the two previous studies focus on dwellings leaving unstudied non-domestic buildings. In addition to that, the study of the use of the thermostat in time seems to be under looked. This could be the result of not being able to capture such data.

In the following, we propose to associate the thermostat behaviour with energy waste.

3.1.5 Human mobility patterns

In recent years, various works have considered the processing of spatiotemporal traces for mining information about how people move [122]. These digital breadcrumbs can be collected from several sources like motion sensors [123] or smart cards [124]. GPS traces are one of the most

popular data sources in this field and social network data has been used for unravelling the goals of this movement [125]. Trajectory pattern mining examples are frequent item mining, trajectory clustering, and graph-based trajectory mining [126]. More novel approaches use historical routes to generate probabilistic models. Examples of these novel approaches are listed below.

Location Prediction

Location prediction is based on the assumption that people follow daily routines and, thus, have only a set of frequently visited locations [127]. This makes regular trips quite predictable due to their high level of repetition [128].

There are two main trends for personal location prediction: (i) using a geometrical approach so as to predict a path in the Euclidean space [129] by applying a mathematical function to the current location and velocity of the target person and (ii) pattern matching solutions that compare the route in progress with a set of mobility models. In that sense, Bayesian networks [130] and hidden Markov models [131] have been some of the applied solutions given high-resolution mobility datasets (e.g., those comprising GPS-based traces).

Regarding social media datasets, the probabilistic model W^4 [132] follows a Bayesian-network approach that forecasts the next location and activity of a user by also taking into account temporal factors. Mobility features selection is also studied [133].

Social Media for Human Mobility

It is possible to classify social media for human mobility under three different categories.

Firstly, processing geotagged tweets in order to create spatial regions depending on their usage (e.g., leisure, home, and work) using visual analytics [134], clustering algorithms [135] or classifiers [136]. Secondly, automatically extracting events (e.g., live shows, earthquakes [137], traffic jams or anomalies, etc.). For instance, [138, 139] creates smart social agendas that can be updated in real time. In a road-traffic monitoring scenario, several works make use of social media data in order to either detect or semantically enrich traffic anomalies. Correlating tweets [140] and using official traffic-management institutions' Twitter accounts [141] in order to detect road-traffic incidents are some examples. The third category includes social media as a new data source for detecting the movement of different kinds of people among places such as dynamic labelling by semantical enrichment of spatiotemporal trajectories [142] and the statement of a worldwide mobility report based geotagged Twitter data [143].

Mobile Crowd Sensing (MCS) Based Mobility Mining Solutions

An important line of research makes use of MCS for mapping activities by collecting the spatiotemporal traces of contributors [144, 145] by composing collaborative maps comprising road networks, bike and hike routes.

Also, works on real time traffic monitoring propose the distributed architectures to keep track or predict road traffic congestions within an area by means of the mobility reports generated by the on-board units of vehicles [146, 147]. As a result, now we can find solutions that combine static and vehicle-mounted and smartphone sensors to detect road traffic in an area [148, 149].

Regions of Interest (ROIs) Detection

Given a collection of spatiotemporal traces, different types of clustering algorithms can be applied to the spatiotemporal traces to uncover the target ROIs. Density-based clustering has been a prominent solution [150]. Distance between GPS points or density connectivity in a two-dimensional Cartesian space are used as features. A different approach makes use of frequency map based spatial-temporal clustering methods [151].

Complex Event Processing (CEP)

CEP is an evolution of the former publish/subscribe model that deals with more complex subscription patterns, so it can be considered a recent technology [152]. Despite CEP's widespread usage, there exists a scarcity of CEP solutions that use spatiotemporal data since only a few works actually propose practical CEP applications [153]. The GPS-based solution [154] formulates a framework to timely detect spatiotemporal relationships between moving entities.

The reviewed studies frequently suffer from some of the following limitations:

- Reliance on GPS traces datasets even though GPS feed is one of the most battery- draining sensors of a mobile device.
- Either they centre on extracting general mobility information related to a particular urban area or use models for every single user as W^4 [132]. In that sense, detecting personal mobility models that also count with the crowd-dynamics could be of interest to come up with personalised but also informed location services.
- The whole available dataset, which sometimes might not be available, is required for preprocessing.
- Moreover, most works using social networks do not fully unlock its potential since they only use the spatiotemporal aspects of the data (check-in posts) but do not consider textual details.
- The anticipation of peoples' activities and locations using geotagged social media documents is scarce.
- The mobility knowledge extracted by the aforementioned solutions focuses on road traffic features so it depends on the road network topology. We plan to capture human dynamics from a wider perspective.
- Only theoretical solutions to define formal event-based information models and architectures for social media processing have been put forward using CEP earlier [155].
- Detecting ROIs related to the movement of people instead of where people tend to remain stopped is missing in the literature.

3.1.6 IoT architectures and projects for smart cities and energy management

Due to the importance of the building sector in energy consumption, it becomes a foremost task to achieve meaningful energy savings that will reduce this energy use in reality. Despite the fact that IoT technologies have been widely used for the realization of the smart building concept, the simple sensorization of buildings is not enough to make a housing stock that consumes fewer energy resources a reality. IoT is also required to properly process, manage and, above all, analyse the energy-related data that would help to develop final energy-aware services targeting the energy efficiency goal.

An overview of both the management of energy data and the implementation of IoT platforms is put forward.

During the last years, some initiatives within the cloud computing domain have been made to intelligently manage energy data of buildings. In that sense, Big Data energy management models have been created ranging from the collection and preprocessing of data to its further analysis and the final exposition to services [156].

From a practical perspective, the Dynamic Demand Response platform [157] makes use of public and private clouds combined with infrastructure and platform as a service for data storage. This platform was extended with Cryptonite, a repository to store sensitive Smart Grid data [158]. Then, different classes of data-driven forecasting models were generated on top of the whole platform with the purpose of carrying out energy prediction among others.

ElasticStream also provides a prototype solution for energy data management and analysis. In this case, the mechanism transfers energy data to a cloud platform for further analysis on the basis of rate changes in the input data streams [159].

The MultiAgent System (MAS) named SAVES (Sustainable multiAgent systems for optimizing Variable objectives including Energy and Satisfaction) defined in [160] is used in [161] regarding actual occupant preferences and schedules, actual energy consumption and loss data measured from a real testbed building at the University of Southern California in order to predict energy consumption at different levels (frequency of prediction and device aggregation). Other works provide energy data management solutions without focusing on analytic aspects. This is the case of the Virtual SCADA architecture for cloud computing (VS-Cloud) that encompasses Cloud Computing for energy data storage [162]. VS-Cloud mainly focuses on the orchestration of components in Smart Grids and the save storage of sensitive data executed actions, incidents or alarms. Therefore, its domain of application is more related to risk management. Similarly, the work in [163] proposes an automation platform for energy monitoring. However, such a platform does not provide any particular feature to support energy data analytics as it focuses more on the definition of control strategies for energy saving.

When it comes to the development of IoT solutions, most of them are just vertical silos that do not support interoperability and with inappropriate models of governance. For that reason,

some architectures and platforms have been developed to lower those barriers.

IoT-A, the IoT Architecture (EU project from 2009 to 2012)[1] defines an Architecture Reference Model (ARM) which ensures the interoperability also scalability requirements and the security and privacy in its design, which are so often neglected. This solution rests upon the creation of an architecture reference model together with an initial set of key building blocks, principles and instructions for the technical design of the protocols, interfaces and algorithms suitable for any IoT system.

Webinos[2] creates an Open Source Platform and software components for the Future Internet in the form of web runtime extensions, to enable web applications and services to be used and shared over a large amount of connected devices in a consistent and secure way.

Buttler[3] platform is a set of enablers and services that supply means for building context-aware applications on smart things. It provides generic APIs to access resources provided by IoT devices and other services such as security, localization, behaviour prediction and context management. Those services enhance the user experience and security. The BUTLER Platform is oriented for IoT devices and applications and provides homogeneous access to the underlying networks.

FI-LAB[4] conforms live instances of FIWARE[5] architecture and generic enablers, for free experimentation. FIWARE is an open Core Platform of the Future Internet, introducing an innovative infrastructure for cost-effective creation and distribution of digital services, providing security guarantees. FI-LAB forms a meeting point between sponsors and application developers.

Nowadays, several analytics software packages are available for buildings. SkySpark[6] is well-known and it mainly runs personal rules that depend on the data collected in a building and identifies non-obvious operational problems. The ability to use artificial intelligence, instead of writing custom programming, to extract knowledge in operational data should be exploded.

Some European projects dedicated to combining IoT platforms and energy management are:

- SINFONIA[7] created a set of measures that include optimisation of the electricity grid and solutions for district heating and cooling in order to set up a large-scale, integrated and scalable energy solutions in mid-sized European cities. SINFONIA allowed the cooperation between cities that belong to the same climate zone with the goal of reducing energy needs to meet people's activity requirements, and the consequent CO_2 emissions in order to guarantee a reliable and progressive transition toward low carbon cities.

- CityPulse[8] provided reliable knowledge extraction techniques that were used to create new

[1] https://cordis.europa.eu/project/rcn/95713/factsheet/es
[2] https://cordis.europa.eu/project/rcn/95713/factsheet/es
[3] https://cordis.europa.eu/project/rcn/101349/factsheet/en
[4] https://account.lab.fiware.org/
[5] https://www.fiware.org/developers/
[6] https://skyfoundry.com/
[7] https://cordis.europa.eu/project/rcn/197825/factsheet/en
[8] https://cordis.europa.eu/project/rcn/109806/factsheet/en

smart city applications. This was done by developing, building and testing a framework for real-time IoT stream processing and large-scale data analytics.

- e-balance[9] goal was the integraton of customers into the smart-grids in order to tackle present and future environmental problems with ICT based solutions, new business models and citizens' behaviour in real world conditions. It focused on investigating the economic and social aspects of the energy efficiency, integrating ICT for decentralized power management providing more autonomy for delivering local decisions including intelligent power generation, consumption.

As above mentioned, several multi-purpose IoT platforms already provide generic solutions to manage IoT data. However, there is a lack of platforms in this field focusing on (1) the household energy domain and (2) providing support for data analytics.

Some energy models only provide a theoretical approach [156]. Also, the aforementioned initiatives do not constitute holistic energy data management and analysis solutions. The platforms do not include explicit features that are necessary for all energy monitoring scenarios like data volatility monitoring and outliers detection to ease the deployment of data mining algorithms and other services over of the stored data.

Another neglected feature by existing IoT platforms is the support of built-in data mining features able to generate new useful knowledge from the collected and stored data [164]. In real IoT deployments, this processing and analysis task has been frequently done by third-party services. However, integrating certain data mining functionalities as built-in features of platforms would provide a great benefit in a wide range of domains, for example quick statistics, easy to generate digests or sanity checks. In that sense, only a few IoT platforms actually include native data analytics features. As a matter of fact, SensorCloud[10] enables a simple interface for common operations like smoothing, filtering and interpolation whereas GroveStreams[11] provides some real-time data analytics mechanisms. However, none of them supports sensor heterogeneity nor follows an open source approach.

3.1.7 IoT architectures and projects for behavioural analysis towards energy efficiency

In order to realize a sustainable energy transition, human behaviour should be considered regarding educational aspects (raising awareness of the benefits) and technologies understanding [165]. Whilst new technologies and materials are available, certain initiatives are required in order to encourage individuals to use them properly.

[9] https://cordis.europa.eu/project/rcn/109806/factsheet/en
[10] https://sensorcloud.com/
[11] https://grovestreams.com/

There is a scarcity of academic works that evaluate how behaviour-change interventions affect energy efficiency. [166]. Organisations have initiated interventions based on holistic programmes including gamification and rewardings for real-live changes [167, 168], comparative feedback and competitions [169], various feedback types toward energy savings [170], prompts, peer-education and dashboards [171].

Several studies try to find how and why the installation of smart metres changed households' electricity use [172]

Some programmes combine these tactics with installing new technological features such as a system that determines activity in homes by integrating motion, door, lighting and temperature sensors which allows adapting the services depending on the behaviour patterns [173]. Also, modelling residents' behaviour by studying presence, lighting and window status sensors can decrease energy consumption by up to one third [174].

There are many European projects that are targeting the integration of technological advances for incentivising behavioural change towards energy efficiency. Some of them are:

ChArGED[12] (CleAnweb Gamified Energy Disaggregation) created a framework that encourages the achievement of energy efficiency and the reduction of wasted energy in public buildings. The framework set up low-cost devices to improve energy disaggregation at several level. Energy waste is targeted by a gamified application that feeds personalized real-time recommendations to individuals. The game is designed in a way that helps users to understand how their actions affect the environment.

enCOMPASS[13] (Collaborative Recommendations and Adaptive Control for Personalised Energy Saving) developed and validated several digital tools that make energy consumption data accessible and understandable for all stakeholders: from residents to building managers and ICT-providers. Those tools were integrated and provide visualisation of energy-related data, energy-saving recommendations that depend on the context, intelligent control and adaptive gamified incentives.

TRIBE[14] (TRaIning Behaviours towards Energy efficiency: Play it!) is based on the development of a serious game implemented through social networks (for information and experience exchange) for the energy sector in which building users adopt energy efficient attitudes. It includes a simulation engine and real time collection data from the ICTs installed in the pilots, enabling a dynamic interaction building-consumer and moving towards a change in players' behaviour. In addition, some tools and guidelines were set up to be used by users and owners of public buildings, including: (1) an initial energy audit and diagnosis, (2) the creation of a virtual pilot similar to the real buildings, (3) an energy efficiency deployment plan based on ICT, (4) a funding plan (5) a user engagement campaign for the detected behaviour change challenges.

[12] http://www.charged-project.eu/
[13] http://www.encompass-project.eu/
[14] http://tribe-h2020.eu/

As can be seen in the literature, there has not been a work that combines the methods, analytics and properties developed for designing an ecosystem that targets at improving energy efficiency through consumers' understanding, engagement and behavioural changes.

Regarding data, they focus either on survey data [172] or on monitored data. The combination of this two ways of data acquisition is not carried out. However, surveys can be of great help at the beginning of an energy efficiency campaign, when data is not available in order to make a profile of the users that can later on be transformed due to the sensed data and performance evaluation of users.

Regarding methodology, they focus either on the competition factor [169] or in the provision of feedback depending on the specific users' internal values [26], which is not updated or verified regarding its behaviour towards the tools.

Within the little number of available platforms, they do not include the possibility of incorporating algorithms. This is a very limiting characteristic since keeps the platforms isolated from the fast evolution of analytic advances.

A unifying platform that gathers all those characteristics is needed in order to achieve good results regarding energy literacy and consequently efficient behaviour towards energy.

3.1.8 Related work summary

Every subsection in this related work ends with a reflection on the missing parts detected on the reviewed literature of the subject. Briefly this includes the application of machine learning techniques for energy consumption forecasting using external predictors, the comparison between black-box and grey-box models and the lack of a complete methodology for multivariate feature selection for energy consumption prediction. Regarding pattern extraction, we found that the set point of HVACs are not fully exploited and human mobility pattern extraction can be improved using social networks data. The collected data for these scenarios is now huge and its volume implies low performance. We found that methods for data reduction do not take into account IoT and Big Data characteristics: variability, volume, velocity, etc. Looking at the platforms, there are no specific ones that address energy management and provide support for data analytics at the same time. The majority of the attempts to do so do not take into account the human behaviour factor, that is also key for the achievement of realistic energy efficiency.

In the following, we expose how our work tackles each of the previously stated gaps, ordered by result as presented in Table 2.1.

3.2 Data analysis in IoT based Smart Environments

In urban environments, there is a wide variety of data sources. A wealth of sensors are distributed around cities, in both indoor and outdoor spaces. This situation has brought new analytics mechanisms and tools that provide insight into the data which allows building powerful systems

and applications in an efficient and collaborative way [175]. Mobile sensors are playing a major role in the development of applications too since they are embedded in our lives as smartphones, smart cards, wearable technology and, in the case of vehicles, on-board sensors. Urban dynamic patterns can be detected and studied thanks to the information that these sensors provide. In the following, we summarise the results of the dissertation. Every subsection refers to a result that is presented in Table 2.1. In the same table, we can see the publications' references to each of the results.

3.2.1 Smart buildings data integration and statistical analysis [R1]

In order to create ML models that forecast energy consumption, it is crucial to understand the relationship between energy consumption and other attributes. Possible influencers or input attributes for energy consumption prediction are weather variables and occupation patterns that refer to the day of the week, the hour of the day, the kind of day, etc. In order to do so, the integration of 3 smart buildings datasets from 2 different buildings formed by collected data from several sources is carried out. Table 3.1 contains a summary of the buildings' information. The first dataset belongs to the TTC Fuente Alamo [15] of the University of Murcia (see Fig. 3.2 left). This data consist of the environmental outdoor observations and the total energy consumption of the building from 1st December, 2014 to 18th February, 2016 in intervals of 8 hours and the origin of the consumption (HVAC, lighting or other electrical equipment) is unknown. In total, 952 observations and 15 variables. Outdoor environmental measures are acquired from The Research Institute of Agriculture and Food Development of Murcia (IMIDA)[16] that provides real time records of weather from several stations across the region of Murcia. The following variables are included in the dataset: temperature (mean, min and max) (°C), humidity (mean, min and max) (%), radiation (mean and max) (w/m^2), wind speed (mean and max) (m/s^2), wind direction (mean) (degrees), precipitation (mm), dew point (°C) and vapour pressure deficit (kPa). Station CA91 with latitude 37.699033 and longitude 1.238044.

The second dataset belongs to the Faculty of Chemistry at the University of Murcia (see Fig. 3.2 right) and it is composed of 5088 observations of 50 attributes that are measured hourly from 2016-02-02 until 2016-09-06. The output attribute is the energy consumption measured in KWh and we have included meteorological measurements from 2 different meteorological stations from IMIDA that are close to the building (MO12 lat:38.007031 long: 1.302564 and MU62 lat: 37.940067 long: 1.134719), hour ahead predictions and other weather related variables provided by Weather Underground[17] web service and also season, day of the week and holiday attributes are inlcuded.

Weather Underground is a web service that through its API provides the following real

[15] https://www.um.es/web/otri/contenido/ctt
[16] http://www.imida.es/
[17] https://www.wunderground.com/

Name & Country	Envelope area (m^2)	Orientation	Coordinates	Cons. year
TTC Fuente Alamo. Spain	3323	South-West	Latitude: 37.724383 Longitude: 1.093324	2004
Chemistry faculty. Spain	1500	South-West	Latitude: 38.020939 Longitude: -1.169722	1944

Table 3.1: Information about the buildings

values: temperature (° C), apparent temperature (° C), dew point (° C), humidity (%), wind speed (m/s), mean sea level pressure (mbar), visibility (km) and precipitations in last hour (mm). We also use *one-hour predictions* for the first six previous attributes, together with *probability of precipitations* (%), *sky cover* (%) and *wind direction* (degrees) .

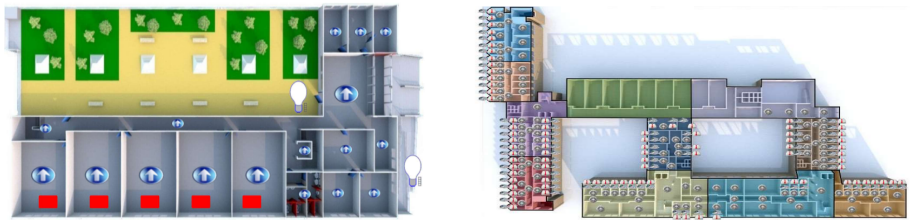

Figure 3.2: 1st floor of the TTC where red labels means energy meter (left) and 2nd floor of the Chemistry Faculty (right)

Finally, we have also monitored the use of HVAC systems in 237 Chemistry Faculty rooms. The dataset consist of 12-minutes aggregated observations regarding room temperature, on/off status and set point from 2015-10-31 until 2017-02-28.

