

MULTIVARIATE STATISTICAL MODELLING AND MONITORING OF SMART BUILDINGS

Llorenç Burgas Nadal

Per citar o enllaçar aquest document:

Para citar o enlazar este documento:

Use this url to cite or link to this publication:

<http://hdl.handle.net/10803/669279>

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



PhD Thesis

Multivariate Statistical modelling and monitoring
of smart buildings

Llorenç Burgas Nadal

2019



DOCTORAL THESIS

**Multivariate Statistical modelling and
monitoring of smart buildings**

Llorenç Burgas Nadal

2019

DOCTORAL PROGRAM in TECHNOLOGY

Supervised by:
Dr. Joquim Melendez
Dr. Joan Colomer

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

Acknowledgments

Aquesta tesi s'ha dut a terme gràcies a la paciència i bons consells dels meus dos directors el Dr. Joaquim Melendez i el Dr. Joan Colomer, als quals agraeixo la seva bona feina, tota la dedicació i paciència.

M'agradaria donar les gràcies a tots els integrants del grup eXiT pel seu suport durant la meva ja llarga estada en aquest grup de recerca. En aquest grup sempre m'he sentit molt ben integrat, també dir que durant aquests anys hem compartit molts bons moments com per exemple dinars de Nadal, calçotades i sopars.

Voldria també donar les gràcies a tota la meva família pel seu suport incondicional durant la durada d'aquesta tesi i també les etapes prèvies de formació acadèmica.

This work was developed with the support of the research group SITES, which was awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016), and the research group eXiT (Control Engineering and Intelligent Systems) of the IiA (Institute of Informatics and Applications) of the Department of Electrical and Electronic Engineering and Automation of the University of Girona.

This thesis has been funded through the competitive grant for doctoral education IFUdG2016 from the University of Girona granted to Llorenç Burgas Nadal. Economic support was also received in the initial years from the following projects:

- ACCUS (Adaptive Cooperative Control in Urban (sub) Systems., ART-010000-2013-2 -333020-1), funded by the Spanish Ministry of Industry, Energy and Tourism and by the JTI ARTEMIS Joint Undertaking of the European Commission
- MESC project (Ref. DPI2013-47450-C2-1-R) funded by the Spanish MINECO within the program aimed at the Challenges of Society.
- CROWDSAVING project (Ref. TIN2016-79726-C2-2-R) funded by the Spanish MINECO within the program aimed at the Challenges of Society.
- HIT2GAP project of the Horizon 2020 research and innovation program under grant agreement N°680708

Last but not least, I would like to specially thank Luis Blanes and all the IRUSE (Informatics Research Unit for Sustainable Engineering) group from the National University of Ireland, Galway (NUI Galway) for helping me during my time in Ireland.

Publications

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

- Burgas, L., Melendez, J., Colomer, J., Massana, J. & Pous, C. (2015). Multi-variate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building. *Energy and Buildings*, 103, 338-351.
Quality index: [JCR IF (2015): 2.973, Q1]
- Burgas, L., Melendez, J., Colomer, J., Massana, J. & Pous, C. (2018). N-dimensional extension of unfold-PCA for granular systems monitoring. *Engineering Applications of Artificial Intelligence*, 71, 113-124.
Quality index: [JCR IF (2017): 2.819, Q1]
- Burgas, L., Melendez, J., Colomer, J., Gamero, F.I. & Herraiz, S. (2019). Integrated Unfold-PCA monitoring application for smart buildings: AHU application example. Submitted to: *Applied Energy*.
Quality index: [JCR IF (2017): 7.9, Q1]

Moreover, during the development of this thesis, the candidate has contributed to other results that complement the work in the compendium. Contributing publications derived from those results are listed below:

Journals

- Massana, J., Pous, C., Burgas, L., Melendez, J. & Colomer, J. (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings*, 92, 322-330.
Quality index: [JCR IF (2015): 2.973, Q1]
- Massana, J., Pous, C., Burgas, L., Melendez, J. & Colomer, J. (2016). Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. *Energy and Buildings*, 130, 519-531.
Quality index: [JCR IF (2016): 4.067, Q1]

- Massana, J., Pous, C., Burgas, L., Melendez, J. & Colomer, J. (2017). Identifying services for short-term load forecasting using data driven models in a Smart City platform. *Sustainable Cities and Society*, 28, 108-117.
Quality index: [JCR IF (2017): 3.073, Q1]

Conferences

- Burgas, L., Melendez, J. & Colomer, J. (2014). Principal component analysis for monitoring electrical consumption of academic buildings. In *6th International Conference on Sustainability in Energy and Buildings*, Energy Procedia 62 (2014) 555–564
- Burgas, L., Melendez, J. & Colomer, J. (2015). Modelling transitions on heating usage in buildings with multivariate statistical monitoring. In *IEEE EUROCON 2015*
- Meléndez, J., Burgas, L., Gamero, F.I., Colomer, J. & Herraiz, S. (2018). Fault detection and diagnosis web service module for energy monitoring in buildings. In *IFAC Volume 51, Issue 10, 2018*, Pages 15-19

Workshops

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

Acronyms

AHU Air Handling Unit

BMS Building Management System

BEMS Building Energy Management System

FDD Fault Detection and Diagnosis

IoT Internet of Things

MPCA Multiway Principal Component Analysis

NOC Normal Operation Conditions

NUI National University of Ireland

PCA Principal Component Analysis

SPE Square Prediction Error

WSN Wireless Sensor Network

List of Figures

6.1	SPE contributions for day 11, March 2013 for a daily monitoring of the whole building	57
6.2	Second and third scores where each point represents a day in a dwelling, Red dots are from dwelling 1 (first floor corner) and blue crosses are from dwelling 18 (third floor non corner)	59

List of Tables

1.1	Main characteristics of statistical and engineering techniques	4
-----	--	---

Contents

List of figures	vii
List of tables	ix
Abstract	xiii
Resum	xvii
Resumen	xxi
1 Introduction	1
1.1 Hypothesis formulation	2
1.2 State of the art	4
1.3 Contributions of this thesis	5
1.4 Structure of the document	6
2 Objective	7
2.1 Vision	7
2.2 Objectives	8
2.3 Journal Contributions	9
3 Multivariate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building	11
4 N-dimensional extension of unfold-PCA for granular systems monitoring	27

5	Integrated Unfold-PCA monitoring application for smart buildings: An AHU application example.	41
6	Main results and discussion	55
6.1	New PCA methodology for building monitoring	55
6.2	Fold-unfold strategy for granular and modular systems monitoring .	56
6.3	Implementing the building monitoring methodology	58
7	Conclusions	61
7.1	Define a data-driven methodology for supervising smart-buildings . .	61
7.2	Extensions to consider granularity and modularity	61
7.3	Validating the methodology	62
7.4	Providing a technological solution	63
7.5	Future Work	63
	Bibliography	65

Abstract

In this modern world, energy has become an absolutely essential good in any developed society. As the technological and scientific advances that we enjoy today have been progressively adopted by society, we can see just how much they need and depend on the generation, storage and distribution of the various kinds of energy that are used nowadays, and which is why the demand for energy is constantly increasing.

Recently, serious concerns about the sustainability of the system have been raised because, unfortunately, not all of the energies we use are renewable. Furthermore, some of the energies in use nowadays can produce unwanted effects, such as, for example, the well-known CO₂ emissions. As a result of such concerns, recently researchers have aimed their efforts at improving the uses we make of energy, improving the efficiency of the various processes linked to energy, and searching for cleaner and more efficient sources. The new regulations and directives, created by both the European Union and various governments around the world, are helping to move towards a more sustainable future. These new regulations encourage the use of renewable energies and try to encourage the improvement of energy uses. This thesis focusses on the uses of the various energies found in buildings and, in particular, controlling and monitoring these uses.

The lack of supervision or control of energy use has often been the cause of many energy losses. These losses are often caused by the complexity that monitoring all the possible uses of the conceivable types of energy requires. Even if we look at this problem only within the framework of buildings, there is clearly a lack of tools. In fact, not having enough versatile and powerful tools has led to mismatches in the multiple uses that are made of the huge amount of energy that we have at our disposal.

In order to reduce mismatches between real and expected consumption, this thesis explores the use of PCA (Principal Component Analysis) as a modelling tool for buildings. PCA is a statistical technique that allows complex systems to be modelled, and, subsequently, to monitor them to detect abnormal behaviours with respect to the conditions initially modelled. The work in this thesis includes

adapting PCA to take full advantage of its potential in buildings. Such adaptation is also verified by applying it to various use cases.

This thesis has been divided into tasks in an attempt to obtain a new methodology and tools with which to monitor buildings. The results obtained as these tasks were carried out, have resulted in three publications that have allowed thesis to be presented as a compendium of articles. In each of the publications, the following contributions have been made to the state of the art.

The first publication [4], takes as its initial point, a variant of PCA, MPCA (Multiway Principal Components Analysis) or Unfold-PCA. This variant of PCA allows batch processes to be analysed. In order to adapt them to the monitoring of buildings, this publication defines the two types of redundancy that we can find in buildings i.e., redundancy in time patterns (daily, weekly, seasonal, annual) and redundancy in physical spaces (room, apartment, plant, building). Considering these redundancies and assimilating them to the temporary redundancy of the batch processes, it is possible to obtain new ways of modelling buildings or parts of buildings, by considering the building as a whole and not as a set of independent systems.

The second publication [1], takes the first publication as its starting point and builds on the concept of the two types of redundancy being useful for modelling. The second publication defines the concept of the granular system. A granular system, in our case a building, is a system that has one or several redundancies in time (granularity) and/or variables/structure (modularity). In order to allow PCA, or other data-mining techniques that need a two-dimensional input, to take advantage of this inherent peculiarity, the data is mathematically defined using a fold and unfold technique (the way of organizing the data in N-dimensional arrays) that allows the structured data required to achieve a model that captures the desired behaviours to be obtained.

The third publication, which has been sent to a magazine, describes the implementation of the tools and methods developed in this thesis as web services included in a web application. This web application allows online PCA models to be used with the historical data of buildings. It also allows, when new data is available, the pre-existent models to be used for monitoring purposes. This tool is integrated into the European project HIT2GAP's platform platform as an external module. This module allows users who are not experts in PCA, to carry out control and detection tasks using the previously created models. These models can be created with the same online tool by users expert in PCA modelling.

While this thesis was being undertaken, and as part of the publications, the correct functioning of the methodology was verified using real data from various buildings. In short, the data used came from three different buildings (both architecturally and in terms of uses and available technical solutions) provided by

the Barcelona Patronat de l'Habitatge, the buildings on the Montilivi campus of the University of Girona and the AHU (Air Handling Unit) of the Alice Perry Building of the National University of Ireland Galway. In fact, the collaboration with the NUI Galway allowed us to have three-month internship and also to request the mention of international doctorate for this thesis.

Resum

En aquest món modern, l'energia ha esdevingut un bé necessari per no dir essencial en tota societat desenvolupada. A mesura que s'han anat fent els avenços tecnològics i científics dels quals gaudim avui en dia, hem pogut anar constatant com aquests depenen en major o menor mesura de la generació, emmagatzematge i distribució de les diverses energies que enguany s'utilitzen. És per tot això que, al llarg dels anys, la demanda d'energia ha anat augmentant.

Darrerament hi comencen a haver serioses preocupacions per la sostenibilitat del sistema, ja que malauradament no totes les fonts d'energia que utilitzem són renovables i també cal tenir en compte que el seu ús pot produir efectes no desitjats, com per exemple, les ja conegudes per tothom emissions de CO₂. Aquesta problemàtica ha fet que molts esforços d'investigadors en els darrers anys hagin anat encaminats a millorar els usos que fem de l'energia, millorar l'eficiència dels diversos processos lligats a l'energia, o també la recerca de noves fonts més netes i eficients. També ha ajudat a avançar cap a un futur més sostenible les noves regulacions i directives, creades tant per part de la Unió Europea com dels diversos governs arreu del món. Aquestes noves regulacions fomenten l'ús d'energies renovables i tracten d'incentivar la millora dels usos que es fan de l'energia. En aquesta tesi ens centrarem en els usos que es fan de les diverses energies que podem trobar en edificis, centrant-nos especialment en el control i el monitoratge d'aquests diversos usos.

Una font de malbaratament d'energia ha estat sovint la manca de supervisió o control dels usos que fem de les energies. Molts cops aquesta deixadesa ha estat provocada per la complexitat que pot arribar a tenir monitoritzar totes les energies que s'utilitzen i tots els possibles usos que se'n fa. Fins i tot si acotem el problema només en el marc dels edificis, apareix una clara manca d'eines. El fet de no tenir eines prou versàtils i potents ha provocat desajustos en els múltiples usos que es fa de la ingent quantitat de fonts d'energia que tenim al nostre l'abast.

Per tal de reduir els desajustos entre el consum real i l'esperat, en aquesta tesi s'explora l'ús de PCA (les sigles en Angles d'Anàlisi de Components Principals) com a eina de modelat per edificis. PCA és una tècnica estadística que permet modelar sistemes complexes i posteriorment monitoritzar-los per detectar comportaments

anòmals respecte a les condicions modelades inicialment. Els treballs d'aquesta tesi inclouen l'adaptació de PCA per poder aprofitar tot el seu potencial en edificis i la verificació de l'adaptació realitzada mitjançant l'aplicació en diversos casos d'ús.

Aquesta tesi s'ha dividit en tasques encaminades a aconseguir una nova metodologia i una nova eina per tal de monitorar edificis. Els resultats obtinguts durant la realització d'aquestes tasques han derivat en tres publicacions que han permès la presentació d'aquesta tesi en format de compendi d'articles. En cada una de les publicacions s'han realitzat les següents aportacions a l'estat de l'art.

A la primera publicació [4], es parteix de la variant de PCA, MPCA (sigles en Anglès de Anàlisi de Components principals Multilineal) o Unfold-PCA que permet analitzar processos batch per adaptar-la a la monitorització d'edificis. En aquesta publicació es defineixen els 2 tipus de redundància que podem tenir presents en edificis. La redundància en patrons temporals (diari, setmanal, estacional, anual) i la redundància en espais físics (habitació, apartament, planta, edifici). Considerant aquestes redundàncies i assimilant-les a la redundància temporal que tenen els processos batch s'aconsegueix obtenir noves maneres de modelar edificis o parts d'aquest considerant l'edifici com un tot i no com a un conjunt de sistemes independents.

La segona publicació [1], parteix de la primera, on es descobreix que els edificis presenten 2 tipus de redundàncies útils per modelar. En aquesta segona publicació es defineix el concepte de sistema granular. Un sistema granular en el nostre cas un edifici, és un sistema que presenta una o diverses redundàncies en el temps (granularitat) i/o en les variables/estructura (modularitat). Per tal permetre PCA, o també altres tècniques de datamining que necessiten una entrada bidimensional, aprofitar aquesta peculiaritat inherent en les dades es defineix matemàticament una tècnica de doblat i desdoblament (forma d'organitzar les dades en arrays N-dimensionals) que permet obtenir les dades estructurades per tal d'aconseguir un model que capturi els comportaments desitjats.

La tercera publicació actualment enviada a revista descriu la implementació de les eines i mètodes desenvolupats en aquesta tesi com a serveis web i inclosos en una aplicació web. Aquesta aplicació web permet totalment en línia crear models PCA amb les dades històriques d'edificis. També permet un cop creat un model el posterior monitoratge de dades noves a mesura que aquestes van arribant. Aquesta eina està integrada com un mòdul extern a la plataforma desenvolupada en el projecte europeu HIT2GAP i permet a usuaris no experts amb PCA portar a terme les tasques de control i detecció utilitzant els models prèviament creats per usuaris experts en modelització.

Durant la tesi i com a part de les publicacions s'ha verificat el correcte funcionament de la metodologia utilitzant dades reals provinents de diversos edificis. A tall de resum, s'ha utilitzat dades de 3 edificis diferents (tant arquitectònicament

com pel que fa a usos i solucions tècniques que disposen) facilitades pel patronat de l'habitatge de Barcelona, els edificis del campus de Montilivi de la Universitat de Girona i l'AHU (de l'Anglès Air Handling Unit, equip de climatització) de l'edifici Alice Perry de la Universitat Nacional d'Irlanda, Galway. La col·laboració amb aquesta Universitat va permetre fer una estada de 3 mesos i sol·licitar la menció de doctorat internacional per aquesta tesi.

Resumen

En este mundo moderno, la energía se ha convertido en un bien necesario por no decir esencial en toda sociedad desarrollada. A medida que se han ido haciendo los avances tecnológicos y científicos de los que disfrutamos hoy en día, hemos podido ir constatando como estos dependen en mayor o menor medida de la generación, almacenamiento y distribución de las diversas energías que hoy en día se utilizan. Es por todo ello que, a lo largo de los años, la demanda de energía ha ido aumentando.

Últimamente surgen serias preocupaciones por la sostenibilidad del sistema, ya que desgraciadamente no todas las fuentes de energía que utilizamos son renovables y también hay que tener en cuenta que su uso puede producir efectos no deseados, como por ejemplo, las ya conocidas por todos emisiones de CO₂. Esta problemática ha hecho que muchos esfuerzos de investigadores en los últimos años hayan ido encaminados a mejorar los usos que hacemos de la energía, mejorar la eficiencia de los múltiples procesos ligados a la energía, o también la búsqueda de nuevas fuentes más limpias y eficientes. También ha ayudado a avanzar hacia un futuro más sostenible las nuevas regulaciones y directivas, creadas tanto por parte de la Unión Europea como de los diversos gobiernos en todo el mundo. Estas nuevas regulaciones fomentan el uso de energías renovables y tratan de incentivar la mejora de los usos que se hacen de la energía. En esta tesis nos centraremos en los usos que se hacen de las diversas energías que podemos encontrar en edificios, centrándonos especialmente en el control y la monitorización de estos usos diversos.

Una fuente de derroche de energía ha sido a menudo la falta de supervisión o control de los usos que hacemos de las energías. Muchas veces esta dejadez ha sido provocada por la complejidad que puede llegar a tener monitorizar todas las energías que se utilizan y todos los posibles usos que se hacen de estas. Incluso si acotamos el problema sólo en el marco de los edificios, aparece una clara falta de herramientas. El hecho de no tener herramientas suficientemente versátiles y potentes ha provocado desajustes en los múltiples usos que se hace de la ingente cantidad de fuentes de energía que tenemos a nuestro alcance.

Con el fin de reducir estos desajustes entre el consumo real y el esperado, en esta tesis se explora el uso de PCA (las siglas en Ingles de Análisis de Componentes

Principales) como herramienta de modelado para edificios. PCA es una técnica estadística que permite modelar sistemas complejos y posteriormente monitorizarlos para detectar comportamientos anómalos respecto a las condiciones modeladas inicialmente. Los trabajos de esta tesis incluyen la adaptación de PCA para poder aprovechar todo su potencial en edificios y la verificación de la adaptación realizada mediante la aplicación en varios casos de uso.

Esta tesis se ha dividido en tareas encaminadas a conseguir una nueva metodología y una nueva herramienta para monitorizar edificios. Los resultados obtenidos durante la realización de estas tareas han derivado en tres publicaciones que han permitido la presentación de esta tesis en formato de compendio de artículos. En cada una de las publicaciones se han realizado las siguientes aportaciones al estado del arte.

En la primera publicación [4], se parte de la variante de PCA, MPCA (siglas en Inglés de Análisis de Componentes principales Multilineal) o Unfold-PCA que permite analizar procesos batch para adaptarla a la monitorización de edificios. En esta publicación se definen los 2 tipos de redundancia que podemos tener presentes en edificios. La redundancia en patrones temporales (diario, semanal, estacional, anual) y la redundancia en espacios físicos (habitación, apartamento, planta, edificio). Considerando estas redundancias y asimilándolas a la redundancia temporal que tienen los procesos batch se consigue obtener nuevas maneras de modelar edificios o partes de este considerando el edificio como un todo y no como un conjunto de sistemas independientes.

La segunda publicación [1], se parte de la primera, donde se descubre que los edificios presentan 2 tipos de redundancias útiles para modelar. En esta segunda publicación se define el concepto de sistema granular. Un sistema granular en nuestro caso un edificio, es un sistema que presenta una o varias redundancias en el tiempo (granularidad) y/o en las variables/estructura (modularidad). Para permitir PCA, o también otras técnicas de datamining que necesitan una entrada bidimensional, aprovechar esta peculiaridad inherente en los datos se define matemáticamente una técnica de doblado y desdoblado (forma de organizar los datos en arrays N-dimensionales) que permite obtener los datos estructurados para conseguir un modelo que capture los comportamientos deseados.

La tercera publicación actualmente enviada a revista describe la implementación de las herramientas y métodos desarrollados en esta tesis como servicios web e incluidos en una aplicación web. Esta aplicación web permite totalmente en línea crear modelos PCA con los datos históricos de edificios. También permite una vez creado un modelo la posterior monitorización de datos nuevos a medida que éstos van llegando. Esta herramienta está integrada como un módulo externo a la plataforma desarrollada en el proyecto europeo HIT2GAP y permite a usuarios no expertos en PCA llevar a cabo las tareas de control y detección utilizando los

modelos previamente creados por usuarios expertos en modelización.

Durante la tesis y como parte de las publicaciones se ha verificado el correcto funcionamiento de la metodología utilizando datos reales provenientes de varios edificios. A modo de resumen, se ha utilizado datos de 3 edificios diferentes (tanto arquitectónicamente como en cuanto a usos y soluciones técnicas que disponen) facilitados por el patronato de la vivienda de Barcelona, los edificios del campus de Montilivi de la Universidad de Girona y el AHU (del Inglés Air Handling Unit, equipo de climatización) del edificio Alice Perry de la Universidad Nacional de Irlanda, Galway. La colaboración con esta Universidad permitió hacer una estancia de 3 meses y solicitar la mención de doctorado internacional para esta tesis.

Chapter 1

Introduction

Nowadays, climate change and global warming are facts, awareness about the problem is growing and governments are starting to move. For example, in the European Union (EU), residential and commercial buildings represent around 40% of all energy use and are responsible for 36% of the EU's total CO₂ emissions. In the face of such evidence, the European Parliament launched the Energy Performance of Buildings Directive (2010) and the Energy Efficiency Directive (2012) [5] as legislative instruments to promote the improvement of the energy performance of buildings within the EU.

Despite these efforts, a gap between the monitoring and real usages still exists in terms of really understanding how energy is being consumed in buildings, identifying major loads and losses, and uncovering the relationships between energy consumption and the activities performed by the users. Enhanced monitoring methods that provide significant information which is useful in terms of understanding energy consumption patterns, are required to perform cost-effective analyses of conservative measures, identify deviations, and help to define and evaluate the new design requirements of buildings.

Energy measuring and monitoring is an essential aspect of understanding energy use. It is necessary to assist energy management activities and to support decision-making based on quantitative and objective information. Thus, actual energy monitoring systems need to be enhanced and evolve towards systems being capable of exploiting the information contained in the variety of data being collected by building management systems (BMS) and other data acquisition systems installed in buildings and facilities, such as weather stations, wireless sensor networks (WSN) or access control systems, among others.

This huge number of data sources and types is also one of the principal challenges in industry's current transformation to the Industry 4.0 paradigm. Integrating,

managing, processing and exploiting data to benefit business is a challenging task. While the Internet of Things (IoT) paradigm provides the infrastructure required for integration and management, data mining provides the background for processing according to the required exploitation goals.

Energy consumption in buildings is related not only to the building's physical characteristics but also to weather, building usages and the interactions among building sub-systems, etc. A simple and easy-to-understand example of such relationships can be found in heating consumption. Heating consumption is related to wall and window insulation, heater efficiency and the distinct elements involved in the heat generation process working properly, but also important correlations with related variables such as outdoor temperature, humidity, sunlight hours and/or the season of the year, among others. Furthermore, occupant behaviour related to routines affect heating demand. Such user-related influencing factors could be, for example, opening and closing windows or blinds, or cooking etc. Correlations with power consumption can also be found with, for example, light bulbs, ovens, fridges, televisions, computers and many other electric devices that produce heat as a residual product of their usage.

Having all these data stored and having techniques capable of gathering correlations among them, will result in models that are more accurate to the reality in buildings. With all this information in the models, detecting deviations in building behaviour should be much easier. Obtaining these models is an affordable task in terms of complexity and time-consuming points of view.

Nowadays, the challenge is to take advantage of multivariate techniques. With this aim, in this thesis a technique used in the batch process industry is adapted to building energy monitoring. Thus, not only does a modelling approach have to gather information about the relationships that exist between consumption and weather or occupancy, but also with regard to dependencies that exist with other variables and factors that can affect energy consumption.

1.1 Hypothesis formulation

Energy consumption in buildings is related not only to the building's physical characteristics but also to weather, building usages and the interactions among building sub-systems, etc. Furthermore, energy consumption in buildings is related to many distinct factors. According to [20], the main factors influencing energy consumption in buildings can be divided into seven categories:

- Climate
- Building characteristics

- User characteristics
- Building services
- Building occupant behaviour and activities
- Social and economic factors
- Indoor environmental quality requirements

Despite energy use is being influenced by so many factors, buildings are not generally monitored taking into account all these factors. Usually buildings are monitored as independent sub-systems (technical parts) without considering any of the interactions between the sub-systems.

The main hypothesis is that using this data to obtain a reference model can improve monitoring strategies significantly adding fault detection and isolation capabilities to smart buildings. Currently, there are many machine learning and statistical methods that can be used for the modelling task. Obtaining these models is an affordable task in terms of complexity and time-consuming points of view. However, not all the techniques are suitable to perform fault detection and diagnosis.