Fig. 3.3 shows the pairwise correlations between all variables involved in the energy consumption forecasting problem. A blue circle indicates positive correlation and a red circle indicates negative correlation. The radius of the circle indicates the magnitude of the correlation. In TTC (left) and focusing on the first row, we see that energy consumption correlates significantly (α = 0.05) and positively (blue circle) with temperature, radiation, wind speed variables, vapour pressure deficit and dew point, and negatively (red circle) with wind direction and humidity variables. This means that we can use safely these variables as inputs of the energy consumption model of our reference building, because they all have a clear impact on the energy consumption except precipitations (crossed out because they are not significant) [7]. Since we collected more attributes for the Chemistry faculty dataset at the right side of Fig. 3.3 we have shown only the relationship between energy consumption and the rest of the variables. In this case, we should exclude from the analysis all variables related to precipitations, dewpoint and sky visibility.

In this preprocessing stage, the study of correlations can be complemented con PCA analysis. PCA is a method that reduces the dimensionality of the data by creating new uncorrelated variables that successively maximize variance. Those new variables are found by solving an

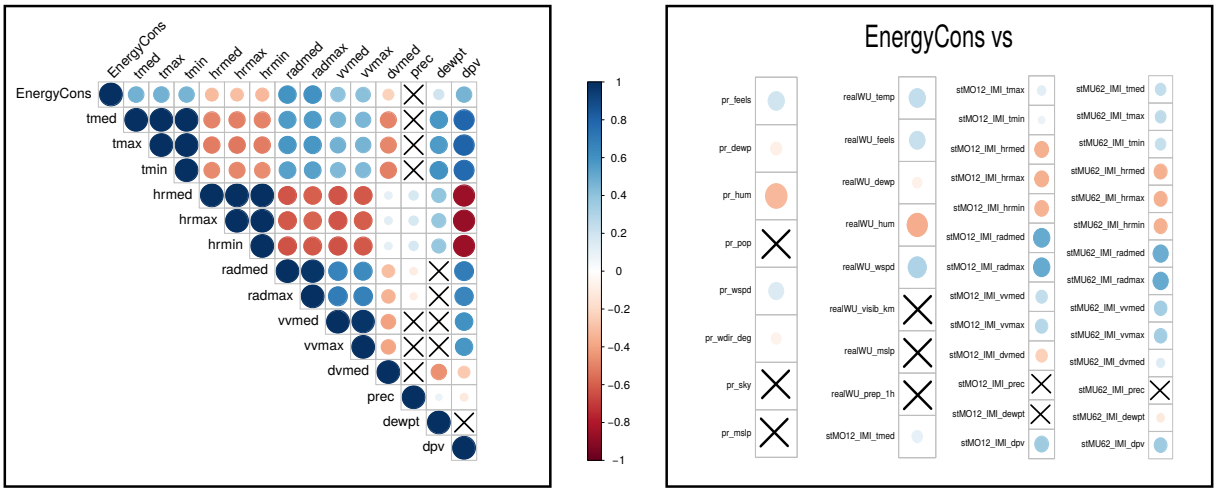

Figure 3.3: Correlation heatmap between consumption and outdoor environmental conditions for both consumption datasets

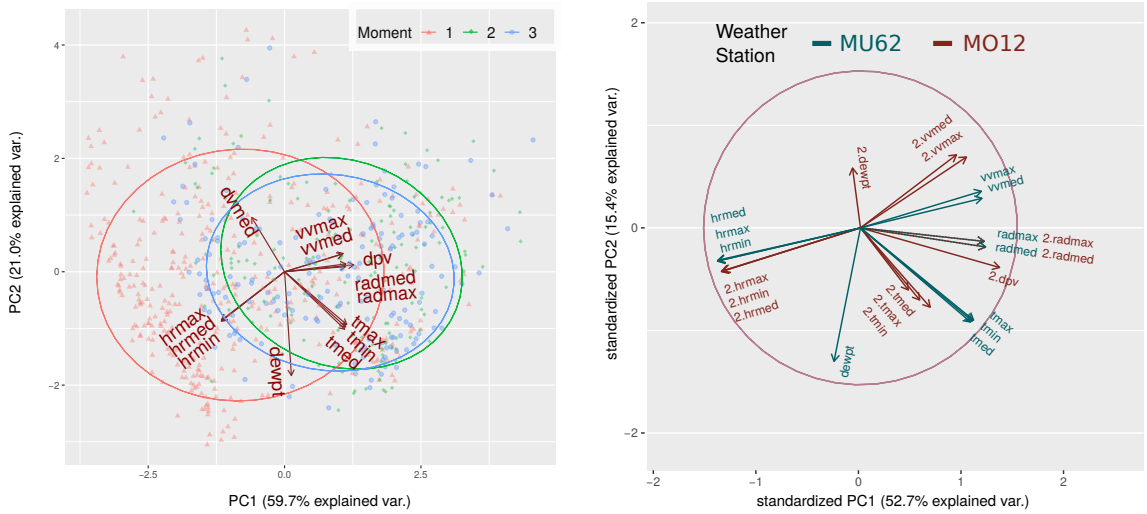

Figure 3.4: Correlation heatmap between consumption and outdoor environmental conditions for both consumption datasets

eigenvalue/eigenvector problem [176]. It is especially important to apply it when inputs might be correlated, as it is our case since dependent variables can hinder ML models' accuracy.

PCA provides intuitive visualizations for exploring relationships in data since subjects are projected into the new dimensions, which are linear combinations of the initial features. In both cases of Fig. 3.4 around 60 % of the variance can be explained with 2 dimensions. The correlation circle information gives a similar result than Fig. 3.3 and also the first dimension in both cases mainly reflects radiation. There are certain parallelisms between PCAs of both datasets, however, something to be noted at the right images is that dew points from different sources are negatively correlated and there exist differences between both sources for wind velocity.

3.2.2 Data representation [R2]

Due to the huge volumes of data that are provided in smart environments, it is of special interest to investigate methods for data reduction. We attempt to aggregate and represent large volumes of data in an efficient and higher-granularity form in order to create sequences of patterns and data segments that occur in large-scale IoT data streams. The contribution of our approach is to do such representation on-the-fly since usually data treatment has to be done very quickly, adapting to unpredictable changes in the data or even without prior knowledge.

Our proposed method is based on splitting time series data into blocks. These blocks can be either overlapping or non-overlapping and they represent subsets of the whole data structure. The method synthesizes independently the information that the blocks contain. It reduces the data points while still preserving their fundamental characteristics (losing as little information as possible). We propose a novel technique using matrix-based data aggregation, DCT and eigenvalues characterization of the time series data. The algorithm is called Blocks of Eigenvalues Algorithm for Time series Segmentation (BEATS).

Transforms, in particular, integral transforms are used to reduce the complexity in mathematical problems. In order to decorrelate the time features and reveal the hidden structure of the time series, they are transformed from the time domain into other domains.

DCT uses cosines obtained from the discretization of the kernel of the Fourier Transform. DCT transfers the series to the frequency domain. Among the four different cosine transformations classified by Wang [177], the second one (i.e. DCT-II) is regarded as one of the best tools in digital signal processing [178] (times series can be regarded as a particular case of signals). Due to its mathematical properties such as unitarity, scaling in time, shift in time, the difference property, and the convolution property, DCT-II is asymptotically equivalent to the KLT where under certain (and general) conditions KLT is an optimal but impractical tool to represent a given random function in the mean square error (MSE) sense. KLT is said to be an optimal transform because:

- It completely decorrelates the signal in the transform domain;
- It minimizes the MSE in bandwidth reduction or data compression;
- It contains the most variance (energy) in the fewest number of transform coefficients; and
- It minimizes the total representation entropy of the sequence.

The details of the proof of the above statements can be found in [178]. Understanding the properties of the DCT, we use it to transform our time series data.

We apply the transformation essentially by using the compression of a stream of square 8x8 blocks, taking reference from the standards in image compression [179] where DCT is widely used (e.g. JPEG). Since 8 is a power of 2, it will ease the performance of the algorithm.

As an illustration, we provide an example in Fig. 3.5. We have divided the time series as blocks of 64 observations that are shown using a dashed red line. If we arrange the first block

into a squared matrix M , we can visualize that the information is spread through the matrix as a heatmap. Intuitively, each 8×8 block includes 64 observations of a discrete signal which is a function of a two-dimensional (2D) space. The DCT decomposes this signal into 64 orthogonal basis signals. Each DCT coefficient contains one of the 64 unique *spatial frequencies* which comprise the *spectrum* of the input series. The DCT coefficient values can be regarded as the relative amount of the spatial frequencies contained in the 64 observations [179].

Let M be the 8×8 input matrix. Then, the transformed matrix is computed as $D = UMU^{\top}$, where U is an 8×8 DCT matrix. U coefficients for the $n \times n$ case are computed as shown in Eq. 3.1:

$$U_{ij} = \begin{cases} \frac{\sqrt{2}}{2} & i,j = 1 \\ \cos\left(\frac{\pi}{n}(i-1)(j-\frac{1}{2})\right) & i,j > 1 \end{cases} \tag{3.1}$$

After applying DCT, the information is accumulated in its upper-left part.

Each of the 64 entries of the matrix D is quantized by point-wise division of the matrices D and Z, where the elements of the quantization matrix Z are integer values ranging from 1 to 255.

Quantization is the process of reducing the number of bits needed to store an integer value by reducing the precision of the integer. Given a matrix of DCT coefficients, we can divide them by their corresponding quantizer step size and round it up depending on its magnitude, normally 2 decimals. If the maximum of the DCT matrix is small, the number of decimals is selected by the operation $|\lfloor \log_{10} max \rfloor - 4|$, where $\lfloor \log_{10} max \rfloor$ returns the position of the first significant figure of the maximum number in the transformed matrix D. This step is used to remove the high frequencies or to discard information which is not very significant in large-scale observations.

The selected matrix Z is the standard quantization matrix for DCT [180]. After the quantization process, a large number of zeroes appears in the bottom-right position of the matrix $Q = \frac{D}{Z}$, i.e. it is a sparse matrix.

We extract the 4×4 upper-left matrix that contains the information of our 64 raw data and compute the eigenvalues.

Using BEATS so far we have significantly reduced the number of points of our time series from 64 to 4 but we have also converted its components into complex numbers. These complex numbers (eigenvalues vector) represent the original block in a lower dimension. This eigenvalues vector is used in BEATS to represent the segments and hence, it is the potential input for the ML models. However, it is not always possible to feed ML algorithms with complex numbers and the eigenvalues could be complex numbers. To solve this problem, we compute the modulus of the eigenvalues and remove the repeated ones (they are presented in pairs so the information would be repeated).

In case that there are no complex numbers in the output of BEATS, we will conserve the first three values, since the latter values are sorted in descending order. This means that we have represented the original 64 observations as three values. In our example, the final representation (modulus of the eigenvalues) consists of $0.1860, 0.0246, 0.0085$.

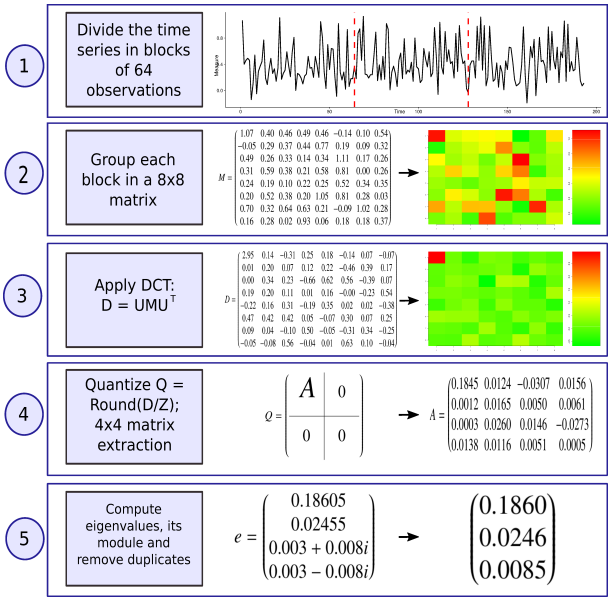

Figure 3.5: BEATS is shown step by step with an example

The BEATS process is summarized in Fig. 3.5.

Big Data BEATS implementation

In contrast to the traditional analysis procedure where data are first stored and then processed in order to deploy models, the major potential of the data generated by IoT is accomplished by the realisation of continuous analytics that allows making decisions in real time.

There are three types of data processing: Batch Processing, Stream Processing and Hybrid Processing.

Batch processing operates over a group of transactions collected over a period of time and reports results only when all computations are done, whereas stream processing produces incremental results as soon as they are ready [181].

Regarding the available Big Data Tools, we have considered Hadoop and Spark Big Data frameworks. Hadoop was designed for batch processing. All data are loaded into HDFS and then MapReduce starts a batch job to process that data. If the data changes the job needs to be run again. It is step by step processing that can be paused or interrupted, but not changed.

Apache Spark allows performing analytical tasks on distributed computing clusters. Spark's real-time data processing capability provides substantial lead over Hadoop's MapReduce and it is essential for online time series segmentation and representation.

The Spark abstraction for a continuous stream of data are called a Discretized Stream or DStream . A DStream is a micro-batch of Resilient Distributed Datasets, RDDs. That means, a DStream is represented as a sequence of RDDs. RDDs are distributed collections that can be operated in parallel by arbitrary functions and by transformations over a sliding window of data (windowed computations).

BEATS adapted to Spark technology

For the online implementation of BEATS we have decided to use pyspark, the Spark Python API that exposes the Spark programming model to Python.

There are many works proposing online time series processing but few of them that have implemented it. In [182] is highlighted that MapReduce is not the appropriate technology for rolling window time series prediction and proposes an index pool data structure.

Pyspark allows us to use the Spark Streaming functionalities that are needed in order to implement BEATS online. BEATS algorithm can be separately applied to windows of the data. Therefore we associate the data received within one window to one RDD, that can be processed in a parallel way.

A suitable type of RDDs for our implementation is key/value pairs. In detail, the key is an identifier of the time series (e.g., sensor name) and the value is the sequence of values of our time series that fall in the window. That way the blocks are exposed to operations that give the possibility to act on each key in parallel or regroup data across the network.

The transformations that we use are:

- Window: used for creating sliding window of time over the incoming data.
- GroupByKey: grouping the incoming values of the sliding window by key (for example, same sensor data).
- Map: The Map function applied in parallel to every pair (key, value), where the key is the time series, values are a vector and the function depends on what has to be done.

3.2.3 Energy consumption prediction [R3]

Energy consumption prediction is a complex task that depends on the context for which the forecast is needed. We have studied several short-term horizons: hourly, 3 moments a day, daily and weekly. For each of them there exists the possibility of predict instances as independent subjects adding temporal characteristics as extra variables (day of the week, hour of the day, etc.) or use ARIMA and temporal models. Also, the possibility to use real or predicted inputs depends on the application:

- Campaign evaluation: it is possible to use real weather and occupation information, but close prior consumption information would bias the experiment. If a model depends on the consumption of the previous day, on the second day of the campaign this model will be biased because previous day consumption is altered by the campaign measures.
- Resources optimisation: it can use prior energy consumption, but no real information regarding the future so occupation, weather and other inputs must be also estimated.

3.2.3.1 The models

Support Vector Regressor (SVR) works in a similar fashion to Support Vector Machine (SVM). Whereas SVM is a classification technique, SVR fits the optimal curve out of which the training data do not deviate more than a small number ϵ. More specifically, during classification the samples that are close to the margin are penalized even if they are correctly classified, whereas in the regression method an acceptable deviation margin of the samples from the prediction curve is set. The free hyperparameter of this model is C, the penalty parameter of the error term. C is the weight of how much the samples inside the margin contribute to the overall error.

Random Forest (RF) is an ensemble learning method in which a group of weak models are combined to form a more powerful model. In RF, multiple regression/classification trees are grown from random with replacement samples. Each tree provides its own prediction and all results are averaged. For each node, m_{try} (hyperparameter) variables are selected at random out of the number of inputs. The best split in these m_{try} is then used to split the node [183].

eXtreme Gradient Boosting (XGB) is built on the principles of gradient boosting and is designed for speed and performance (extreme). XGB generates a prediction by means of an ensemble of weak prediction models that, in our case, are decision trees. The concept is to sequentially build the model by fitting a weak prediction model on the weighted training data set, in which the higher weights are assigned to samples that were previously difficult to predict. The free hyperparameters that are adjusted in this model are the maximum depth limit of the number of nodes in the tree, the minimum number of samples required to split an internal node and the learning rate by which the contribution of each tree is shrunk.

The **Artifical Neural Networks** (ANN) here used are Multilayer Perceptron (MLP) and Bayesian Regularized Neural Networks (BRNN). ANN consist of three layers: input (contains independent variables), hidden, and output layers. The hidden layer contains activation functions and it calculates the weights of the variables in order to explore the effects of predictors. In the output layer, results are presented with an estimation error. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output. However, because of the big number of connections, overfitting may occur. Regularization techniques with the backpropagation training are used in order to have a smoother network that is less likely to overfit BRNNs are more robust than standard back-propagation nets.

GAUSS with RBF Kernel (GAUSS). A Gaussian process is a random process where any point x is assigned a random variable $f(x)$ and where the joint distribution of a finite number of these variables is itself Gaussian, that is: $p(f|X) = \mathcal{N}(f|\mu, \mathcal{K})$, where $\mathcal{K}$ is a positive definite kernel function. The kernel function returns a measure of similarity between two points that also encodes how similar its realisations should be. If points x_i and x_j are considered to be similar by the kernel the function values at these points, $f(x_i)$ and $f(x_j)$, can be expected to be similar too. Here, we have used the squared exponential kernel, also known as RBF kernel:

$k(x_i, x_j) = \sigma_f^2 exp(-\frac{1}{2l^2}(x_i - x_j)^T(x_i - x_j))$. The length parameter l controls the smoothness of the function and σ_f the vertical variation. [184]

Prophet is a modular regression model with interpretable parameters that can be adjusted with domain knowledge about the time series [185]. Prophet conducts an automatic optimisation procedure for forecasting time-series data by fitting a non-linear additive model and generates uncertainty intervals. Three main model components compose the model: trend, seasonality, and holidays: $y(t) = g(t) + s(t) + h(t) + \epsilon_t$, where

- g(t): represents non-periodic components using piecewise linear model with automatic change point selection or logistic growth curve trend.
- s(t): trend factor that represents periodic changes. Time series often have multi-period seasonality as a result of the human behaviours they represent. This part relies on Fourier series to provide a flexible model of periodic effects.
- h(t): effects of holidays (a list provided by the user). Holidays often do not follow a periodic pattern, so their effects are not well modeled by a smooth cycle.
- ϵ_t : error which will be assumed to follow a normal distribution.

3.2.3.2 Model assessment energy consumption

The RMSE and MAE [186] are two of the most common metrics used to measure accuracy for continuous variables and they are appropriate for model comparisons because they express average model prediction error in the units of the variable of interest as can be seen by their definition in the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \qquad MAE = \frac{\sum_{i=1}^{n}|y_i - \bar{y}_i|}{n}$$

where y_i is the real consumption, $\bar{y}_i$ is the predicted consumption and n is the number of observations.

However, in order to compare energy consumption prediction within works that do not use the same dataset or the same values of energy to be predicted it is not useful to compare such metrics whose output depends on the magnitude of the output data.

For that reason, we complement the information with the coefficient of variance of the RMSE. The Coefficient of Variation of the RMSE (CVRMSE) is a non-dimensional measure calculated by dividing the RMSE of the predicted energy consumption by the mean value of the actual energy consumption. For example, a CVRMSE value of 5% would indicate that the mean variation in actual energy consumption not explained by the prediction model is 5% of the mean value of the actual energy consumption [187]. CVRMSE has often been used in energy prediction studies [188]. Similarly, the Mean Average Prediction Error (MAPE) metric has been used in a wide

number of electricity prediction studies [189, 190]. It expresses the average absolute error as a percentage. They are calculated as follows:

$$CVRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2}}{\bar{y}} \times 100, \qquad MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \bar{y_i}}{y_i}| \times 100.$$

3.2.3.3 Hourly predictions

We were able to control an office based on an occupation schedule in order to test how the inclusion of external weather predictions would influence accuracy on energy consumption. In this case, we computed each hour's consumption prediction in a horizon of 24h.