Principal Component Analysis (PCA) is a well-known technique used in the statistical control of industrial processes and it is known for its ability to define models based on correlations. PCA gathers information about the relationships among the variables in a projection space in terms of correlated information. Non-correlated information falls into the residual space. Two statistics (Hotelling's T^2 , and SPE) can be defined in those subspaces. These statistics are easily interpretable to check the adequacy of monitored data to the previously learned normal operation model (fault detection). The method is theoretically sound and allows identifying the variables affecting those detected faults in terms of contributions (fault diagnosis).

However, there is the necessity to adapt PCA data model to the nature of data being gathered in buildings. Buildings present some time repeatability patterns (e.g. daily and weekly) and the same variables are usually measured in different parts of the building resulting in a kind of modularity. This repeatability and modularity add redundancy that can be exploited when models are created giving different monitoring capabilities.

Thus, this hypothesis suggests the traditional PCA method can be extended to consider repeatability and modularity in building data. This requires new pre-processing methods to organise data, according to data redundancy and monitoring goals. This PCA extension will allow to create more precise models and will enhance fault detection and diagnosis of smart buildings.

Statistical	Engineering
Depends on the availability and representativeness of historic data.	Detailed building information, including constructive parameters and materials, is needed.
Resulting models can be biased according to available data.	Accuracy depends on the quality of parameters.
Can model user's behaviour if appropriate variables are included.	Neglects user energy relationships or simply are included as assumptions.
Sample specific (data only represents the site and time where it was acquired).	Data is also required for validation.

Table 1.1: Main characteristics of statistical and engineering techniques

1.2 State of the art

In the state of the art of building energy modelling and monitoring, many methodologies can be found. Building energy modelling methodologies can be categorized by generalizing, ([18]) into two large groups: top-down methodologies and bottom-up methodologies. The starting point for top-down techniques is analysing energy consumption but not attempting to detail causes or end-uses. Instead, these techniques mainly focus on the cause-effect relationship between long persistent changes (in buildings) and consumption. On the other hand, bottom-up approaches are generally based on identifying contributions related to end energy uses in order to build an aggregated energy model. Two distinct strategies can be differentiated in this second group: statistical (or data-driven) and engineering (or based on first principles) approaches (Table 1.1). An extended review of the techniques addressing both approaches for building energy modelling can be found in [8] and [21]. Table 1.1 briefly summarizes the general weaknesses and strengths for each group of techniques according to [8] and [21]. This thesis comes under the data-driven methods group.

Modelling is essential for energy monitoring, since good modelling increases system knowledge and allows the reference model required for any assessment task to be established. Particular cases addressing modelling of large public buildings are studied in [15]. Methodologies to improve the adjustment and calibration of tools to support monitoring are studied in [7] and a solution based on evidence is proposed in [5]. Energy efficiency models for urban environments and buildings are usually calibrated with hourly data [17] and typologies of days and seasons are used to in-

roduce corrections. Recently, modelling improvements have been supported by the deployment of wireless sensors (i.e. [14]). Along that line, the case studies analysed in [16] serve to propose recommendations when monitoring the energy performance of buildings.

Data driven methods are specified for modelling in several scenarios but especially when mathematical models do not exist because they are incomplete or imprecise, or when, due to dimensionality or complexity, it is impossible to apply other techniques. Many data-driven techniques exist within two large groups, namely computation or statistical methods. Statistical methods assume that a probabilistic behaviour exists in the data, and this behaviour is generally assumed to be the normal conditions. The aim of statistical methods is then to detect the variations among these normal conditions.

Principal Component Analysis is a data-driven statistical approach, commonly used for dimension reduction that, for example in [13] it has been used in a dwelling energy data dimension reduction. However, this method can also be extended to modelling multivariate systems, for example, air-handling units ([12]) or chilling plants ([9]). PCA has also been proposed for clustering in the heating evaluation of school buildings ([6]), applied as a feature selection technique in [11] or to analyse seasonal variations in electricity use in office buildings in ([10]).

This general and state of the art on energy modelling and monitoring is completed in the following chapters, with more specific state of the art focused on the achievements proposed in every contribution. Thus, Chapter 3, corresponds to the first contribution of this thesis and presents the PCA and MPCA state of the art for monitoring, Chapter 4 includes an extension of unfolding techniques, so the state of the art included in the chapter analyses alternatives in this field. Finally, Chapter 5 extends the state of the art in the specific field of modelling and monitoring HVAC and air handling units.

1.3 Contributions of this thesis

In this thesis we extend and generalize the uses of PCA to take advantage of all periodicity on uses and possible granularity of buildings to propose a multivariate monitoring method that exploits such characteristics. PCA has been selected because it offers a balance between dimensionality and complexity, thanks to its solid theoretical principles and its adequacy in terms of problem formulation. A particular extension known as Unfold Principal Component Analysis (Unfold PCA) or Multiway Principal Component Analysis (MPCA), commonly used in batch process industries ([19]), has been studied and adapted for modelling and monitoring energy consumption in buildings. This selection is based on the fact that buildings are

usually operated following pseudo-periodic patterns (e.g. daily, weekly patterns) and interest resides in analysing such patterns and dependencies, with variables being monitored during the building operation period, instead of considering only instantaneous relationships. Moreover, the method considers that there are inter-dependencies among multiple entities (i.e. dwellings or buildings), that allows the method be extended to the monitoring of residential buildings composed of multiple dwellings or to an entire neighbourhood.

Adapting PCA allows energy monitoring systems with the following capabilities to be enhanced:

- Description of dependencies among energy consumption and other monitored variables.
- Detection of faults in sensors and reconstruction of sensor readings when data is missing or corrupted.
- Rapid detection of emergent behaviour.
- Forecasting of energy consumption based on independent variables.
- Robust monitoring in the presence of data errors or missing values.
- Creation of simple control charts to monitor multiple variables at a glance.
- Rapid identification and isolation of variables involved in abnormal consumption patterns.
- Modelling of relationships amongst dwellings in residential buildings or neighbourhoods.

1.4 Structure of the document

This document has been structured into eight chapters.

The introduction covers the hypothesis formulation, the state of the art, the general contributions and this short explanation on the structure of this thesis. Following the introduction, Chapter 2 presents the objectives of the thesis, including the tasks required to achieve the objectives and the work executed throughout. As this thesis is presented as a compendium, Chapters 3, 4 and 5 make up the research articles and these are followed by Chapter 6 which presents the main results and discussion of the work presented in each of those articles. Chapter 7 presents the general conclusions of the thesis and future works, this chapter is followed by the bibliography.

Chapter 2

Objective

The main goal in this thesis is to develop a new methodology to improve the exploitation of data gathered in smart buildings for modelling and monitoring energy performance. Thus, by achieving a better understanding of the demand curves of energy consumptions in buildings, an improvement in efficiency can be reached.

2.1 Vision

Currently, building monitoring techniques consider buildings as a set of sub-systems, and monitoring each sub-system without taking into account the other sub-systems, the building structure, usages or user interactions. Monitoring considering buildings as a set of sub-systems implies assuming that these sub-systems are independent, but this assumption is not true. Any building must be considered as a system itself. In buildings many correlations among all of its sub-systems and users exist. For example, electric plug demands can seem independent a priori with climate generation, but this is not always true, plug demand can indicate user presence in a specific room of the building or even heat can be generated as an unwanted product of electric equipment equipment (computers, fridges, light bulbs, etc.) usage.

Another situation where actual monitoring strategies can be improved is when talking about blocks of flats. Dwellings in a block of flats have interactions among them (for example, heat is transferred through walls, even through those well insulated), and while this situation may be obvious, it is not usually taken into account in the monitoring steps. Areas and equipment in common also affect all the dwellings (for instance, corridor temperatures or impulsion and return temperatures in the case of central heating, etc.) and so it makes sense to have a methodology capable of managing all these data sources effectively in order to build accurate models that take such considerations into account.

Another aspect of buildings which is not considered in their monitoring is that there are patterns in the data that are almost always repeated, and which can be classified into two types. The first are the repetitive patterns that result from users habits. For instance, in residential buildings behaviours such as sleeping, working, eating or taking a shower etc., usually show a temporal pattern. Timetables can even be found when describing non-residential buildings as well. The second pattern type is to be found within the structural organization of the building itself. Any building can be divided into floors, rooms or sections, but it can also be grouped into city neighbourhoods. Therefore, developing a methodology capable of dealing with these two repetitive pattern types, will open up the opportunity to perform comparisons between the distinct parts of a building or neighbourhood, or between the days, weeks or seasons in buildings, thus providing the methodology users with valuable information.

2.2 Objectives

The main goal of this thesis is to better understand energy consumption in buildings and provide better modelling and monitoring techniques for buildings so as to improve efficiency.

The general objective in this thesis can be divided into three smaller objectives or tasks that must be achieved separately:

1. Define a data-driven methodology for supervising smart buildings. This method will exploit the redundancy provided by multiple sensors in order to build a reference model from the data gathered during normal operating conditions. This model will then serve as the reference with which to detect abnormal behaviours and faults during monitoring. Once this has been detected, the method will provide isolation capabilities by considering the redundancy in the model.
2. Provide extensions of the previous method to consider periodicity in the consumption patterns. The objective is not only to cope with temporal repetitions (similar daily and weekly profiles, for example) but also to consider possible repetitiveness in the spatial structure of buildings (e.g. multiple apartments, storeys, or rooms with similar monitoring structures)
3. Validate the proposed methodology with different typologies of buildings and subsystems.

These three scientific objectives are completed with a fourth technological objective. This last objective is useful to exploit and validate the methodology.

4. Provide a technological solution for the applicability of the proposed method by implementing the methodology as web services and developing a final web application for the end users.

2.3 Journal Contributions

This thesis is presented as a compendium of articles, the distinct tasks of the general objective and the technological one have been carried out and submitted to three different journal publications.

The first journal paper [4] focuses on sub-objectives one and three. This first work proposes a general methodology, like PCA, to be applied in a building scope. In this first work, a PCA-based methodology is used to build data-driven models capable of exploiting the information contained in the data records being collected by building monitoring systems (BMS) and building energy management systems (BEMS) during normal operational conditions. These models are then exploited as part of the monitoring tasks to evaluate changes in energy consumption behaviour. Thus, the model representing the normal operational conditions of a building can be easily used to detect and diagnose deviations from the modelled behaviour (i.e. faults, over-consumption, efficiency losses, etc.) or to evaluate the effectiveness of energy conservation measures. The work aims to extend the methodology not only to buildings but also to consider communities (residential buildings, social housing buildings and neighbourhoods, for example) by gathering information about possible relationships in the variables being monitored. The energy demands of a central heating system can be influenced by both weather conditions and individual household occupancy, for example. In concordance with objective three, a use case example using residential dwelling data is presented. In this use case, real data from a multi-dwelling residential block located in down-town Barcelona is studied and modelled in two distinct ways (time-wise and entity-wise).

The second journal paper [1] focuses mainly on sub-objectives two and three. This work enlarges and mathematically defines the Unfold-PCA technique to deal with multi-dimensional data. In this paper, a method for correctly exploiting all the potential of granular and/or modular systems by using N-dimensional data arrays is introduced. These N-dimensional data arrays are transformed into the suitable two-dimensional matrices required to perform statistical processing. Here, the focus is on pre-processing data, using a non-unique folding-unfolding algorithm in a way that allows different statistical models to be built in accordance with the monitoring requirements. The fold and unfold method was initially designed as an extension of the Unfold Principal Component Analysis (Unfold-PCA or Multiway PCA), applied to 3D arrays, to deal with N-dimensional matrices. However, this data pre-treatment

method is general enough to be applied to other multivariate monitoring strategies. In concordance with objective three, two examples in the area of energy efficiency illustrate the application of the method for modelling. Both examples show how a unique data set, folded and unfolded in different ways, offers different modelling capabilities. Moreover, one of the examples is extended to exploit real data. In this case, real data collected over a two-year period from a multi-dwelling social-housing building located in down-town Barcelona (Catalonia) has been used.

The third paper, which has been submitted to a first quarter journal, focusses on the last sub-objective, implementing the thesis works as a tool, but also contains a further use case example in accordance with objective three. This third study presents the implementation of the previous work in a service orientated web application. This application was implemented as part of the HIT2GAP European project and integrates the Unfold-PCA methodology into the HIT2GAP core as an FDD module. The application is not only able to build models using the historical data available to the core, but is also able to perform online monitoring and fault isolation for new data using the previously created NOC models. This third work presents a complete example for AHU (Air Handling Units) monitoring in a real use case by using the online data from the Alice Perry building at NUI Galway.

Not included in the compendium, but also related to the objectives of the thesis, two conference publications are presented. These congress papers present the methodology of the thesis being applied to two different scenarios.

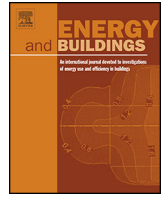
- At the 6th International Conference on Sustainability in Energy and Buildings, the modelling of the energy use of the University of Girona's Montilivi Campus is presented,[3].
- In IEEE EUROCON 2015 a methodology for detecting heating/cooling usage transitions in a social-housing building in Barcelona is presented, [2].

Chapter 3

Multivariate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building

In this chapter, Principal Component Analysis is introduced in the building energy monitoring scope. A use case example over real data is also presented. This publication has been published in the following paper:

Paper published in the **Energy and Buildings**
Volume: 103, Pages: 338-351, Published: June 2015
DOI: [10.1016/j.enbuild.2015.06.069](https://doi.org/10.1016/j.enbuild.2015.06.069)
JCR IF (2015): 2.973
Q1(6/61) - Construction and building technology
Q1(6/126) - Civil engineering



Multivariate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building



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

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

ARTICLE INFO

Article history:

Received 13 March 2015

Received in revised form 26 June 2015

Accepted 29 June 2015

Available online 3 July 2015

Keywords:

Principal Component Analysis

Unfold-PCA

MPCA

BEM

Building energy monitoring

Data mining

ABSTRACT

A complete methodology for energy building monitoring based on Principal Component Analysis (PCA) is proposed. The method extends the Unfolding or Multiway Principal Component Analysis (MPCA) used in statistical batch process control in terms of building and neighbourhood monitoring. Relationships between energy consumption and independent variables such as weather, occupancy or any other variables that are significant for monitoring can be gathered in a model using the proposed methodology. Historic data are used to obtain a reference model that will be used for monitoring. Two unfolding strategies are proposed (time-wise and entity-based) offering complementary views of the building or of the community under consideration. The first, time-wise unfolding, is suitable for detecting behavioural changes over time, whereas entity-wise unfolding allows the identification of entities, e.g. dwellings in a building, that behave substantially differently from others over a period of time. Two simple statistics, T^2 and SPE , are used to define two monitoring charts capable of detecting abnormal behaviours and, furthermore, the isolation of variables that mainly explain such a situation. The paper presents the theoretical background, followed by the methodological principles. The results are illustrated by a case study.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, residential and commercial buildings account for around 40% of final energy use and are responsible for 36% of the European Union's total CO₂ emissions. In order to reduce energy consumption, great efforts are being made to develop and apply European directives [1] that act as an incentive with regard to the efficient use of energy, and the better performance of buildings. However, a gap still exists in the area of energy monitoring in terms of facing the challenge of understanding how energy is being consumed, identifying major loads and losses, and discovering relationships between energy consumption and the activities performed by users. Enhanced monitoring methods, providing significant information that is useful in terms of understanding energy consumption patterns, are required to perform cost-effective analyses of conservative measures, identify irregularities, and help to define and evaluate the new design

requirements of buildings. Energy measuring and monitoring is an essential aspect of understanding energy uses, and assists energy management activity supporting decision making based on quantitative and objective information. Thus, the enhancement of actual energy monitoring systems is necessary, and they have to evolve towards systems capable of exploiting information contained in the variety of data being collected by building management systems (BMS) and other data acquisition systems installed in buildings, facilities and/or neighbourhoods, such as weather stations, wireless sensors networks (WSN) or access control systems, among others.

The purpose of this paper is to propose a general methodology to automatically build data-driven energy models capable of exploiting the information contained in data records being collected with regard to a building (or a neighbourhood) during normal operational conditions, and exploit it as part of the monitoring tasks to evaluate changes in energy consumption behaviour. Thus, a model representing the normal operational conditions of a building can be easily used to detect and diagnose deviations from the modelled behaviour (faults, over-consumption, efficiency losses, etc.) or to evaluate the effectiveness of energy conservation measures. The work aims to extend the methodology not only to buildings but also to consider communities (residential buildings, social buildings and neighbourhoods, for example) by gathering information about

* Corresponding author.

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

possible relationships in the variables being monitored. The energy demands of a central heating system can be influenced by both the weather conditions and by individual household occupancy, for example. The idea is to take advantage of multivariate techniques, as used in the process industry, and adapt them to building energy monitoring. Thus, a modelling approach has to gather information about the relationships that exist between consumption and weather or occupancy, but also with regard to dependencies that exist with other variables and factors that can affect energy consumption. According to [2] the main factors influencing energy consumption in buildings can be divided in seven categories:

- Climate
- Building characteristics
- User characteristics
- Building services
- Building occupant's behaviour and activities
- Social and economic factors
- Indoor environmental quality requirements

The proposed method has to be able to deal with observation vectors describing the variables and factors related to these seven areas. Since the number of variables to be included in the model could be large, a method capable of identifying correlations among variables and, at the same time, compressing redundant information into lower dimensional spaces, is required. Principal Component Analysis (PCA) has been selected to deal with the trade-off between dimensionality and complexity, because of its solid theoretical principles and its adequacy in terms of problem formulation. A particular extension known as Unfold Principal Component Analysis (Unfold PCA) or Multiway Principal Component Analysis (MPCA), commonly used in batch process industries [3], has been studied and adapted for modelling and monitoring energy consumption in buildings. This selection is based on the fact that buildings usually are operated following pseudo-periodic patterns (e.g. daily, weekly patterns) and interest resides in analysing such patterns and dependencies, with variables being monitored during the building operation period, instead of considering only instantaneous relationships. Moreover, the method considers the existence of interdependencies among multiple entities (i.e. dwellings or buildings), that allows the extension of the method to the monitoring of residential buildings composed of multiple dwellings or to a neighbourhood.

PCA is a projection technique that allows the representation of dependencies among variables in a lower dimension space defined by orthogonal components. Thus, the method gathers information about relationships among variables in this projection space in terms of correlated information, whereas non-correlated information falls in the residual space. Two statistics – Hotelling's T^2 , and SPE (square prediction error) – defined in those subspaces in the form of projection and residual spaces, respectively, are used to model the adequacy of the new data with respect to the PCA model (obtained from historical data). Thus, large deviations of these statistics (over a statistical threshold) are used to detect emergent behaviours when monitoring new observations. Moreover, when this happens, the proposed methodology allows the identification of the variables responsible for such large variation by applying contribution analysis. The adaptation of these statistical principles allows the enhancement of energy monitoring systems with the following capabilities:

- Description of dependencies among energy consumption and other monitored variables.
- Detection of faults in sensors and the reconstruction of sensor readings when data are missing or corrupted.
- Rapid detection of emergent behaviour.

- Forecasting of energy consumption based on independent variables.
- Robust monitoring, in the presence of data errors or missing values.
- Creation of simple control charts to monitor multiple variables at a glance.
- Rapid identification and isolation of variables involved in abnormal consumption patterns.
- Modelling of relationships among dwellings in residential buildings or neighbourhoods.

In the next section, a review of the state of the art introduces the interest in this field. The paper follows with a description of the PCA background in terms of modelling and monitoring, including fault detection and diagnosis capabilities. Then, the Unfold-PCA extension is considered, emphasizing how different unfolding strategies can be used to offer different monitoring views of a building or of a neighbourhood. Finally, a case study focusing on energy monitoring of social buildings is presented to illustrate the benefits of the approach.

2. Related works

This section illustrates the interest in data-driven modelling, and how these paradigms can be used to extend energy monitoring capabilities. The section does not pretend to cover all the methods described in the literature, but focuses on those that contributed to a definition of new energy monitoring paradigms for buildings. Building energy modelling methodologies can be categorized [4] in terms of two big groups: top-down and bottom-up methodologies. Top-down modelling techniques start from the analysis of energy consumption and do not try to detail causes or end-uses. They mainly focus on the cause-effect relationship between long persistent changes (in buildings or neighbourhoods) and with regard to consumption. On the other hand, bottom-up approaches are generally based on the identification of contributions related to energy end-uses in order to build an aggregated energy model. Two distinct strategies can be differentiated in this second group: statistical (or data-driven) and engineering (or based on first principles) approaches. An extended review of techniques addressing both approaches for building energy modelling can be found in [5] and [6]. Table 1 briefly summarizes the general weaknesses and strengths of each group of techniques.

Modelling is essential for energy monitoring, since it increases system knowledge and allows the establishing of the reference model required for any assessment task. Particular cases addressing the monitoring of large public buildings are studied in [7]. Methodologies to improve the adjustment and calibration of tools to support monitoring are studied in [8] and a solution based on evidences is proposed in [1]. Energy efficiency models for urban environments and buildings are usually calibrated with hourly data [9] and typologies of days and seasons are used to introduce corrections. Recently, modelling improvements have been supported by the deployment of wireless sensors (i.e. [10]). In this context, case studies analysed in [11] have served to offer recommendations when monitoring energy performance in buildings. On the other hand, the use of Principal Component Analysis (PCA) is not new in this area; however the way in which it is used differs substantially from the method proposed in this paper. Thus, PCA has been proposed simply as a reduction technique in [12] dealing with dwelling energy data, air-handling units [13], chilling plants [14], or for human activities' outlier detection [15]. PCA has also been proposed for clustering in the heating evaluation of school buildings [16], applied as a feature selection technique in [17] or to analyse seasonal variations in electricity use in office buildings [18].

Table 1
Main characteristics of statistical and engineering techniques.

Statistical	Engineering
Depends on the availability and representativeness of historic data	Detailed building information, including constructive parameters and materials, is needed
Resulting models can be biased according to available data	Accuracy depends on quality of parameters
Can model users behaviour if appropriate variables are included	Neglects user energy relationships or simply are included as assumptions
Sample specific (data only represent the site and time where it was acquired)	Data are also required for validation

Also, the authors of this paper prove the benefits of PCA in terms of analysing the energy profiles of eight buildings in a university campus [19]. In this work we extend and generalize these uses to take advantage of the ability of PCA in modelling the variables involved in residential buildings by considering influences between variables and between apartments.

3. Background

In this section the theoretical background necessary to create a PCA-based model and to exploit it for monitoring activities is presented. When variables being monitored include energy consumption (i.e. electricity, gas, heating, etc.) and other variables that presumably are related to them (i.e. indoor and outdoor temperature, occupancy, etc.), the models obtained using this methodology are able to capture the relationships between them. Multivariate statistical models obtained with the use of historical data during normal operating conditions (NOC) are used as reference models. Then, the exploitation of these models for monitoring purposes is proposed. Two statistics (T^2 and SPE) are used to check the adequacy of new observations with regard to this reference model. A graphical representation of these statistics, named SPE and T^2 control charts, is enough to enhance monitoring systems with the introduction of mechanisms to detect emerging behaviours (sensor fault detections, leakages, overconsumption or other irregularities) and to isolate the variables that are most commonly related to such abnormal operating conditions.

3.1. Principal Component Analysis modelling

Principal Component Analysis (PCA) is a multivariate statistical technique that allows the identification of correlations among variables represented by a set of observations stored in a matrix. Usually data describing normal operating conditions are organized in a bi-dimensional matrix, where columns represent the variables (m) and rows the observations (n) at a given time instant with regard to these variables. Thus, a single observation is a row vector of length m . Data in this matrix, X , are pre-treated to have a zero mean and unit variance (standardization), giving the same importance to all the variables in terms of variability. PCA aims to model correlations among variables and to transform them into a set of linearly uncorrelated variables called principal components. The transformation matrix, P , is obtained by applying a singular value decomposition of the covariance matrix of X , and selecting the largest r eigenvalues. The eigenvalues of such a decomposition correspond to the variance captured with regard to each principal component. Eigenvectors (columns of P) define the orientation of these principal components, ordered according to the decreasing order of eigenvalues (variance explained in each direction) and they are orthonormal (linearly uncorrelated) [20]. Fig. 1 represents this transformation with a three dimensional ($m=3$) data set. An observation $X_j=(x_1, x_2, x_3)$ (on the right) is transformed into a principal component space defined with $r=2$ components $T_j=(t_1, t_2)$, on the right (the third component is discarded because it represents a

small amount of variance). Thus, the original data matrix, X , can be rewritten in the following way [21]:

$$X = TP^T + \tilde{X} \quad (1)$$

where T is known as the ‘score’ matrix and contains the projection of the n observations (X) in the principal component subspace, or projection subspace (r components), and is defined by the transformation matrix P . The residual or error matrix $\tilde{X}(n \times m) = X - TP^T$ represents the projection error for each observation, due to the fact that only the first r components have been used instead of m .

Based on this subspace decomposition, two indices or statistics can be used to define the quality of an observation with respect to the model: T^2 (Hotelling’s T^2 , in the projection space) and SPE (square prediction error, in the residual space). The former represents a kind of square distance (Mahalanobis’ distance) between an observation and the centre of the model measured in the projection space (see Fig. 2 left). The latter is related to the square distance of the observation to the projection hyperplane, and consequently is an indicator of how the observation fits the model structure (see Fig. 2 right).

Hotelling’s T^2 and square prediction error can be computed for a single observation, x , using the following expressions [21]:

$$T^2 = \sum_{i=1}^r \frac{t_i^2}{\lambda_i} \quad (2)$$

$$SPE = \sum_{i=1}^m (x_i - \hat{x}_i)^2 \quad (3)$$

T^2 and SPE, respectively, are the values of those statistics corresponding to the observation x (with m components, x_i , $i=1 \dots m$); t_i is the i th component of the score vector t corresponding to the observation x , and λ_i represents its associated eigenvalue.

3.2. PCA monitoring

Monitoring consists of the continuous comparison of acquired variables with respect to the pre-defined normal operating conditions of a system. The strategy proposed in this work consists of obtaining this reference model based on PCA methodology using observations (m -dimensional) collected during the normal operating conditions (NOC) of a building. Thus, the model is being represented by the loading matrix P obtained in such normal operating conditions, and T^2 and SPE statistics of a given observation will represent the fitness of this observation to the model. Thus, for any new observation, x_{new} , is easy to check its adequacy in terms of the model by simply projecting it, using the loading matrix (P) obtained during normal operating conditions ($t_{new}=x_{new}P$), computing the associated statistics (T_{new}^2 and SPE_{new}), and comparing them with appropriate thresholds defined during normal operating conditions.

Eq. (4) can be used to compute the T^2 threshold when NOC data are used to compute the model.

$$T_{r,n,\alpha}^2 = \frac{r(n-1)}{n-r} F_{r,n-r,\alpha} \quad (4)$$

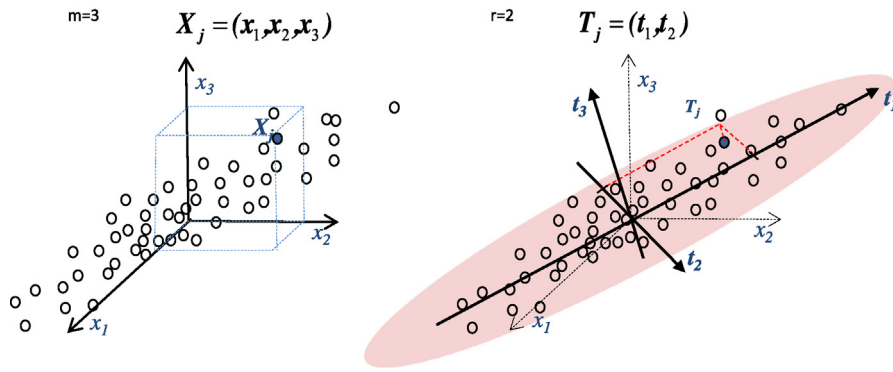


Fig. 1. Data representation in the original space (left) and in the projection subspace (right).

where n is the number of observations used to build the PCA model, r is the number of principal components retained in the model, F is Fisher distribution function and α the desired confidence level.

For the SPE threshold, as was proposed by [22], Eq. (5) can be used.

$$SPE_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \quad (5)$$

where c_α is the standard normal deviate corresponding to the upper $(1 - \alpha)$ percentile. θ_i can be computed using Eq. (6) and h_0 can be computed using Eq. (7).

$$\theta_i = \sum_{j=r+1}^m \lambda_j^i \quad \text{for } i = 1, 2, 3 \quad (6)$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \quad (7)$$

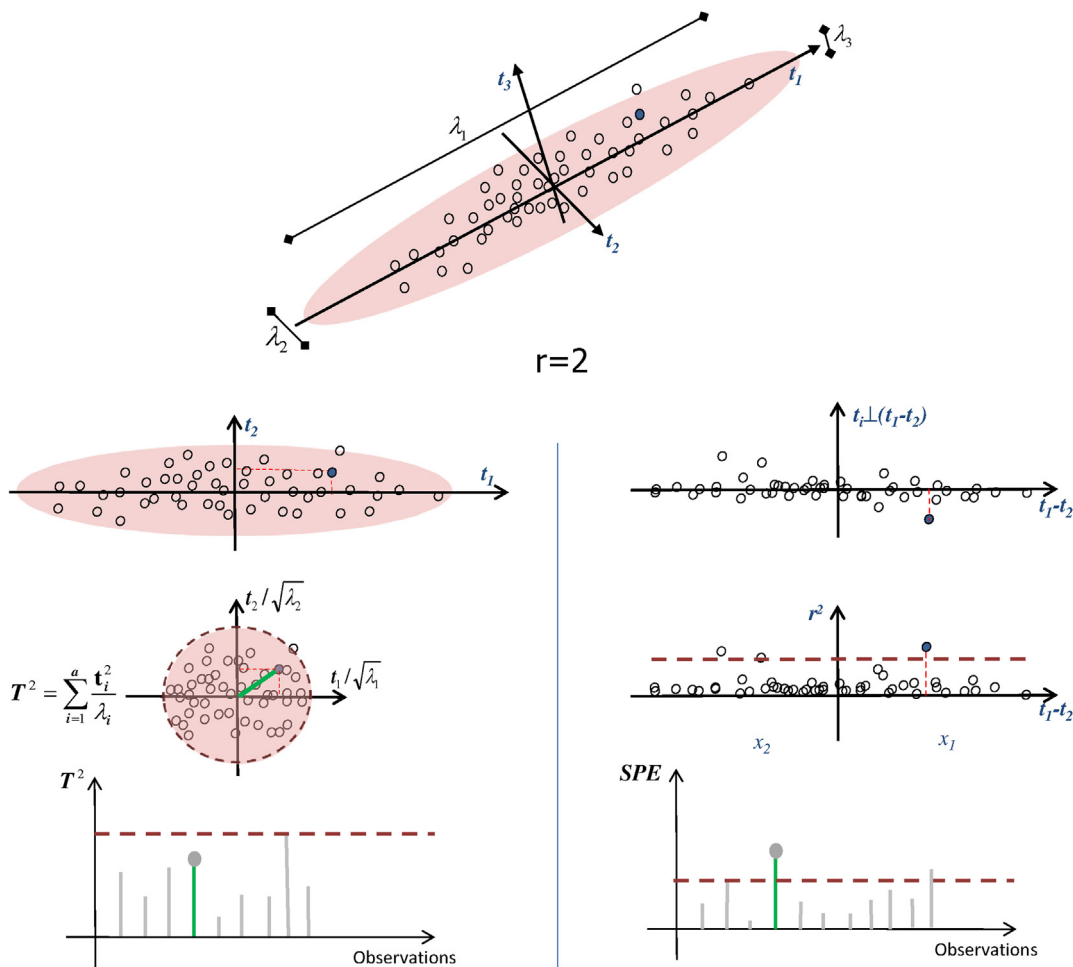


Fig. 2. Visual representation of Hotelling's T^2 (left) and SPE (right).

where r is the number of principal components retained in the model, m is the total number of principal components or original variables, and λ_j corresponds to the j eigenvalues not retained in the PCA model.

Overpassing these thresholds (one or both of them) will provide evidence of a deviation with respect to the NOC model. Thus, simple control charts (plots representing T^2 and SPE for each new observation) based on these two indices can be used to monitor building behaviour and detect deviations from the model. New observations, projected with the P matrix, obtained during NOC, are represented with the corresponding T^2 and SPE indicators and compared with respect to the NOC thresholds. Values larger than the thresholds (one or both of them) provide evidence of a deviation in terms of the system behaviour with respect to NOC (see Figs. 7 and 13 as a T^2 chart example, and for a SPE chart example see Figs. 8 and 14).

3.3. Diagnosis based on contributions

The second step in monitoring is associated with the isolation of the origin of the deviation by identifying (isolation and quantification) which variables in the observation vector explain this deviation from the expected situation. This is achieved by analysing the contribution of the original variables to the statistics that detected the deviations (either T^2 or SPE). The decomposition of these statistics into contributions of the variables in the original space can be computed according to [23]. The contribution of the variable number j of an observation x to its T^2 can be calculated according to the following expression derived from Eq. (2):

$$CONT^{T^2}(x) = \sum_{i=1}^r \frac{t_i}{\lambda_i} (p_{i,j} x_j) \quad (8)$$

where t_i is the i th component of the score vector t corresponding to the observation x and λ_i represents its associated eigenvalue, $p_{i,j}$ is the loading corresponding to the variable number j and x_j its observation. In addition, the contributions of each original variable to the SPE statistic are obtained simply by subtracting the components of the projection and the residual spaces:

$$CONT_j^{SPE}(x) = (x_j - \hat{x}_j)^2 \quad (9)$$

with \hat{x} being the back projection of a single observation x to the original space ($\hat{x} = tP^T$) and x_j and \hat{x}_j are the respective components. Thus, when an observation is out of control (overpasses a predefined threshold), the original variables that mainly contribute to this situation can be identified by simply analysing the magnitude of the contributions to the associated statistic. Causes of such deviation can be diverse and embrace sensor faults, disturbances or changes in the building operational conditions (occupancy, modifications, weather conditions, etc.) among others (see Figs. 9 and 15 for a T^2 contribution chart example).

4. Multi-view monitoring: Unfold-PCA

The previously described method is useful for monitoring buildings and infrastructures as a whole. The method allows the monitoring of a large number of variables, embracing not only consumption, but other variables from the BEMS/BMS or other external sources, e.g. sensor networks, weather stations, etc. However, an extension is required to deal with large buildings (e.g. malls, hotels, housing buildings, offices, etc.) or communities (e.g. neighbourhoods, residential districts, industrial or business parks, etc.) that incorporate multiple entities, that may be affected by interactions among them, and that require both an overall monitoring of the infrastructure and an individualized monitoring of each sub-entity.

With this aim, an extension of the previous method is proposed under the assumption that every sub-entity is being monitored

with the same subset of J variables (e.g. energy consumption, occupancy, indoor temperature, etc.). Assuming that the L entities are being monitored, and that other N variables (e.g. weather stations, overall consumption, energy production, etc.) can be of interest for the whole facility, the total number of variables involved in the monitoring project is $N + LJ$. Additionally, the method allows considering differences between entities (e.g. surface, usage, activity, etc.) by defining each one by a subset of M parameters.

Thus, the data model for a building consisting in L entities will be represented by three matrices: a three dimensional matrix, ($L \times J \times K$), containing information from the J variables of the L entities during K time instants, a second matrix ($K \times N$) with information from the N global variables, and a third matrix containing M fixed attributes for each entity ($L \times M$). This basic data model is represented in Fig. 3a). The matrix lengths L , M , N and J are invariant during the whole process, whereas the time dimension K is expected to vary, depending if we are in the modelling or monitoring step. During the modelling step, historical data are used, and consequently a large data set is expected, whereas during monitoring, a single observation ($K = 1$) will be evaluated at each time instant.

This data model has the complexity of dealing with a 3D data matrix, but at the same time it offers the possibility to study data correlation in different directions. The unfold-PCA, or Multi-way Principal Component (MPCA) approach proposed in [21] for monitoring batch processes, shows how historical data of several executions of a batch organized in a 3D matrix defined by variables, time and batches, can be unfolded into a 2D matrix in order to consider correlations among process variables at different time instants (batch wise unfolding). Following a similar principle, adapted to the data model we have presented, two interesting unfoldings of the 3D matrix can be followed for building monitoring:

- *Entity-wise unfolding*, consists of reorganizing the 3D matrix containing monitored variables of entities into a 2D matrix, where each row represents a single entity and columns contain information with regard to variables during the whole observation period (JK columns), resulting in a $L \times JK$ matrix (see Fig. 3). This unfolding is useful when it comes to capturing dependencies among variables over time that are common for all entities. Consequently it is also appropriate to identify those entities that behave significantly differently during the observation period K . When additional information, defined by a list of M parameters (e.g. entity dimensions, socio-economic data of householders, etc.), is available for each entity ($L \times M$ matrix), this can be included in the model by appending it to the unfolded matrix, resulting in an $L \times (JK + M)$ matrix as depicted in Fig. 3c.
- *Time-wise unfolding*, similarly to the previous one, the 3D matrix can be unfolded into a 2D matrix by aligning information from variables of all the entities in a single row, preserving the temporal reference. This unfolding results in a $K \times LJ$ matrix with as many rows as observation time instants, K , and as many columns as the number of variables times the number of entities (LJ) (see Fig. 3). This unfolding is useful when it comes to monitoring the community as a whole, allowing modelling dependencies among variables within an entity and from different entities as well. In case of additional variables reporting general information (weather stations, general consumption, power generation, etc.) organized in a $K \times N$ matrix, this information can be appended to the model, resulting in a $K \times (LJ + M)$ 2D matrix as represented in Fig. 3b).

We can observe that both unfolding procedures result in a 2D matrix. Therefore, the general methodology presented in the previous section can now be applied with regard to modelling

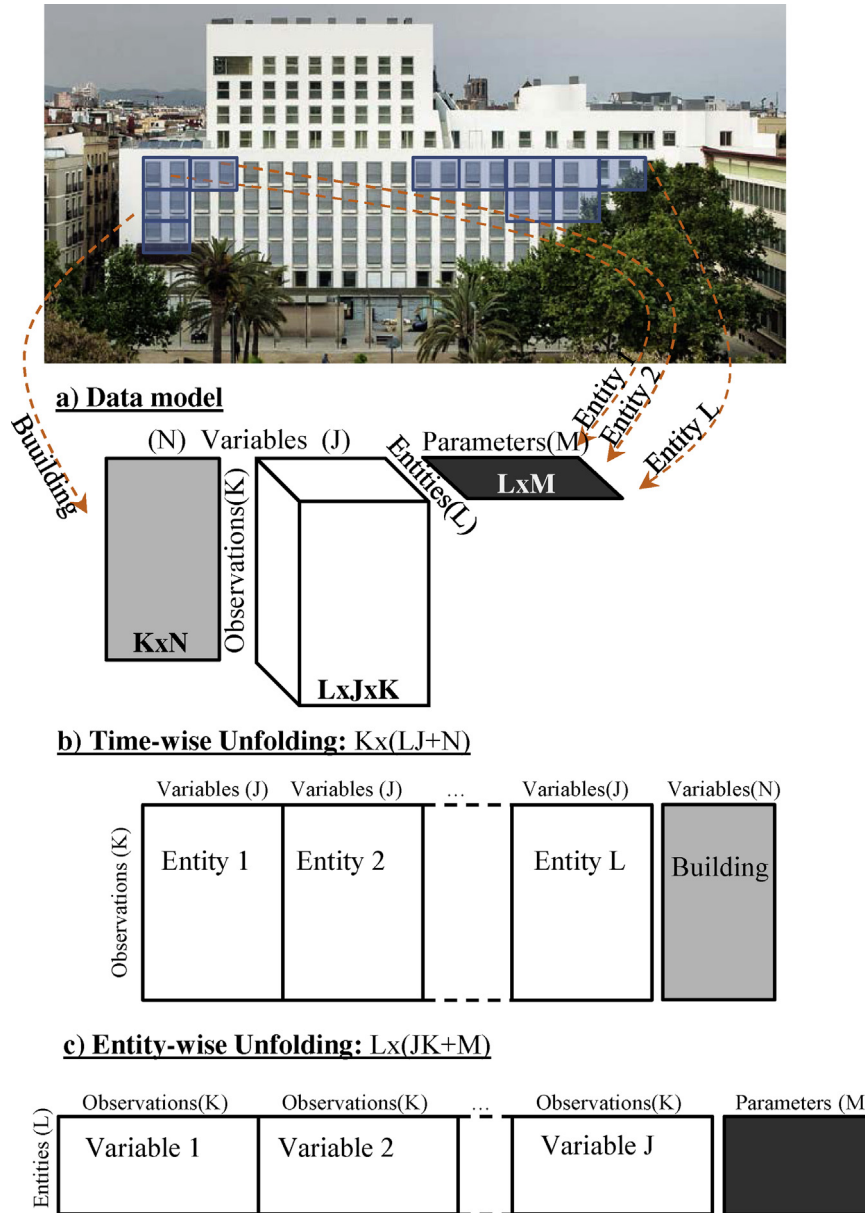


Fig. 3. Data model and 3D matrix unfolding: (a) data model, (b) time-wise unfolding, (c) entity-wise unfolding.

and monitoring. Thus, historical data organized according to the time-wise unfolding 2D matrix can be used for modelling the building when we are interested in continuous monitoring over time, whereas entity-wise data organization will be used to capture the common behaviour of entities and, at the same time, identify those that behave differently during an observation time window.

In the following section the benefits of this multi-view approach are highlighted in terms of a real scenario. This consists of a social housing building with $L=96$ entities where $J=3$ variables (heating energy, hot water volume and hot water energy) have been measured during 636 days. This data set is completed with information from $N=11$ variables from a weather station, gathered during the same period of time. A winter model has been created with $K=201$ observations using the $201 \times (96 \times 3 + 11) = 201 \times 299$ time-wise unfolded matrix and involving the application of PCA method outlined in Section 3. This model has been used as reference for continuous monitoring by successively projecting new observations described by a 1×299 vector. A second view of the same building

is given as a result of entity-wise unfolding. The whole historic data have been used in this case, resulting in a $96 \times (3 \times 636) = 96 \times 1908$ matrix. Since entities are quite similar (similar surface, tenants, and age) no additional attributes have been appended ($M=0$). The resulting model has been used to identify those dwellings that presented significantly different behaviour during the observation period.

5. Case study: multi-view monitoring of a social housing building

5.1. Motivation and end-user benefits

This case study is motivated by the need to develop enhanced monitoring tools capable of automatically detecting irregular consumption patterns, and providing accurate information with respect to the variables involved. Energy managers, or energy service companies (usually in charge of many buildings) do not

have enough time to analyse energy consumption in terms of looking for irregularities. Instead, building energy monitoring systems (BEMS) should include these capabilities to automatically identify and report such situations.

The energy system of a social housing building has been chosen to exemplify the methodology and to highlight the benefits of the method. The building has the particularity of having a central heating system that is used to provide both hot water and heating to common spaces and to 96 dwellings, each of which is rented by one or two tenants. The operation and maintenance of the energy system are outsourced to an energy service company who has to guarantee the quality of supply and comfort in the dwellings and, at the same time, to ensure the most energy efficient conditions. This implies a continuous monitoring of the system to detect deviations and to evaluate the effect of energy conservation measures.

The multi-view approach presented in the previous section (time-wise and entity-wise unfolding) has been implemented in this real scenario, showing the benefits derived from the direct application of the methodology in the domain. In particular, from the perspective of the energy manager, the time-wise unfolding view has been demonstrated to be effective in terms of automatically detecting behavioural changes, such as alerting the manager when an apartment is empty or occupied, detecting sensor or metering faults, detecting water over-consumption (possibly produced by leakages or abnormal occupancy), among others. On the other hand, the entity-wise view has been demonstrated to be more informative for people who deal with tenants, and in terms of their social condition, since it allows an analysis of the whole facility as a set of dwellings, instead as a single entity, and allows the manager to identify those dwellings that behave significantly differently during specific observation periods.

5.2. Scenario description

The social housing building used for the test is located in down-town Barcelona (Catalonia). It has a centralized hot water production system for 96 dwellings (1 or 2 occupants each), a car park, and common areas. Hot water production is provided by two Remeha GAS 310 Kw units combined with several thermal solar panels and supplies both hot water and heating to dwellings. Hot water radiators are used for heating, and each tenant controls the comfort temperature with a thermostat. Thus, three variables are being monitored for each dwelling (heating energy (kWh), hot water volume (l) and hot water energy (kWh)). Daily data gathered between 2012 and 2014 (March) have been made available for the study.

Fig. 4 represents the evolution of these three variables for a single dwelling during the whole period. The solid line corresponds to the daily consumption, and the shadowed areas represent three times the standard deviation around the daily mean value (represented by zero), computed with values from all the apartments.

Fig. 5 represents the value of the same variables in a day for the whole set of dwellings. The solid line represents variables, whereas shadowed area correspond to 3 standard deviation area around the mean value (represented by the zero reference) with regard to each variable.

Data have been cleansed (days with empty registers were deleted), resulting in a data set of 636 days, standardized (zero mean and unit variance) and formatted into the $L \times J \times K$ 3D matrix (Fig. 3) containing $J=3$ variables of $L=96$ entities during $K=636$ time instants, one for each day, resulting in a $(96 \times 3 \times 636)$ matrix. Additionally, weather information in the period was provided by the Catalan public weather agency (MeteoCat) and consists of $N=11$ variables, summarized in Table 2, that has been organized into a (636×11) matrix.

Table 2
Weather variables summary.

TM (°C)	Mean daily temperature
TX (°C)	Maximum daily temperature
TN (°C)	Minimum daily temperature
PPT24h (mm)	Daily precipitation
HRM (%)	Mean daily humidity
RS24h (MJ/m ²)	Global irradiation
VVM10 (m/s)	Mean daily wind velocity
DVM10 (°C)	Mean daily wind direction
VVX10 (m/s)	Maximum daily wind speed
DVX10 (°C)	Maximum daily wind speed direction
PM (h Pa)	Mean daily atmospheric pressure

5.3. Monitoring based on time-wise unfolding

In this first scenario, time-wise unfolding of the 3D matrix will be used to create the building energy model. This approach will be useful in terms of finding relationships between monitored variables, detecting sensor faults, and errors, or reconstructing missing values.

5.3.1. Data matrix construction

The original 3D data matrix $(96 \times 3 \times 636)$ is time-wise unfolded, resulting in a (636×288) matrix, where rows represent time instants (days) and columns the variables of every dwelling. Then, the matrix is completed by appending weather variables (636×11) matrix, resulting in a final data matrix, X , of dimensions 636×299 .

5.3.2. Model construction

A simple visual inspection of the variables evidences differences between the winter and summer seasons. The heating energy variable shown in Fig. 4 indicates that in summer it falls to zero, whereas in winter there is a large variability. Since the objective of the model is to gather relationships between variables, we decided to create specific models for winter and summer. For simplicity only the winter model will be described, but the same procedure would be replicated for the summer model. Thus, 201 winter days between 1st January 2012 and 28th February 2013 were selected to create the winter model, resulting in a (201×299) matrix instead of the original one. The procedure described in Section 3.1 has been followed to obtain the loading matrix, P (only $r=6$ principal components have been retained, explaining a 80.52% of total variance). Thresholds for both statistics, T^2 and SPE , have been computed according to Eqs. (4) and (5), respectively, with a confidence factor $\alpha = 95\%$. Projection, in terms of the 3 first principal components (percentiles in brackets for each component indicate the variance captured in each direction) of the 201 observations used to create the winter model are depicted in Fig. 6 as dots. Every dot represents a day, and the centre of the grey-shadowed ellipsoid corresponds to the mean daily behaviour. Distance (in fact, the weighted distance) of dots with respect to this centre is measured with T^2 (only 3 out of 6 components are represented).

5.3.3. Control charts: monitoring and detection of emergent behaviour

Statistics T^2 and SPE , computed with Eqs. (2) and (3), respectively, are used for monitoring. The respective control charts with the corresponding thresholds previously computed (horizontal solid line) are depicted in Figs. 7 and 8, respectively. Points near zero in the T^2 chart represent days with measurements close to the centre of the model (normal behaviour), whereas large values of T^2 (over the threshold) correspond to days with measurements indicating values away from the average. Thus, values larger than the threshold can be considered as indicating suspicions of a change in the system's behaviour.

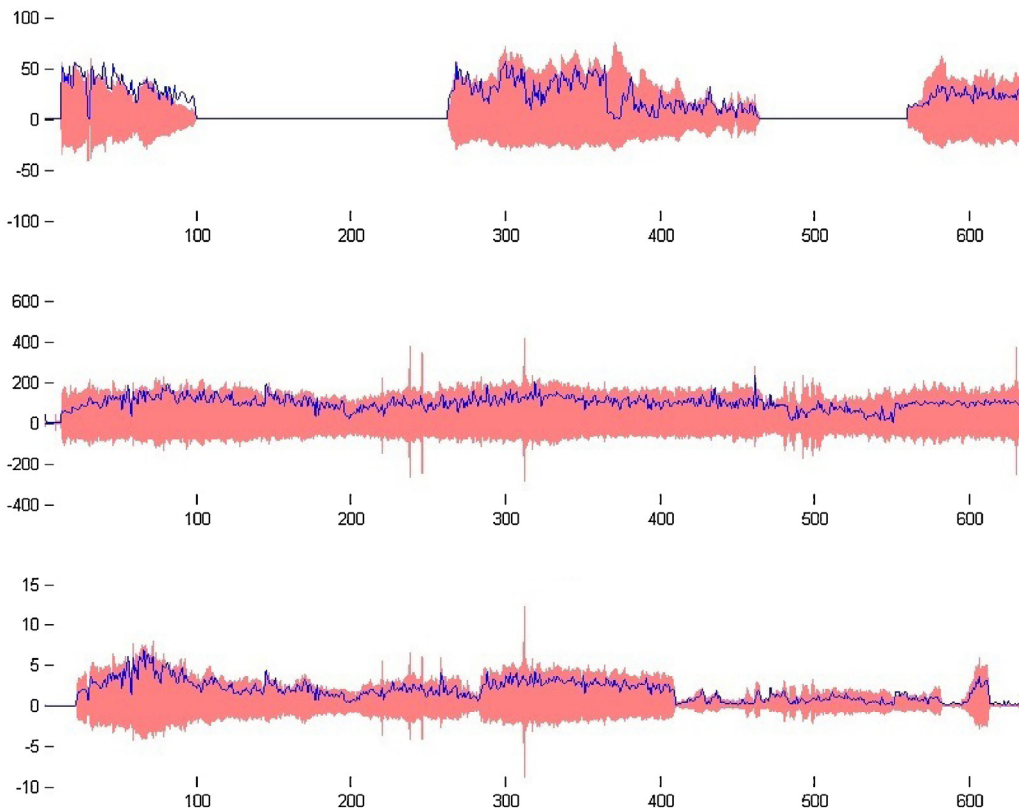


Fig. 4. Monitored variables in a dwelling: heating energy (kWh) (top), hot water volume (L) (middle), and hot water energy (kWh) (bottom).