Data was collected from 10 June to 14 August 2017. In this period, up to 4 workers were working in a normal schedule from 9:00 to 17:00. Equipment was turned on and off and the settings of the conditioning system were modified remotely.

We have tested the Prophet software, which was previously used in other disciplines and found that it could be excellent for experiments that require the prediction of pertinent variables. Our contribution has also been to prove that this package is an ideal "soft" addition to the infrastructure.

There was a positive correlation between the predicted external temperature (1 day before) and the measured external temperature r = 0.9, p-value < 0.01 so it is feasible to anticipate and predict the real outdoor conditions. A significant regression equation was found (F(1,1582) = 6285, p-value < 0.01, R^2 = 0.8. The real temperature is equal to $8.59 + 0.66 \times (predicted)$ ºC.

Energy consumption in buildings has several characteristics appropriate for the Prophet algorithm and thus should perform well for energy prediction: strong multiple human-scale seasonality (such as day of the week and the time of year), important holidays that occur at irregular intervals that are known in advance and a certain random component.

We evaluated two scenarios in order to asses the inclusion of temperature forecasts.

- Model 1: Previous energy consumption, previous occupation and future occupation with a known pattern and schedule (RMSE = 286.73 KWh, CVRMSE = 10.2 %).

- Model 2: Model 1 + outdoor temperature values with corrected temperature predictions (RMSE = 268.56 KWh, CVRMSE = 9.5 %).

Out of working times, consumption stays always the same. For that reason, we have computed metrics for the working hours and we can see that CVRMSE is better using Model 2 than Model 1, justifying the inclusion of weather forecasts on the modelling.

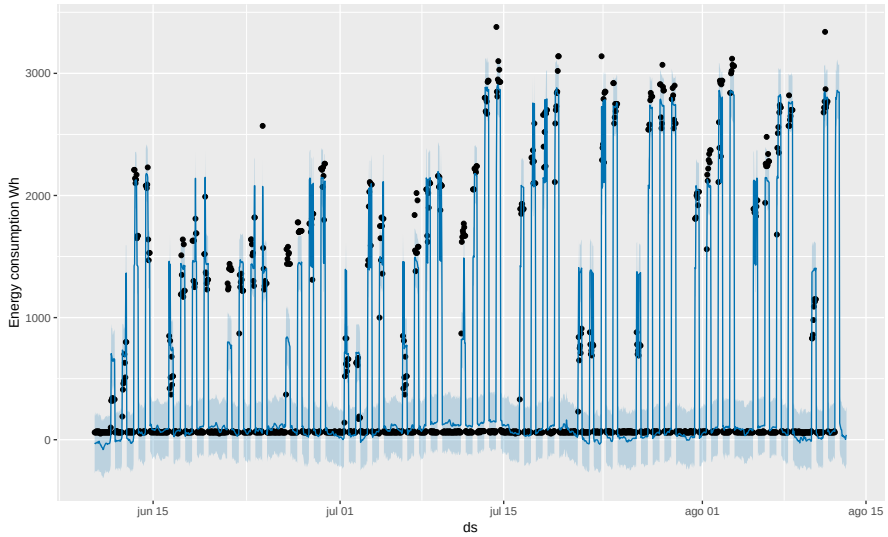

Figure 3.6: 24h predictions performed with the fitter model (blue line) and the true values (black dots) with Model 2)

3.2.3.4 Moment of the day predictions

We have displayed an outline based on basic and logic usability estimations of the building and we have included their Consumption Range (CR) and Consumption Mean (CM):

- Moment 1: holidays, weekends, nights (22h-6h). CR = [3.578, 14.1] KWh, CM = 7.904 KWh
- Moment 2: regular mornings (6h-14h). CR = [26.01, 86.19] KWh, CM = 54.27 KWh.
- Moment 3: regular afternoons (14h-22h). CR = [6.357, 53.290], CMs = 31.48 KWh.

The visual differences that are noticeable at Fig. 3.7 were confirmed with a Kruskall Wallis H[191] test. There is a significant difference between groups ($H(2) = 547.7$, p-value < 0.01). An analysis of the differences by pairs performing the post-hoc Wilcoxon test [191], determines that it is possible to divide data in those moments. This reasoning leads us to suggest three different models corresponding to the just mentioned partitions.

We considered 8 different observations for each environmental input (one per hour). Also, we created two new variables for every attribute by taking their mean and median. Just to clarify the considered inputs, for situation 1 and, for example, temperature, we will have 11 attributes: temperature at 6 AM, at 5 AM, ... at 22 PM, mean of temperature (from 6AM to 22PM) and median of temperature.

After training the models using several combinations of inputs we achieved the best results using the day of the week, month, season, mean temperature and mean humidity with RF algorithm for moment 1 (mtry = 4, RMSE = 1 KWh) and moment 3 (mtry = 2, RMSE = 3.87 KWh) and BRNN for moment 2 (number of neurons = 2, RMSE = 7.08 KWh) as can be seen in Table 3.2. All these values represent between a 12.09% and a 12.86% of error (CVRMSE).

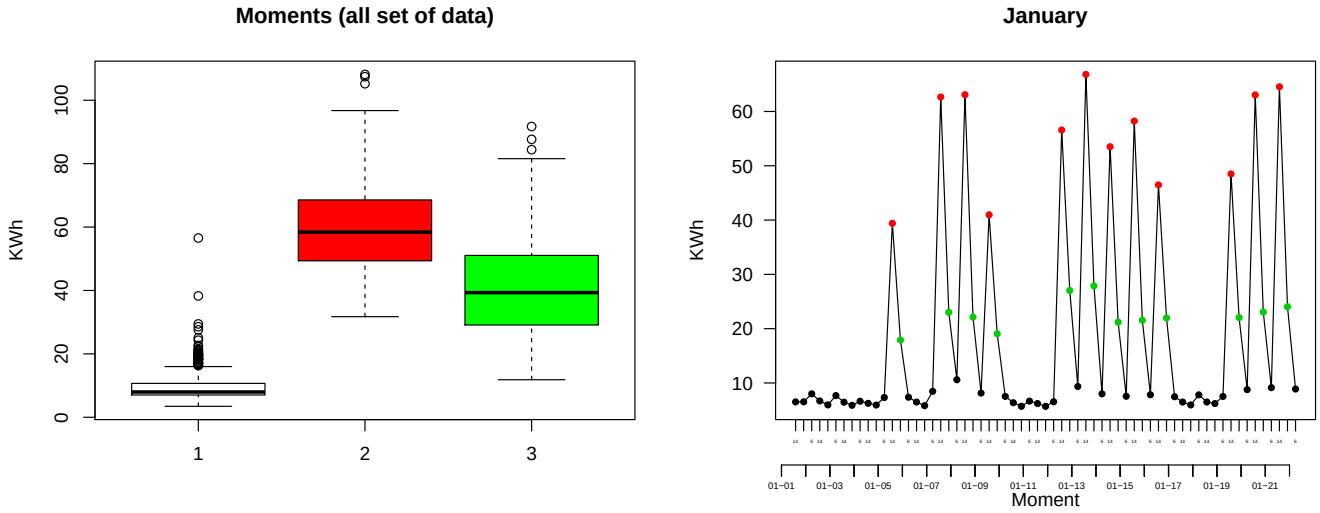

Figure 3.7: Boxplot of the energy consumption by moments considering all data (left); and, the time series of the energy consumption by moments during January (right)

Moment	Technique	Best Parameter	RMSE (KWh)	CVRMSE (%)	R^2
1	Gauss	σ = 0.1	1.1	13.43	0.57
	MLP	size = 34	1.1	13.46	0.55
	SVR	cost = 4	1.09	13.26	0.58
	BRNN	neurons = 3	1.1	13.47	0.55
	RF	mtry = 4	1	12.18	0.65
2	Gauss	σ = 0.1	7.76	14.1	0.67
	MLP	size = 37	1.56	15	0.68
	SVR	cost = 1	4.26	13.4	0.71
	BRNN	neurons = 2	7.08	12.86	0.75
	RF	mtry = 2	7.48	13.59	0.72
3	Gauss	σ = 0.1	4.5	14.07	0.67
	MLP	size = 37	4.81	15.03	0.69
	SVR	cost = 1	4.20	13.14	0.73
	BRNN	neurons = 5	4.31	13.45	0.73
	RF	mtry = 2	3.87	12.09	0.76

Table 3.2: Results obtained for each moment

Having trained and tested 5 different models, it is necessary to find statistical evidence that the selected one outperforms better not just in a punctual way. Our 10-fold cross-validation with 5 times repetition strategy generates a set of 50 measurements for each model. In Figure 3.8 (left) it is displayed the model's performance for situation 3, where red crosses show the median of RMSE for each model. For every situation, the Friedman test, that is the nonparametric alternative for repeated measures (within subjects) Analysis Of Variance (ANOVA)is significant (p-value < 0.05), and looking for corrected pairwise differences, we find that RF is the only one

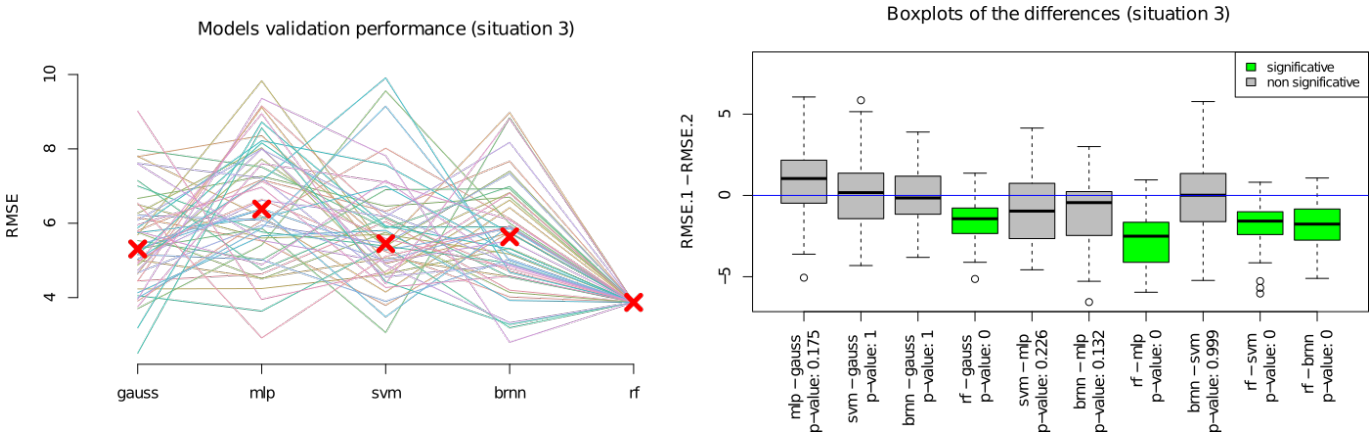

Figure 3.8: Models validation performance (left) and Pairwise differences between models performance (right)

that differs from the others. Between the other models' performances, there aren't significant differences as can be observed in Figure 3.8 (right), where boxplots of the differences are coloured in green when the differences are significatively different from zero.

3.2.3.5 Daily and weekly predictions

Our interest lies in the weekly quantification of energy use. However, daily dynamics are useful since there are patterns that can be found depending on the day of the week. Our model predicts daily energy consumption and then computes the metrics in an aggregated manner, so the global quantification takes place on a weekly basis.

The data that is used in order to build and train our baseline corresponds to 1 year's worth of data from a whole building, from February 2016 to February 2017 and we will compare a grey-box and black-box methodologies.

In order to make use of grey-box models, the set of outputs and inputs have to be defined together with the topology of the system. The most common mathematical representation of lumped parameter models is the state-space representation. The general form for time-invariant models can be written as shown in Eq. 3.2

$$\begin{cases} x'(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{3.2}$$

where x is a vector concerning the states of the model, in our case the temperatures, x' is the derivative (rate of change) of the states, A is a characteristic matrix of the model, B defines the effect of the inputs in the model, and u are the inputs, in our case the outside temperature and gains in electric. In this formulation, y represents the variables that are measured, in our case electricity, C is the identity matrix; and D is zero in all cases for this work. Using this

		Models					
		SVR	**RF**	XGB	TWT	Gauss	Grey
Daily	CVRMSE	12.4	**9**	11	14.9	17.45	33.57
	MAPE	7.2	**6**	7.3	12.3	15.01	43.02
Weekly	CVRMSE	6.4	**5**	6.2	11.1	16.3	19.53
	MAPE	5.2	**4.5**	5.5	9.4	12.3	15.48

Table 3.3: Metrics for energy consumption forecasting

formulation, every time a solution has to be evaluated, the built-in GNU Octave function **lsim** was used.

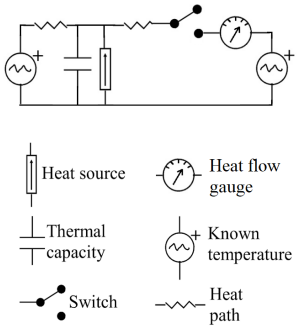

Figure 3.9: Dual-mode RC network

We have considered a dual-mode RC-network as the one shown in Fig. 3.9 and previously introduced by Ramallo-González on [192]. These grey-box models have been largely used in the past for building energy simulation. The reader is referred to [66], [193] and [192] for quantitative evidence of the accuracy of this kind of models.

Our black-box methodology is highly versatile with respect to the input data since it allows the addition of variables with minimal effort. We create the method in a constructive manner by relating the 24 temperature values of each day with the energy consumption of the building.

The subject building has several features that are typical of educational buildings: the load on weekends is substantially lower than that of weekdays and there might be also differences among weekdays. In these terms, we used ANOVA in order to determine whether there were differences between the consumption on the different days of the week (p-value = 0.001 < 0.05). After carrying out a **post-hoc** test we concluded that Fridays could be considered to behave differently to the other days of the week, which could be owing to lower occupation. Having attained this knowledge, we considered it necessary to add a dichotomous variable that indicates the kind of day of the week. Weekend and holiday consumption is estimated using the mean of previous weekends and holidays.

The algorithms that were found to be relevant for use within our black-box methodology are: SVR, RF and XGB.

The prediction metrics are summarized in Table 3.3. The first three methods: SVR, RF, XGB

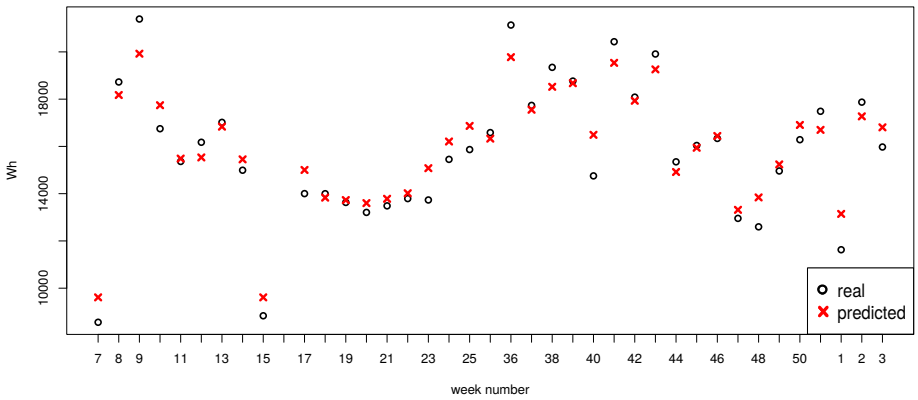

Figure 3.10: Weekly predictions using RF and real consumption

belong to the group of black-box models and are blind to the physics of the problem. We also have the TWT, the GAUSS and the Grey-box model, the last of which contains information regarding the physical phenomenon of the model topology. As can be seen, they return the best results when compared to the Gaussian method, which is applied in a more traditional manner, i.e., by relating the instantaneous consumption measurement with the instantaneous inputs measurements and also with our grey-box model approach.

Of the three black-box methods, RF is that which stands out since it attained a CVRMSE of 9 and 5 % and a MAPE of 6, and 4.5 % for the daily and weekly predictions respectively. We have plotted the weekly consumption in Fig. 3.10 .

3.2.4 Feature selection [R4]

In our framework for energy prediction we have devised a methodology that consists of: database transformation, FS, regression, decision making and forecasting. Next, each step is described separately.

After transforming the database, it is necessary to define the search strategy and the evaluator. We have configured a multitude of different methods for FS, both filter and wrapper, univariate and multivariate. For multivariate wrapper FS methods, we evaluate attribute sets by using a learning scheme with cross-validation and a performance measure. For univariate wrapper FS methods, we evaluate the worth of an attribute by using a user-specified classifier, cross-validation and a performance evaluation measure. The FS and classification processes have been executed in batch mode.

We considered for this research five wrapper and three filter FS methods, five multivariate and three univariate as it is shown in Table 3.4.

3.2.4.1 Strategies

As multivariate FS methods, we used the probabilistic strategy MultiObjectiveEvolutionary-Search [194] (two objectives: performance metric and attribute subset cardinality), and the

Database #Id.	Type of FS method	Search strategy	Evaluator	Acronym
#1	Wrapper Multivariate	MultiObjectiveEvolutionarySearch	RF (MAE)	MOES-RF-MAE
#2	Wrapper Multivariate	MultiObjectiveEvolutionarySearch	RF (RMSE)	MOES-RF-RMSE
#3	Wrapper Multivariate	MultiObjectiveEvolutionarySearch	kNN (RMSE)	MOES-kNN-RMSE
#4	Wrapper Multivariate	MultiObjectiveEvolutionarySearch	LR (MAE)	MOES-LR-MAE
#5	Wrapper Univariate	Ranker	RF (RMSE)	RANKER-RF-RMSE
#6	Filter Multivariate	GreedyStepwise	ConsistencySubsetEval	GS-CFSSE
#7	Filter Univariate	Ranker	ReliefFAttributeEval	RANKER-RFAE
#8	Filter Univariate	Ranker	PrincipalComponents	RANKER-PCA

Table 3.4: Proposed FS methods for energy time series forecasting

deterministic strategy GreedyStepwise [195]. The two most popular Multi Objective Evolutionary Algortihms (MOEA) are Elitist Pareto-based MOEA for diversity reinforcement (ENORA) [196], on which our team is intensively working over the last decade and NSGA-II (elitist non-dominated sorting genetic algorithm) [197]. In [198] is statistically shown that ENORA performs better than NSGA-II in terms of hypervolume [199, 200] for regression tasks, for which we have decided to use ENORA in this work. GreedyStepwise performs a greedy forward or backward search stopping when the addition or deletion of any of the remaining attributes results in a decrease in evaluation, thus, it has no backtracking capability.

For univariate FS, we used Ranker method [201] that ranks attributes by their individual evaluations is used.

3.2.4.2 Evaluators

For the wrapper methods, we used RF, k-Nearest Neighbors (kNN) and LR, with the metrics RMSE and MAE in order to test the features subsets. Those methods offer a good compromise between performance and computational time.

We considered the multivariate filter evaluator ConsistencySubsetEval [202], which scores a subset of features as a whole, by projecting the training instances according to the attribute subset. As univariate filter evaluators we used RelieffAttributeEval [203] and PrincipalComponents [204]. *RelieffAttributeEval* evaluates the value of an attribute by repeatedly sampling an instance and examining the value of the given attribute for the nearest instance of the same and different class. *PrincipalComponents* performs dimensionality reduction by choosing enough eigenvectors to account for some percentage of the variance in the original data (default 95%), and then transforms back to the original space.

3.2.4.3 Regression

After FS we obtained 8 features subsets. The used predictive algorithms, evaluated with cross-validation, were: RF, kNN, LR, SVRs [205] , GAUSS. Table 3.5, shows the evaluation in 10-fold cross-validation, 3 repetitions (a total of 30 models with each regression algorithm in each database), for the metrics *RMSE*, *MAE*. We have used a corrected paired t-test in order to test whether results are statistically significantly different to *#1* . A mark $*$ denotes that it is worse,

		#1	#2	#3	#4	#5	#6	#7	#8	TD
RMSE	RF	12.6685	12.9133	13.3814 ∗	14.4111 ∗	15.4203 ∗	18.7996 ∗	13.9174 ∗	36.7834 ∗	21.3455 *v*
	kNN	17.7612	20.8112 ∗	17.2423	25.0447 ∗	25.8680 ∗	25.7792 ∗	22.7562 ∗	37.8315 ∗	29.5936 ∗
	LR	19.3960	18.3234 *v*	18.5017 *v*	18.1808 *v*	18.6416	22.0898 ∗	18.1092 *v*	53.6083 ∗	17.7597 *v*
	SVR	20.0636	18.8337 *v*	18.9237 *v*	18.6714 *v*	18.9697 *v*	23.0016 ∗	18.8051 *v*	55.5770 ∗	18.2458 *v*
	GAUSS	21.9231	21.7133	24.7083 ∗	19.4114 *v*	18.8832 *v*	22.1160	18.3440 *v*	54.6361 ∗	17.8482 *v*
MAE	RF	5.8264	6.0012 ∗	6.2785 ∗	6.8621 ∗	7.5242 ∗	9.0191 ∗	6.4675 ∗	23.3071 ∗	12.6164 ∗
	kNN	8.8796	10.0150 ∗	8.6797	13.1307 ∗	13.3098 ∗	12.7038 ∗	11.0927 ∗	17.4372 ∗	14.3159 ∗
	LR	11.2708	10.1276 *v*	10.2363 *v*	9.7287 *v*	10.1738 *v*	13.2454 ∗	10.5091 *v*	38.3269 ∗	10.6126 *v*
	SVR	10.0410	9.0122 *v*	9.0806 *v*	8.9702 *v*	9.0669 *v*	11.5835 ∗	8.9962 *v*	36.7292 ∗	8.9314 *v*
	GAUSS	15.3332	15.1402 *v*	17.9369 ∗	12.0857 *v*	10.6294 *v*	13.3943 *v*	11.0307 *v*	38.8031 ∗	10.9028 *v*
CPU time	RF	0.9474	1.0349 ∗	0.7792 *v*	1.3432 ∗	0.9714	0.5708 *v*	0.7802 *v*	1.6078 ∗	3.0609 ∗
	kNN	0.0005	0.0005	0.0000	0.0000	0.0005	0.0000	0.0016	0.0000	0.0005
	LR	0.0042	0.0109	0.0026	0.0125 ∗	0.0089	0.0063	0.0115	0.0057	4.2172 ∗
	SVR	31.9255	29.0380 *v*	26.4307 *v*	62.5958 ∗	87.1901 ∗	75.5521 ∗	141.1615 ∗	9.0995 *v*	1626.4151 ∗
	GAUSS	115.4714	115.3620	115.4219	115.5542	110.7714 *v*	110.4302 *v*	110.5990 *v*	110.6505 *v*	114.0536

Table 3.5: RMSE, MAE and CPU time(in seconds) with 10-fold cross-validation (3 repetitions)

Input attribute	Rank	Importance
Lag_energy-1	1	7.398
Lag_stMO12_IMI_radmax+0	2	1.337
holiday	3	0.367
Lag_energy-3	4	0.357
ArtificialTimeIndex	5	0.302
Lag_stMO12_IMI_radmed-3	6	0.273
Lag_pr_feels-2	7	0.248
Lag_pr_temp-2	8	0.172

Table 3.6: Selected attributes with *MOES-RF-MAE* (database #1) and their ranks.

a mark v denotes a statistically better result, and no mark denotes no statistically meaningful difference.

3.2.4.4 Decision making

Looking at Table 3.5 we see that the best results have been obtained with the FS method *MOES-RF-MAE* (database *#1*) when *RandomForest* is used as regression algorithm, which shows statistically significant differences with respect to the rest and *MOES-RF-MAE* is also superior, with statistically significant differences except for the FS method *MOES-RF-RMSE*. With respect to the UserCPU_Time_training performance metrics, its results are acceptable in comparison to the rest of the methods. We can then choose the FS method MOES-RF-MAE and the database #1 for the final forecasting process.

Table 3.6 shows the selected attributes with MOES-RF-MAE and their rank and importance for each of the datasets. An attribute is evaluated by measuring the impact of leaving it out from the full set.

	1 step		2 steps		3 steps		Average	
	#1	**TD**	**#1**	**TD**	**#1**	**TD**	**#1**	**TD**
MAE	10.994	26.758	20.465	34.777	32.749	49.7	21.403	37.079
RMSE	16.051	36.556	28.768	45.079	44.834	59.821	29.884	47.152
# Instances	1526		1525		1524		-	

Table 3.7: Evaluation on test data with RF - database #1 and TransformedDatabase (TD)

3.2.4.5 Forecasting

Finally, we analyse the prediction ability of the forecaster obtained with the selected attributes. First, we train the model on the data, and then it is applied to make a forecast at each time point by stepping through the data. These predictions are collected and summarized, using MAE and RMSE metrics.

Table 3.7 with the databases *#1* and *TransformedDatase* (with all lagged variables and all overlay variables), on test data (30%). The reduced database *#1* improves the 1,2,3-steps-ahead predictions using the database without performing FS (TD). Using the averages of the steps-ahead predictions we see that with our methodology MAE is improved by 42.28% and RMSE by 36.62%, evaluated on test data.

3.2.4.6 Comparison with other methods proposed in the literature

For current and future comparisons with further research, the hourly *CVRMSE* was 20 % and we have also averaged it per day obtaining a daily *CVRMSE* = 11 % for the 1-step case.

Multivariate ARIMA: we have used the traditional time series method *ARIMA* with exogenous regressors [206]. Results are much worst than using out ML oriented approach. Using our selected features, mean *MAE* is 119 and mean *RMSE* is 126. This results are way worse than ours but still better than using all variables with *ARIMA*, for which *MAE* increases between 35 and 55 KWh and *RMSE* increases between 37 and 58 Kwh.

3.2.4.7 Analysis of results and discussion

As expected, wrapper show better performance than filter FS methods, and multivariate show better performance than univariate FS methods. Multivariate methods can identify interaction amongst features simultaneously, specially wrapper-based FS methods [207]. To reduce the computational time of multivariate wrapper FS methods (NP-hard), deterministic search strategies, such as *GreedyStepwise*, can be used but hidden and basic interactions could be missed due to the way the search space is traversed [208]. Probabilistic search techniques, such as *MultiObjectiveEvolutionarySearch*, can overcome these difficulties by allowing to generate new subsets in different locations of the search space guided by a metaheuristic. In the thesis, we propose to use a multivariate wrapper FS method where the search strategy is based on multi-objective

evolutionary computation, thus intrinsically overcoming the problem of interactions between features.

For wrapper FS methods, the RF evaluator has proven more effective than kNN and LR based evaluators. SVR and GAUSS are discarded as evaluators for wrapper methods because of their excessive computational time. Run time of RF is acceptable, and this method is not very sensitive to the variation of its parameters.

MAE has shown better behaviour than *RMSE* as metric performance in evaluators for wrapper FS methods since *MOES-RF-MAE* (database *#1*) produces better results than the method *MOES-RF-RMSE* (database *#2*) (see Table 3.5) when evaluated on cross-validation with *RandomForest* using the *RMSE* metric (12.6685 vs. 12.9133, an improvement of 1.9%). This improvement can also be observed in Table 3.5 when both databases are evaluated with the *MAE* metric (5.8264 vs. 6.0012, an improvement of 2.91% in this case).

3.2.5 HVAC patterns [R5]

Each terminal unit has a remote controller that facilitates the interaction of the user with the conditioning system. The user can turn on an off the room unit at their will, but cannot program the operation based on a timer. Also, the user can control the set point temperature. This means that the user can change at any time the thermostat control of the unit to any value between 16 and 29-degree C at their will. Each room has also a wall-mounted screen that shows the temperature of the room captured by the machine, the set point (thermostat) temperature and the fan operation mode. Every 12 minutes the following data was gathered: room temperature (ºC), on/off status, set point (ºC).

The intention is to create virtual areas comprising several building space areas, finding patterns in the HVAC use and consequently in the energy-related use and defining these virtual areas according to such information to optimise the content of information.

To do so, we aggregate each attribute per energy device daily. We can represent each device as a time series and with this, it is possible to fit a model or find a clustering algorithm that groups every attribute of the time series finding some distinctions between them

- Interaction frequency to turn it on/off
- Interaction frequency to change the set point
- Daily hours of operation (how many hours the machine is on)
- Average and standard deviation of the daily set point preferences

We have defined two ways to find patterns: based on the interaction of people with the machines in order to change the set point and based on the temperature preferences.

Fig. 3.11 (left) shows the histogram of interactions of the users with their controllers normalised by hours of use. Due to the skewness of the data, we then applied a logaritmic transformation.

We wanted to investigate the yearly fluctuations of data that could be found in thermostat values because on the ASHRAE Standard 55 for thermal environmental conditions and on [209] is stated that preferences differ throughout the year. In order to smooth the data yet imposing the yearly periodicity we fit a sinusoidal function with a fixed wavelength of 365 days. The equation was:

$$T_{set}(time) = a + b\sin(\frac{2\pi time}{356} + \lambda),$$

where a is the constant term, b is the yearly swing, and λ is the phase or lag.

We plotted the kernel density functions of the set point changes per hour, grouping the curves by people who used the machines for more than the hours given in the legend of Fig. 3.11 (right), showing a bimodal nature. As a result, two groups of users can be defined: one that interacts with the controller often (a change every week) and also consumes less and another that tends to not to interact with the set point (1 or 2 changes in the whole period) and are higher consumers.

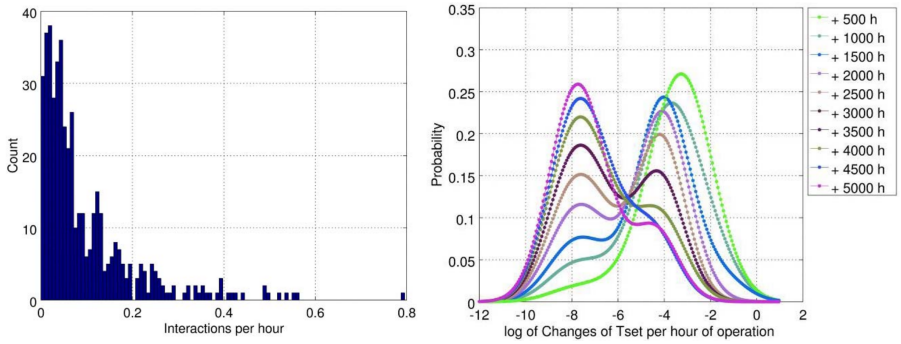

Figure 3.11: Changing frequency (left) and (right)

According to ASHRAE, the bounds of comfort for summer and winter are 25-27, 21-24 respectively. To evaluate the number of hours that the systems were pushed outside the comfort ranges, and how much, the integral of the area defined by the curve representing the set point temperature and the upper or lower bound were calculated. This provided with an indicator of overheating or over-cooling on Kelvin-hour, Kh a measurement well recognized on the Building Physics community.

It was seen that users tend to tolerate high temperatures much more than cold temperatures and use values of the thermostat for cooling that are close to the upper bound of the ASHRAE comfort range. The mean of overheating was 292.7 Kh, which is substantial. However, in the case of overcooling the figures were more prominent: a mean of 866.4 Kh. Those users that interact with the machines less often have registered the larger overcooling and no particular relationship between the number of interactions with the controls and overheating was found.

We also propose another way to classify the rooms based on the clustering of the raw set point time series. In this case, we used several algorithms: Hierarchical clustering, longitudinal K-means, DBSCAN and Spectral clustering.

Once every device is assigned to a cluster or virtual area, the mean of the elements of each cluster is computed in order to get an average measurement. Finally, each generated cluster is stored as an instance of a virtual energy area. Those virtual sensors were useful in order to:

- Send specific actions by checking the cluster that the user/room belongs to,
- Improve energy consumption prediction aggregating by cluster,
- Detecting outlier data which implies device failures [210],
- Reduce the monitored number of HVAC in order to obtain similar results in the analysis.

3.2.6 Human mobility patterns at macro and micro levels [R6]

Wearable devices are equipped with sensors like GPS that allow capturing a large amount of high-resolution digital traces which are instrumental for the mobility mining discipline, focusing on giving insight into the spatiotemporal trajectories of people. Looking at a macro level we try to detect regions with a high density of human transit. At the same time, many social network sites have included location-based capabilities into their smartphone's applications so most of the data belonging to those sites can be geotagged. This can be used to recognise human mobility models and patterns at both micro and macro levels.

3.2.6.1 Dense Transit Areas identification with GPS (macro)

We propose to characterise the flow of people inside regions using the online aggregation of the spatiotemporal traces. Our mechanism allows discovering Dense Transit Areas (DTA) that represent a spatial region of a city that is visited by a large number of citizens' routes in real time. Such monitoring aims to detect relevant changes of human mobility within regions that can be signs of events of interest, like unplanned demonstrations or serious traffic problems.The present system supports two different modes of execution, DTA discovery (generating areas) and DTA monitoring (controlling if areas remain as such).

DTA discovery

Routes are composed by the GPS sensor tuples (x, y, t) where (x, y) are latitude-longitude coordinates at instant t. The route is delivered to the central server and also stored in the local personal routes repository that keeps the last routes covered by the user within the mobile client. Distinguishing between low speed (walking) and high speed (vehicle) routes.

The spatial region under study is divided into squared cells and subcells. The route density is calculated in each of them considering: length of the cell, average speed and spatial length of the route and the coming and outgoing side of the route. Two areas are merged if the transit

information they represent is strongly related (similar speed and direction). Once a consistent set of DTAs is generated (when the ratio between new DTAs and the number of routes is below a threshold) DTAs monitoring starts.

DTA monitoring

This phase focuses on controlling the evolution of mobility features of the DTAs to early detect potential mobility shifts using a reliable subset of participants. For each person, the probability of visiting a DTA within a timestamp is calculated in the case that they have initiated a route nearby. For each period, the subset of users providing the best coverage of the detected DTAs is then used. For them, the ongoing route's sequence is stored. The similarity between the current and the historical state of each DTA is computed and if either the speed or direction similarity is below a predefined threshold during consecutive sampling periods then the algorithm infers that the human dynamics inside the DTA under review have changed so we go back to the DTA discovery phase.

Evaluation

In order to test our system, we have used the GeoLife dataset (GL) [211], a public collection of human trajectories produced by 178 users carrying different GPS feeds for over three years in Beijing city (China).

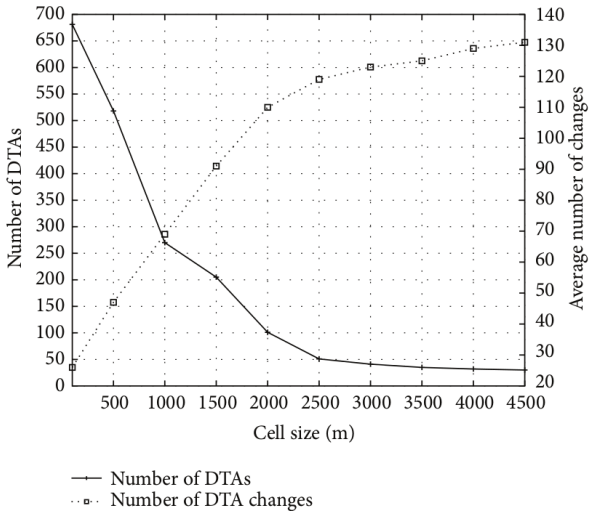

Figure 3.12: Number of DTAs and average number of changes per DTA with respect to the cell size

In Fig. 3.12 we can see that setting a small cell size generates a large number of DTAs and when increasing cell size DTAs become more sensitive to changes. We also analysed the accuracy of the approach by comparing the capability of the system to detect variations of speed and directional features with respect to the DSSIM function [212] -that measures the dissimilarity between two spatiotemporal trajectories during a time period- and the event-based mechanism proposed in [213] to detect abnormally high or low speeds of moving objects in real time. The

results achieved by our proposal vary according to the DSSIM function. Our system had a higher precision for large DSSIM values that indicate very evident differences between current and historical routes. When DSSIM ranged between 0.8 and 1.0, our proposal achieved acceptable precision results, about 0.8. However, for DSSIM values below 0.6 ours system's precision decays. This is because certain differences of the routes do not imply changing their incoming or outgoing sides in the DTA which are the features used by the proposal to detect transit changes. An strengh of our method is that it des not need the whole sequence of timestamped locations to operate whereas the others do.

3.2.6.2 Human routes identification with social media (macro and micro)

The previous approach does not take into account the activity level of the users within each detected area. Here we propose two ways for predicting the movement of the population within the city on an online fashion using the social network Twitter[18]. Those are a graph-based and a cluster-based approach.

Graph-based

A user's graph is generated and updated on the basis of his geotagged documents ("tweets") gathered on composing a hierarchy that represents the mobility flows of a large spatial region. The graphs include frequent locations ("landmarks") and frequent topics used for predicting the next meaningful location, or landmark, that will be visited by a person. The system architecture is compound by a client-side that runs on the mobile device of each user for detecting personal landmarks and a server-side responsible for composing collective ones.

For preprocessing, CEP is used. A CEP system is useful for the timely detection of situations of special significance that cannot be directly handled by humans. It consists of a palette of asynchronously interconnected Event Processing Rules (EPRs), defined by expert knowledge. Our defined CEP events are *tweet*: a raw tweet wit textual content and metadata (user nick, timestamp, and geotagging) as attributes; *tweet with topic*: includes the most probable topics that it refers to using a bag of words; *landmark*: indicates if a tweet has been written inside any personal or collective landmark; and *route*: represents a completed route as a sequence of personal or collective landmark. Retweets, URL links, mentions to other users, and stop words are deleted.

Route composition

The landmark comprising each new tweet with topic event including spatial region and its frequently associated activities is detected. Personal and collective landmarks are spatial regions with a high density of tweets related to one or more activities. As a result, a slightly modified version of the online landmark discovery algorithm (LDM) [214] has been applied to *tweets* locations for landmark detection.

Graph Generation

[18] `https://twitter.com/`

Personal and collective mobility graphs using the completed routes' sequences are generated on the fly. Both graphs encode the statistical information from the routes as a directed multigraph where each vertex represents a unique landmark. The personal mobility graph is updated when a route is completed. If all elements (landmark{activity}) are already vertices of the graph and there is a route identifier whose edges connect these elements in the same order, its identifier is extracted. Otherwise, a new identifier is generated. Then, the frequency attributes of each edge associated with this identifier is incremented or created. We perform a similar task also with the collective routes and create two graphs (working days and weekends).

Location prediction

Each time a new landmark event is appended, the route is delivered to the Local Predictor Maker (LPM) as Fig. 3.13 shows. LPM focuses on forecasting the next landmark (spatial region and associated topic) covered by the ongoing route. The algorithm detects the historical routes that best fit the ongoing route. This detection is done by searching the maximum set of edges that connect the visited landmarks in the same order. If target user statistics do not provide a good prediction then the algorithm makes use of the collective statistics.