Fig. 7 represents the evolution of T^2 for all winter data, with the first 201 dots corresponding to the projection of the days used to build the model, whereas crosses from 202 to the end (367) represent new winter data (gathered between 2013 and 2014), that are not used in the construction of the model. From the picture we can observe that the general behaviour in 2012 follows the model, with only few of the measures falling slightly over the threshold, including a single one that presents a large value (observation number 93, 30th October 2012) over 100, what is consistent

with the 0.95 confidence factor. On the other hand, the projection of new observations (from 202 onwards) present a couple of groups of observations that move away from the threshold. The larger of them has been marked for further analysis (11th of March 2013).

Similarly, *SPE* chart allows monitoring of how far the observation is from the projection space (*SPE* is computed with principal components not retained in the model – components from $r + 1$ to m). Thus, large values of *SPE* mean indicate that the observation

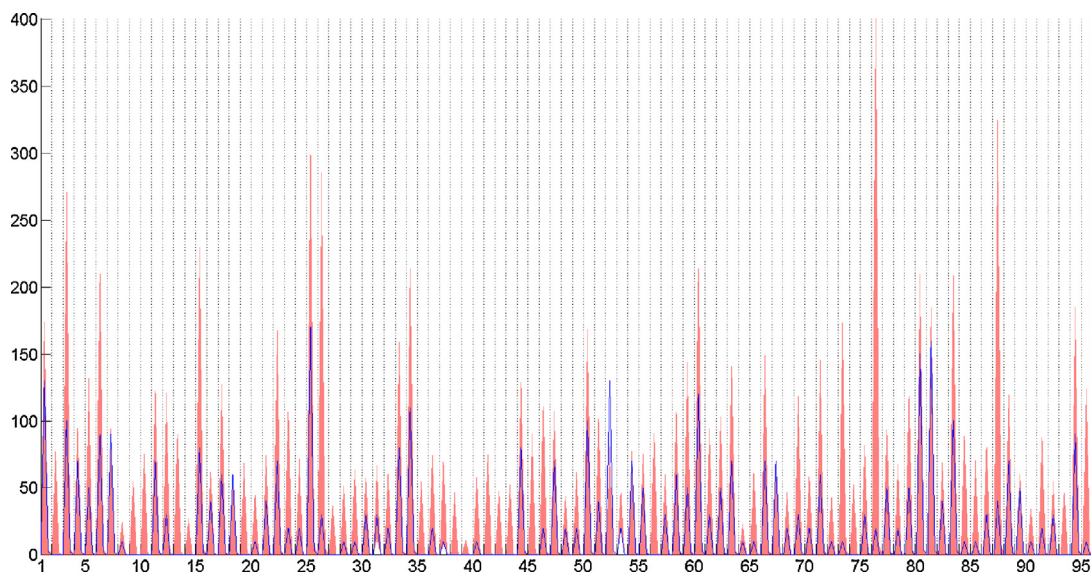


Fig. 5. One day values of variables for all the dwellings. From left to right heating energy (kWh), hot water volume (L) and hot water energy (kWh), successively repeated for the 96 dwellings.

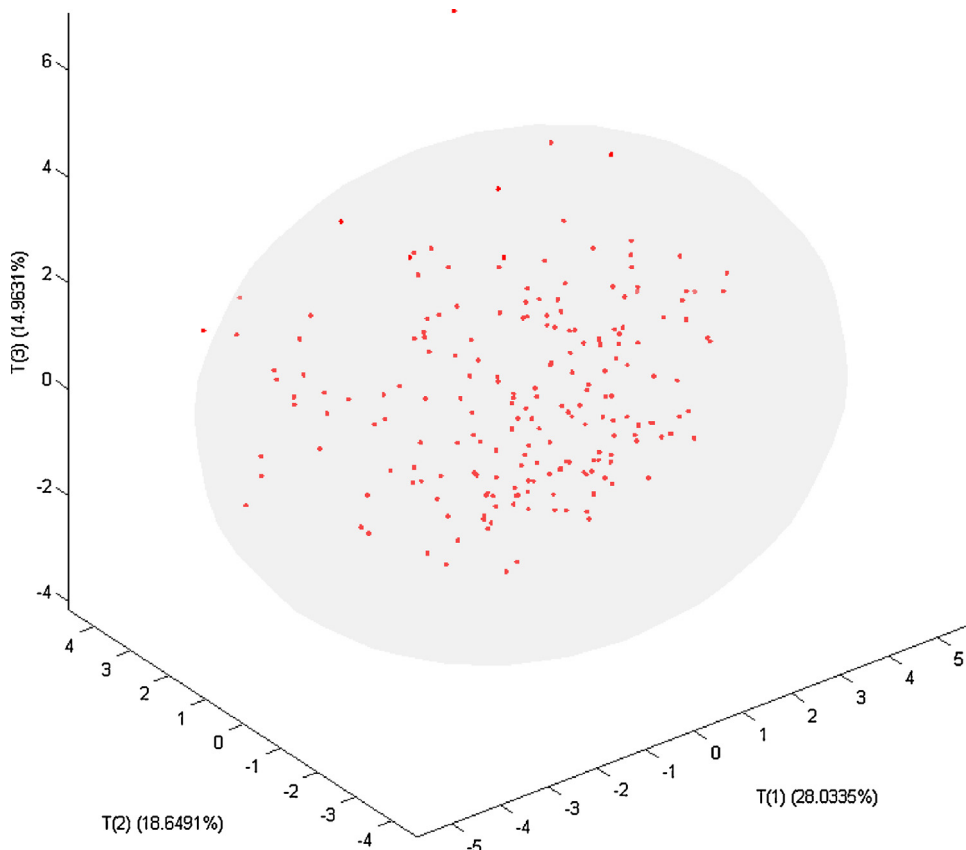


Fig. 6. Scores of the first three principal components: every dot represents a single day.

does not fit the model, and values larger than the statistical threshold will evidence structural changes in the system (i.e. faults). Fig. 8 represents the evolution of *SPE* for winter data (model and new data). Large values of *SPE* indicate dates on which correlation between the variables did not follow the model. This happens, for example, when dwellings become empty (with no consumption, causing *SPE* to fall out the limit) or as a result of sensor faults, for example. From Fig. 8 we can observe that new data (from 202 onwards) present several groups of observations that exceed the threshold (two of them, 11th March 2013 and 26th

October 2013, have been marked in the figure for further analysis). Observe that the first one (observation number 212, 11th March 2013) also presented a large value of T^2 (Fig. 9), which suggests the need for some special situation that is discussed in the next subsection.

5.3.4. Contribution analysis: explaining deviations

When a deviation is detected (large deviation in T^2 or *SPE* charts), contribution analysis (Section 3.3) can help to interpret the origin of such a situation with respect to the magnitudes of the measured

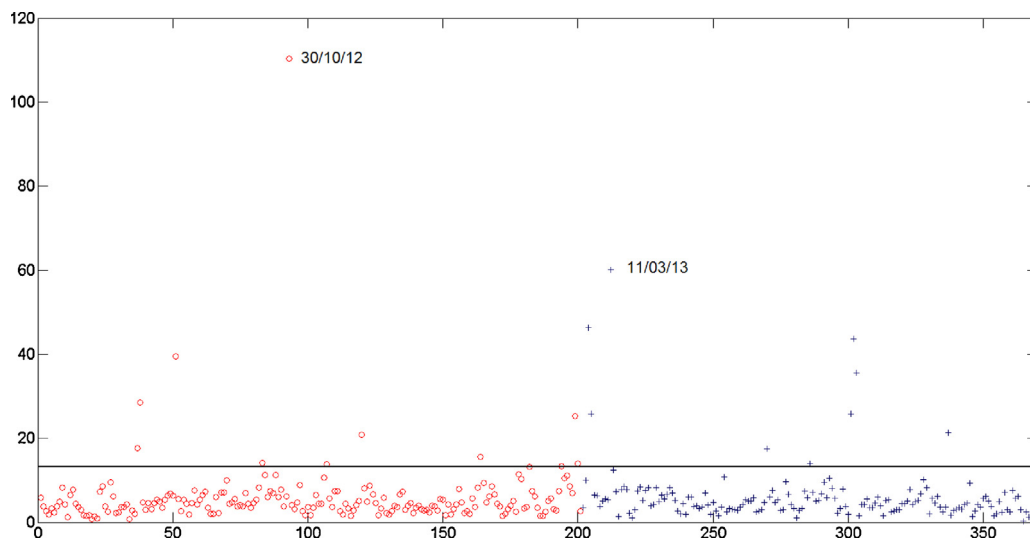


Fig. 7. T^2 chart: dots – on the left – represent days used in constructing the model; crosses – on the right – represent new observations.

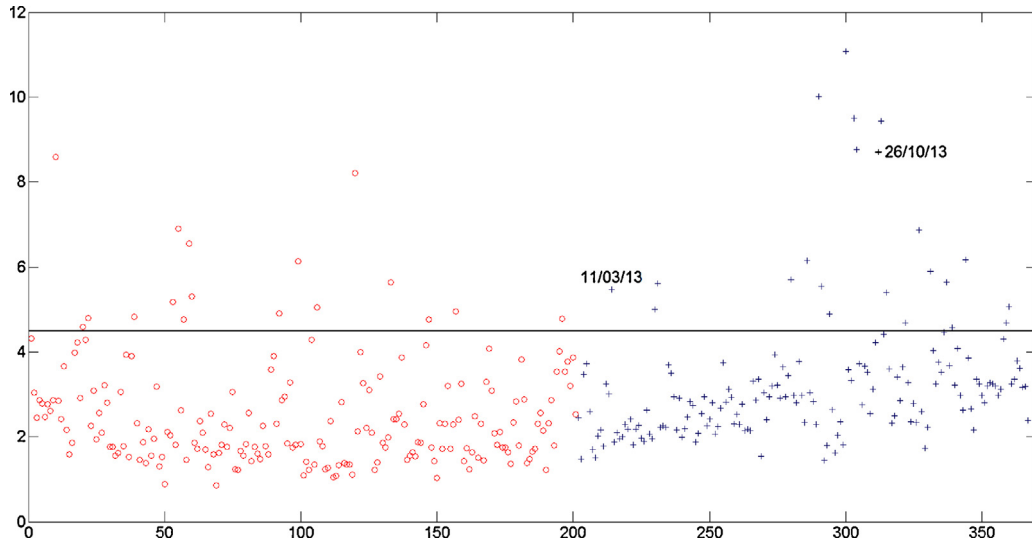


Fig. 8. SPE chart: dots – on the left – represent days used in constructing the model; crosses – on the right – represent new observations being monitored.

variables. Eqs. (8) and (9) have been used to compute T^2 contributions and SPE contributions, respectively. In order to facilitate the explanation, contributions obtained with such expressions are presented graphically. Thus, Fig. 9 represents T^2 contributions for the date 11th March 2013, and Figs. 10 and 11 correspond to SPE contributions for the dates 11th March 2013 and 26th October 2013.

The magnitude of the contributions of the original variables to the statistics (T^2 or SPE) are represented by solid bars, whereas the thin solid line indicates the three sigma margin of contributions computed with observations included in the model. The variables in the pictures are organized as follows: from left to right, heating energy, hot water volume and hot water energy, successively repeated for the 96 dwellings, and on the right the 11 weather variables.

In Fig. 9 we can see that the variables causing the T^2 out of control associated with the date 11th March 2013 are (marked on top with a cross) mainly the 5th weather variable (humidity), and heating energy for dwellings 19, 28, 40, 52, 62 and 73. The hot water volume and hot water energy for dwelling 45 also present a high contribution for that date. As this observation presents a deviation for SPE

the associated contributions have been analysed (Fig. 10). We can see that the variables causing the SPE deviation are mainly two variables corresponding to hot water volume and hot water energy for dwelling 45, and in a minor contribution to heating energy for some dwellings (7, 16, 21, 43, 46, 55, 67, 78, 79, 82 and 89). Looking at the original variables for dwelling 45, we observe that the values for hot water volume and hot water energy were higher than usual at 230 L in a day, when the mean consumption for this dwelling was around 100 L per day. On the other hand, heating energy (7, 16, 21, 43, 46, 55, 67, 78, 79, 82 and 89) had very low values, or zero in some cases, whereas dwellings (19, 28, 40, 52, 62 and 73) had higher values than usual.

On the other hand, SPE contribution analysis for 26th November 2013 (observation number 312), which only presented the large SPE index (see Fig. 8) performed in Fig. 11 reveals that many variables associated with heating energy (first variable of dwellings 8, 13, 21, 30, 32, 33, 39, 40, 42, 43, 44, 45, 46, 47, 49, 70, 72, 75, 81 and 83) presented large values and, at the same time, weather variables associated with temperature (mean and maximum) are over 3σ what can be explained by an unusually cold day.

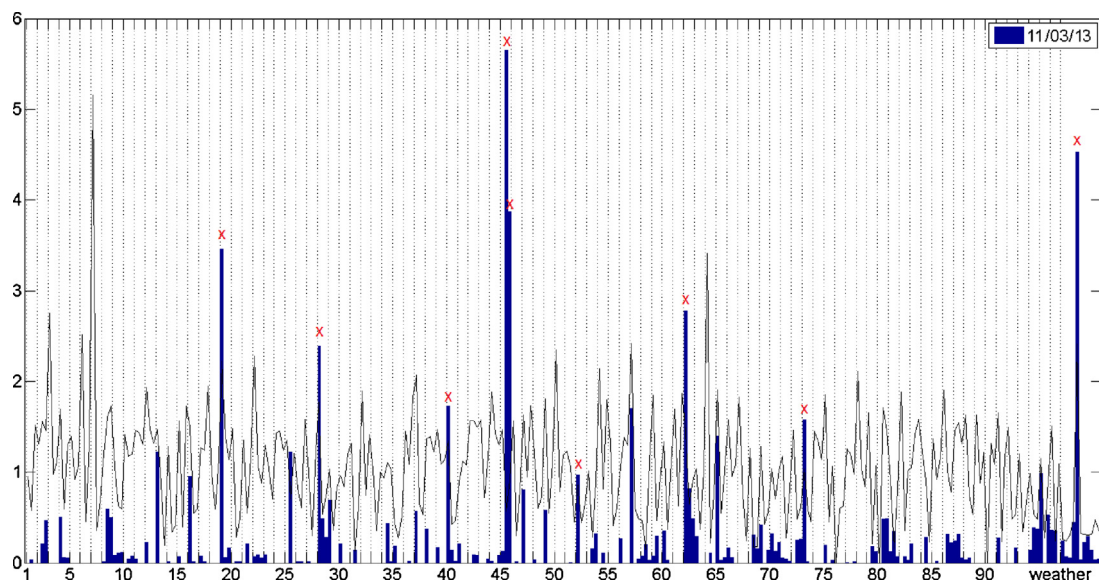


Fig. 9. T^2 contributions: bars represent the weight of a variable in a particular day (11th March 2013), black line represents 3σ given by model data.

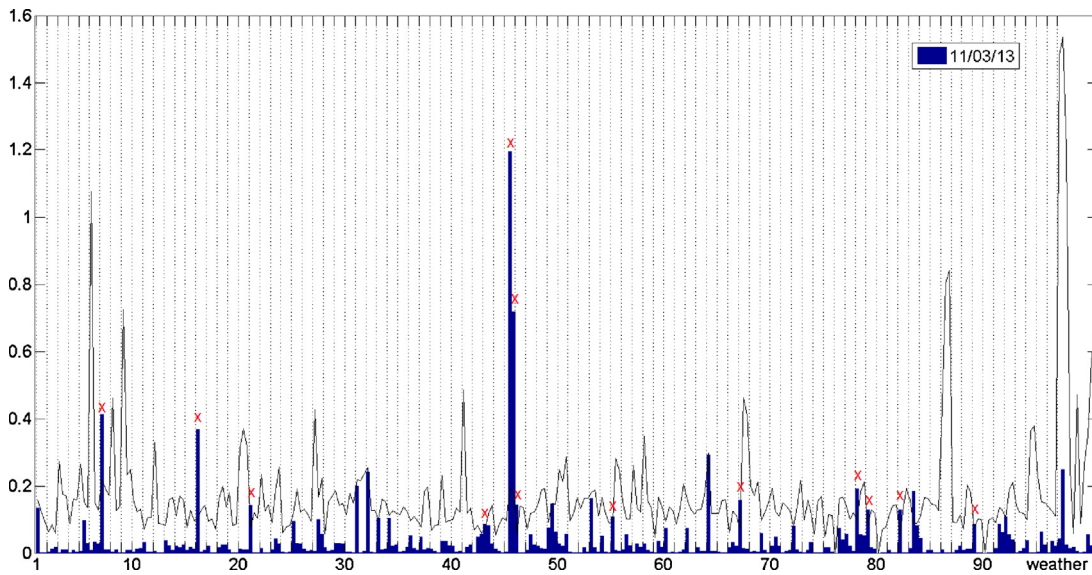


Fig. 10. SPE contributions: bars represents the weight of a variable in a particular day (11th March 2013), black line represents 3σ given by model data.

5.4. Monitoring based on entity-wise unfolding

In this second scenario, an entity-wise unfolding will be applied in order to unfold the 3D matrix into a 2D matrix to perform PCA. This approach aims to model the overall behaviour of dwellings. This will be useful in modelling dwellings based on their energy consumption patterns, along the observation period and, at the same time, finding those that present dissimilar uses.

5.4.1. Model construction

From the 3D matrix described at the beginning of this section ($96 \times 3 \times 636$), an entity-wise unfold has been applied, resulting in a (96×1908) matrix. Now each row contains all the data for a single entity (dwelling) during the whole observation period, and each column contains the values of one variable for one time instant for all the entities. In this scenario, weather information is not considered since it is the same for all the considered observations (apartments, entities).

The procedure described in Section 3.1 has been followed to obtain the loading matrix from the unfolded matrix, P . Only $r=6$ principal components have been retained, explaining 57.95% of the total variance. The thresholds for both statistics, T^2 and SPE, have been computed according to Eqs. (4) and (5), respectively, with a confidence factor of $\alpha = 95\%$.

5.4.2. Model exploitation: understanding dwelling behaviours

Once the model is constructed, observations can be projected in the principal component space. Now, every dot represents a dwelling during the whole observation period. Fig. 12 represents this projection, considering only the first three principal components (percentiles in brackets for each component indicate the variance captured in each direction). The centre of the grey-shadowed ellipsoid corresponds to the mean behaviour of the 96 dwellings. Again, T^2 (only 3 out of 6 components are represented) gives a measure of how far (distance) each entity (dots in Fig. 12) is from the centre. Thus, we can observe that the majority of dots are grouped, and only few of them are somewhat more dispersed.

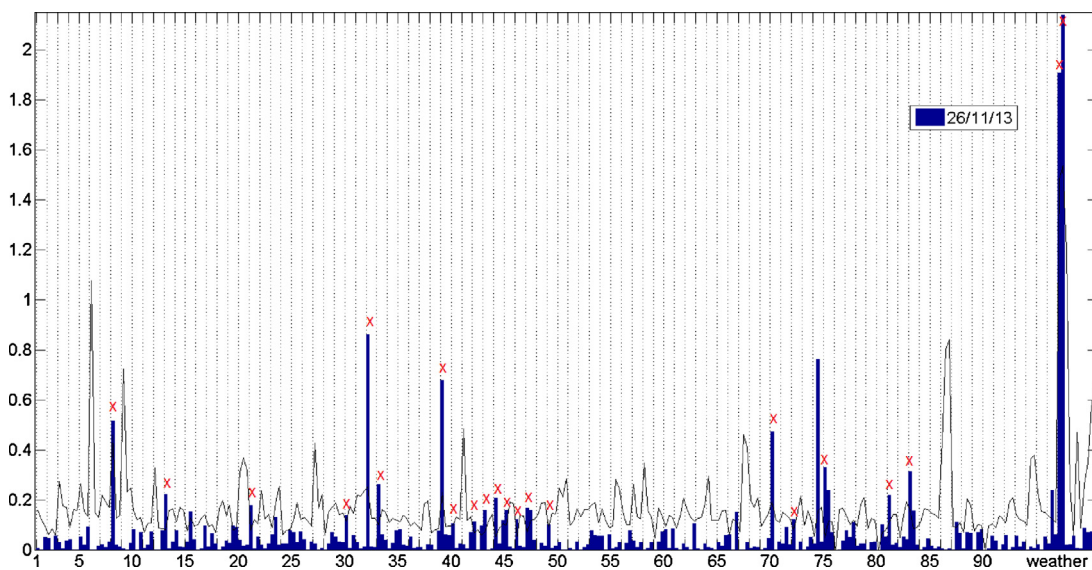


Fig. 11. SPE contributions (26th November 2013).

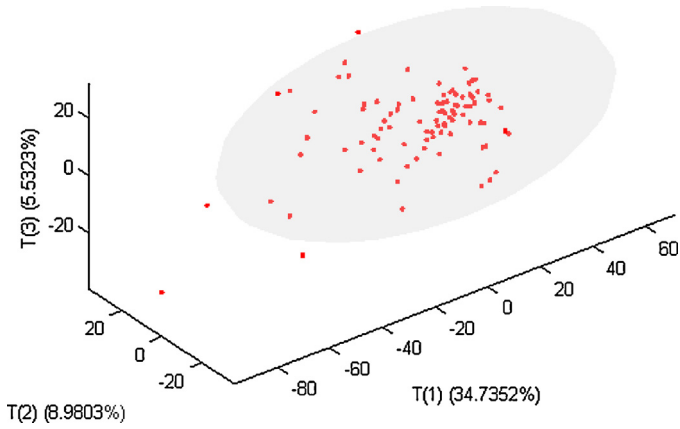


Fig. 12. Scores of the first three principal components, every dot representing a dwelling.

Presumably, these will be quickly identified by the T^2 index when analysing the T^2 chart (Fig. 13). Observe in the chart that the majority of observations are close to zero, and only eight of them (entity number: 2, 3, 25, 60, 73, 76, 80 and 83) are beyond the statistical threshold.

On the other hand, SPE , will be useful in terms of detecting such entities that the correlation among the variables is substantially different from those expressed by the model, and consequently they fall far from the projection space. Since only heating and water (and hot water) are being monitored, substantial differences are not expected among entities when analysing the SPE chart (Fig. 14). However, a large variability can be observed in this statistic (the range below the statistical threshold spans to 1400) representing uncorrelated (among entities) variability. Remember that only 57.95% of the original variability has been retained by the model (6 principal components). Consequently, the analysis of the SPE chart will not be very significant in the analysis of this second view of the building.

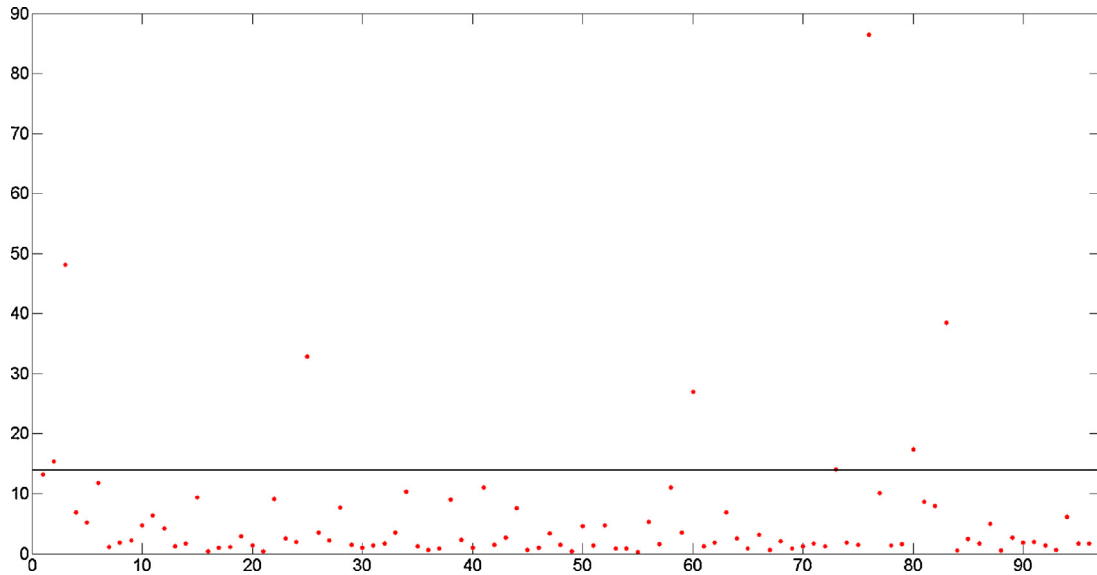


Fig. 13. T^2 chart: dots represent dwellings. Solid line: T^2 threshold.

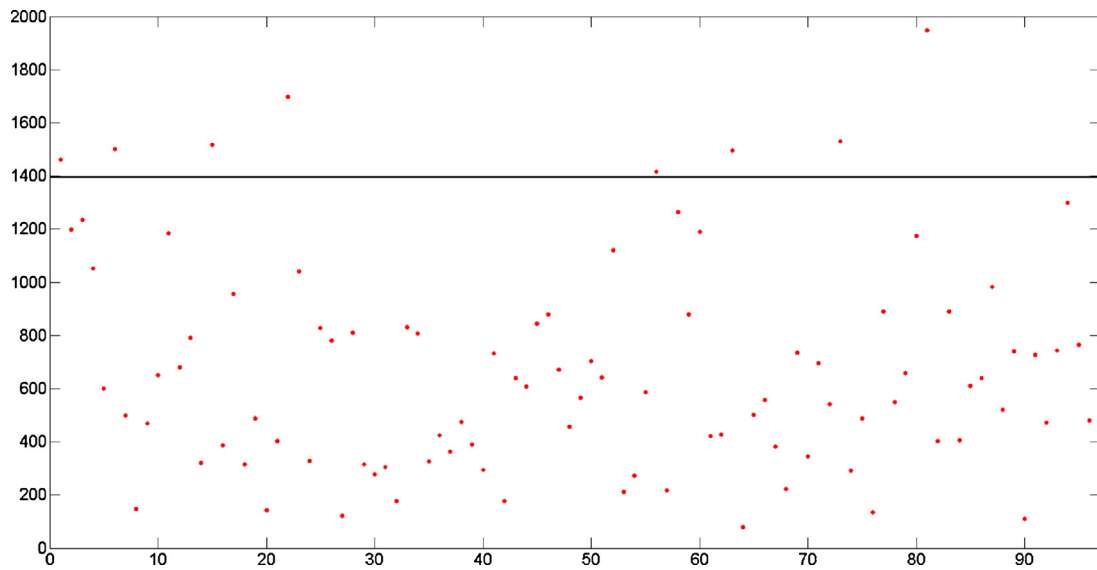


Fig. 14. SPE chart: dots represent dwellings. Solid line: SPE threshold.