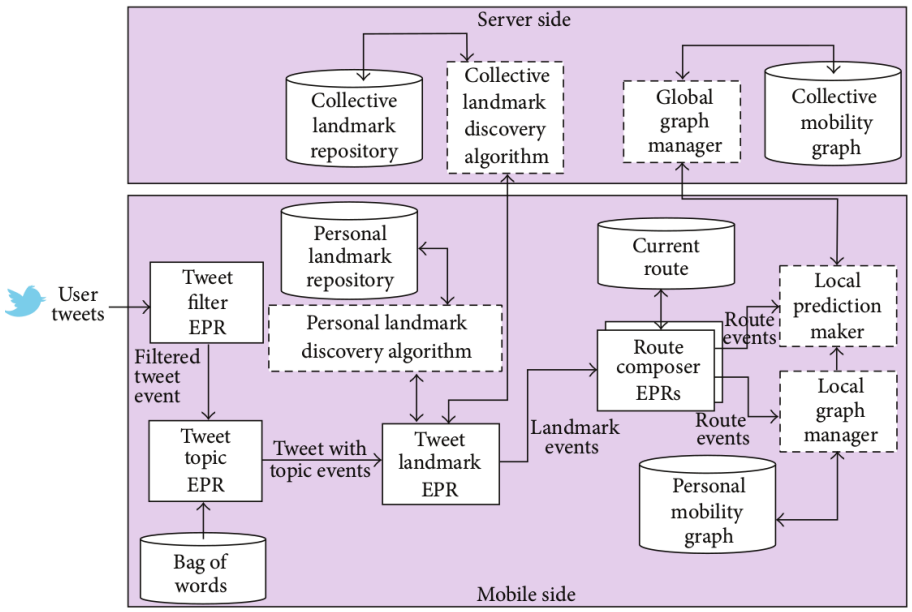

Figure 3.13: System architecture. The components that are not EPRs are depicted as dashed boxes

Evaluation metrics

The prediction rate (PR) and the prediction error (PE) metrics are used. PR counts the number of routes for which at least one landmark is provided as a prediction when a new element is appended (coverage). PE is the average of all distance deviations across each prediction (deviation from the actual next landmark $real_l$). Since each landmark may be associated with an activity, the distance between the predicted ($pred_a$) and the real activity ($real_a$) must be also

considered so we have used the Semantic-Hierarchical Similarity (SHS) [214].

$$PR = \frac{\#routeswithprediction}{\#routes}, \qquad PE = w_1 \times (1 - \frac{dist(real_l, pred_l}{dmax}) + w2 \times SHS(real_a, pred_a)$$

where $dmax$ is the maximum distance between two points in the dataset's spatial region and w_1, w_2 are adjustable coefficients.

Cluster-based

We cluster social-media documents with the fuzzy c-means (FCM) algorithm for different predefined time slots. Those are: early morning (0am-8am), moring (8am-12am), evening (12am-4pm), late (4pm-9pm), night (9pm-0am). For such clustering, the spatiotemporal features of the documents are considered. On the basis of these clusters, we define different profiles to predict the movement of the population within the city. Furthermore, the levels of social media activity have been measured for each cluster. As we will see, a correlation exists among the detected clusters, their associated activity and the prediction levels of its related movement.

For preprocessing, we aggregate tweets that have been posted in similar locations and time in order to avoid the disturbance of the real prediction with the same user's tweets which do not represent a real movement in space-time dimension.

Clustering

The FCM clustering algorithm [215] is applied to each of the five time-slots. The result is a membership matrix between all tweets and all clusters. For each time-slot, the cluster with the highest membership is selected to be the representative one for the user in that timeslot. As a result, five pairs of centroid-time slots are obtained which represent the usual movement of the user during the day across the time slots. Users are classified depending on the average posts per day, users are classified into three levels: low, medium, high. The activity level of each cluster is measured in order to enrich the information about the users and discover the kind of users in each cluster.

Movement prediction between clusters and time slots

For this task, the percentage of users in each time slot is calculated using the representative cluster for each time slot and user. Then, using the information between the pairs of clusters in consecutive time slots (e.g. from early morning to morning) we obtain the percentage of users that flow from one cluster to another. This information, combined with the activity level for each cluster provides a global and precise vision of the behaviour of the population studied in the total area. Using the information between the pairs of clusters in consecutive time slots (e.g. from early morning to morning) we obtain the percentage of users that flow from one cluster to another.

Experiments and results

In order to evaluate the proposals we used the Twitter Crawling API targeting Madrid city during 82 days using 181581 tweets from 41008 users.

Figure 3.14 (left) shows the collective landmarks generated by the system. We can see that the higher concentration of collective landmarks corresponds to spatial areas with a high density

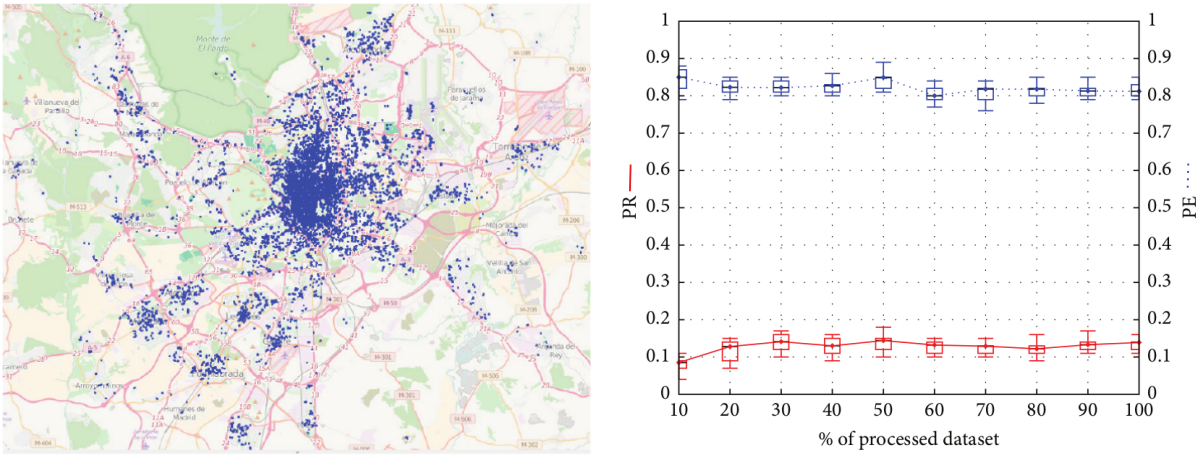

Figure 3.14: Collective landmarks (left) and metrics evolution

of human movement like the city centres. The evolution of the PR and PE of the system can be seen in Fig. 3.14 (right). There we appreciate that the PE remained more or less flat with no significant variance. This way, when only 10% of the dataset was processed our proposal was capable of achieving around 0.85 PE, thanks to the spatial distribution of the tweets. Users tend to post tweets located relatively close among them, limiting the locations that should be considered by the system and, thus, it is more likely to correctly predict the next location of a user.

Fig. 3.15 depicts the heat map of resulting digital traces of the datasets in the early morning and morning time slots and also the flow of people within time slots. Unsurprisingly, remarkable density of tweets exists in the centre of the city whereas the tweets in the suburbs are more spread.

3.2.7 IoT-based Big Data architecture for smart cities [R7]

The previously described methods need a common environment to interact with. In that sense, it is necessary to create an IoT-based platform to share the large volumes of heterogeneous information and to manage all interoperability aspects and enable the integration of the ML techniques above described.

The IoT platform is compliant with the FIWARE architecture, a key initiative of the Future Internet Public–Private Partnership (PPP) to create a well-aligned set of open enablers to receive, process, contextualize and publish IoT data from and for smart cities including from city-wide information to dwelling specific data[19]. In particular, the Orion Context Broker (OCB)[20] and the COMET[21] modules are used in order to store in a NoSQL repository the historical data, that are

[19] `https://www.fiware.org/`
[20] `https://fiware-orion.readthedocs.io/en/master/`
[21] `https://fiware-sth-comet.readthedocs.io/en/latest/`

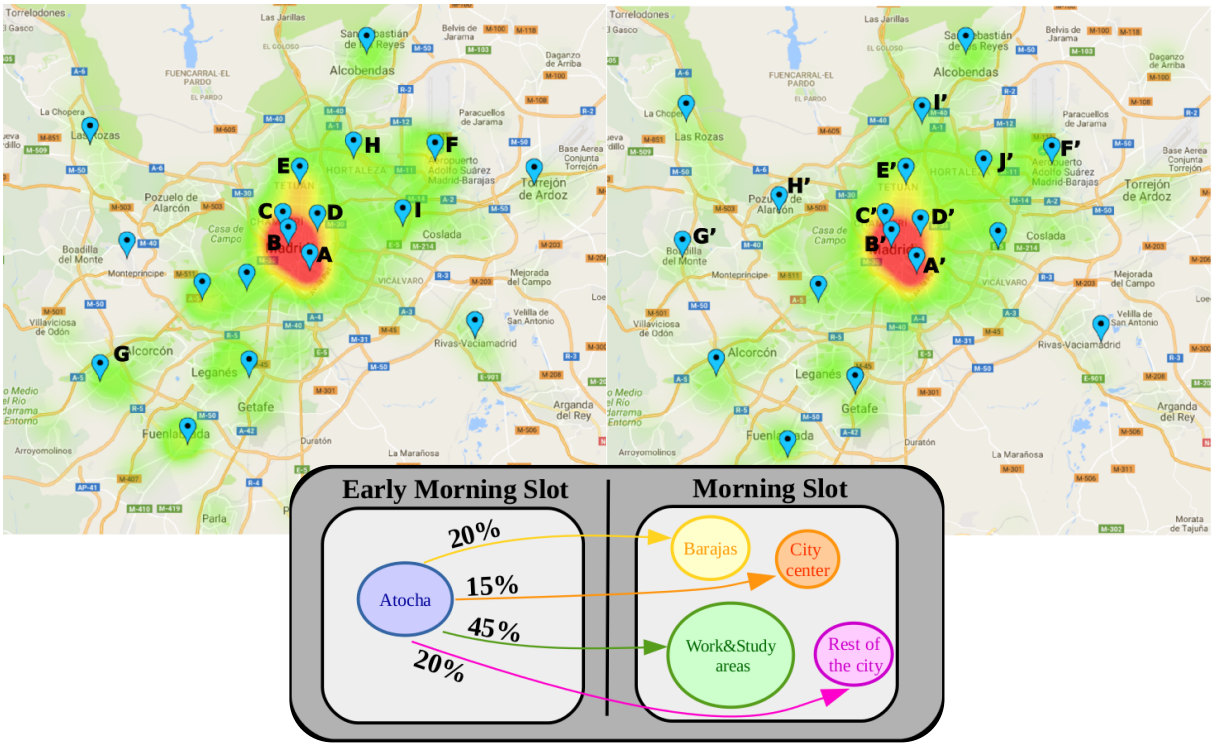

Figure 3.15: Heatmaps of clusters and movement prediction between early morning and morning slots

the measurements from the several data sources.

We started by defining a model compliant with the NGSI information model accepted in the FIWARE ecosystem that follows an entity–attribute approach. Each entity represents real or virtual elements of interest and has a type that allows defining type-based hierarchies. In this way, an entity has its own defined attributes and the inherited ones from its ancestors. Among its components there are three key entity groups related to the energy ecosystem of a building by means of NGSI entities:

- *Building* entity: the operational (opening hours, building use, etc.) and architectonic (fabrics, orientation, etc.) details of the building are the attributes of this entity.

- The *Spacial region* entity serves to link buildings with similar energy usage patterns because they are located in similar geographic regions.

- The *Building space area* entity represents the inner structure of a building (e.g., classrooms, corridors, etc.).

  Introducing data related to the previous entities facilitates the transfer of information between platforms and Building Information Modeling (BIM).

- Entities referring to the energy-related sensors: building sensor, power meter, and HVAC entities. Each entity includes the set of attributes monitored by the corresponding sensor along with other metadata (e.g., location, timestamp). The clean version of these entities refers to the filtered data.

- Entities that represent sensors outside the building that may provide useful information. As Fig. 3.16 shows, this is defined by means of the *external sensor* and *weather conditions* entities.

Finally, only the entities in gray in Fig. 3.16 have instances stored in ORION and COMET as we will see later.

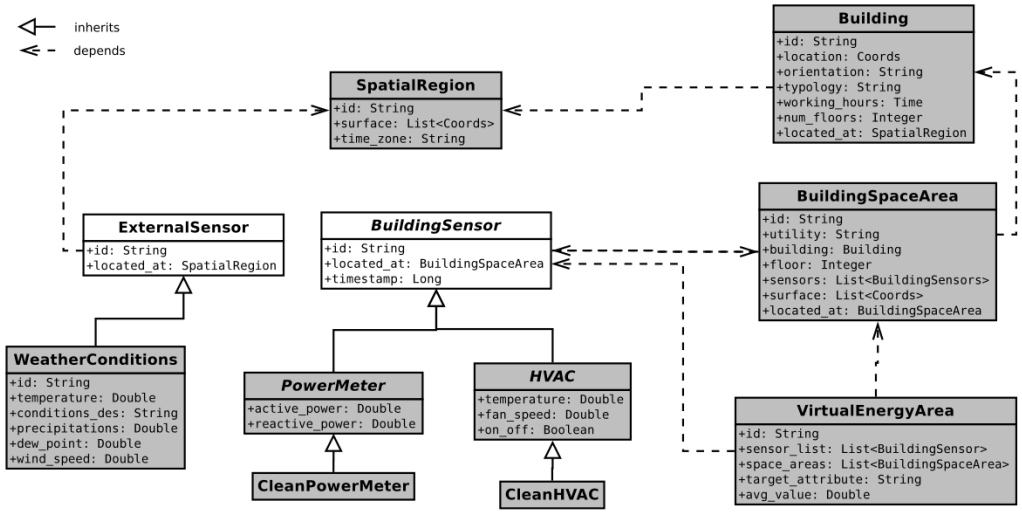

Figure 3.16: Information Model

The proposed platform that was created for creating applications and services for smart cities, and that especially covers the household energy domain and the provision of support for data analytics: sensorisation, homogenisation and storage, analytics and services is composed by several layers.

- Sensorisation layer

  This layer is in charge of connecting physical devices or actuators that provide data to the platform. Then it maps the collected data to the NGSI entities of the information model using the FIWARE IoT Agent enabler[22] and sends the mapped information to the next layer.

- Homogenisation and storage layer

[22] https://fiware-tutorials.readthedocs.io/en/latest/iot-agent/index.html

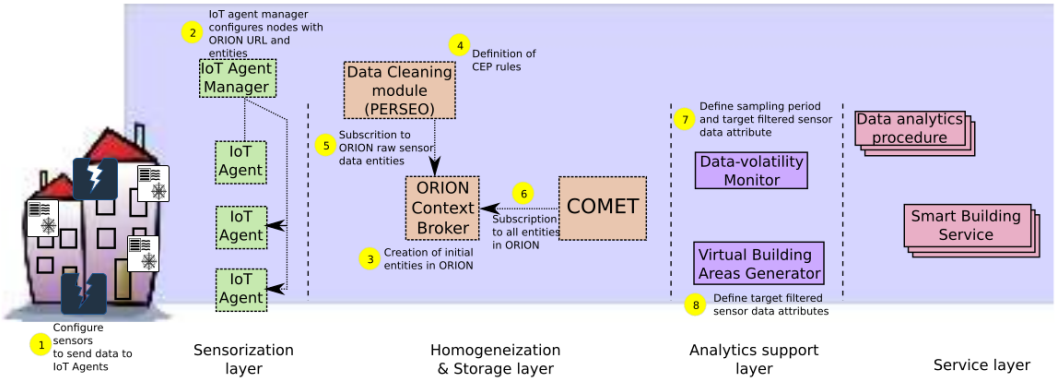

Figure 3.17: IoTEP workflow

This layer addresses the heterogeneity of the incoming data and contains real time data cleaning stage which ensures the quality of the data collected. Firstly, OCB implements a publish-subscribe store providing data access and the IoT Agents in the sensorization layer update the sensor entities' attributes in real time with the new readings from the devices. Secondly, the COMET enabler supports the access to historical time series data and incorporates an ad-hoc API to retrieve raw historical sensor data along with several built-in simple aggregation functions.

- Analytics support layer

  The third layer embraces all the functionalities of the platform to provide support for data mining services that can run on top of the platform. In particular, we have included predictive ML algorithms (including the algorithms described in 3.2.3), an energy data volatility detector and a virtual entities generator (including the algorithms described in 3.2.5). The data volatility monitor detects either abnormal energy consumption related to building spaces or an abnormal temperature setting related to HVACs. An alarm is triggered when the current averaged value of an attribute differs substantially from a recent historic rate of change.

- Service layer

  This layer serves as an interface between the IoTEP and the user, that could be anything from a building services manager to the back end of a smartphone application.

  Smart-building services can be nested at this level of the IoTEP platform, and will allow features such as advanced HVAC predictive control, home automation, fuel poverty evaluation, sick building syndrome diagnostics, risk situations for vulnerable people (as in heat waves), smart tariff strategies, manage emergencies, saving energy and many others including results visualisation. These actions can either involve managers or be automatically set.

3.2.8 IoT mechanisms to provide personalized energy management and awareness services by analysing behavioural aspects related to energy efficiency [R8]

The combination of IoT technologies, data modelling, management and fusion, Big Data analytics, and personalized recommendation mechanisms has resulted in an open and extensible architectural approach able to exploit in a homogeneous, efficient and scalable way the vast amount of energy, environmental, and behavioural data collected in energy efficiency campaigns and lead to the design of energy management and awareness services targeted to the occupants' lifestyles. It is named ENTROPY platform.

Two actors are needed to start an energy efficiency campaign : managers and end users. Potential campaign managers are: administrators, energy efficiency experts, data scientists and behavioural scientists. They are responsible for setting up sensor data monitoring, data analysis, and personalized recommendation delivery processes, that will depend on the kind of building and the kind of users engaged. They are able to define the set of buildings along with their division in subareas and their characteristics (surface, working hours, location, etc.). Next, sensors per area are assigned and queries are designed. This data is used as input for data mining and analysis processes (specifying which algorithm, the required input and the desired output variable). The campaign administrator should be also in charge of undertaking corrective actions in general.

End users may consist of citizens, students, academic personnel, employees, etc. As a starting point, they get a profile by means of a questionnaire regarding its personality, work engagement, energy conservation habits, and game interaction preferences. Through the ENTROPY mobile applications and serious games they get information regarding energy consumption and environmental parameters in the areas that they have activities at and receive personalized recommendations and requests for action. At the end of an energy efficiency campaign, they fill in an evaluation questionnaire, targeting at measuring the perception of behavioural change, as well as any changes with regards to their gaming profile.

A high-level view of the ENTROPY energy-aware IT ecosystem architectural approach is provided at Fig. 3.18.

The IoT management and data aggregation layer is responsible for IoT nodes registration, sensor activation, management and data aggregation and cleaning functionalities at the edge part of the infrastructure.

The data representation and fusion layer represents data based on a set of defined semantic models [216], supports a set of data fusion mechanisms over active data streams and then stores them in the Big Data repository based on MongoDB. Upon the activation of a new sensor data stream, the manager denotes the mapping between the monitored sensor metric with the relevant parameter in the semantic model, supporting the unified access to the collected data.

The Energy Semantic Model is similar to the one described in 3.2.7 where entities related to

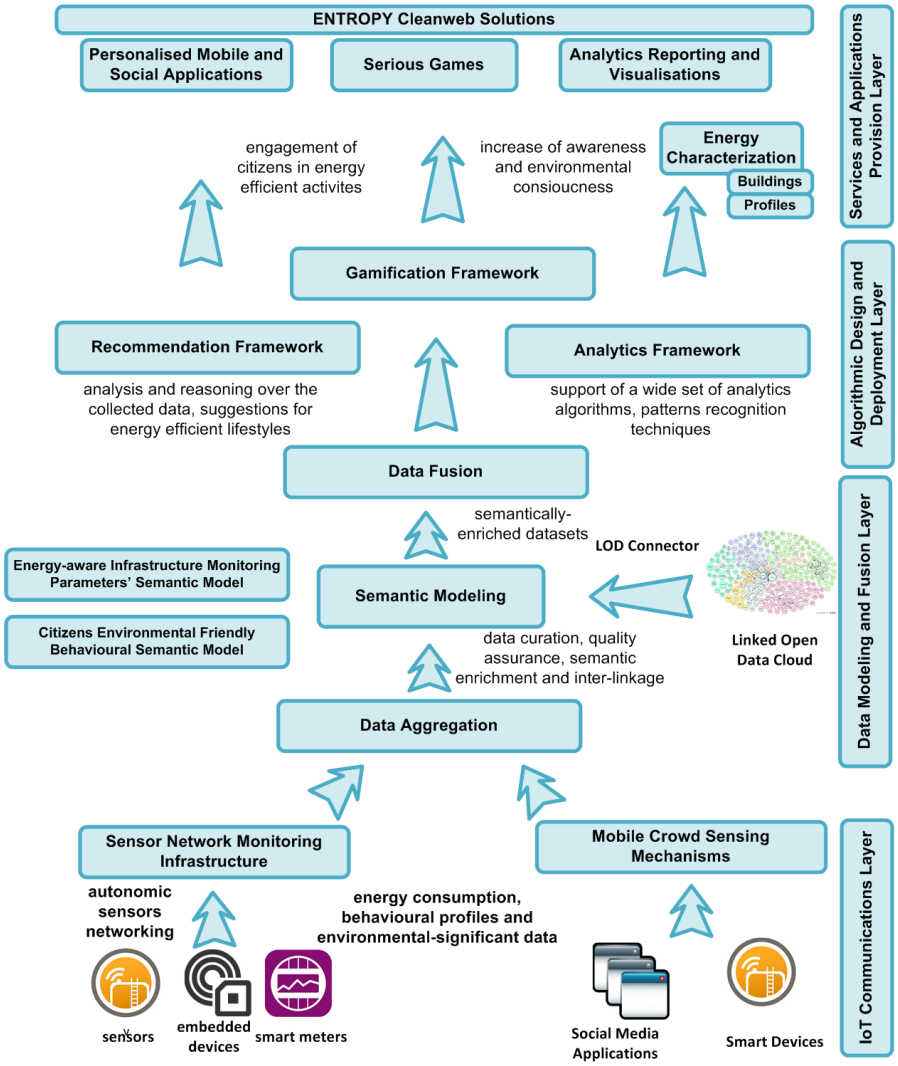

Figure 3.18: Entropy platform architecture

building areas, building spaces (windows, doors, etc.), equipment and outside sensors are defined and related to each other.

The Behavioural Semantic Model facilitates the categorization of users and the provision of personalized content and recommendations for achieving behavioural change. The main concepts regard the Agent and the Recommendation. An Agent can be a Person or a Group to whom personalized recommendations can be sent. It has several characteristics: personality traits such as extroversion and agreeableness; work engagement characterized by vigour, dedication, and absorption; gaming preferences such as socializer, free spirit, achiever and prize preferences such as rewards, badges, points, etc. Recommendations can have the form of a Message, a Quiz/Challenge, an Action whose result contributes to the elimination of a certain energy waste cause or a Question that leads to the collection of crowd-sensing feedback (e.g., comfort level).

The smart energy management services layer is responsible for providing advanced analytics

and recommendations to end users, as well as incorporating learning techniques for continuously exploiting the produced output by each service. These mechanisms work in a complementary fashion since the produced output from an analysis process can trigger the provision of a new recommendation.

A rule consists of a condition element which connects a context change with specific target user group criterion. When a rule is fired, the recommendation engine selects the set of target users based on their filters (location, responsive at the proposed actions through the personalized recommendations, etc) and creates a personalized recommendation for each of them by using the defined recommendation template. A produced recommendation contains the target user, the related content, the measurement attributes that are involved in the creation of the recommendation, the possible reward for the completion of the recommendation, as well as the validation method for it.

This layer also includes the support of a set of Big Data mining and analysis techniques towards the extraction of energy and behavioural analytics. An analysis process is based on the selection of an analysis template and the selection of the queries to be executed for providing the input datasets (training and/or evaluation datasets). Each analysis template represents a specific algorithm and provides the user with the flexibility to adjust the relevant configuration parameters. A set of initial algorithms are considered, however, the overall implementation facilitates the incremental addition of further analysis mechanisms.

The interconnection of the platform components with the analysis toolkits is based on the OpenCPU system for embedded scientific computing that provides a reliable and interoperable HTTP API for data analysis based on the R Project for Statistical Computing [217]. In the case of large-scale data processing and the need for a Big Data analysis framework, the Apache Spark engine is used, where the analysis process is realized in a set of worker nodes, each one of which is hosting an Apache Spark OpenCPU Executor[23].

The end user applications layer is responsible for the design of personalized mobile applications and web-based serious games able to take advantage of the set of services provided by the lower layers.

Some indicators for the evaluation of the different aspects of the platform are: energy savings and users behavioural change.

- Indicators of energy savings: savings at users' level, savings at areas' level, savings at buildings' level and savings extrapolation.
- Indicators of users' interactions: changes on behaviour based on data, self-awareness of change, changes in participants' perceived norm, changes on participants' personal values.
- Indicators of users' behavioural change: indicators of users' improvements on energy literacy, results from surveys and results from games.

[23] `https://github.com/onetapbeyond/opencpu-spark-executor`

There is a total of four buildings in the three pilots in which the platform is tested.

- A positive behaviour change was reported by the participants as they perceived it. Overall, the ENTROPY intervention led to an improvement on all the behavioural parameters we recorded, and especially on the participants' self-determination to conserve energy at work (+61.36%) and, also notably on the strength of their energy-saving competence (+14.35%), negative attitude towards saving energy at work (-13.77%), energy-saving at work as a habit (+13.60%), and the intention to save energy at work (+11.23%)
- Too much interaction does not lead to better results.

  The best results were obtained in UMU, where the users' interaction with the apps was mid-level. In POLO, on the other hand, where the average interaction with the apps was by far the highest (compared to UMU and HESSO), the strength of behaviour change was not the highest. It was higher than in HESSO, where minimum interaction took place, but lower than in UMU, where the interaction was a fair amount but not as much as in POLO. A fair amount of interaction with the apps was the optimum remedy to effect behaviour change. Too little led to lower behavioural change, whereas too much interaction from a point on was in lines with "bombarding" the users with content from the apps, and overall led to less behaviour change.

The campaigns used a series of automatic triggers that allowed to identify specific behaviours that were not optimal in terms of energy use. Based on these triggers, the platform was able to tackle behaviours that resulted in energy waste. Also and in parallel with this more focused actions, energy literacy was improved in the participants as means of background-permanent improvement of the behaviour in terms of energy use. As the sensors continued providing information during and after the campaigns, it was possible to identify how effective the campaigns were. In addition to that, data from the interaction of the users was captured. This was of great use, as one can consider that engagement and effectiveness of energy behaviour advice can only work together, and engagement is driven by a variety of factors, some of which scape from the scope of this project.

Through several campaigns we obtained a 19.7% of savings in heating, 12.3% in cooling and 16.16 % in electricity in average for the 3 pilots.

3.3 Lessons Learned

The main driver of this work is real data. Data coming from sensors and Internet of Things devices capture the real dynamics of the environment and it has been shown that its analysis can provide the improvement of services and the creation of new ones.

In this thesis, we have developed a set of analytical tools that intend to provide cities with intelligence by means of data analysis, including Machine Learning techniques.

Realistic scenarios in which energy consumption prediction is useful are provided. Also, the development of methodologies for the proper use of algorithms depending on the forecasting needs, improving accuracy with respect to previous works. The best algorithm in accuracy and time was RF in the majority of the scenarios. We have also studied ways to preprocess inputs, including a thorough Feature Selection methodology that reduces computational time. We found multiple objective optimisation using MAE to be the most suitable tool for Feature Selection in energy consumption forecasting scenarios.

One of the main consumers in buildings are Heating, Ventilating and Air Conditioning systems. We were able to characterise people behaviour towards their use, finding two groups depending on how often they change their set point. Those that interact little with the controller use the equipment for more hours in the year. In contrast, those users that interact with the machines less often registered the larger overcooling. We were also able to define virtual areas depending on the temperature preferences using DBSCAN clustering. These results may serve well when defining actions towards users regarding controllers and thermal preferences.

Another very important factor in energy efficiency is human mobility. Our efforts on human mobility focused on finding Dense Transit Areas with GPS at a macro level and identifying routes and predicting location using social networks at both micro and macro level, using collective behaviour in personal models. These efforts are mainly applicable for intelligent public transports and traffic forecasting, event detection, and urban planning in general but can also be considered as a first step towards the linking of human mobility, occupation and efficiency in buildings.

The deployed platforms allowe to fuse data and support the integration of data mining procedures for the provision of final services for energy data mining. Data pre-processing, clustering and forecasting are integrated in the platform. This enables the development of more sophisticated energy-aware services. For example, through the ENTROPY mobile applications and serious games, users get information regarding energy consumption and environmental parameters in the areas that they have activities at and receive personalized recommendations and requests for action. Those are big steps towards a more efficient energy-literate society. Together with algorithms, we have studied human patterns and also we have provided educational tools that make possible the achievement of better services. This work favours not only the emergence of smart cities but also smarter citizens. Automation processes are of great interest for some scenarios, however, they can lead to misunderstanding of the surrounding environment and can also lead to lack of reaction under failures. At the end, combining behavioural analytic with technological advances and ML we aim to use our resources sustainably, specifically energy.

Finally, even though the majority of the thesis results are applications, the fundamental properties of Internet of Things data have also been investigated. A method for time series data representation named BEATS was developed. Data is segmented and represented in order to extract their key characteristics in lower-dimensionality. Time series are segmented and reduced at high rates when using overlapping windows. The independence between blocks that our

algorithm provides is one of its most important features, that it can be applied in a distributed, online manner. BEATS also presents other qualities such as adapting to drifts and low latency.

3.4 Conclusions and Future Work

The work on this thesis is focused on two problems: real data analysis and its use for improving energy efficiency. It was our goal to find and fill some of the gaps that prevent us from using the precious information hidden in data. The combination of fine data management with ML serves to extract knowledge that can be used for many problems in smart cities, especially improving and creating services regarding efficiency in buildings.

Using statistical analysis and FS techniques we have found that the most important variables for the problem of energy consumption prediction were temperature, radiation, occupation and previous values of consumption. In the collection of data, we see that radiation predictions are not always available so its use is restricted. Additionally, occupation is not always available and it is difficult to predict it accurately, so we proposed to analyze the patterns by differentiating between working hours and days and non-working days. The results show that in there is an important difference between mornings and afternoons and also between all days of the week and Fridays in the studied buildings. Regarding occupation, we have also studied mobility patterns using wearable devices and social media. The extracted patterns could further be applied in the estimation of buildings occupation and its relation with consumption.

We have analysed several scenarios for energy consumption prediction and we have found that RF is an outstanding method in many of them, very appropriate due to its easy parallelisation. In order to obtain a horizon of predictions, multivariate time series methods also work well.

Results also show well-defined usage patterns of HVAC, one of the main contributors for energy consumption in buildings. The findings support the fact that there are two kinds of HVAC users: those who interact a lot with the set point and those who do not. The former appear to be higher consumers. These findings have been used to design strategies for energy consumption reduction.

We observed that the data that have been collected from real scenarios have several characteristics that pose challenges: their volume and their temporal nature. In that sense, we have also developed also a segmentation and representation algorithm called BEATS that transforms the data so that it provides similar amounts of information in a more compact manner. We have proved its effectiveness in classification and clustering problems using real data.

Finally, we developed IoT architectures that integrate all the steps developed in this thesis from the collection to the analysis of the data and even the provision of personalised services. Those services were designed for the improvement of energy efficiency in smart buildings targeting behavioural change and have proven to be useful for reducing energy consumption and improving energy literacy.

The following are some future research options that could be developed using the results of this thesis as a starting point both on the field of energy and data analysis:

- Going further than smart buildings, towards smart grids: integrating the created IoT platforms and predictive models into the novel smart grid scenarios

  The connection of the monitoring and management IoT platforms to the plethora of energy-demanding devices, energy generating local systems and energy exchange platforms between all kinds of buildings, prosumers and companies needs to be accomplished. This targets ultimately the automated effective orchestration and guides the actuation and decision-making at all levels of the energy system: Transmission and Distribution System Operators (TSOs, DSOs), Energy Services Companies (ESCOs), prosumers and consumers.

- Cross-building knowledge transfer by using ML time series prediction techniques

  All methods aiming to predict energy consumption that we have studied require labelled data, such as historical data. Such labels and datasets are not always available and it takes a long time and effort to collect, clean and manage them. At this point, data are employed in these forecasting methods in a non-adaptable way since they are completely static and thus without considering future events or changes which can occur in the network and the infrastructure of the buildings.

  It is not always economically feasible or possible time-wise to develop an IoT infrastructure in all buildings. In that sense, the need of unsupervised methods for energy consumption prediction is evident, especially after showing what it is possible to do with this information.

  Reducing the total time required for the analytic procedure is the key to scaling the deployment of energy efficiency projects in general, and reducing overall costs [218]. It is for this reason that a transfer learning approach should be considered in future studies in order to reduce the quantity of data that needs to be collected to create a reasonable building model.

- Finding further scenarios and ways to apply BEATS

  The development of the time series representation method BEATS is an important achievement of this work that can be further explored in the following ways:

  - Adding 3-dimensional (3D) data to the possible inputs and studying the modification of BEATS by substituting Discrete Cosine Transform by its 3D version.
  - Considering multi-sensor data. So far, BEATS is applied on each sensor in order to represent data in lower dimensions.
  - Studying the possibility to apply BEATS for dimensionality reduction by using the obtained eigenvectors to project the data in a similar fashion than to Principal Component Analysis (PCA).

- Connecting mobility results with energy consumption

  The investigation of the effects of urban mobility on energy consumption of specific urban areas is another line that can be derived from this thesis. That is, developing a deeper understanding of whether a similar spatial dependency exists in human mobility as an indicator for urban consumption of energy.

- Extend the context in which to apply the developed methods

  The application of the forecasting, feature selection and time series representation methods that have been developed in this thesis can be of great interest in further contexts than smart buildings and energy. Also, the developed platforms can serve for the connection between analysis and services in other areas such as smart agriculture, water management, etc.

  For example, smart agriculture scenarios are emerging and for the discovery of useful trends and patterns, there is a need to work on large sets of data obtained across multiple farms. After collecting more data and measurement about the production: soil quality, irrigation levels, weather, presence of insects and pests, its fusion using our IoT platform and the analysis through techniques here developed can serve of great help for the realisation of a smarter agriculture and livestock farming.

CHAPTER 4

PUBLICATIONS COMPOSING THE PHD THESIS

4.1 BEATS: Blocks of Eigenvalues Algorithm for Time Series Segmentation

Abstract

The massive collection of data via emerging technologies like the Internet of Things (IoT) requires finding optimal ways to reduce the observations in the time series analysis domain. The IoT time series require aggregation methods that can preserve and represent the key characteristics of the data. In this paper, we propose a segmentation algorithm that adapts to unannounced mutations of the data (i.e., data drifts). The algorithm splits the data streams into blocks and groups them in square matrices, computes the Discrete Cosine Transform (DCT), and quantizes them. The key information is contained in the upper-left part of the resulting matrix. We extract this sub-matrix, compute the modulus of its eigenvalues, and remove duplicates. The algorithm, called BEATS, is designed to tackle dynamic IoT streams, whose distribution changes over time. We implement experiments with six datasets combining real, synthetic, real-world data, and data with drifts. Compared to other segmentation methods like Symbolic Aggregate approXimation (SAX), BEATS shows significant improvements. Trying it with classification and clustering algorithms it provides efficient results. BEATS is an effective mechanism to work with dynamic and multi-variate data, making it suitable for IoT data sources. The datasets, code of the algorithm and the analysis results can be accessed publicly at: `https://github.com/auroragonzalez/BEATS`

Title	BEATS: Blocks of Eigenvalues Algorithm for Time Series Segmentation
Authors	Aurora González-Vidal, Payam Barnaghi, and Antonio F. Skarmeta
Type	Journal
Journal	IEEE Transactions on Knowledge and Data Engineering
Impact factor (2018)	3.857
Rank	Q1
Publisher	IEEE
Volume	30
Issue	11
Pages	2051-2064
Year	2018
Month	March
ISNN	1041-4347 (Print), 1558-2191 (Electronic)
DOI	10.1109/TKDE.2018.2817229
URL	https://ieeexplore.ieee.org/document/8319952/
State	Published
Author's contribution	The PhD student, Aurora González Vidal, is the main author of the paper

4.2 A methodology for Energy Multivariate Time Series Forecasting in Smart Buildings based on Feature Selection

Abstract

The massive collection of data via emerging technologies like the Internet of Things (IoT) requires finding optimal ways to reduce the created features that have a potential impact on the information that can be extracted through the machine learning process. The mining of knowledge related to a concept is done on the basis of the features of data. The process of finding the best combination of features is called feature selection. In this paper we deal with multivariate time-dependent series of data points for energy forecasting in smart buildings. We propose a methodology to transform the time-dependent database into a structure that standard machine learning algorithms can process, and then, apply different types of feature selection methods for regression tasks. We used Weka for the tasks of database transformation, feature selection, regression, statistical test and forecasting. The proposed methodology improves MAE by 59.97% and RMSE by 40.75%, evaluated on training data, and it improves MAE by 42.28by 36.62%, evaluated on test data, on average for 1-step-ahead, 2-step-ahead and 3-step-ahead when compared to not applying any feature selection methodology.

Title	A methodology for Energy Multivariate Time Series Forecasting in Smart Buildings based on Feature Selection
Authors	Aurora González-Vidal, Fernando Jiménez and Antonio Skarmeta-Gómez
Type	Journal
Journal	Energy and Buildings
Impact factor (2018)	4.495
Rank	Q1
Publisher	Elsevier
Volume	196
Pages	71-82
Year	2019
Month	August
ISNN	0378-7788
DOI	doi: 10.1016/j.enbuild.2019.05.021
URL	https://www.sciencedirect.com/science/article/pii/S0378778818338775
State	In Press, Accepted Manuscript
Author's contribution	The PhD student, Aurora González Vidal, is the main author of the paper

4.3 Commissioning of the Controlled and Automatized Testing Facility for Human Behavior and Control (CASITA)

Abstract

Human behavior is one of the most challenging aspects in the understanding of building physics. The need to evaluate it requires controlled environments and facilities in which researchers can test their methods. In this paper, we present the commissioning of the Controlled and Automatized Testing Facility for Human Behavior (CASITA). This is a controlled space emulation of an office or flat, with more than 20 environmental sensors, 5 electrical meters, and 10 actuators. Our contribution shown in this paper is the development of an infrastructure-Artificial Intelligence (AI) model pair that is perfectly integrated for the study of a variety of human energy use aspects. This facility will help to perform studies about human behavior in a controlled space. To verify this, we have tested this emulation for 60 days, in which equipment was turned on and off, the settings of the conditioning system were modified remotely, and lighting operation was similar to that in real behaviors. This period of commissioning generated 74.4 GB of raw data including high-frequency measurements. This work has shown that CASITA performs beyond expectations and that sensors and actuators could enable research on a variety of disciplines related to building physics and human behavior. Also, we have tested the PROPHET software, which was previously used in other disciplines and found that it could be an excellent complement to CASITA for experiments that require the prediction of several pertinent variables in a given study. Our contribution has also been to proof that this package is an ideal “soft” addition to the infrastructure. A case study forecasting energy consumption has been performed, concluding that the facility and the software PROPHET have a great potential for research and an outstanding accuracy.

Title	Commissioning of the Controlled and Automatized Testing Facility for Human Behavior and Control (CASITA)
Authors	Ignacio Rodríguez-Rodríguez, Aurora González-Vidal, Alfonso P. Ramallo-González and Miguel Ángel Zamora
Type	Journal
Journal	Sensors
Impact factor (2018)	3.031
Rank	Q1
Publisher	MDPI
Volume	18
Issue	9
Year	2018
Month	September
ISSN	1424-8220
DOI	doi:10.3390/s18092829
URL	https://www.ncbi.nlm.nih.gov/pubmed/30150573
State	Published
Author's contribution	The PhD student, Aurora González Vidal, contributed to the methodology, software, analysis and research and writing the publication

4.4 Applicability of Big Data Techniques to Smart Cities Deployments

Abstract

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

Title	Applicability of Big Data Techniques to Smart Cities Deployments
Authors	M. Victoria Moreno, Fernando Terroso-Sáenz, Aurora González-Vidal, Mercedes Valdés-Vela Antonio F. Skarmeta, Miguel A. Zamora and Victor Chang
Type	Journal
Journal	IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS
Impact factor (2018)	7.377
Rank	Q1
Publisher	IEEE
Volume	13
Issue	2
Pages	800 - 809
Year	2017
Month	April
ISNN	1551-3203 (Print), 1941-0050 (Electronic)
DOI	10.1109/TII.2016.2605581
URL	https://ieeexplore.ieee.org/abstract/document/7558230/
State	Published
Author's contribution	The PhD student, Aurora González Vidal, contributed to the methodology, software, analysis and research and writing the publication

4.5 An open IoT platform for the management and analysis of energy data

Abstract

Buildings are key players when looking at end-use energy demand. It is for this reason that during the last few years, the Internet of Things (IoT) has been considered as a tool that could bring great opportunities for energy reduction via the accurate monitoring and control of a large variety of energy-related agents in buildings. However, there is a lack of IoT platforms specifically oriented towards the proper processing, management and analysis of such large and diverse data. In this context, we put forward in this paper the IoT Energy Platform (IoTEP) which attempts to provide the first holistic solution for the management of IoT energy data. The platform we show here (that has been based on FIWARE) is suitable to include several functionalities and features that are key when dealing with energy quality insurance and support for data analytics. As part of this work, we have tested the platform IoTEP with a real use case that includes data and information from three buildings totalizing hundreds of sensors. The platform has exceed expectations proving robust, plastic and versatile for the application at hand.