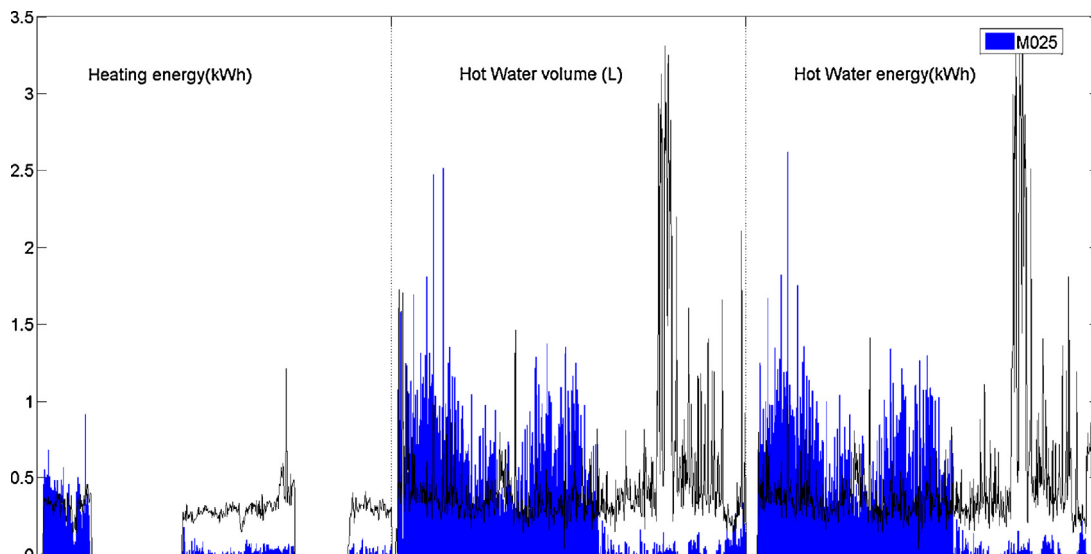


Fig. 15. T^2 contributions for dwelling 25: thin bars represents daily contributions for each variable. Solid line: 3σ threshold.

5.4.3. Contribution analysis: explaining deviations

Contribution analysis is applied to explain why a dwelling might perform slightly differently from the model. In order to exemplify it, T^2 contributions for a single apartment (dwelling 25) are represented in Fig. 15. The graph represents the evolution of the contributions of the three monitored variables for dwelling 25 during the whole observation period (the contribution for each day). We can observe that the variables representing hot water consumption are over the statistical threshold during the first half observation period (corresponding to year 2012). On the other hand, during 2013 and 2014, these contributions return to normal (under the 3σ threshold). In order to refine the analysis, the original variables have been analysed. It resulted that during this period (2012) a high daily consumption was reported (150–250 L per day with a peak consumption of 400 L). Some possible causes could be assigned to a problem in the water installation (water leakage) or an over-occupation of the dwelling.

6. Discussion and conclusions

The paper presents an adaptation of a multivariate process monitoring method used in the batch process industry, to assist the monitoring of communities (buildings or neighbourhoods) with multiple dwellings, instrumented by a common same sensor set. The novelty of the method resides in the ability to provide a multi-view analysis of a building (replicable for a neighbourhood) by applying the same methodology to the matrices that result after applying two different unfolding procedures (time-wise and entity-wise). The method also considers how to append additional information to the unfolded matrices in order to enrich the PCA models. Data from a social housing building consisting of 96 dwellings and a weather station have been used as an illustrative case study.

T^2 and SPE indices have been presented as control charts to facilitate the monitoring task on the basis of statistical threshold trespassing. Thus, large values with regard to these statistics reveal deviations from the reference model obtained from historic data representing normal operational conditions. The causes of deviation range from sensor faults to consumption habit changes or operational issues. Usually, data collected during 1 year are enough to represent all the variability of normal operating situations, including seasonal, weekly and daily variations. Enhanced

monitoring based on PCA is completed with these control charts for fault detection and a contribution analysis for both statistics (T^2 and SPE) is proposed to identify variables responsible for such deviation. This is a simple method that can be used to disaggregate these indices into the corresponding magnitude of the original variables (sensors). This strategy allows the identification of the variable (or sets of them) responsible for the out of control situation by simply analysing those with larger contributions.

The limitation of the traditional PCA monitoring method with regard to capturing relationships among multiple instances (entities) and among variables over time, has been overcome by proposing a new data model for building monitoring and two unfolding strategies. The data model consists of a 3D matrix, to represent information from multiple entities being monitored with a common set of variables, plus a couple of 2D matrices to include, on the one hand, additional sensors affecting the overall infrastructure and, on the other hand, particular attributes to characterize each entity.

Two unfolding strategies (time-wise and entity-wise) to convert this data model into a 2D matrix, making it suitable for PCA, have been proposed. The time-wise unfolding, in fact, corresponds to a classical PCA monitoring method and allows capture relationships among variables inter- and intra-entity. Thus, new data being collected continuously from sensors are monitored all together by simply projecting them into the reference model and computing the associated SPE and T^2 index, allowing the user to isolate faults in sensors or detect unusual dwelling consumption patterns (leakages, abnormal occupancy, etc.) among other irregularities. On the other hand, entity-based unfolding allows a complementary analysis of the same data for a specific period of time. This approach allows the user to capture correlations, not only among variables in an entity, but also among variables over time during the observation period. This is particularly useful in terms of identifying dwellings that behave significantly differently from others during the observation period (whole or part) e.g. those more exposed to adverse weather conditions.

The illustrative use case, relating to a social housing building, allowed us to demonstrate the benefits of the method and, in particular, to detect deviations using both unfolding approaches isolating the variables (and time instants) that significantly contributed to explaining such situations. The paper includes an analysis of a couple of situations for illustrative purposes. However, it would be

simple to systematize such a study to automatically (or periodically) generate reports to document the detected deviations in order to assist maintenance and to ensure energy efficient policies.

Acknowledgements

This work has been developed within the project Plataforma para la monitorización y evaluación de la eficiencia de los sistemas de distribución en Smart Cities, ref. DPI2013-47450-C2-1-R and project ACCUS (Adaptive Cooperative Control in Urban (sub) Systems., ART-010000-2013-2 -333020-1), funded by the Spanish Ministry of Industry, Energy and Tourism and by the JTI ARTEMIS Joint Undertaking of the European Commission. Appreciation is given for the data provided by the Patronat de l'habitatge de Barcelona with the collaboration of AITEL. Data from Meteocat were also used.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.enbuild.2015.06.069>

References

- [1] European Parliament and the Council of 25 October 2012 on energy efficiency. Directive 2012/27/EU, 2012.
- [2] Z. Yu, B.C.M. Fung, F. Haghghat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, *Energy Build.* 43 (June (6)) (2011) 1409–1417.
- [3] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Comparing alternative approaches for multivariate statistical analysis of batch process data, *J. Chemom.* 13 (3–4) (1999) 397–413.
- [4] G. Lukas, V. Swan, I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, *Renew. Sustain. Energy Rev.* 13 (October (8)) (2009) 1819–1835.
- [5] M. Kavgić, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, *Build. Environ.* 45 (July (7)) (2010) 1683–1697.
- [6] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, *Renew. Sustain. Energy Rev.* 16 (August (6)) (2012) 3586–3592.
- [7] K. Park, Y. Kim, S. Kim, K. Kim, W. Lee, H. Park, Building energy management system based on smart grid, in: 2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC), IEEE, 2011, October, pp. 1–4.
- [8] M. Hajdukiewicz, M. Keane, B. O'Flynn, W. O'Grady, Formal calibration methodology for CFD model development to support the operation of energy efficient buildings, in: Tenth International Conference for Enhanced Building Operations, Kuwait, 2010.
- [9] P. Raftery, M. Keane, J. O'Donnell, Calibrating whole building energy models: an evidence-based methodology, *Energy Build.* 43 (September (9)) (2011) 2356–2364.
- [10] B. O'Flynn, E. Jafer, R. Špinar, Development of Miniaturized Wireless Sensor Nodes Suitable for Building Energy Management and Modelling, ECPPM, Ireland, 2010.
- [11] A. Costa, P. Rafferty, M. Keane, J. O'Donnell, Energy monitoring systems value, issues and recommendations based on five case studies, in: Clima Conference, Antalya, Turkey, 2010.
- [12] D. Ndiaye, K. Gabriel, Principal component analysis of the electricity consumption in residential dwellings, *Energy Build.* 43 (February (2–3)) (2011) 446–453.
- [13] S. Li, J. Wen, A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform, *Energy Build.* 68 (2014, January) 63–71.
- [14] J.C. Lam, K.K.W. Wan, K.L. Cheung, An analysis of climatic influences on chiller plant electricity consumption, *Appl. Energy* 86 (June (6)) (2009) 933–940.
- [15] S. Mahmoud, A. Lotfi, C. Langensiepen, User activities outliers detection; integration of statistical and computational intelligence techniques, *Comput. Intell.* (2014).
- [16] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias, Using principal component and cluster analysis in the heating evaluation of the school building sector, *Appl. Energy* 87 (June (6)) (2010) 2079–2086.
- [17] J.C. Lam, K.K.W. Wan, S.L. Wong, T.N.T. Lam, Principal component analysis and long-term building energy simulation correlation, *Energy Convers. Manage.* 51 (January (1)) (2010) 135–139.
- [18] J.C. Lam, K.K.W. Wan, K.L. Cheung, L. Yang, Principal component analysis of electricity use in office buildings, *Energy Build.* 40 (January (5)) (2008) 828–836.
- [19] L. Burgas, J. Melendez, J. Colomer, Principal component analysis for monitoring electrical consumption of academic buildings, *Energy Proc.* 62 (2014) 555–564.
- [20] E. Russell, L.H. Chiang, R.D. Braatz, *Data-Driven Methods for Fault Detection and Diagnosis in Chemical Processes*, vol. 49, Springer, London, 2000.
- [21] P. Nomikos, J.F. MacGregor, Monitoring batch processes using multiway principal component analysis, *AIChE* 40 (1994) 1361–1375.
- [22] J.E. Jackson, G.S. Mudholkar, Control procedures for residuals associated with principal component analysis, *Technometrics* 21 (3) (1979) 341–349.
- [23] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, *Int. J. Adapt. Control Signal Process.* 19 (May (4)) (2005) 213–246.

Chapter 4

N-dimensional extension of unfold-PCA for granular systems monitoring

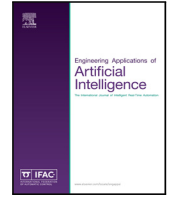
In this chapter, there is an explicit extension and its mathematical formalization over the Unfold-PCA methodology to deal with N-dimensional data arrays. This methodology enables new modelling and monitoring opportunities in building scope but also has applications to other scopes and techniques as this fold and unfold methodology is basically a data preprocessing methodology. Also a couple of use cases are presented. This publication has been published in the following paper:

<p>Paper published in the Engineering Applications of Artificial Intelligence Volume: 71, Pages: 113-124, Published: March 2018 DOI: 10.1016/j.engappai.2018.02.013 JCR IF (2017): 2.819 Q1(13/86) - Multidisciplinary engineering Q1(32/132) - Artificial intelligence, Computer science</p>



Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

N-dimensional extension of unfold-PCA for granular systems monitoring

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

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



ARTICLE INFO

Keywords:

Principal component analysis
Unfold-PCA
MPCA
Building energy monitoring
Data mining
Statistical process monitoring

ABSTRACT

This work is focused on the data based modelling and monitoring of a family of modular systems that have multiple replicated structures with the same nominal variables and show temporal behaviour with certain periodicity. These characteristics are present in many systems in numerous fields such as the construction or energy sector or in industry. The challenge for these systems is to be able to exploit the redundancy in both time and the physical structure.

In this paper the authors present a method for representing such granular systems using N-dimensional data arrays which are then transformed into the suitable 2-dimensional matrices required to perform statistical processing. Here, the focus is on pre-processing data using a non-unique folding–unfolding algorithm in a way that allows for different statistical models to be built in accordance with the monitoring requirements selected. Principal Component Analysis (PCA) is assumed as the underlying principle to carry out the monitoring. Thus, the method extends the Unfold Principal Component Analysis (Unfold-PCA or Multiway PCA), applied to 3D arrays, to deal with N-dimensional matrices. However, this method is general enough to be applied in other multivariate monitoring strategies.

Two of examples in the area of energy efficiency illustrate the application of the method for modelling. Both examples illustrate how when a unique data-set folded and unfolded in different ways, it offers different modelling capabilities. Moreover, one of the examples is extended to exploit real data. In this case, real data collected over a two-year period from a multi-housing social-building located in down town Barcelona (Catalonia) has been used.

1. Introduction

One of main challenges in industry's current transformation to the Industry 4.0 paradigm is to integrate, manage, process and exploit process data to benefit business. While the internet of Things (IoT) paradigm provides the infrastructure required for integration and management, data mining provides the background for processing according to the required exploitation goals. This paper focuses on the goal of such monitoring and assumes that a multivariate data mining technique is used for that purpose. In fact, the paper assumes that Principal Component Analysis (PCA) is the underlying principle to perform the monitoring and it focuses on the problem of organizing data to apply PCA. This method is also general enough to be applied to other multivariate monitoring strategies.

PCA is a well-known multivariate statistical technique which is not only widely used for dimensional reduction, but also for modelling and monitoring continuous processes based on observations provided by sensors (Russell et al., 2000; Edward Jackson and Mudholkar,

1979). PCA helps to control the processes by using the Hotelling's T^2 and SPE indices to provide charts to detect and analyse faults. The isolation of those faults is made with the contribution analysis (Kourti, 2005). However, as many other statistical methodologies, PCA requires a 2D matrix organization of data where columns represent variables and rows observations. Thus, models obtained with this technique gather correlations between the variables according to the observations (conveniently organized into rows) and assume independence between them. In monitoring applications, these observations usually refer to a single time instant (continuous processes). However, variations of PCA for monitoring include extensions for batch process monitoring based on Multiway PCA (MPCA, Nomikos and MacGregor, 1994) and other variants to address real-time (R-PCA, Yu et al., 2017), and outlier detection in an IoT context (Peter He et al., 2017).

The Multiway approach extends the concept of single instant observations to observations that have a temporal extension (typically the duration of the execution of the batch process) and consequently,

* Corresponding author.

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

<https://doi.org/10.1016/j.engappai.2018.02.013>

Received 1 February 2017; Received in revised form 2 February 2018; Accepted 19 February 2018

Available online 20 March 2018

0952-1976/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

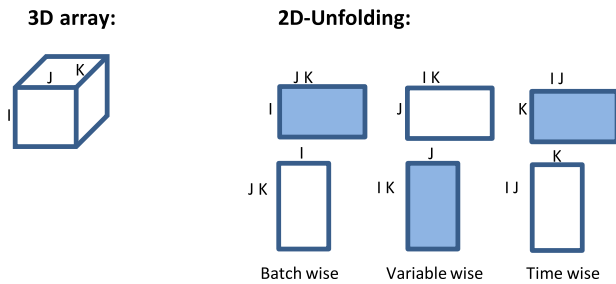


Fig. 1. Graphic representation of all the unfolding possibilities of a 3D matrix.

observations, instead of simple rows, are represented by 2D arrays (variables \times samples acquired during the batch execution) and by adding one new dimension to the historic data structure, it now becomes a 3D matrix. Thus, the dimensions of this 3D matrix, containing the historic data of a batch process, are defined by the number of variables being monitored in the process, J , the number of samples acquired at each execution of the batch process, K , and the number of executions included as historic data, I . Again, the I observations represented by these 2D arrays ($I \times K$) containing the data for the monitored variables of a complete execution of a batch process, are assumed to be independent.

Independent of how complex the observations are, the fundamental principles of PCA do not change, but reorganizing (unfolding) the data under study (i.e. to be modelled) into a 2D matrix is required. This implies that, in the case of batch processes, an unfolding preprocessing of the data is required to convert the 3D matrix into a 2D array before applying PCA. This unfolding process is not unique and, depending on how it is done, the interpretations of the results after applying PCA can differ substantially. Thus, there are six known possible combinations to unfold a 3D matrix into a 2D matrix, (see Fig. 1) and not all of them provide interpretable results. (NB: in fact, for the PCA purposes, there are only really three combinations, because half of them are simply the other half transposed.) In batch process monitoring (Nomikos and MacGregor, 1994) variable-wise unfolding ($I \times J \times K \rightarrow IK \times J$, observations in rows are all the samples acquired during the execution of batches) and the batch-wise unfolding ($I \times J \times K \rightarrow K \times IJ$, where observations in rows represent completed batches and number of columns extends to the variables at every time instant, IJ , during the execution of a batch) are commonly used. In other domains, such as monitoring energy in housing buildings for example, time-wise unfolding can also be meaningful (see, for instance, Burgas et al., 2015) to identify singularities in the power consumption of dwellings.

However, there are situations where 3D arrays are not suitable for organizing historic observations and higher dimensional data arrays, or hypercubes, need to be used instead. The need to analyse and model this complex data as a whole, requires developing of a clear methodology to manage the folding/unfolding procedures (as well as other preprocessing measures) for N -dimensional arrays to make them suitable for building interpretable and exploitable PCA models. This occurs, for example, when observations contain not only information from continuous sensors, but also images or spectroscopic information evolving through time where tensor-based dimension reduction techniques are used (Lu et al., 2008; Chen and Shapiro, 2009). A similar situation transpires when considering processes, or systems in a general way, with multiple replicated structures being monitored with the same set of nominal variables (e.g. solar fields and wind farms, injection and assembly lines, cavities in a mould, inkwells in offset industrial printers, power consumers in a grid, or monitoring stores in a mall or rooms in a hotel, etc.). A new challenge appears, one that consists of monitoring not only every subsystem, but also the interactions between them and through time.

Consequently, this requires monitoring tools to be developed that are not only capable of automatically detecting the significantly differently operating elements in any subsystem (e.g. sensor faults, faulty

components, performance reduction, misbehaviour detection, etc.) but that also monitor the interactions between these elements and detect any emergent behaviours. By considering modular replication as a new dimension in the data structure this analysis can be carried out, but first requires the adequate pre-treatment and management of the data. Similarly, when an operating continuous system presents a repetitive or periodic behaviour through time, this introduces a degree of redundancy that can be exploited when monitoring. This happens, for instance, in many systems that operate 24/7, but accommodate this operation accordingly due to, for example, shifts, power prices, seasons, solar illumination, etc. Examples of systems with this kind of pseudo-periodic temporal pattern (daily, weekly, seasonally, etc.) are, again, solar fields and wind farms, process industries, or hotels and tertiary buildings affected by daily variations. Such repetitive operations allow models to be built that can then be used as references for monitoring on different time scales or granularity (hourly, daily, weekly, etc.). An example of a multivariate analysis considering this temporal pattern in academic buildings is presented by the authors in Burgas et al. (2014).

Thus, organizing data into multi-dimensional arrays (usually dimension higher than four) is required for data from large systems built on the principle of repetitive modularity and periodic behaviour. This paper aims to provide a method for constructing multivariate models that will monitor such systems as a whole and allow MPCA methodology to deal with N -dimensional arrays. Because the methodology proposed is focused on a previous stage of the PCA modelling itself, then it can be useful not only for PCA modelling and monitoring, but also for other Data Mining tools, such as PLS (Partial least squares). Therefore, this work focuses on the pre-processing stage and, in particular, analysing the significance of the models obtained once specific unfolding strategies have been applied.

This introduction is followed by a background section that includes related work. Following on from that, the methodology to deal with N -dimensional arrays is introduced and the procedure to follow before applying PCA is explained step-by-step. The paper then describes an example of the application and a complete, real exploitation use case is depicted to illustrate the different models that can be obtained from an initial data set and their interpretation and use for monitoring purposes. The paper ends with a section devoted to conclusions and future work.

2. Background and related work

PCA is a method that allows linear dependencies between the variables of a system to be modelled (Russell et al., 2000; Edward Jackson and Mudholkar, 1979). Data gathered during normal operating conditions (NOC) is usually used to obtain a reference model in a new space of lower dimensionality (for instance, waste-water treatment plants as in Aguado and Rosen, 2008). Once the system has been modelled, the new observations projected onto the model's subspace can be used to verify its consistency. Usually two statistics, Hotelling's T^2 and SPE (Square Prediction Error), both defined in the model subspace, are used as the bounds of the model to check if any new observations fall inside or outside the model's thresholds. Hotelling's T^2 indicates how far an observation is from the centre of the model and SPE specifies to what extent the correlations mismatch the ones modelled. Those falling outside the model are considered faulty. Optionally, by using a contribution analysis it is possible to isolate the variables responsible for the deviation outside the statistical thresholds (Kourti, 2005). Currently, there are variations of PCA such as R-PCA (Recursive principal component analysis) in Yu et al. (2017) for sensor outlier detection or monitoring (Peter He et al., 2017) in an IoT scope, that meet the challenges that real-time presents. A complete comparison and study of PCA and its variations can be found in Camacho et al. (2008a, 2008b) and González-Martínez et al. (2014).

However, PCA itself, as with many other data modelling and mining techniques, operates over two-dimensional data matrices organized as $observations(rows) \times variables(columns)$. Some extensions of PCA (for

batch process monitoring for example), known as Multiway PCA (MPCA) described in Nomikos and MacGregor (1994), were defined to deal with three dimensional arrays. Multiway PCA can deal with batch processes (e.g. sequencing batch reactors (SBRs) in waste-water treatment facilities Haimi et al., 2016), thus allowing redundant information stored in the historic data bases of the batch process containing cyclical executions to be exploited. Each complete execution constitutes an observation and supposes adding a new dimension into the input data to be used for modelling and monitoring. Thus, a single observation becomes a 2D matrix containing a set of time series describing the evolution of every variable during the execution of the batch, instead of a single vector containing the samples of variables at a single time instant.

This temporal repetitiveness can be found in other domains. For instance, the power demands of a building present repetitive daily patterns affected by occupancy and weather conditions (Burgas et al., 2014). In Burgas et al. (2015), the same authors extended this approach to deal with multi-entity systems such as buildings (e.g. malls, hotels, housing buildings, offices, etc.) or communities (e.g. neighbourhoods, residential districts, industrial or business parks, etc.), dealing with up to 4D arrays and offering a multi-view monitoring approach for housing buildings when applying different unfolding processes.

However, PCA is not the only methodology available to deal with multivariate data. Other multi-way decomposition approaches that have been conceived for batch processes have their origins in PARAFAC (Harshman, 1970; Chang and Carroll, 1970), Tucker (1966). A survey of previous multi-way decompositions including PARAFAC (or CAN-DECOMP), Tucker and two-way PCA, is reviewed in Bro (1997). The survey discusses the similarities, constraints and links between them and notes that while a data-set that can be modelled adequately with PARAFAC can also be modelled by Tucker3 or two-way PCA, PARAFAC requires fewer degrees of freedom. On the other hand, Kiers (1991) says that two-way PCA will always fit better than a PARAFAC or Tucker3 model, except in extreme cases where they may all fit equally well. The suitability of the three methods for batch processes is analysed in Westerhuis et al. (1999). None of the studies, however, propose a method to systematically organize and unfold data.

In the following sections the authors formalize and extend the unfolding methodology (Nomikos and MacGregor, 1994) to deal with N -dimensional arrays, taking into account the repeatability and granularity (formal definition in Bettini et al., 1998) of modular systems. Working with folded N -dimensional data-sets allows for all the characteristics of the data to be preserved and for new modelling opportunities to be derived from the redundancy of data.

3. Methodology

3.1. Granular monitoring of multi-entity systems

This work focuses on pre-treating and organizing multidimensional data for monitoring, especially in the case of systems that present repetitive behaviour and/or structures. The method is applied to multi-entity or modular systems (e.g. housing buildings) where every single entity (e.g. a dwelling) is being monitored by the same nominal set of variables (e.g. power consumption, interior temperature, water consumption, occupancy, etc.). To exploit the method's potential, it is expected that there is some kind of interactions between these units (e.g. heat transfer through walls, shared areas, central heating, etc.). The method is general enough to consider multiple levels of modularity in a way that, for a given level, the monitoring variables in a module contain repetitions of those in the level immediately inferior. Thus, in the previous example of a dwelling, this can be defined as the lower level of modularity where five variables are being monitored. A second level could be a floor divided into four dwellings (20 sensors) and a third level could be defined by the whole six-storey building with four dwellings on each floor (i.e. 120 variables in total). Thus, in the initial set of variables,

the dimension J is 120 variables long, although this can be split into three levels of modularity, resulting in a 3D array ($J_1 \times J_2 \times J_3$) of $5 \times 4 \times 6$.