Title	An open IoT platform for the management and analysis of energy data
Authors	Fernando Terroso-Saenz, Aurora González-Vidal, Alfonso P. Ramallo-González and Antonio F. Skarmeta
Type	Journal
Journal	Future Generation Computer Systems
Impact factor (2018)	5.768
Rank	Q1
Publisher	ELSEVIER
Volume	92
Pages	1066 - 1079
Year	2019
Month	March
ISNN	0167-739X
DOI	10.1016/j.future.2017.08.046
URL	https://www.sciencedirect.com/science/article/pii/S0167739X17304181?via%3Dihub
State	Published
Author's contribution	The PhD student, Aurora González Vidal, contributed to the methodology, research and writing the publication

4.6 Providing Personalized Energy Management and Awareness Services for Energy Efficiency in Smart Buildings

Abstract

Considering that the largest part of end-use energy consumption worldwide is associated with the buildings sector, there is an inherent need for the conceptualization, specification, implementation, and instantiation of novel solutions in smart buildings, able to achieve significant reductions in energy consumption through the adoption of energy efficient techniques and the active engagement of the occupants. Towards the design of such solutions, the identification of the main energy consuming factors, trends, and patterns, along with the appropriate modeling and understanding of the occupants' behavior and the potential for the adoption of environmentally-friendly lifestyle changes have to be realized. In the current article, an innovative energy-aware information technology (IT) ecosystem is presented, aiming to support the design and development of novel personalized energy management and awareness services that can lead to occupants' behavioral change towards actions that can have a positive impact on energy efficiency. Novel information and communication technologies (ICT) are exploited towards this direction, related mainly to the evolution of the Internet of Things (IoT), data modeling, management and fusion, big data analytics, and personalized recommendation mechanisms. The combination of such technologies has resulted in an open and extensible architectural approach able to exploit in a homogeneous, efficient and scalable way the vast amount of energy, environmental, and behavioral data collected in energy efficiency campaigns and lead to the design of energy management and awareness services targeted to the occupants' lifestyles. The overall layered architectural approach is detailed, including design and instantiation aspects based on the selection of set of available technologies and tools. Initial results from the usage of the proposed energy aware IT ecosystem in a pilot site at the University of Murcia are presented along with a set of identified open issues for future research.

Title	Providing Personalized Energy Management and Awareness Services for Energy Efficiency in Smart Buildings
Authors	Eleni Fotopoulou, Anastasios Zafeiropoulos, Fernando Terroso-Sáenz, Umutcan Simsek Aurora González-Vidal, George Tsiolis, Panagiotis Gouvas Paris Liapis, Anna Fensel and Antonio Skarmeta
Type	Journal
Journal	Sensors
Impact factor (2018)	3.031
Rank	Q1
Publisher	MDPI
Volume	17
Issue	9
Year	2017
Month	September
ISNN	1424-8220
DOI	10.3390/s17092054
URL	https://www.ncbi.nlm.nih.gov/pubmed/28880227
State	Published
Author's contribution	The PhD student, Aurora González Vidal, contributed to the development of the set of data mining mechanisms, to the problem contextualisation and the expriments

Publications

[1] Aurora Gonzalez-Vidal, Payam Barnaghi, and Antonio F Skarmeta. Beats: Blocks of eigenvalues algorithm for time series segmentation. *IEEE Transactions on Knowledge and Data Engineering*, 30(11):2051–2064, 2018.

[2] Aurora Gonzalez-Vidal, Fernando Jimenez, and Antonio F Skarmeta. A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy and Buildings*, 2019.

[3] Ignacio Rodríguez-Rodríguez, Aurora González Vidal, Alfonso Ramallo González, and Miguel Zamora. Commissioning of the controlled and automatized testing facility for human behavior and control (casita). *Sensors*, 18(9):2829, 2018.

[4] M Victoria Moreno, Fernando Terroso-Sáenz, Aurora González-Vidal, Mercedes Valdés-Vela, Antonio F Skarmeta, Miguel A Zamora, and Victor Chang. Applicability of big data techniques to smart cities deployments. *IEEE Transactions on Industrial Informatics*, 13(2):800–809, 2017.

[5] Fernando Terroso-Saenz, Aurora González-Vidal, Alfonso P Ramallo-González, and Antonio F Skarmeta. An open iot platform for the management and analysis of energy data. *Future Generation Computer Systems*, 2017.

[6] Eleni Fotopoulou, Anastasios Zafeiropoulos, Fernando Terroso-Sáenz, Umutcan Şimşek, Aurora González-Vidal, George Tsiolis, Panagiotis Gouvas, Paris Liapis, Anna Fensel, and Antonio Skarmeta. Providing personalized energy management and awareness services for energy efficiency in smart buildings. *Sensors*, 17(9):2054, 2017.

[7] Aurora González-Vidal, Victoria Moreno-Cano, Fernando Terroso-Sáenz, and Antonio F Skarmeta. Towards energy efficiency smart buildings models based on intelligent data analytics. *Procedia Computer Science*, 83:994–999, 2016.

[8] Aurora González-Vidal, Alfonso P Ramallo-González, and Antonio Skarmeta. Empirical study of massive set-point behavioral data: Towards a cloud-based artificial intelligence that democratizes thermostats. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 211–218. IEEE, 2018.

[9] Aurora González-Vidal, Alfonso P Ramallo-González, Fernando Terroso-Sáenz, and Antonio Skarmeta. Data driven modeling for energy consumption prediction in smart buildings. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4562–4569. IEEE, 2017.

[10] Tomás Vantuch, Aurora González Vidal, Alfonso P Ramallo-González, Antonio F Skarmeta, and Stanislav Misák. Machine learning based electric load forecasting for short and long-term period. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, pages 511–516. IEEE, 2018.

[11] Aurora González-Vidal, Victoria Moreno, and Antonio F. Skarmeta. Model-free approach based on iot data analytics for energy efficiency in smart environments. In *2016 European Conference on Networks and Communications EuCNC*, 2016.

[12] M Victoria Moreno, Fernando Terroso-Sáenz, Aurora González-Vidal, and Antonio F Skarmeta. Data analytics in smart buildings. *Building Blocks for IoT Analytics*, page 167, 2016.

[13] Fernando Terroso-Sáenz, Jesús Cuenca-Jara, Aurora González-Vidal, and Antonio F Skarmeta. Human mobility prediction based on social media with complex event processing. *International Journal of Distributed Sensor Networks*, 12(9):5836392, 2016.

[14] Fernando Terroso-Sáenz, Aurora González-Vidal, and Antonio F. Skarmeta. Towards anticipate detection of complex event processing rules with probabilistic modelling. *International Journal of Design & Nature and Ecodynamics*, 11(3):275–283, 2016.

[15] Fernando Terroso-Sáenz, Mercedes Valdes-Vela, Aurora González-Vidal, and Antonio F Skarmeta. Human mobility modelling based on dense transit areas detection with opportunistic sensing. *Mobile Information Systems*, 2016, 2016.

[16] Jesus Cuenca-Jara, Fernando Terroso-Saenz, Mercedes Valdes-Vela, Aurora Gonzalez-Vidal, and Antonio F Skarmeta. Human mobility analysis based on social media and fuzzy clustering. In *2017 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 2017.

[17] Fernando Terroso-Sáenz, Victoria Moreno, Aurora González-Vidal, Miguel Ángel Zamora-Izquierdo, and Antonio F. Skarmeta. Internet de las cosas y gamificación aplicados a eficiencia energética en edificios. In *III Congreso EECN Edificios Energía Casi Nula 2016*, pages 25–30, 2016.

[18] Eleni Fotopoulou, Anastasios Zafeiropoulos, Fernando Terroso, Aurora Gonzalez, Antonio Skarmeta, Umutcan Şimşek, and Anna Fensel. Data aggregation, fusion and recommendations for strengthening citizens energy-aware behavioural profiles. In *2017 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 2017.

References

[19] Accenture Strategy. # smarter2030: Ict solutions for 21st century challenges. *The Global eSustainability Initiative (GeSI), Brussels, Brussels-Capital Region, Belgium, Tech. Rep*, 2015.

[20] Evangelos Theodoridis, Georgios Mylonas, and Ioannis Chatzigiannakis. Developing an iot smart city framework. In *IISA 2013*, pages 1–6. IEEE, 2013.

[21] Dave Evans. The internet of things: How the next evolution of the internet is changing everything. *CISCO white paper*, 1(2011):1–11, 2011.

[22] Rajendra K Pachauri, Myles R Allen, Vicente R Barros, John Broome, Wolfgang Cramer, Renate Christ, John A Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, et al. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC, 2014.

[23] Jane Marceau. Innovation in the city and innovative cities introduction, 2008.

[24] María Victoria Moreno Cano, José Santa, Miguel Angel Zamora, and Antonio F Skarmeta Gómez. Context-aware energy efficiency in smart buildings. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*, pages 1–8. Springer, 2013.

[25] EC. Benchmarking smart metering deployment in the eu-27 with a focus on electricity. *Report from the Commission*, 2014.

[26] N. Mogles, I. Walker, A. P. Ramallo-González, S. Lee, J.and Natarajan, J. Padget, E. Gabe-Thomas, T. Lovett, S. Ren, G.and Hyniewska, E. O'Neill, R. Hourizi, and D. Coley. How smart do smart meters need to be? *Building and Environment*, (125):439–450, 2017.

[27] Raghunath Nambiar, Rajesh Shroff, and Shane Handy. Smart cities: Challenges and opportunities. In *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, pages 243–250. IEEE, 2018.

[28] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.

[29] Monthly energy review. *U.S. Energy Information Administration*, September, 2017.

[30] CABA. Intelligent buildings and theimpact of the internet of things. *Landmark Research Project. Executive Summary*, pages 1–18, 2017.

[31] Elham Delzendeh, Song Wu, Angela Lee, and Ying Zhou. The impact of occupants' behaviours on building energy analysis: A research review. *Renewable and Sustainable Energy Reviews*, 80:1061–1071, 2017.

[32] Rishee K Jain, Rimas Gulbinas, John E Taylor, and Patricia J Culligan. Can social influence drive energy savings? detecting the impact of social influence on the energy consumption behavior of networked users exposed to normative eco-feedback. *Energy and Buildings*, 66:119–127, 2013.

[33] Anca-Diana Barbu, Nigel Griffiths, and Gareth Morton. Achieving energy efficiency through behaviour change: what does it take. *European Environment Agency (EEA), Copenhagen*, 2013.

[34] Maarten De Groote, Jonathan Volt, and Frances Bean. Is europe ready for the smart buildings revolution? *Building Performance Institute Europe. Retrieved July*, 7:2017, 2017.

[35] Jennifer King and Christopher Perry. *Smart buildings: Using smart technology to save energy in existing buildings*. Amercian Council for an Energy-Efficient Economy, 2017.

[36] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[37] Muhammad Fahim Uddin, Navarun Gupta, et al. Seven v's of big data understanding big data to extract value. In *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, pages 1–5. IEEE, 2014.

[38] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007:1–16, 2012.

[39] David L Hall and James Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[40] Meisong Wang, Charith Perera, Prem Prakash Jayaraman, Miranda Zhang, Peter Strazdins, RK Shyamsundar, and Rajiv Ranjan. City data fusion: Sensor data fusion in the internet of things. *International Journal of Distributed Systems and Technologies (IJDST)*, 7(1):15–36, 2016.

[41] Carnot Institutes. White paper: Smart networked objects and internet of things. *Information Communication Technologies and Micro Nano Technologies alliance, White Paper,(Jan. 2011)*, 2011.

[42] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453. Society for Industrial and Applied Mathematics, 2013.

[43] Marvin Weinstein, F Meirer, A Hume, Ph Sciau, G Shaked, R Hofstetter, Erez Persi, A Mehta, and David Horn. Analyzing big data with dynamic quantum clustering. *arXiv preprint arXiv:1310.2700*, 2013.

[44] Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U Khan. Big data reduction methods: a survey. *Data Science and Engineering*, 1(4):265–284, 2016.

[45] Rubén García-Pajares, José M Benítez, and GI Sainz-Palmero. Frasel: a consensus of feature ranking methods for time series modelling. *Soft Computing*, 17(8):1489–1510, 2013.

[46] Selwyn Piramuthu. Evaluating feature selection methods for learning in data mining applications. *European journal of operational research*, 156(2):483–494, 2004.

[47] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[48] H Burak Gunay, Weiming Shen, and Guy Newsham. Data analytics to improve building performance: A critical review. *Automation in Construction*, 97:96–109, 2019.

[49] Marti Frank, Hannah Friedman, Kristin Heinemeier, Cory Toole, David Claridge, Natascha Castro, and Philip Haves. State-of-the-art review for commissioning low energy buildings: Existing cost/benefit and persistence methodologies and data, state of development of automated tools and assessment of needs for commissioning zeb. *NISTIR*, 7356:2007, 2007.

[50] Sool Yeon Cho. *The persistence of savings obtained from commissioning of existing buildings*. PhD thesis, 2008.

[51] Fateme Fahiman, Sarah M Erfani, Sutharshan Rajasegarar, Marimuthu Palaniswami, and Christopher Leckie. Improving load forecasting based on deep learning and k-shape clustering. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4134–4141. IEEE, 2017.

[52] Christos N Schizas, Soteris A Kalogirou, and Costas Neocleous. Artificial neural networks in modelling the heat-up response of a solar steam generating plant. 1996.

[53] SA Kalogirou, CC Neocleous, and CN Schizas. Building heating load estimation using artificial neural networks. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, volume 8, page 14, 1997.

[54] Soteris A Kalogirou. Applications of artificial neural-networks for energy systems. *Applied Energy*, 67(1):17–35, 2000.

[55] D MacKay. Bayesian non-linear modeling for the 1993 energy prediction competition. *Maximum Entropy and Bayesian Methods*, pages 221–234, 1993.

[56] Fazli Wahid and Do-Hyeun Kim. Prediction methodology of energy consumption based on random forest classifier in korean residential apartments. 2015.

[57] Yangyang Fu, Zhengwei Li, Hao Zhang, and Peng Xu. Using support vector machine to predict next day electricity load of public buildings with sub-metering devices. *Procedia Engineering*, 121:1016–1022, 2015.

[58] Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1342–1351, 1998.

[59] Douglas J Leith, Martin Heidl, and John V Ringwood. Gaussian process prior models for electrical load forecasting. *Probabilistic Methods Applied to Power Systems*, pages 112–117, 2004.

[60] Sunil Mamidi, Yu-Han Chang, and Rajiv Maheswaran. Improving building energy efficiency with a network of sensing, learning and prediction agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 45–52. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[61] Minoru Kawashima, Charles E Dorgan, and John W Mitchell. Hourly thermal load prediction for the next 24 hours by arima, ewma, lr and an artificial neural network. Technical report, American Society of Heating, Refrigerating and Air-Conditioning Engineers . . . , 1995.

[62] X. Yan. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Company Pte Limited, 2009.

[63] Coskun Hamzacebi and Huseyin Avni Es. Forecasting the annual electricity consumption of turkey using an optimized grey model. *Energy*, 70:165–171, 2014.

[64] ASHRAE Fundamentals Handbook. American society of heating, refrigerating and air-conditioning engineers. *Inc.: Atlanta, GA, USA*, 2017.

[65] G Burnand. The study of the thermal behaviour of structures by electrical analogy. *British Journal of Applied Physics*, 3(2):50, 1952.

[66] Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, 2011.

[67] Alfonso P Ramallo-González, Matthew Brown, Elizabeth Gabe-Thomas, Tom Lovett, and David A Coley. The reliability of inverse modelling for the wide scale characterization of the thermal properties of buildings. *Journal of Building Performance Simulation*, pages 1–19, 2017.

[68] Gianluca Serale, Massimo Fiorentini, Alfonso Capozzoli, Daniele Bernardini, and Alberto Bemporad. Model predictive control (mpc) for enhancing building and hvac system energy efficiency: Problem formulation, applications and opportunities. *Energies*, 11(3):631, 2018.

[69] Sama Aghniaey, Thomas M Lawrence, Javad Mohammadpour, WenZhan Song, Richard T Watson, and Marie C Boudreau. Optimizing thermal comfort considerations with electrical demand response program implementation. *Building Services Engineering Research and Technology*, 39(2):219–231, 2018.

[70] Yuehong Lu, Shengwei Wang, Yongjun Sun, and Chengchu Yan. Optimal scheduling of buildings with energy generation and thermal energy storage under dynamic electricity pricing using mixed-integer nonlinear programming. *Applied Energy*, 147:49–58, 2015.

[71] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *Industrial Informatics, IEEE Transactions on*, 7(3):381–388, 2011.

[72] Yang-Seon Kim and Jelena Srebric. Improvement of building energy simulation accuracy with occupancy schedules derived from hourly building electricity consumption. *Ashrae Transactions*, 121(1):353–361, 2015.

[73] Dean Abbott. *Applied predictive analytics: principles and techniques for the professional data analyst*. John Wiley & Sons, 2014.

[74] Eamonn J Keogh and Michael J Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *KDD*, 98:239–243, 1998.

[75] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. *Data mining in time series databases*, 57:1–22, 2004.

[76] Hafzullah Aksoy, Abdullah Gedikli, N Erdem Unal, and Athanasios Kehagias. Fast segmentation algorithms for long hydrometeorological time series. *Hydrological Processes*, 22(23):4600–4608, 2008.

[77] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040. ACM, 2006.

[78] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, pages 1–35, 2013.

[79] Yuelong Zhu, De Wu, and Shijin Li. A piecewise linear representation method of time series based on feature points. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 1066–1072. Springer, 2007.

[80] Eamonn J Keogh and Michael J Pazzani. A simple dimensionality reduction technique for fast similarity search in large time series databases. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 122–133. Springer, 2000.

[81] Ngoc Thanh Nguyen, Bogdan Trawiński, Radosław Katarzyniak, and Geun-Sik Jo. *Advanced methods for computational collective intelligence*, volume 457. Springer, 2012.

[82] Houtao Deng, George Runger, Eugene Tuv, and Martyanov Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.

[83] Mustafa Gokce Baydogan and George Runger. Time series representation and similarity based on local autopatterns. *Data Mining and Knowledge Discovery*, 30(2):476–509, 2016.

[84] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[85] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, 15(2):107–144, 2007.

[86] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with cote: the collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, 2015.

[87] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[88] Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[89] Asha Gowda Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.

[90] Amir Ahmad and Lipika Dey. A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, 26(1):43–56, 2005.

[91] StevenL. Salzberg. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, 1994.

[92] Hai-Xiang Zhao and Frédéric Magoulès. Feature selection for predicting building energy consumption based on statistical learning method. *Journal of Algorithms & Computational Technology*, 6(1):59–77, 2012.

[93] Oscar Utterbäck. Feature selection methods with applications in electrical load forecasting. *Master's Theses in Mathematical Sciences*, 2017.

[94] Sven F Crone and Nikolaos Kourentzes. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. *Neurocomputing*, 73(10-12):1923–1936, 2010.