The term granular monitoring refers to the possibility of organizing observations on different levels of temporal detail and performing monitoring accordingly. Thus, in batch process monitoring it is easy to distinguish the minimum two levels of granularity (or multi-trajectories), i.e. sampling time and batch (time series acquired during the execution of the batch). Some continuous systems also present this kind of repetitive behaviour. For example, fed-batch reactors or any other calendar operated system that has repetitive behaviour on daily, weekly or yearly time scales. In all of these systems, the sampling time defines the lowest level of granularity and the longest repetition periods define the highest. For a given level, the information contained in a single observation (a granule) does not overlap with any of the other observations on that same level. However, it does, of course, contain multiple observations from an inferior level (for a formal definition of the time granularity concept, the interested reader is referred to Bettini et al., 1998).

Imagine in the previous housing building example, that data sensors gathered data hourly (sampling time) for three years. This will result in a total of $I = 26\,208$ observations (samples acquired every hour). An accurate observation of daily and weekly shapes should show that they present repetitive behaviour that can be analysed on the following four granularity levels: hour, day, week and year. Thus, the initial set of hourly observations ($I = 26\,208$) can now be reorganized into four levels of granularity: I_1 , hours a day; I_2 , days a week; I_3 , weeks a year; I_4 , available years of historic data. The initial data-set defined by the 2D matrix ($I \times J$), can in fact be organized into an N dimensional array ($I_1 \times I_2 \times I_3 \times I_4 \times J_1 \times J_2 \times J_3$ with $N = 4 + 3 = 7$), resulting in an array size of $24 \times 7 \times 52 \times 3 \times 5 \times 4 \times 6$.

The next section details the correspondence between elements in both 2D and N -dimensional arrays and shows different ways to unfold this into a new 2D matrix with a different data distribution suitable for monitoring.

3.2. Basic pre-processing operations: folding, standardization, merging and unfolding

Acquisition systems usually gather information sequentially, resulting into long 2D matrices where columns represent every sensor installed and rows contain dated values acquired at every time-stamp. For this work, the initial 2D matrix is called X and is assumed to contain I observations (rows) of J variables (columns). The objective is to transform this matrix into a new 2D matrix, X' , with dimensions $J' \times I'$ ($J \neq J'$ and $I \neq I'$), suitable for PCA. This PCA suitable matrix is obtained after reordering observations and variables conveniently to observe the system at the convenient granularity and modularity level defined by the monitoring goals.

Folding is the procedure that will be used to reorganize the data into this N -dimensional *folded* array, \underline{X} , by considering system granularity and modularity. Specific dimensions in the N -folded array will correspond to different granularity levels, allowing the data acquired at different sampling times to be merged by simply appending matrices in the correct dimension (same granularity). Additionally, a standardization procedure, one which avoids variables with larger magnitudes and variation range dominating, must be applied to make the data suitable for PCA.

Thus, to perform this transformation of X into X' there are four basic operations to carry out: folding, unfolding, standardization and merging.

1. **Folding.** This is the procedure that allows the original 2D matrix to be transformed into an N -dimensional folded array, \underline{X} , in such a way that granularity and modularity are consistently represented.

2. **Standardization.** This is data centring (zero mean) and equalization in terms of variance (unit variance in all the columns). The purpose is to avoid variables with large variances and bias dominating.
3. **Merging** (Optional). This is only required when the original data is split into several arrays with sampling times on different time scales or distinct modularity. It consists of appending two distinct X' matrices (when possible) to add more information to the models at certain levels of modularity/granularity.
4. **Unfolding.** This is the procedure that reshapes the folded \underline{X} array into the best bi-dimensional matrix X' , according to the monitoring goals.

These operations are analysed in detail in the following subsections:

3.3. Folding

Folding is the transformation of the original 2D matrix (I observations $\times J$ variables) into an N -dimensional folded array, \underline{X} , in such a way that granularity and modularity are consistently represented. At this point, that there are other arrays to be merged is not considered (this issue will be discussed further) and it is assumed that the original X matrix contains equally sampled data that has been aligned without blanks.

If the system presents M levels of modularity and L levels of granularity, then it is possible to fold it into an N dimensional, \underline{X} array, with $N = L + M$. Since granularity and modularity are defined in a context of repeatability, the length of the observations and the grouping of the variables at a given level will be fixed and define a dimension of \underline{X} . These dimensions are labelled as I_l , with $l = 1 \dots L$ and J_m , with $m = 1 \dots M$, respectively, and the product of their sizes equals the size of the original 2D matrix: $\prod I_l = I$ and $\prod J_m = J$. Observe that if only the lowest levels of granularity and modularity are considered ($L = M = 1$), this will result in the original 2D matrix ($I_1 \times J_1 = I \times J$).

The main problem when performing the folding procedure, is to have control over how the samples are reorganized to facilitate applying the pre-processing algorithms to the most convenient data organization. To establish a clear correspondence between elements in X and \underline{X} Eqs. (1) and (2) have been established, where any observation in the X matrix ($x_{i,j}$) is mapped to an observation in the folded array ($x_{i_1 \dots i_L, j_1 \dots j_M}$), where i_l (with $l = 1 \dots L$) and j_m (with $m = 1 \dots M$) represent the coordinates of the sample in \underline{X} .

$$i_l = \left\lfloor \frac{i-1}{\prod_{p=0}^{l-1} I_p} \right\rfloor \% I_l + 1 \tag{1}$$

$$j_m = \left\lfloor \frac{j-1}{\prod_{p=0}^{m-1} J_p} \right\rfloor \% J_m + 1. \tag{2}$$

The symbol for the remainder operator of the division performed between the left and right arguments is %, and the square brackets with missing upper bars is the symbol to represent the integer part of the division inside. Where $i = 1 \dots I$ and $j = 1 \dots J$ are the coordinates of the observation ($x_{i,j}$) in the original X matrix; i_l and j_m are the coordinates of the element in the I_l or J_m dimension of the folded array \underline{X} ; I_p and J_p are the length, or number of elements, in the I_p and J_p dimension of the folded matrix. $I_0 = 1$, $J_0 = 1$ and non-existent dimensions (due to nomenclature when distinct X are folded for merging later) must be considered to be 1.

To exemplify this relationship, suppose an X matrix with $I = 100\ 800$ observations and $J = 110$ variables where three levels of granularity are identified in time ($L = 3$) and two levels of modularity in variables ($M = 2$) where $I_1 = 60$ (min), $I_2 = 24$ (h), $I_3 = 70$ (days), $J_1 = 11$ and $J_2 = 10$. The X matrix can be folded into an \underline{X} array with $N = L + M = 3 + 2 = 5$ dimensions. Eqs. (1) and (2) have been used to find the correspondence

between any $x_{i,j}$ and the corresponding ($x_{i_1, i_2, i_3, j_1, j_2}$) resulting, for this particular case, in the following five corresponding Eqs. (3)–(7).

$$i_1 = \left\lfloor \frac{i-1}{1} \right\rfloor \% 60 + 1 \tag{3}$$

$$i_2 = \left\lfloor \frac{i-1}{60 * 1} \right\rfloor \% 24 + 1 \tag{4}$$

$$i_3 = \left\lfloor \frac{i-1}{24 * 60 * 1} \right\rfloor \% 70 + 1 \tag{5}$$

$$j_1 = \left\lfloor \frac{j-1}{1} \right\rfloor \% 11 + 1 \tag{6}$$

$$j_2 = \left\lfloor \frac{j-1}{11 * 1} \right\rfloor \% 10 + 1 \tag{7}$$

where i_1, i_2, i_3, j_1 and j_2 are the corresponding indices of the element $x_{i,j}$ in the 5-dimensional matrix \underline{X} .

3.4. Unfolding

The unfolding procedure consists of reshaping a folded N -dimensional array, \underline{X} , into a bi-dimensional one, X' , adequate for PCA modelling purposes. Depending on the unfolding process chosen, distinct X' matrices can be obtained. Observe that for an N -dimensional data matrix, the number of unfolding possibilities doubles according to the following expression:

$$\sum_{k=1}^{k=N-1} \binom{N}{k} = \sum_{k=1}^{k=N-1} \frac{N!}{k!(N-k)!} \tag{8}$$

Thus, for $N = 3, 4, 5, 6$, and 7 , the unfolding possibilities are 6, 14, 30, 62 and 126, respectively. The unfolding possibilities double (plus two) each time N increases a unit. Notice that half of the unfolding possibilities is the transposition of the other half, so the progression is divided by two. However, many combinations appear for large N . For example, Fig. 2 represents the 14 unfolding possibilities for a 4D matrix of lengths I, J, K, L . However, not all these unfoldings make sense in monitoring applications, and so the most appropriate ones must be chosen according to the monitoring goals that have been set. The possible X' matrices that are obtained after unfolding \underline{X} have different correlation structure and consequently the meaning of the PCA analysis changes. Being able to choose the appropriate unfolding process for each monitoring purposes is a critical point. Some indications to help decide which dimensions in \underline{X} will be unfolded as columns or rows are the following:

- The dimension associated to the original variables (corresponding to the lower level of modularity, J_1), should always be placed in the columns' group (always part of J'). Unfolding results where J_1 are considered part of the set of rows to be analysed (part of I') make no sense from a monitoring point of view. Therefore, half of the unfolding possibilities (those with J_1 placed in the rows' group) should be discarded.
- PCA will find linear correlations between the J' variables, explaining the variations in the I' observations. So, dimensions susceptible to holding correlations of interest in our system should be considered for being placed in the variables set (part of J'), i.e. in the classical batch approach, where I, J and K dimensions are defined, only the Batch-Wise $I' \times J' = I \times (JK)$ and Variable-Wise $I' \times J' = (IK) \times J$ unfolding make sense.
- The order (position) of the row and column elements is not relevant in terms of modelling, as the results will be the same, but it is highly recommendable to choose a meaningful organization to easily visualize and understand the model results. This is especially important for the composition of J' dimension (variables).

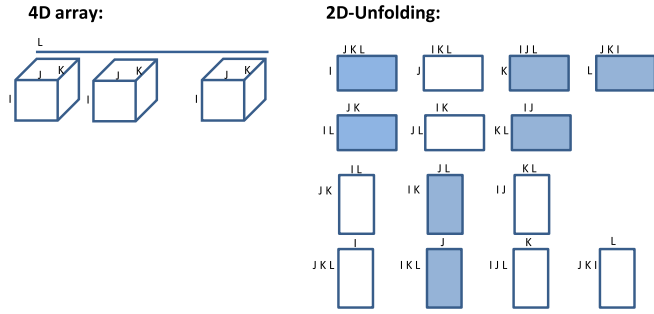


Fig. 2. Graphic representation of all the unfolding possibilities of a 4D matrix.

- Consider re-sampling or data aggregation, when the lower level of granularity is not required, thus reduces the computational cost of creating models and reduces the influence of noise has.
- As PCA studies the correlations between variables, uncorrelated variables can be avoided and computational costs reduced.

Considering this hints the user should be able to choose the best unfolding for his problem and formulating the correspondence between elements in both matrices \underline{X} and X' .

Thus, given an N -dimensional array \underline{X} and a desired unfolding structure $I' \times J'$ (*Rows* \times *Columns*), the correspondence between an element $(x'_{i',j'})$ in the X' matrix with the corresponding element $(x_{i_1 \dots i_{L'} j_1 \dots j_{M'}})$ in \underline{X} is given by the mapping Eqs. (9) and (10). Where i'_l (with $l = 1 \dots L'$) and j'_m (with $m = 1 \dots M'$) represent the coordinates of the sample in the N -dimensional folded array \underline{X} reordered according to the final organization of data in rows and columns of X' required. Thus, the notation $(x_{i'_1 \dots i'_{L'} j'_1 \dots j'_{M'}})$ represents the same element $(x_{i_1 \dots i_{L'} j_1 \dots j_{M'}})$ once this reordering of the coordinates has taken place in such a way that the first L' coordinates will be unfolded as rows describing observations and the last M' will be unfolded as columns in the final matrix X' .

$$i' = i'_1 + \sum_{l=2}^{L'} \left((i'_l - 1) \prod_{p=1}^{l-1} I'_p \right) \quad (9)$$

$$j' = j'_1 + \sum_{m=2}^{M'} \left((j'_m - 1) \prod_{p=1}^{m-1} J'_p \right) \quad (10)$$

where i'_l (j'_m) is the index of the element in the l th (m th) dimension assigned to the rows' (columns) group, I'_p (J'_m) is the length, or number of elements in that dimension, L' (M') is the number of dimensions in the rows' (columns) group and i' (j') the corresponding index in the final unfolded matrix X' .

To exemplify the correspondence given by the previous equations, suppose that the desired transformation is from a 5-dimensional array $(I_1 \times I_2 \times I_3 \times J_1 \times J_2)$ into a 2D matrix distributed as $(J_2 I_1) \times (J_1 I_2 I_3) = (I'_1 I'_2) \times (J'_1 J'_2 J'_3)$ with sizes $(I_1 = 60, I_2 = 24, I_3 = 70, J_1 = 11 \text{ and } J_2 = 10)$. The correspondence equations that links any element $x_{i_1 i_2 i_3 j_1 j_2}$ to the corresponding $x_{i' j'}$ are given by the Eqs. (11) and (12).

$$\begin{aligned} i' &= i'_1 + (i'_2 - 1) * I'_1 \\ i' &= j_2 + (i_1 - 1) * J_2 \\ i' &= j_2 + (i_1 - 1) * 10 \end{aligned} \quad (11)$$

$$\begin{aligned} j' &= j'_1 + (j'_2 - 1) * J'_1 + (j'_3 - 1) * J'_2 * J'_1 \\ j' &= j_1 + (i_2 - 1) * J_1 + (i_3 - 1) * I_2 * J_1 \\ j' &= j_1 + (i_2 - 1) * 11 + (i_3 - 1) * 264. \end{aligned} \quad (12)$$

Thus, a given element in \underline{X} , represented by $x_{10,20,30,4,5}$ the i' and j' indices of the corresponding $x_{i' j'}$, will be computed with Eqs. (13) and

$$i' = 5 + (10 - 1) * 10 = 95 \quad (13)$$

$$j' = 4 + (20 - 1) * 11 + (30 - 1) * 264 = 7869. \quad (14)$$

Therefore, the element $x_{10,20,30,4,5}$ in the 5D array will be reallocated as the element $x'_{95,7869}$ in the unfolded 2D matrix X' .

3.5. Standardization

PCA requires variables being centred and with similar variance. To guarantee this, a standardization procedure should be applied. Standardization will consist of obtaining data with zero mean and unit variance. The procedure is simple: for each variable, its mean (μ) and standard deviation (σ) are obtained, once every sample has been standardized by subtracting μ and dividing by σ , as in expression (15). For the sake of simplicity, x is used for the standardized value and x^o for the original data.

$$x = \frac{x^o - \mu}{\sigma}. \quad (15)$$

In classical 3D unfold-PCA, depending on how μ and σ are obtained (which dimension is considered as the sample), the literature purposes four main standardization procedures known as Continuous Scaling (CS), Auto-Scaling (AS), and Group-Scaling (GS) and Block Scaling. In Continuous Scaling (Esbensen et al., 1987), μ and σ are obtained for each variable during all the time instants (observations). Then, according to the methodology proposed, this is equivalent to performing it at the initial step, that is from X :

$$x_{ij} = \frac{x^o_{ij} - \mu_j}{\sigma_j} \quad (16)$$

with

$$\mu_j = \frac{\sum_{i=1}^{I} x^o_{ij}}{I} \quad (17)$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{I} (x^o_{ij} - \mu_j)^2}{I - 1}}. \quad (18)$$

When the initial data-set presents distinct granularity (different sampling times, for example) or not all the modules have the same degree of replication (for instance, the existence of common or global variables) the initial data-set must be divided into homogeneous subsets. The resulting subsets have to be able to be represented consistently as the initial matrix X . After performing the previously described folding/unfolding procedures, obtaining a set of matrices X' with the same granularity will then be possible. The following must be considered:

- The same nomenclature must be followed in the unfolded matrices (e.g. if I_1 = seconds in one folding then it cannot be I_1 = hours in another)
- Unfolded dimensions that result in rows in X' must be consistent in granularity (sampling), units and order.

So, an X' matrix coming from a $(I_1 J_2) \times (AnyGroup)$ unfolding procedure can be added to any X' matrix coming from a $(I_1 J_2) \times (AnyOtherGroup)$ unfolding procedure if I_1 represents the same time instants and the same frequency in both matrices and J_2 represents the same variables or entities.

A detailed explanation of the merging procedure can be found in Camacho et al. (2008b). Since the merging is performed with unfolded matrices, it is the same, independently of the dimension of the folded matrix.

In Auto-Scaling (Westerhuis et al., 1999), μ and σ are obtained for each variable at each time instant of the batch (observations are now the time series during the batch). Thus, according to the proposed

methodology, this will be equivalent to performing it after unfolding, in the matrix X' :

$$x'_{i'j'} = \frac{x'_{i'j'o} - \mu_{j'}}{\sigma_{j'}} \quad (19)$$

with

$$\mu_{j'} = \frac{\sum_{i'=1}^{I'} x'_{i'j'o}}{I'} \quad (20)$$

$$\sigma_{j'} = \sqrt{\frac{\sum_{i'=1}^{I'} (x'_{i'j'o} - \mu_{j'})^2}{I' - 1}} \quad (21)$$

Finally, Group Scaling and Block Scaling are used when data consist of several groups or blocks of variables with some given uniform feature (i.e. unit of measure). Different groups have different features. Group and Block Scaling are performed by scaling each group or block by the same standard deviation (i.e. the grand mean of their standard deviations). Following the methodology proposed, an extension of Group or Block Scaling can be defined by allowing for the possibility of obtaining the standard deviation from different unfold matrices (X'') and, once standardized, going back to the initial data format.

3.6. Merging

When the initial data-set presents distinct granularity (different sampling times, for example) or not all the modules have the same degree of replication (for instance, common or global variables exist), it must be divided into homogeneous subsets and the resulting subsets must be able to be represented consistently as the initial matrix X . After performing the folding/unfolding procedures previously described, it will now be possible to obtain a set of matrices X' with the same granularity. The following considerations should be taken into account:

- The same nomenclature must be followed in the unfolded matrices (e.g. if $I_1 =$ seconds in one folding then cannot be $I_1 =$ hours in another)
- Unfolded dimensions that result rows in X' must be consistent in granularity (sampling), units and order.

So, an X' matrix coming from a $(I_1 J_2) \times (AnyGroup)$ unfolding can be added to any X' matrix coming from a $(I_1 J_2) \times (AnyOtherGroup)$ unfolding if I_1 represents the same time instants and the same frequency in both matrices and J_2 represents the same variables or entities.

A detailed explanation about the merging procedure can be found in Camacho et al. (2008b). Since this is performed with unfolded matrices, it is the same, independently of the dimension of the folded matrix.

4. Application example

To illustrate the methodology, a case study monitoring a parabolic trough solar power plant is presented. In this case, the granularity in both the monitored variables and time can be used to reach new modelling options. In the following sections these options are introduced and the proposed methodology is followed.

4.1. Data information

On the one hand, the plant being monitored (Fig. 3) consists of four identical solar fields, each with 50 parallel loops composed of four solar collector assemblies. To generate electricity, the collectors capture the solar radiation by heating a fluid to drive a turbine connected to an electrical generator. In each collector assembly, three variables are measured at the same frequency (transfer fluid temperature, volumetric flow rate, and solar irradiation). Moreover, three production plant variables are provided (power, transfer fluid temperature and volumetric flow rate).

Table 1
Dimensions summary for field, production and meteorological data.

Dimension		Field	Production	Weather
I_1	Hours	24	24	24
I_2	Days	360	360	360
J_1	Variables	3	3	3
J_2	Collectors	4	(none)	(none)
J_3	Parallel loop	50	(none)	(none)
J_4	Solar field	4	(none)	4

On the other hand, as solar plant generation is highly correlated with weather, four weather stations, (one for each solar field), provide three weather variables (temperature, wind, humidity) at an hourly rate.

To summarize, the system has three variables that are replicated at every collector assembly, three global variables from the plant and three global weather variables.

4.2. Data folding

Since three different data sources are available, and to later merge the unfolded matrices, a common nomenclature must be established. For time dimensions, I_1 is used for hours and I_2 for days in the three data-sets. J_1 is always used for the measured variables that are different for each data-set. Then, the J_2 dimension will be used for the collector assemblies of each parallel loop, J_3 for parallel loops of each solar field and J_4 for the solar fields of the power plant. The sizes of each dimension for each data source are indicated in Table 1. Note that the size of J_1 is, by coincidence, the same for the three data sources, but this condition is not really needed to later merge the unfolded matrices.

Eq. (1) has been applied to the three data-sets known as $X1$, $X2$ and $X3$ matrices, to obtain three folded arrays $\underline{X1}$, $\underline{X2}$ and $\underline{X3}$. Thus, $\underline{X1}$ results in a 6D $(I_1 \times I_2 \times J_1 \times J_2 \times J_3 \times J_4)$ array of $24 \times 360 \times 3 \times 4 \times 50 \times 4$ for the collected power plant data, $\underline{X2}$ a 3D $(I_1 \times I_2 \times J_1)$ array of $24 \times 360 \times 3$ for the production data and $\underline{X3}$ a 4D $(I_1 \times I_2 \times J_1 \times J_4)$ array of $24 \times 360 \times 3 \times 4$ for the weather data. According to the proposed methodology, these three N -dimensional arrays are suitable to be unfolded and then used for data-based modelling of the power plant.

4.3. Unfolding

Depending on the objective, the three \underline{X} matrices can be unfolded in several ways following the indications in Section 3.4. Since the high dimensional matrix is $\underline{X1}$, which contains the largest amount of data, this will be used as the basis for the unfolding. Next, $\underline{X2}$ and $\underline{X3}$ will be optionally added by using the merging procedure described in Section 3.6. Considering that $\underline{X1}$ is a 6D matrix, according to Eq. (8) and the associated constraints, there will be up to 31 meaningful unfolding options. To show the value of some of these possibilities, two different modelling objectives are defined: monitoring and benchmarking.

4.3.1. Unfolding for monitoring

The most common modelling objective is to monitor the whole system to detect faults and for diagnostic purposes. This corresponds to a classical data-based monitoring and is achieved by placing I_1 and I_2 in the Rows' group and the rest of dimensions in the Columns' group. In this way, the unfolded $X1'$ $(I_1 I_2) \times (J_1 J_2 J_3 J_4)$ matrix is the same as the original one, $X1$. All the variables measured at each time instant (in this case hourly) are continuously monitored for fault detection and diagnosis tasks.

In addition, daily monitoring can be reached by placing only I_2 in the Rows' group and the rest of dimensions in the Columns' group $(I_2) \times (I_1 J_1 J_2 J_3 J_4)$ for modelling. In this way, when monitoring, all the measurements obtained during a day are used as inputs. This allows, as in batch processes, the repetitiveness (in this case daily) of the data to be considered to perform more accurate fault detection and diagnosis tasks, albeit only once a day. In both monitoring versions, weather and

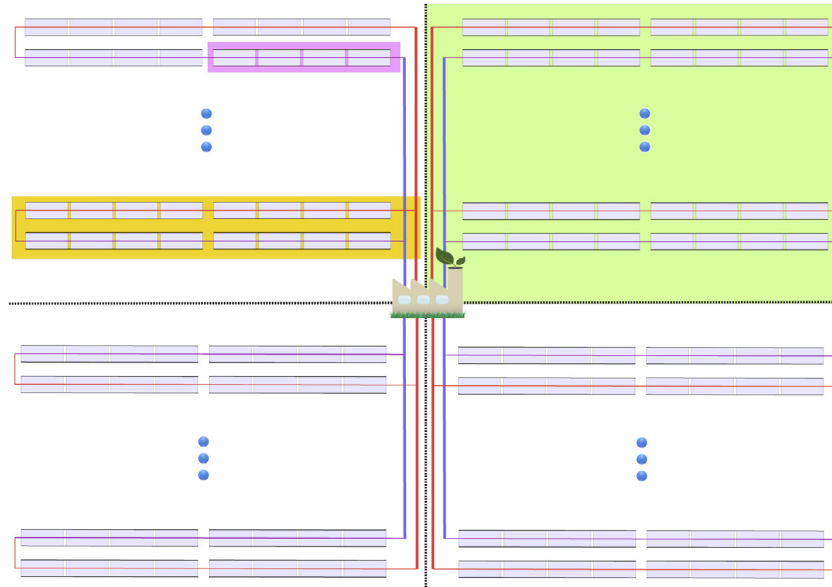


Fig. 3. Schema of monitored parabolic trough solar power plants. One solar field is marked in green, one loop in yellow and one solar collector assembly in pink. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

production variables can be added by using the merging procedure. Weather data can be unfolded $(I_2) \times (I_1 J_1 J_4)$ or $(I_1 I_2) \times (J_1 J_4)$ for daily and hourly monitoring, respectively. In this way, the correlation between the measured process variables and the weather variables is also modelled and then used for monitoring. Likewise, the production data matrix can be unfolded $(I_2) \times (I_1 J_1)$ or $(I_1 I_2) \times (J_1)$ and merged for modelling. This will allow production plant variables to be used for monitoring.

As a numerical example, consider that a daily monitoring of the whole plant, including weather and production data, is going to be carried out. According to Table 1, the historical data matrix used for modelling will be $360 \times (24 * 3 * 4 * 50 * 4 + 24 * 3 + 24 * 3 * 4) = 360 \times 57960$. Then, with the monitoring system running on-line, the 57960 measurements obtained during a day, will be the input to obtain the daily diagnostic of the plant. If the goal is an hourly on-line monitoring using only plant and production data, the matrix with the historical data used for modelling will be $(360 * 24) \times (3 * 4 * 50 * 4 + 3) = 8640 \times 2403$. Then, the 2403 measurements obtained hourly will be the input of the on-line monitoring system.

Moreover, individual monitoring can be done for each specific part of the plant (i.e. each solar field, loop or collector assembly). In this case, models can be built by either taking advantage of the information gathered from the whole plant or from only a specific part. For example, for online daily monitoring of a loop, it is clear that the $3 * 4 * 24 = 288$ measurements obtained each day in the loop should be used. However, at the modelling stage there are several possibilities. Directly, the X' matrix can be built from the dimension $(I_2) \times (I_1 J_1 J_2)$ and the size 360×244 in such a way that only the historical data of that loop is used for modelling. However, other options are to build X' from the dimension $(I_2 J_3) \times (I_1 J_1 J_2)$ and the size $360 * 50 \times 244 = 18000 \times 244$, or from $(I_2 J_3 J_4) \times (I_1 J_1 J_2)$ and $360 * 50 * 4 \times 244 = 72000 \times 244$, thus obtaining a unique model for all the loops in the same solar field or for the whole plant, respectively. In a similar way, specific models can be built for hourly monitoring, and for collector assembly or solar field monitoring. In all the models proposed in this paragraph, both the weather and the production data-sets can be merged. Then, when monitoring, the daily weather and/or production data should be used, and will likely obtain better results.

4.3.2. Unfolding for benchmarking

In the same way that the granularity of the process behaviour can be useful for monitoring, the modularity of the process structure, based

on historical data, could be used for benchmarking. In this case, time dimensions I_1 and I_2 are not in the rows of the unfolded matrix but in the columns. Depending on the dimensions put in the rows of the unfolded matrix, solar fields, loops or collector assemblies can be compared.

In the case of solar fields, X' will have the dimension $(J_4) \times (I_1 I_2 J_1 J_2 J_3)$ and a size of $4 \times (360 * 24 * 3 * 4 * 50) = 4 * 5184000$, meaning that the 4 solar fields will be compared according to the 5184000 measurements obtained for each one during the year. Since weather data is different for each solar field this can be merged, resulting a merged unfolded matrix of size $4 \times (360 * 24 * 3 * 4 * 50 + 24 * 360 * 3) = 4 * 5261760$.

In the case of parallel loops, X' will have the dimension $(J_3 J_4) \times (I_1 I_2 J_1 J_2)$ and a size of $(4 * 50) \times (360 * 24 * 3 * 4) = 200 \times 103380$. this means that the 200 parallel loops of the plant will be compared depending on the 103380 measurements collected at each one during the year. Moreover, in this case, the four solar fields can be considered as independent, so four unfolded matrices with the dimension $(J_3) \times (I_1 I_2 J_1 J_2)$ and size 50×103380 can be built to obtain four different models that compare only the parallel loops in each solar field.

In the case of collector assemblies, X' will have the dimension $(J_2 J_3 J_4) \times (I_1 I_2 J_1)$ and a size of $(4 * 50 * 4) \times (360 * 24 * 3) = 800 \times 25920$. This means that, the 800 collector assemblies of the plant will be compared depending on the 25920 measurements obtained at each one during the year. As in the previous case, the four solar fields can be considered as independent, so four unfolded matrices of the dimension $(J_2 J_3) \times (I_1 I_2 J_1)$ and the size 200×25920 can be built to obtain four different models to compare only the collector assemblies of each solar field.

In previous benchmarking examples, the production data cannot be merged since it is common to all the elements compared. This means that, if it were to be used, a number of identical measurements would have to be added at the end of each row. Something which makes no sense for benchmarking tasks. For the same reason, weather data should only be merged in the case of solar fields.

Finally, thanks to the folding and unfolding methodology proposed, some more sophisticated benchmarking possibilities can be analysed for better understand the plant. For example, consider that the structure of the four solar fields is identical and the unfolding is done to obtain an X' of the dimension $(J_3 J_2) \times (J_1 J_4 I_1 I_2)$. In this case, the matrix of the size $(50 * 4) \times (360 * 24 * 3 * 4) = 200 \times 103380$ will be useful to analyse the influence the location of the collector assemblies within the solar field

Table 2
Dwelling variables.

<i>Sani</i> (kWh)	Heating energy for Hot water for sanitary use
<i>Heat</i> (kWh)	Heating energy
<i>Cold</i> (kWh)	Cooling energy

Table 3
Production plant generation variables.

<i>Glsa</i> (kWh)	Energy for heating water for sanitary use in dwellings
<i>HRF</i> (kWh)	Energy for radiant floor heating in dwellings
<i>HFan</i> (kWh)	Energy for fan-coils heating in common areas
<i>CRF</i> (kWh)	Energy for radiant floor cooling in dwellings
<i>CFan</i> (kWh)	Energy for fan-coils cooling in common areas
<i>Gas</i> (kWh)	Gas consumption
<i>Ele</i> (kWh)	Electric consumption
<i>Slr</i> (kWh)	Solar generation

has. Similar models can be built for collector assemblies with respect to the parallel loop and/or for parallel loops with respect the solar field.

5. Exploitation example

To better illustrate the benefits of the proposed methodology, another case study using real data from a social building is presented. The building is located in downtown Barcelona (Catalonia) and consists of 32 separated dwellings, common areas and a common generation plant which is used for heating and cooling. Three modelling options derived from the different unfolding strategies from the same initial data-set and defined according to the monitoring goals will be illustrated. PCA has been used as the statistical monitoring strategy following the same principles as in Tucker (1966) (see the reference for further details on applying PCA for multi-housing building monitoring). In the next subsection the data structure will be introduced. Then, following the proposed methodology, several models built from the same initial data will be shown.

5.1. Data information

The social building consists of 32 dwellings. Dwellings are small apartments between 35.58 and 41.24 m² each, and each dwelling has its own kitchen, bathroom and one bedroom. Radiant floors heat and cool the apartments and as each dwelling has their own thermostat, the occupants can set the temperature according to their needs. The variables monitored in each dwelling are summarized in Table 2. All the variables being monitored are sampled hourly. The building has a single generation plant that serves the whole building and includes a solar field to generate hot water and three 110 kW Brotje Heizung Ecotherm Plus WGB condensation boilers. The generation plant provides hourly data on consumption and generation. Table 3 summarizes the variables monitored in the production plant.

Weather information during the period is also available through the Catalan public weather agency MeteoCat, and consist of the 11 variables summarized in Table 4. These variables present a sample time of 1 day.

Therefore, the system has three variables that are replicated in every dwelling, eight common energy variables from the generation plant and 11 weather variables.

5.2. Data folding

To be able to later merge the unfolded matrices produced any of the three data sets, a common nomenclature must first be established. The dimension I_1 is used for Hours, I_2 for Days, J_1 for Variables and J_2 for Dwellings. The sizes of each dimension for each data-set are indicated in Table 5.

Table 4
Summary of weather variables.

TM (°C)	Mean daily temperature
TX (°C)	Maximum daily temperature
TN (°C)	Minimum daily temperature
$PPT24$ h (mm)	Daily precipitation
HRM (%)	Mean daily humidity
$RS24$ h (MJ/m ²)	Global irradiation
$VVM10$ (m/s)	Mean daily wind velocity
$DVM10$ (°)	Mean daily wind direction
$VVX10$ (m/s)	Maximum daily wind speed
$DVX10$ (°)	Maximum daily wind speed direction
PM (hPa)	Mean daily atmospheric pressure

Table 5
Summary of dimensions for dwelling, generation and meteorological data.

Dimension		Dwelling	Production	Weather
I_1	Hours	24	24	(none)
I_2	Days	621	621	621
J_1	Variables	3	8	11
J_2	Dwellings	32	(none)	(none)

Finally, Eq. (1) has been applied to the three subsets considered as X matrices, to obtain three folded matrices \underline{X} . Thus, \underline{X} results in a 4D ($24 \times 621 \times 3 \times 32$) matrix for dwelling data, a 3D ($24 \times 621 \times 8$) matrix for generation data and a 2D (621×11) matrix for weather data.

The three distinct data sources present different granularity and spatial receptivity. Thus, the first and second present granularity on two levels (hour and day), while the weather data-set only has information on a daily level. Similarly, the dimension corresponding to variables in the first data-set has two levels of modularity (sensors or variables and dwellings), whereas the other two data-sets only have one (variables).

Re-sampling or aggregating variables collected at hourly rates to a daily frequency could produce losses of significant information. Instead of this, by applying the proposed methodology all the data sources are retained and used. They have been folded according to previous structures and adequately unfolded further to be merged when possible.

The main information is provided by the data from Dwellings. This data source is then the basis of the proposed models, and the two data-sets will be used as complementary information sources when needed by applying the merging operation.

5.3. Unfolding

The three \underline{X} matrices can be unfolded in several ways, depending on the monitoring goals and by following the indications in Section 3.4. In this application example, the following objectives were defined:

- Daily monitoring of the whole building
- Identify dwellings that behave similarly
- Daily monitoring of individual dwellings

The following subsections introduce three real use cases, where the corresponding unfolding and merging are described, and some results on using the methodology with these three distinct modelling scenarios are presented.

5.3.1. Unfolding for daily monitoring of the building

In the first use case, the aim is to model the building for supervision purposes to find sensor faults, leakages, poor configurations of the system, etc. This model aims to help the building manager easily obtain information about the building by using simple control charts like dashboard, and performing fault detection daily (I_2). The goal of the model is to explain the differences in the building's daily performance. Consequently, the unfolding is done by placing (I_2) in the Rows' group while I_1 , J_1 and J_2 are placed in Columns' group. Thus, an initial model with the dwelling data unfolded as (I_2) \times ($I_1 J_1 J_2$) is obtained and

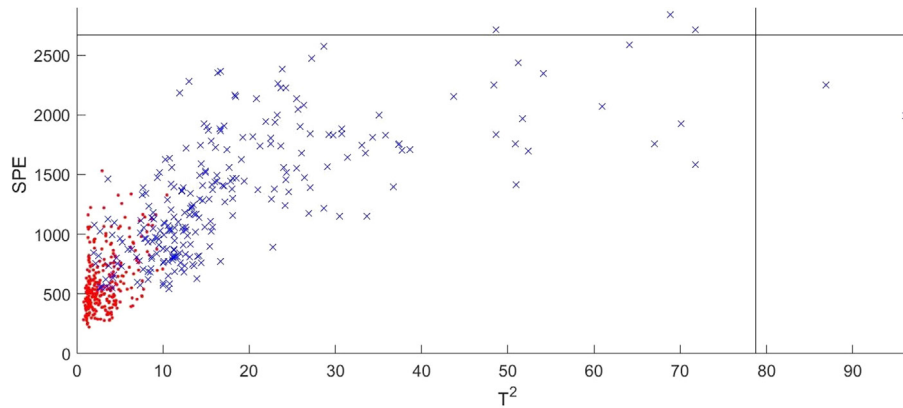


Fig. 4. Hotelling’s T^2 index vs. SPE index for daily monitoring using only dwellings data, each red point represents a summer day and each blue cross a winter day.

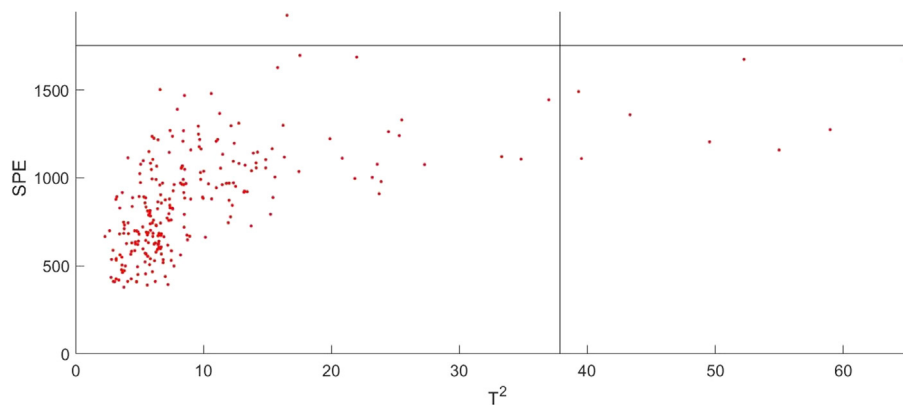


Fig. 5. Hotelling T^2 index vs. SPE index, for daily monitoring using a winter model and only dwellings data, each point represents a day in the system (only dwellings).

studied. Later, production plant data is unfolded as $(I_2) \times (I_1 J_1)$ and merged, and finally weather data is unfolded as $(I_2) \times (J_1)$ and merged to obtain more precise models by advantage of the correlations between all these variables.

5.3.1.1. *Modelling only with dwelling data.* A first result that can be easily obtained by building the model with the whole data set, is that the winter (blue dots) and summer (red dots) behaviour is totally different (see Fig. 4). This change in the behaviour is obvious, because in winter heating is consumed, whereas in summer this consumption is in cooling. Therefore, winter and summer models must be obtained and analysed separately for more accurate results. In this use case, only the winter models are shown.

Once the winter model has been obtained, classic PCA monitoring charts are then used to detect faulty days (for example, days where the correlations between distinct dwellings change from the ones modelled or days with abnormal magnitudes). Later, when a faulty day is detected, contribution analysis can be used to discriminate the variables causing the fault.

As an example of the monitoring charts provided by PCA, the Hotelling’s T^2 vs. SPE graphic is shown in Fig. 5 and Scores in Fig. 6.

Fig. 5 shows some days that surpass the limits. Such days are those that do not follow the normal behaviour modelled by PCA. Generally, days falling over Hotelling’s T^2 are magnitude faults and those falling over SPE are correlation faults.

Fig. 6 shows the score space (grey ellipsoid) and the location of each modelled day (a red point). In the Score space some groups can be found. These groups can usually be associated with distinct consumption patterns. In our case, one group with autumn and spring days can be found (generally located at the positive side of the first score ($T(1)$),

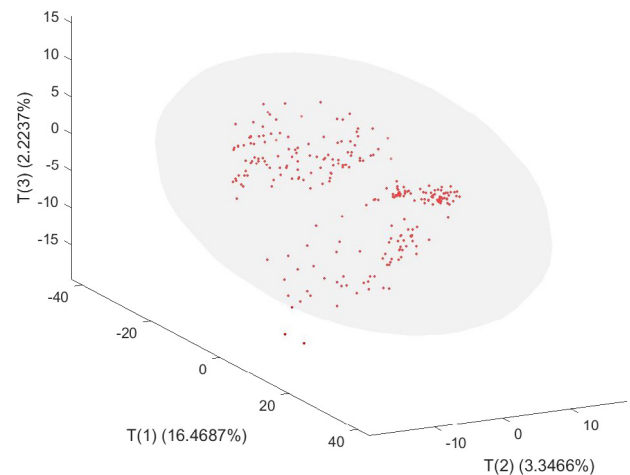


Fig. 6. Three first scores graph marked as T(Number of score)(% of variance), for daily monitoring using a winter model and only dwellings data, each point represents a day in the system (only dwellings).

grouped and near the centre of the model), whereas winter days are more dispersed.

5.3.1.2. *Modelling with dwelling data merged with production plant and weather data.* According to the methodology and by merging the data from the production plant using the merging procedure, it is possible to attain the same control charts (Figs. 5 and 6), but these now include

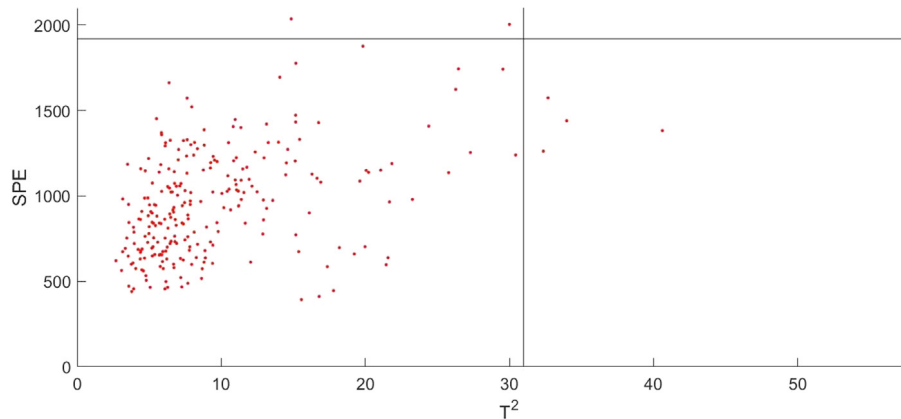


Fig. 7. Hotelling's T^2 index vs. SPE index, for daily monitoring using a winter model and dwellings+production data, each point represents a day in the system (only dwellings and production plant).

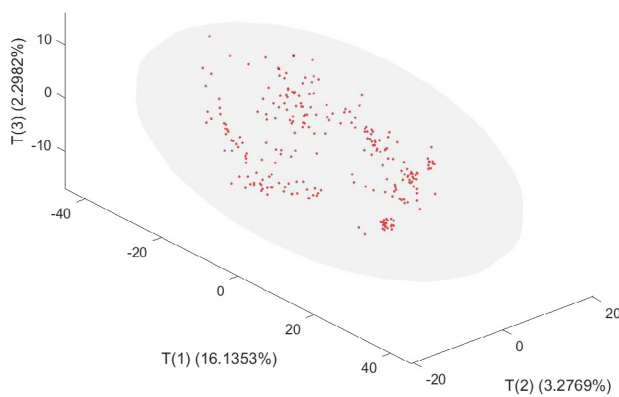


Fig. 8. Three first scores graph marked as T(Number of score)(% of variance), for daily monitoring using a winter model and dwellings+production data, each point represents a day in the system (only dwellings and production plant).

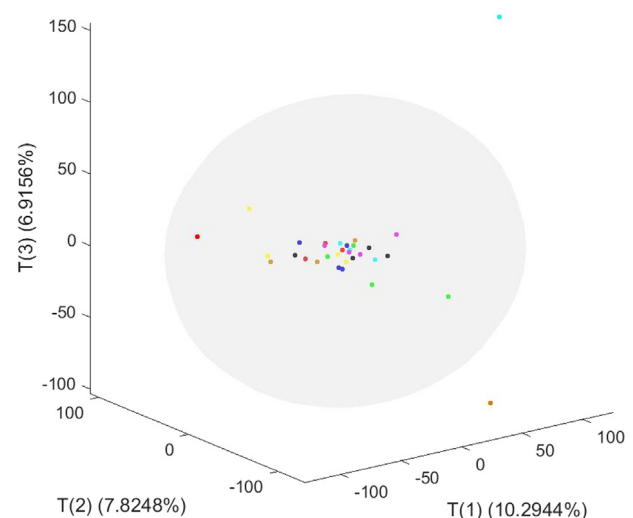


Fig. 9. Three first scores graph marked as T(Number of score)(% of variance), for identify dwellings similarities and using dwellings data, each point represents a dwelling coloured for orientations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

production plant consumption and generation. The resulting control charts are shown in Fig. 7 for Hotelling's T^2 vs. SPE chart and scores in Fig. 8. The model now includes the data from dwellings and production. By introducing this data, it is now possible to detect poor configurations or errors in the production area.

In Fig. 7, some small changes occur when including the production data, limits are now a little bit lower, some of the previous days near the limits now fall inside the control area while others fall outside. These small changes are due to the information from the production plant calendar and the production settings have been indirectly introduced to the model. Days with errors in the production area or poor configurations are not present during the monitored period, so there are no great changes in the control chart.

On the other hand, in Fig. 8 it is now possible to see distinct groups within the previous autumn and spring group (groups are also located in the positive side of the first score ($T(1)$) axis as in the previous model), these groups are caused by the distinct production configurations.

Finally, weather data is added using the methodology's merge procedure. The model now includes the data from dwellings, production and weather. Consequently, the control charts Hotelling's T^2 vs. SPE and scores will also include the merged data. In this case, weather does not introduce any new detail into the model since the production plant gathers correlated behaviours. However, it can be used to differentiate faults from extreme behaviours caused by weather variations.

5.3.2. Unfolding for identify dwellings similarities

In this second use case, the aim is to benchmark the consumption of the dwellings. Thus, the model does not aim to monitor dwellings on a

daily scale, but rather give global information to find the similarities and differences between them. This information can be useful for understanding the system and managing energy more efficiently. Using this model, it is possible to attain information about suspicious or abnormal user behaviours, similarities between users, and also information about the building itself, for instance, finding relationships between the consumption of dwellings that have similar locations (orientation, floor, etc.). Thus, in this case (J_2) is placed in Rows' group and the rest of the dimensions in Columns', resulting in the unfolding structure (J_2) \times ($I_1 I_2 J_1$). Note that weather and production plant matrices cannot be appended as they do not have the J_2 dimension.

The scores chart obtained from this model is shown in Fig. 9. Apartments on the corners (two external sides) or with poor orientation tend to have behaviours distant from the centre of the model.

In a similar plot, coloured according to floor (Fig. 10), it can be seen how the first floor presents the most distant behaviour to the centre of the model. This is because of the influence of the facilities located on the ground floor. Meanwhile, the second and third floors, except for a few outliers (probably due to the habits of the occupants), present similar and statistically normal behaviours. Finally, the fourth floor also presents a behaviour more distant from the centre of the model. This last

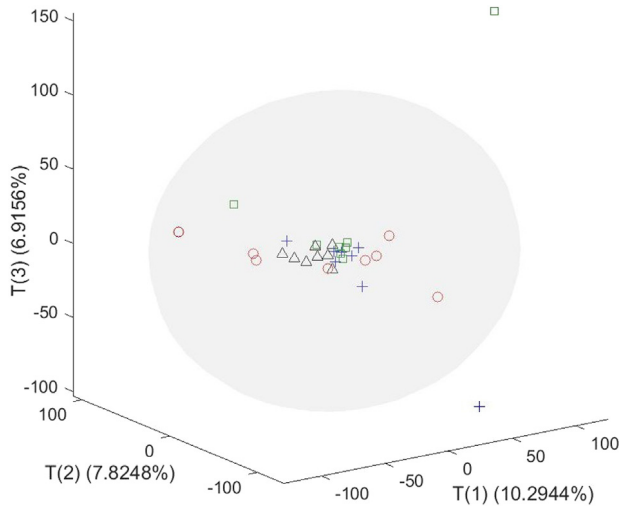


Fig. 10. Three first scores graph marked as T(Number of score)(% of variance), for identify dwellings similarities and using dwellings data, each marker represents a dwelling differentiated by floors. Red circle first floor, blue cross second floor, green square third floor and grey triangle for fourth floor.

behaviour can be explained by the influence the roof isolation has on their consumption.

5.3.3. Unfolding for daily monitoring of dwellings

This third use case, aims to monitor the building, but explains every dwelling separately albeit without losing sight of the whole building. Traditionally, modelling would be done separately for each dwelling, but here the same methodology will be applied without losing the information about the rest of the building. Initial data was divided into distinct X matrices, preserving the singularities described in the previous step (folding). Since the aim is to explain the differences between dwellings, (J_2) is placed in Rows and (I_2) is also placed in Rows to preserve the daily resolution. The others will be reorganized as Columns, thus obtaining $(I_2 J_2) \times (I_1 J_1)$ as the desired unfolding. Note that in this third use case, as in the second one, the unfolding can only be reached using the 4D matrix of the dwellings. Weather and production plant matrices do not have the J_2 dimension so they cannot be merged.

Once unfolded, as in the first example, the model is focused only on winter so cooling production and consumption variables are deleted. Also, the non-winter days are avoided when building the model.

By plotting their scores, this model allows how dwellings behave on a day scale to be compared. See Fig. 11 which shows the behaviour of a first-floor corner apartment (dwelling 1 in red dots) in comparison to a third floor non-corner apartment (dwelling 18 in blue crosses). The corner dwelling presents a larger variability and more outliers than the non-corner one over the winter period observed.

6. Conclusions

In this paper a new methodology to deal with N -dimensional data for monitoring through PCA models has been presented. First, it is assumed that, from the monitoring point of view, multidimensional is caused by data modularity (repetition of variables) and granularity (periodicity in time). From the point of view of granularity, the method deals with the possibility of organizing detailed observations on different levels and performing monitoring accordingly. In a similar way, the method is general enough to consider multiple levels of modularity in a way that, for a given level, the monitoring variables in a module contain repetitions of those contained in the level immediately inferior. To guide users when choosing their desired unfolding data organization that does not lose any information and respects the original data structure, the

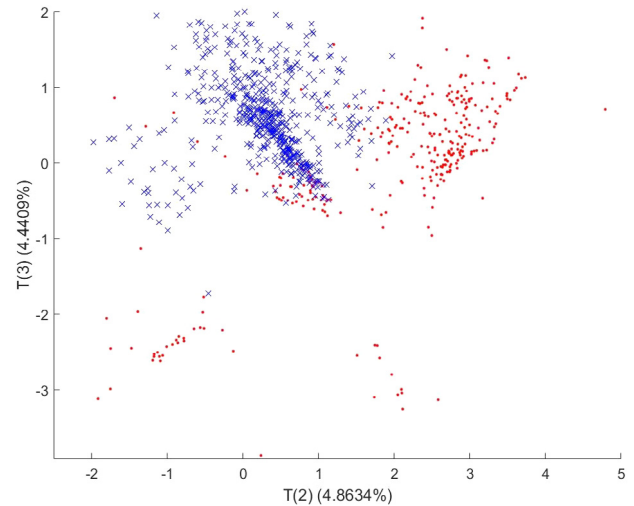


Fig. 11. Two first scores graph marked as T(Number of score)(% of variance), for daily monitoring of each dwelling using a winter model and dwellings data. Each point represents a day in a dwelling. Red dots are from dwelling 1 (first floor corner) and blue crosses are from dwelling 18 (third floor non corner).

methodology provides a step-by-step explanation of the process to be applied before applying PCA. It also includes standardizations and the possibility of merging data-sets with different granularity or modularity.