[95] Youqiang Sun, Jiuyong Li, Jixue Liu, Christopher Chow, Bingyu Sun, and Rujing Wang. Using causal discovery for feature selection in multivariate numerical time series. *Machine Learning*, 101(1-3):377–395, 2015.

[96] Shohei Hido and Tetsuro Morimura. Temporal feature selection for time-series prediction. In *2012 21st International Conference on Pattern Recognition (ICPR 2012)*, pages 3557–3560. IEEE, 2012.

[97] O García-Hinde, Vanessa Gómez-Verdejo, Manel Martínez-Ramón, Carlos Casanova-Mateo, J Sanz-Justo, Silvia Jiménez-Fernández, and Sancho Salcedo-Sanz. Feature selection in solar radiation prediction using bootstrapped svrs. In *Evolutionary Computation (CEC), 2016 IEEE Congress on*, pages 3638–3645. IEEE, 2016.

[98] Daniel O'Leary and Joel Kubby. Feature selection and ann solar power prediction. *Journal of Renewable Energy*, 2017, 2017.

[99] Cong Feng, Mingjian Cui, Bri-Mathias Hodge, and Jie Zhang. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy*, 190:1245–1257, 2017.

[100] Rubén García Pajares, Jose M Benítez, and Gregorio Sáinz Palmero. Feature selection for time series forecasting: A case study. In *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*, pages 555–560. IEEE, 2008.

[101] Eija Ferreira. Model selection in time series machine learning applications. *PhD Thesis. University of Oulu Graduate School. University of Oul*, 2015.

[102] Ling Tang, Chenghao Wang, and Shuai Wang. Energy time series data analysis based on a novel integrated data characteristic testing approach. *Procedia Computer Science*, 17:759–769, 2013.

[103] Irena Koprinska, Mashud Rana, and Vassilios G Agelidis. Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems*, 82:29–40, 2015.

[104] Francisco Martínez-Álvarez, Alicia Troncoso, Gualberto Asencio-Cortés, and José C Riquelme. A survey on data mining techniques applied to electricity-related time series forecasting. *Energies*, 8(11):13162–13193, 2015.

[105] Cyril Goutte. Note on free lunches and cross-validation. *Neural Computation*, 9(6):1245–1249, 1997.

[106] Richard De Dear and Gail Schiller Brager. The adaptive model of thermal comfort and energy conservation in the built environment. *International journal of biometeorology*, 45(2):100–108, 2001.

[107] Anna Carolina Menezes, Andrew Cripps, Dino Bouchlaghem, and Richard Buswell. Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. *Applied energy*, 97:355–364, 2012.

[108] Ian Richardson, Murray Thomson, and David Infield. A high-resolution domestic building occupancy model for energy demand simulations. *Energy and buildings*, 40(8):1560–1566, 2008.

[109] Jean Rouleau, Alfonso Ramallo-González, and Louis Gosselin. Towards a comprehensive tool to model occupant behaviour for dwellings that combines domestic hot water use with active occupancy. In *Proceedings of the 15th IBPSA Conference, San Francisco, CA, USA*, pages 7–9, 2017.

[110] Nataliya Mogles, Ian Walker, Alfonso P Ramallo-González, JeeHang Lee, Sukumar Natarajan, Julian Padget, Elizabeth Gabe-Thomas, Tom Lovett, Gang Ren, Sylwia Hyniewska, et al. How smart do smart meters need to be? *Building and Environment*, 125:439–450, 2017.

[111] H Burak Gunay, William O'Brien, Ian Beausoleil-Morrison, and Jayson Bursill. Development and implementation of a thermostat learning algorithm. *Science and Technology for the Built Environment*, 24(1):43–56, 2018.

[112] John Goins and Mithra Moezzi. Linking occupant complaints to building performance. *Building Research & Information*, 41(3):361–372, 2013.

[113] Frédéric Haldi and Darren Robinson. On the behaviour and adaptation of office occupants. *Building and environment*, 43(12):2163–2177, 2008.

[114] H Burak Gunay, William O'Brien, Ian Beausoleil-Morrison, and Sara Gilani. Development and implementation of an adaptive lighting and blinds control algorithm. *Building and Environment*, 113:185–199, 2017.

[115] Sarah Darby. Smart metering: what potential for householder engagement? *Building Research & Information*, 38(5):442–457, 2010.

[116] Tom Hargreaves, Michael Nye, and Jacquelin Burgess. Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. *Energy policy*, 38(10):6111–6119, 2010.

[117] Directive (eu) 2018/844 of the european parliament and of the council amending directive 2010/31/eu on the energy performance of buildings and directive 2012/27/eu on energy efficiency. *Official Journal of the European Union*, L156:75–91, 2018.

[118] Maitreyee Dey, Manik Gupta, Mikdam Turkey, and Sandra Dudley. Unsupervised learning techniques for hvac terminal unit behaviour analysis. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–7. IEEE, 2017.

[119] Maitreyee Dey, Soumya Prakash Rana, and Sandra Dudley. Smart building creation in large scale hvac environments through automated fault detection and diagnosis. *Future Generation Computer Systems*, 2018.

[120] Michelle Shipworth, Steven K Firth, Michael I Gentry, Andrew J Wright, David T Shipworth, and Kevin J Lomas. Central heating thermostat settings and timing: building demographics. *Building Research & Information*, 38(1):50–69, 2010.

[121] Rune Vinther Andersen, Bjarne W Olesen, and Jørn Toftum. Modelling occupants' heating set-point prefferences. In *Building Simulation 2011: 12th Conference of International Building Performance Simulation Association*, 2011.

[122] Yu Zheng and Xiaofang Zhou. *Computing with spatial trajectories*. Springer Science & Business Media, 2011.

[123] Tian Guo, Zhixian Yan, and Karl Aberer. An adaptive approach for online segmentation of multi-dimensional mobile data. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pages 7–14. ACM, 2012.

[124] Nicholas Jing Yuan, Yingzi Wang, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. Reconstructing individual mobility from smart card transactions: A space alignment approach. In *2013 IEEE 13th International Conference on Data Mining*, pages 877–886. IEEE, 2013.

[125] Huiji Gao and Huan Liu. Mining human mobility in location-based social networks. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 7(2):1–115, 2015.

[126] Miao Lin and Wen-Jing Hsu. Mining gps data for mobility patterns: A survey. *Pervasive and Mobile Computing*, 12:1–16, 2014.

[127] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6:8166, 2015.

[128] Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 381–390. ACM, 2012.

[129] Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, and Christian S Jensen. Path prediction and predictive range querying in road network databases. *The VLDB Journal*, 19(4):585–602, 2010.

[130] John Krumm, Robert Gruen, and Daniel Delling. From destination prediction to route prediction. *Journal of Location Based Services*, 7(2):98–120, 2013.

[131] Disheng Qiu, Paolo Papotti, and Lorenzo Blanco. Future locations prediction with uncertain data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–432. Springer, 2013.

[132] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613. ACM, 2013.

[133] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th international conference on data mining*, pages 1038–1043. IEEE, 2012.

[134] Natalia Andrienko, Gennady Andrienko, Georg Fuchs, and Piotr Jankowski. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, 15(2):117–153, 2016.

[135] Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237–245, 2014.

[136] Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, page 6. ACM, 2013.

[137] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and data engineering*, 25(4):919–931, 2012.

[138] Ruchi Parikh and Kamalakar Karlapalem. Et: events from tweets. In *Proceedings of the 22nd international conference on world wide web*, pages 613–620. ACM, 2013.

[139] Alan Ritter, Oren Etzioni, Sam Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[140] Calin Railean, Monica Borda, and Alexandra Moraru. Complex event processing in social media. *Acta Technica Napocensis*, 55(3):10, 2014.

[141] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. Real-time detection of traffic from twitter stream analysis. *IEEE transactions on intelligent transportation systems*, 16(4):2269–2283, 2015.

[142] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. Semantic annotation of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1253–1263. International World Wide Web Conferences Steering Committee, 2015.

[143] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[144] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[145] Shane B Eisenman, Emiliano Miluzzo, Nicholas D Lane, Ronald A Peterson, Gahng-Seop Ahn, and Andrew T Campbell. Bikenet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks (TOSN)*, 6(1):6, 2009.

[146] Linjuan Zhang, Deyun Gao, Weicheng Zhao, and Han-Chieh Chao. A multilevel information fusion approach for road congestion detection in vanets. *Mathematical and Computer Modelling*, 58(5-6):1206–1221, 2013.

[147] Jiafu Wan, Jianqi Liu, Zehui Shao, Athanasios Vasilakos, Muhammad Imran, and Keliang Zhou. Mobile crowd sensing for traffic prediction in internet of vehicles. *Sensors*, 16(1):88, 2016.

[148] Ármin Petkovics and Károly Farkas. Efficient event detection in public transport tracking. In *2014 International Conference on Telecommunications and Multimedia (TEMU)*, pages 74–79. IEEE, 2014.

[149] Shaohan Hu, Lu Su, Hengchang Liu, Hongyan Wang, and Tarek Abdelzaher. Smartroad: a crowd-sourced traffic regulator detection and identification system. In *Proceedings of the 12th international conference on Information processing in sensor networks*, pages 331–332. ACM, 2013.

[150] Bryce W Sharman and Matthew J Roorda. Analysis of freight global positioning system data: clustering approach for identifying trip destinations. *Transportation Research Record*, 2246(1):83–91, 2011.

[151] Zhenhui Li, Jiawei Han, Bolin Ding, and Roland Kays. Mining periodic behaviors of object movements for animal and biological sustainability studies. *Data Mining and Knowledge Discovery*, 24(2):355–386, 2012.

[152] Opher Etzion, Peter Niblett, and David C Luckham. *Event processing in action*. Manning Greenwich, 2011.

[153] Sebastian Stipkovic, Ralf Bruns, and Jürgen Dunkel. Pervasive computing by mobile complex event processing. In *2013 IEEE 10th International Conference on e-Business Engineering*, pages 318–323. IEEE, 2013.

[154] Lu-An Tang, Yu Zheng, Jing Yuan, Jiawei Han, Alice Leung, Chih-Chieh Hung, and Wen-Chih Peng. On discovery of traveling companions from streaming trajectories. In *2012 IEEE 28th International Conference on Data Engineering*, pages 186–197. IEEE, 2012.

[155] Andreas Bauer and Christian Wolff. An event processing approach to text stream analysis: basic principles of event based information filtering. In *Proceedings of the 8th acm international conference on distributed event-based systems*, pages 35–46. ACM, 2014.

[156] Kaile Zhou, Chao Fu, and Shanlin Yang. Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, 56:215–225, 2016.

[157] Yogesh Simmhan, Saima Aman, Alok Kumbhare, Rongyang Liu, Sam Stevens, Qunzhi Zhou, and Viktor Prasanna. Cloud-based software platform for big data analytics in smart grids. *Computing in Science & Engineering*, 15(4):38, 2013.

[158] Alok Kumbhare, Yogesh Simmhan, and Viktor Prasanna. Cryptonite: a secure and performant data repository on public clouds. In *2012 IEEE Fifth International Conference on Cloud Computing*, pages 510–517. IEEE, 2012.

[159] Atsushi Ishii and Toyotaro Suzumura. Elastic stream computing with clouds. In *2011 IEEE 4th International Conference on Cloud Computing*, pages 195–202. IEEE, 2011.

[160] Laura Klein, Jun-young Kwak, Geoffrey Kavulya, Farrokh Jazizadeh, Burcin Becerik-Gerber, Pradeep Varakantham, and Milind Tambe. Coordinating occupant behavior for building energy and comfort management using multi-agent systems. *Automation in construction*, 22:525–536, 2012.

[161] Nan Li, Jun-young Kwak, Burcin Becerik-Gerber, and Milind Tambe. Predicting hvac energy consumption in commercial buildings using multiagent systems. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 30, page 1. IAARC Publications, 2013.

[162] Cristina Alcaraz, Isaac Agudo, David Nunez, and Javier Lopez. Managing incidents in smart grids a la cloud. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science*, pages 527–531. IEEE, 2011.

[163] M Moreno, Benito Úbeda, Antonio Skarmeta, and Miguel Zamora. How can we tackle energy efficiency in iot basedsmart buildings? *Sensors*, 14(6):9582–9614, 2014.

[164] Julien Mineraud, Oleksiy Mazhelis, Xiang Su, and Sasu Tarkoma. A gap analysis of internet-of-things platforms. *Computer Communications*, 89:5–16, 2016.

[165] Linda Steg, Goda Perlaviciute, and Ellen van der Werff. Understanding the human dimensions of a sustainable energy transition. *Frontiers in psychology*, 6:805, 2015.

[166] Sam C Staddon, Chandrika Cycil, Murray Goulden, Caroline Leygue, and Alexa Spence. Intervening to change behaviour and save energy in the workplace: A systematic review of available evidence. *Energy Research & Social Science*, 17:30–51, 2016.

[167] Kathy Kuntz, Rajan Shukla, and I Bensch. How many points for that? a game-based approach to environmental sustainability. *Proceedings of the American Council for an*

Energy-Efficient Economy Summer Study on Energy Efficiency in Buildings, 7:126–137, 2012.

[168] Brian Orland, Nilam Ram, Dean Lang, Kevin Houser, Nate Kling, and Michael Coccia. Saving energy in an office environment: A serious game intervention. *Energy and Buildings*, 74:43–52, 2014.

[169] Graham N Dixon, Mary Beth Deline, Katherine McComas, Lauren Chambliss, and Michael Hoffmann. Using comparative feedback to influence workplace energy conservation: A case study of a university campaign. *Environment and Behavior*, 47(6):667–693, 2015.

[170] Andreas Kamilaris, Jodi Neovino, Sekhar Kondepudi, and Balaji Kalluri. A case study on the individual energy use of personal computers in an office setting and assessment of various feedback types toward energy savings. *Energy and Buildings*, 104:73–86, 2015.

[171] Stephanie N Timm and Brian M Deal. Effective or ephemeral? the role of energy information dashboards in changing occupant energy behaviors. *Energy Research & Social Science*, 19:11–20, 2016.

[172] Liga Poznaka, Ilze Laicane, Dagnija Blumberga, Andra Blumberga, and Marika Rosa. Analysis of electricity user behavior: case study based on results from extended household survey. *Energy Procedia*, 72:79–86, 2015.

[173] Brian Thomas and Diane Cook. Activity-aware energy-efficient automation of smart buildings. *Energies*, 9(8):624, 2016.

[174] Tianzhen Hong, Sarah C Taylor-Lange, Simona D'Oca, Da Yan, and Stefano P Corgnati. Advances in research and applications of energy-related occupant behavior in buildings. *Energy and buildings*, 116:694–702, 2016.

[175] Li Da Xu, Wu He, and Shancang Li. Internet of things in industries: A survey. *IEEE Transactions on industrial informatics*, 10(4):2233–2243, 2014.

[176] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[177] Zhongde Wang. Fast algorithms for the discrete w transform and for the discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(4):803–816, 1984.

[178] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[179] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[180] Alan C Bovik. *The essential guide to image processing*. Academic Press, 2009.

[181] Martin Hirzel, Henrique Andrade, Bugra Gedik, Gabriela Jacques-Silva, Rohit Khandekar, Vibhore Kumar, Mark Mendell, Howard Nasgaard, Scott Schneider, Robert Soulé, et al. Ibm streams processing language: Analyzing big data in motion. *IBM Journal of Research and Development*, 57(3/4):7–1, 2013.

[182] Lei Li, Farzad Noorian, Duncan JM Moss, and Philip HW Leong. Rolling window time series prediction using mapreduce. In *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*, pages 757–764. IEEE, 2014.

[183] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[184] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[185] S Taylor and B Letham. Prophet: Automatic forecasting procedure. *R package version 0.2*, 1, 2017.

[186] C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005. cited By (since 1996)149.

[187] T Agami Reddy, Namir F Saman, David E Claridge, Jeff S Haberl, W Dan Turner, and Alan T Chalifoux. Baselining methodology for facility-level monthly energy use-part 1: Theoretical aspects. In *ASHRAE transactions*, pages 336–347. ASHRAE, 1997.

[188] Franklin L Quilumba, Wei-Jen Lee, Heng Huang, David Y Wang, and Robert L Szabados. Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Transactions on Smart Grid*, 6(2):911–918, 2015.

[189] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127:1–10, 2014.

[190] Richard E Edwards, Joshua New, and Lynne E Parker. Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, 49:591–603, 2012.

[191] Jeremy Miles Andy Field and Zoe Field Niblett. *Discovering Statistics Using R*. Sage Publications Ltd, 1st edition, 2012.

[192] Alfonso P Ramallo-González, Matthew Brown, and David A Coley. Identifying the ideal topology of simple models to represent dwellings.

[193] DA Coley and JM Penman. Second order system identification in the thermal response of real buildings. paper ii: recursive formulation for on-line building energy management and control. *Building and Environment*, 27(3):269–277, 1992.

[194] F Jiménez, G Sánchez, JM García, G Sciavicco, and L Miralles. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing*, 2016.

[195] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 2 edition, 2003.

[196] Fernando Jiménez, Gracia Sánchez, and José M Juárez. Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. *Artificial intelligence in medicine*, 60(3):197–219, 2014.

[197] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[198] F Jiménez, G Sánchez, JM García, G Sciavicco, and L Miralles. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing*, 234:75–92, 2017.

[199] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation*, 8(2):173–195, 2000.

[200] E. Zitzler, L. Thiele, M. Laumanns, C.M. Fonseca, and V. Grunert da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7:117–132, 2002.

[201] J. Novakovic. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21(1), 2016.

[202] Huan Liu and Rudy Setiono. A probabilistic approach to feature selection - a filter solution. In *Proceedings of the 13th International Conference on Machine Learning (ICML)*, volume 96, pages 319–327, 1996.

[203] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning*, ML92, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[204] Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[205] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[206] Henrik Spliid. *Multivariate ARIMA and ARIMA-X Analysis*. CRAN, 2017. License GPL-2, Version 2.2, RoxygenNote 5.0.1.

[207] Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang. A hybrid both filter and wrapper feature selection method for microarray classification. *CoRR*, abs/1612.08669, 2016.

[208] Luis Carlos Molina, Lluís Belanche, and Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 306–313, 2002.

[209] J Fergus Nicol and Michael A Humphreys. Adaptive thermal comfort and sustainable thermal standards for buildings. *Energy and buildings*, 34(6):563–572, 2002.

[210] Balakrishnan Narayanaswamy, Bharathan Balaji, Rajesh Gupta, and Yuvraj Agarwal. Data driven investigation of faults in hvac systems with model, cluster and compare (mcc). In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 50–59. ACM, 2014.

[211] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[212] Elias Frentzos, Kostas Gratsias, and Yannis Theodoridis. Index-based most similar trajectory search. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 816–825. IEEE, 2007.

[213] Fernando Terroso-Saenz, Mercedes Valdes-Vela, and Antonio F Skarmeta-Gomez. A complex event processing approach to detect abnormal behaviours in the marine environment. *Information Systems Frontiers*, 18(4):765–780, 2016.

[214] Fernando Terroso-Sáenz, Mercedes Valdés-Vela, Francisco Campuzano, Juan A Botia, and Antonio F Skarmeta-Gómez. A complex event processing approach to perceive the vehicular context. *Information Fusion*, 21:187–209, 2015.

[215] Robert Babuška. *Fuzzy modeling for control*, volume 12. Springer Science & Business Media, 2012.

[216] Umutcan Şimşek, Anna Fensel, Anastasios Zafeiropoulos, Eleni Fotopoulou, Paris Liapis, Thanassis Bouras, Fernando Terroso Saenz, and Antonio F Skarmeta Gómez. A semantic approach towards implementing energy efficient lifestyles through behavioural change.

In *Proceedings of the 12th International Conference on Semantic Systems*, pages 173–176. ACM, 2016.

[217] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[218] Jessica Granderson, Phillip N Price, David Jump, Nathan Addy, and Michael D Sohn. Automated measurement and verification: Performance of public domain whole-building electric baseline models. *Applied Energy*, 144:106–113, 2015.