The application example demonstrates how, by applying the methodology to a single set of data from a parabolic trough solar power plant, many different models can be obtained. These models can have many different purposes, including monitoring or even benchmarking the plant.

The exploitation example, using real data from a social building located in down-town Barcelona (Catalonia), shows the possibilities the proposed methodology has. From same data it is possible to reach distinct unfolding (and then PCA models) that offer different monitoring points of view for the same system (the building). The three different use cases show how different models are obtained and how both classical and new monitoring possibilities are achieved.

Acknowledgements

This work has been carried out by the research group eXIT (<http://exit.udg.edu>), funded through the following projects: MESC project (Ref. DPI2013-47450-C21-R) and its continuation CROWDSAVING (Ref. TIN2016-79726-C2-2-R), both funded by the Spanish Ministerio de Industria y Competitividad within the Research, Development and Innovation Program oriented towards the Societal Challenges, and also the project Hit2Gap of the Horizon 2020 research and innovation program under grant agreement N680708. The author Llorenç Burgas would also like to thank Girona University for their support through the competitive grant for doctoral formation IFUdG2016.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.engappai.2018.02.013>.

References

- Aguado, Daniel, Rosen, Christian, 2008. Multivariate statistical monitoring of continuous wastewater treatment plants. *Eng. Appl. Artif. Intell.* 21 (7), 1080–1091.
- Bettini, Claudio, Dyreson, C.E., Evans, W.S., Snodgrass, R.T., Wang, X.S., 1998. A glossary of time granularity concepts. In: *Temporal Databases: Research and Practice*, Vol. 1399. pp. 406–413.

- Bro, Rasmus, 1997. Parafac. tutorial and applications. *Chemometr. Intell. Lab. Syst.* 38 (2), 149–171.
- Burgas, Llorenç, Melendez, Joaquim, Colomer, Joan, 2014. Principal component analysis for monitoring electrical consumption of academic buildings. *Energy Procedia* 62, 555–564.
- Burgas, Llorenç, Melendez, Joaquim, Colomer, Joan, Massana, Joaquim, Pous, Carles, 2015. Multivariate statistical monitoring of buildings. case study: Energy monitoring of a social housing building. *Energy Build.* 103, 338–351.
- Camacho, José, Picó, Jesús, Ferrer, Alberto, 2008a. Bilinear modelling of batch processes. Part I: Theoretical discussion. *J. Chemom.* 22 (5), 299–308.
- Camacho, José, Picó, Jesús, Ferrer, Alberto, 2008b. Bilinear modelling of batch processes. Part II: A comparison of PLS soft-sensors. *J. Chemom.* 22 (10), 533–547.
- Chang, I., Carroll, J.D., 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of and eckart-young decomposition. *Psychometrika* (35), 283–319.
- Chen, Jiun-Hung, Shapiro, Linda G., 2009. PCA vs. tensor-based dimension reduction methods: An empirical comparison on active shape models of organs. *IEEE Eng. Med. Biol. Soc.* 2009, 5838–5841.
- Edward Jackson, J., Mudholkar, Govind S., 1979. Control procedures for residuals associated with principal component analysis. *Technometrics* 21 (3), 341–349.
- Esbensen, K., Wold, S., Geladi, P., Öhman, J., 1987. Multi-way principal components-and pls-analysis. *Chemometrics* 1 (1), 41–56.
- González-Martínez, Jose Maria, Camacho, Jose, Ferrer, Alberto, 2014. Bilinear modeling of batch processes. Part III: Parameter stability. *J. Chemom.* 28 (1), 10–27.
- Haimi, Henri, Mulas, Michela, Corona, Francesco, Marsili-Libelli, Stefano, Lindell, Paula, Heinonen, Mari, Vahala, Riku, 2016. Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. *Eng. Appl. Artif. Intell.* 52, 65–80.
- Harshman, Richard A., 1970. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. In: *UCLA Working Papers in Phonetics*, 16(10): 1–84.
- Kiers, Henk A.L., 1991. Hierarchical relations among three-way methods. *Psychometrika* 56 (3), 449–470.
- Kourti, Theodora, 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. *Internat. J. Adapt. Control Signal Process.* 19 (4), 213–246.
- Lu, Haiping, Plataniotis, Konstantinos N., Venetsanopoulos, Anastasios N., 2008. MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural Netw.* 19 (1), 18–39.
- Nomikos, P., MacGregor, J.F., 1994. Monitoring batch processes using multiway principal component analysis. *AIChE* 40 (8), 1361–1375.
- Peter He, Q., Wang, Jin, Shah, Devarshi, Vahdat, Nader, 2017. Statistical process monitoring for iot-enabled cybermanufacturing: Opportunities and challenges. *IFAC-PapersOnLine* 50 (1), 14946–14951. 20th IFAC World Congress.
- Russell, E., Chiang, L.H., Braatz, R.D., 2000. *Data-Driven Methods for Fault Detection and Diagnosis in Chemical Processes*, Vol. 49. Springer, London.
- Tucker, LedyardR, 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31 (3), 279–311.
- Westerhuis, Johan A., Kourti, Theodora, MacGregor, John F., 1999. Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemom.* 13 (3–4), 397–413.
- Yu, T., Wang, X., Shami, A., 2017. Recursive principal component analysis based data outlier detection and sensor data aggregation in iot systems. *IEEE Internet Things J.* PP (99), 1–1.

Chapter 5

Integrated Unfold-PCA monitoring application for smart buildings: An AHU application example.

In this chapter, the implementation of a web application based on the Unfold-PCA methodology is described and a demonstration over a real AHU data from NUIG Alice Perry Building. This publication has been published in the following paper:

Paper is under review in the **Applied Energy**

Reference: APEN-D-19-04710

JCR IF(2017): 7.9

Q1(8/97) - Energy and Fuels

Q1(4/137) - Chemical Engineering

Chapter 6

Main results and discussion

This thesis contains three distinct main contributions. The first encompasses adapting an existing methodology (PCA) to monitor buildings, the second formulates a fold and unfold strategy to provide new modelling options for granular and modular (building) systems, while the third and final centres on providing a monitoring tool and validating the methodology. The results for each of these are discussed separately in the following sections.

6.1 New PCA methodology for building monitoring

A PCA-based approach has been validated as a suitable technique for modelling building behaviours which have multiple sensors. This achievement has been included in Chapter 3 of this thesis and is one of the articles making up this compendium [4]. The method has proven capable of gathering the influence different factors (represented by acquired variables) have on energy demand profiles.

The proposed method extends the capabilities of PCA to provide a complete monitoring strategy according to the following steps:

1. A reference model of the normal operating conditions of the building is obtained with historic data. The method allows the data to be organised in two ways, which offers two complementary views for monitoring.
2. Monitoring takes advantage of this reference model to detect uncommon deviations. Two unique and complementary control charts are required during monitoring, (independent of the number of variables being monitored), based on the statistics T^2 and SPE.

3. Fault detection: the quality of an observation with respect to the reference model is based on a statistical decision criteria that consists of evaluating a threshold that represents the coverage the distribution of these statistics (T^2 and SPE) has during normal operating conditions.
4. Fault isolation: once a fault has been detected, the methodology proposed allows for the variables and time instants responsible for this out-of-control situation to be identified. This method is known as contribution analysis.

The proposed PCA monitoring strategy has been validated in a social-housing building located in down-town Barcelona. In this real-life application example, the building has 96 dwellings and three variables per apartment that have been monitored daily over an (approximately) two-year period along with external weather data containing eleven available variables. In applying this monitoring strategy, previously-defined and distinct abnormal conditions are able to be detected.

For example, when modelling the whole building in a daily monitoring scenario, the PCA methodology can detect (as can be observed in the SPE chart in Figure 6.2) that variables corresponding to hot water volume and hot water energy for dwelling 45, and also in much lower measures in dwellings (7, 16, 21, 43, 46, 55, 67, 78, 79, 82 and 89) present abnormal values that day. Looking at the original variables for dwelling 45, the values for hot water volume and hot water energy can be confirmed as being 203L per day, which is higher than the usual mean consumption for this dwelling which is around 100L per day.

6.2 Fold-unfold strategy for granular and modular systems monitoring

After observing that the behaviour of consumption and other monitored variables presented some kind of repeatability in the time patterns, we also notice that many buildings have certain modularity that was reflected in the existence of replications of groups of sensors according to the organisation of technical subsystems (e.g. temperatures, lighting, partial consumptions, etc.) of the building's structure. This observation suggested that the previous unfolding method from 3D matrices could be extended to any data organisation for data in N-dimensional matrices. For this reason it was necessary to formulate a method to systematise the folding and unfolding methodology of these new N-dimensional organisation of data from buildings. This contribution is included in Chapter 4 as one of the articles of this compendium [1].

In Chapter 4, and assuming that buildings can be granular systems and/or modular systems, the two concepts of granular system and modular system are formally

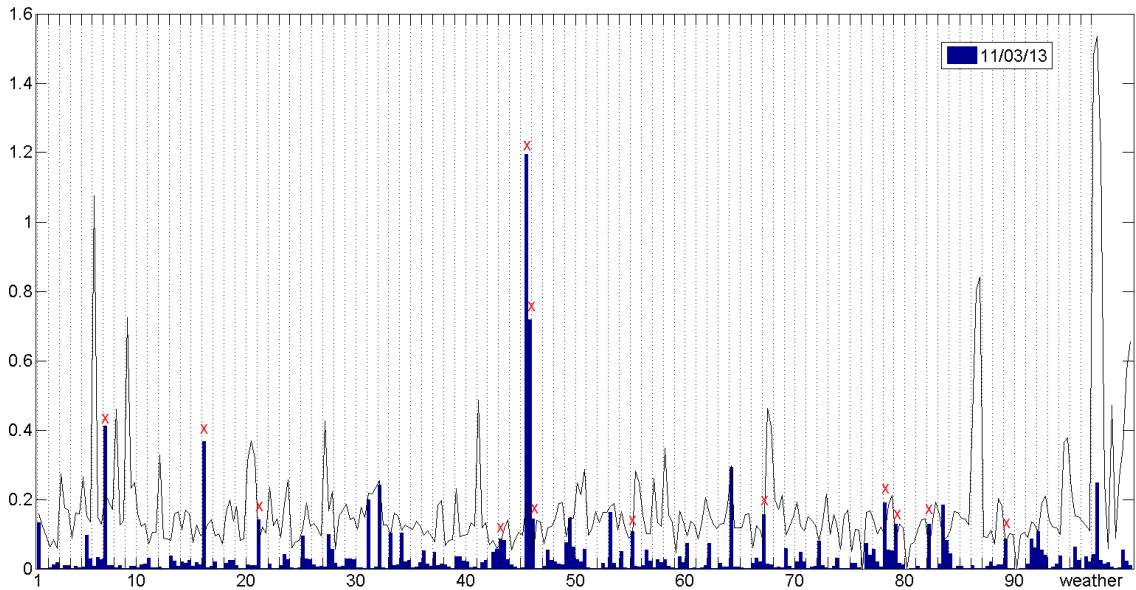


Figure 6.1: SPE contributions for day 11, March 2013 for a daily monitoring of the whole building

defined as:

- A granular system that refers to the possibility of organizing observations on different levels of temporal detail (daily, weakly, season, year, etc.).
- A modular or multi-entity system that is a system where several identical systems with interactions among them can be found (rooms, dwellings, etc.).

These kinds of systems presenting granular and/or modular characteristics (buildings) can be represented more accurately in N-dimensional arrays but PCA, like many other algorithms, needs two-dimensional data matrices as input. To perform a more accurate modelling and posterior monitoring of such systems, a folding and unfolding methodology is mathematically formulated. The fold step is used for organising a 2D data input, usually directly from sensors, into a N-Dimensional array. Then, the unfold step is used to reshape this N-Dimensional array to a new 2D matrix with data in the correct order to reach the desired model using PCA. By applying the proposed methodology, new modelling and monitoring possibilities emerge in granular and modular systems.

The proposed folding and unfolding methodology has been validated in a multi-apartment social-housing building in down-town Barcelona. In this real application example, the building has 32 almost identical apartments with three variables per

apartment being monitored, as well as central heating production with eight variables and an external weather station providing eleven variables. Three distinct monitoring goals have been defined for this data:

- Daily monitoring of the whole building.
- Identifying dwellings that behave similarly.
- Daily monitoring of individual dwellings.

Following the folding and unfolding methodology, three different 2D matrices offering new different monitoring points of view are achieved. For example, daily monitoring of the individual dwellings can be carried out without losing the whole-building perspective. See Figure 6.2, where each point represents a day in a dwelling. The red dots are from dwelling 1 (first floor corner) and the blue crosses are from dwelling 18 (third floor non corner).

Furthermore, in Chapter 4 a simulated example demonstrates how, by applying the methodology to a single set of data from a parabolic trough solar power plant, data can be folded into a 6D data matrix and then unfolded into 31 distinct unfolds to reach 31 meaningful modelling options. Data from the production area or weather data can also be added.

6.3 Implementing the building monitoring methodology

Finally, to conclude the theoretical work in this thesis, the previous achievements have been implemented in a common framework in order to validate the usability with different building typologies. From this implementation (see Chapter 5), the following results can be highlighted:

- The application outlined for modelling, monitoring and fault isolation on buildings is programmed and tested.
- The application is suitable for non PCA expert use to monitor buildings.
- The module and methodology support continuous monitoring and can be used to implement monitoring tools required by energy management procedures such as the energy management procedures ISO50001 or ISO 50006:2014 and can also be used to support audits (EN16247-1, ISO 50002) and performance measure and verification protocols (IPMVP, ISO 50015:2014).

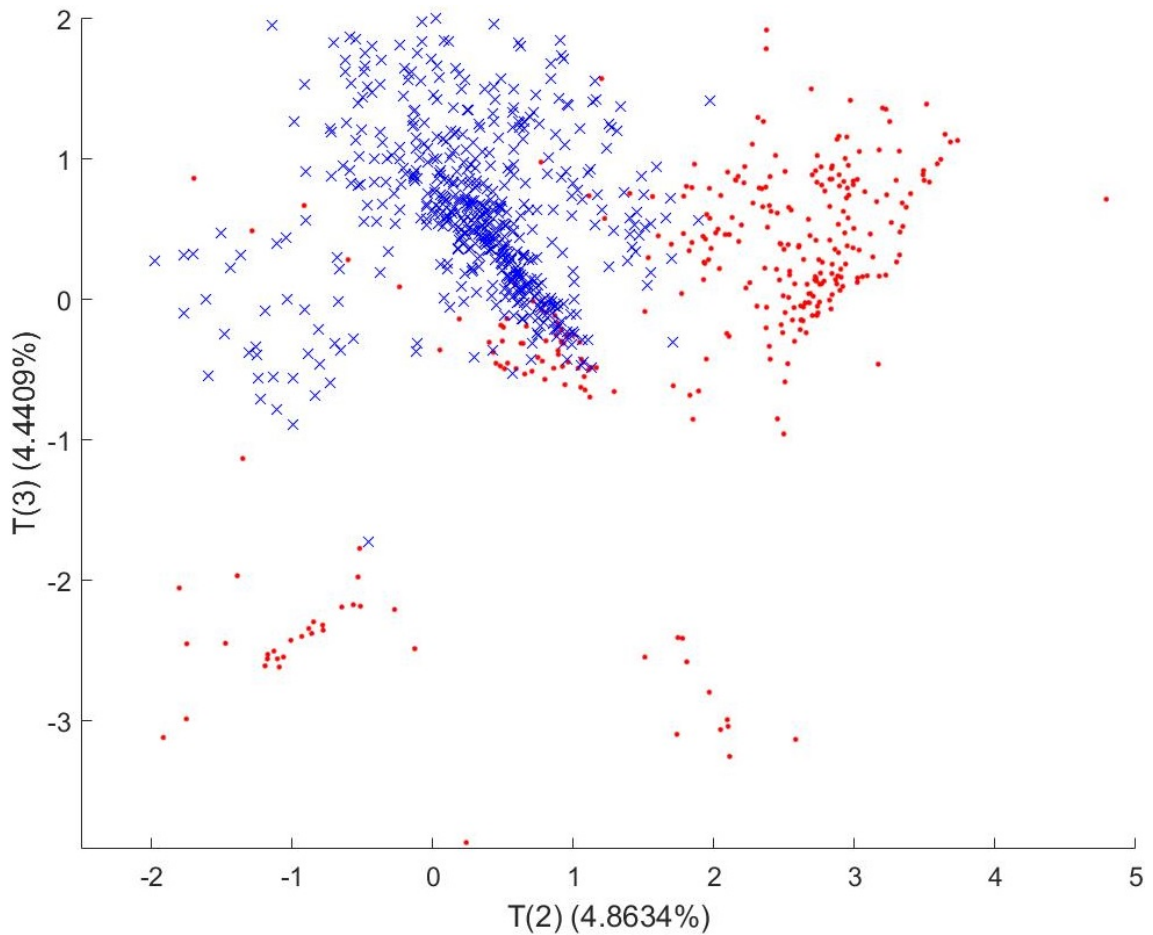


Figure 6.2: Second and third scores where each point represents a day in a dwelling, Red dots are from dwelling 1 (first floor corner) and blue crosses are from dwelling 18 (third floor non corner)

- An application example on a pilot site in the HIT2GAP project is presented. In the application example, a lecture theatre in the Alice Perry Building at NUI Galway is monitored using the Unfold-PCA methodology and the online application.

Chapter 7

Conclusions

In this thesis, PCA has been introduced and successfully adapted to model, monitor, and isolate abnormal behaviour in buildings as granular systems. As was initially observed, buildings generally present granular and/or modular characteristics and many factors influence their consumption levels. In order to achieve the goal of improving the exploitation of data gathered in smart buildings by providing better modelling and monitoring tools, four separated objectives were defined. The following sections describe the conclusions drawn for each sub-objective.

7.1 Define a data-driven methodology for supervising smart-buildings

The first sub-objective was achieved and presented mainly in Chapter 4, where a PCA-based modelling and monitoring methodology for dealing with a building's data is presented. PCA means an enormous set of data coming from the buildings can be summarised into two simple monitoring charts Hotelling's T^2 and SPE statistics based. Furthermore, when a detection is made, the variables causing the misbehaviour can be isolated by means of Hotelling's T^2 and SPE Contributions and graphically visualized.

7.2 Extensions to consider granularity and modularity

The second sub-objective was achieved in Chapter 5, where a folding and unfolding methodology is defined and mathematically formulated to take advantage of the in-

herent granularity and modularity of buildings. By using this folding and unfolding methodology, the granular behaviours and modularities in buildings can be used to reach new modelling and monitoring capabilities. The granularity and modularities are rarely exploited by current building monitoring techniques, as they usually consider buildings as a set of uncorrelated sub-systems. The use of these granularities and modularities for modelling brings valuable information and new modelling capabilities (as seen in Chapter 5).

7.3 Validating the methodology

Covering the third sub-objective of validating the methodology in different scenarios, many use cases of the PCA based methodology have been presented in Chapters 3, 4 and 5 and are summarized in the following bullet points:

- In Chapter 4, a real use case from a social-housing building located in down-town Barcelona was presented. This building has a centralized hot water production and consumption system for 96 dwellings and user programmable thermostats. The use case is made using three variables for each building and common variables such as weather. In the use case, two distinct models are used: one for modelling dwellings and another for comparing dwellings.
- In Chapter 5 an example of exploitation of real data from a social-housing building located in down-town Barcelona was presented. In this example, three real use cases are studied using the same data, including the new modelling capabilities. In the three use cases, the same data from a social-housing building with 32 dwellings spread out over four identical floors (eight identical dwellings per floor), as well as central heating production and weather variables are used. The monitoring capabilities explored were:
 - daily monitoring of the whole building
 - identifying dwellings that behave similarly
 - daily monitoring of the dwellings but considering the interactions among them
- In Chapter 5, a theoretical use case is also studied on a parabolic trough solar power plant. In this use case, how 31 meaningful modelling options can be obtained from a unique data base by applying a fold and unfold strategy is demonstrated.
- In Chapter 6, applying the methodology as an AHU monitoring system is presented. In this use case, real data from a multi-purpose lecture theatre

located in the Alice Perry Building at NUI Galway (one of the pilot sites in the HIT2GAP project) is used.

Also, during the duration of this thesis and presented as conference publications two different scenarios of the use of the PCA based methodology were presented.

- In [3], a use case using real data for modelling the energy usage of the University of Girona Montilivi Campus is presented.
- In [2], a use case using real data uses the methodology to detect heating/cooling transitions in a social-housing building in Barcelona is described.

7.4 Providing a technological solution

The final technical sub-objective was achieved and presented in Chapter 6 where the implementation as an end-user online web application of parts of the theoretical work included in this thesis. This web application makes it easy for non PCA experts to monitor buildings and also simplifies the model creation tasks by providing online modelling and monitoring tools.

7.5 Future Work

In this thesis, PCA has been introduced and successfully adapted to model, monitor, and isolate abnormal behaviour in buildings as granular systems. As seen PCA is able to perfectly deal with buildings data and in conjunction with the purposed folding and unfolding methodology new points of view can be reached.

This thesis is mainly focused on building's scope but the purposed methodology is general enough to be applied in other scopes where granular systems can be found. From now on, new research efforts will be focused in finding new application fields of the folding and unfolding methodology. Following the example included in this thesis of Solar power plant monitoring and the ongoing research made in the IMAQUA and RESOLVD projects, where the methodology is being applied in water distribution networks and electric distribution networks.

By studying the use of the folding and unfolding methodology in conjunction with other datamining techniques. This is possible because the folding and unfolding methodology is a data pre-processing method for dealing with multidimensional data with two dimensional techniques. For example some tests are carried out using Case Based Reasoning in conjunction with PCA in the IMAQUA project.

Finally, another idea that came up during this thesis that should be studied in a future is the possibility of scaling data directly over the N-dimensional folded data matrix being able to perform a mean form subtraction in each of the dimensions according to the modelling needs. And the possibility of including the information used to fold and unfold the data to present better contribution analysis.

Bibliography

- [1] L. Burgas, J. Melendez, J. Colomer, J. Massana, and C. Pous. N-dimensional extension of unfold-PCA for granular systems monitoring. *Engineering Applications of Artificial Intelligence*, 71, 2018.
- [2] Llorenç Burgas, Joan Colomer, and Joaquim Melndez. Modelling transitions on heating usage in buildings with multivariate statistical monitoring. In *Proceedings - EUROCON 2015*, 2015.
- [3] Llorenç Burgas, Joaquim Melendez, and Joan Colomer. Principal component analysis for monitoring electrical consumption of academic buildings. In *Energy Procedia*, volume 62, pages 555–564, 2014.
- [4] Llorenç Burgas, Joaquim Melendez, Joan Colomer, Joaquim Massana, and Carles Pous. Multivariate statistical monitoring of buildings. Case study: Energy monitoring of a social housing building. *Energy and Buildings*, 103:338–351, 2015.
- [5] European Parliament and the Council of 25 October 2012 on energy efficiency. Directive 2012/27/EU , 2012.
- [6] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, and P. Patargias. Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy*, 87(6):2079–2086, June 2010.
- [7] Magdalena Hajdukiewicz, Marcus Keane, B O’Flynn, and W O’Grady. Formal calibration methodology for CFD model development to support the operation of energy efficient buildings. In *Tenth International Conference for Enhanced Building Operations, Kuwait*, 2010.
- [8] M. Kavgic, a. Mavrogianni, D. Mumovic, a. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic. A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment*, 45(7):1683–1697, July 2010.

- [9] Joseph C. Lam, Kevin K.W. Wan, and K.L. Cheung. An analysis of climatic influences on chiller plant electricity consumption. *Applied Energy*, 86(6):933–940, June 2009.
- [10] Joseph C. Lam, Kevin K.W. Wan, K.L. Cheung, and Liu Yang. Principal component analysis of electricity use in office buildings. *Energy and Buildings*, 40(5):828–836, January 2008.
- [11] Joseph C. Lam, Kevin K.W. Wan, S.L. Wong, and Tony N.T. Lam. Principal component analysis and long-term building energy simulation correlation. *Energy Conversion and Management*, 51(1):135–139, January 2010.
- [12] Shun Li and Jin Wen. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. *Energy and Buildings*, 68:63–71, January 2014.
- [13] Demba Ndiaye and Kamiel Gabriel. Principal component analysis of the electricity consumption in residential dwellings. *Energy and Buildings*, 43(2-3):446–453, February 2011.
- [14] B O’Flynn, Essa Jafer, and R Špinar. Development of miniaturized wireless sensor nodes suitable for building energy management and modelling. In *ECPPM, Ireland*, 2010.
- [15] Kyunggyu Park, Y Kim, Seonmi Kim, K Kim, Wookhyun Lee, and Hwachoon Park. Building Energy Management System based on Smart Grid. In *2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC)*, pages 1–4. Ieee, October 2011.
- [16] Andrea Costa Rafferty Paul, Marcus Keane, James O’Donnell. Energy Monitoring Systems value, issues and recommendations based on five case studies. In *Clima conference, Antalya, Turkey*, 2010.
- [17] Paul Raftery, Marcus Keane, and James O’Donnell. Calibrating whole building energy models: An evidence-based methodology. *Energy and Buildings*, 43(9):2356–2364, September 2011.
- [18] Lukas G. Swan and V. Ismet Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, 13(8):1819–1835, October 2009.
- [19] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics*, 13(3-4):397–413, 1999.

-
- [20] Zhun Yu, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, 43(6):1409–1417, June 2011.
- [21] Hai-xiang Zhao and Frédéric Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, August 2012.