UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA CHIMICA E DEI PROCESSI INDUSTRIALI

Tesi di Laurea Magistrale in
Ingegneria Chimica e dei Processi Industriali

DATA ANALYSIS THROUGH MULTIVARIATE STATISTICAL TECHNIQUES: AN INDUSTRIAL APPLICATION

Relatore: Prof. Massimiliano Barolo
Correlatore: Dott . Riccardo De Luca
Correlatore: Dott . Mario Cammarota

Laureanda: DOINA JIGNEA

ANNO ACCADEMICO 2019-2020

Abstract

This Thesis has the objective of proposing an industrial procedure tailored to the company Unox S.p.A. for data analysis through advanced statistical techniques. Multivariate statistical methodologies are used to analyse the data collected from the ovens during cooking and washing processes. Process variables are analysed for both process understanding and data-driven design. The design goal is to create a simple and intuitive user-machine interface in order to improve the user experience. First, cooking and washing modes are analysed in order to: *(i)* identify the least frequently used ones for future elimination or integration; *(ii)* propose further development and improvement of the most used ones (cooking types 1 and 2, selected in 85% of the cases). Then, we proposed a multivariate statistical approach for both process monitoring and predictive maintenance through exemplificative case studies. Process monitoring consists in real-time tracking of process condition through online measurements of process variables, whereas predictive maintenance is a maintenance strategy consisting in data-based equipment failure prediction. Process monitoring has been implemented for a specific oven through PCA method. The obtained process model allowed to describe 93% of data variability and to detect anomalies during cooking processes with respect to normal operating conditions in order to ensure the final product quality. Predictive maintenance has been implemented through a PLS-DA model based on process variables. Specific equipment failures have been predicted by identifying abnormal patterns in the process variables for two case studies: failures of the gasket and the core probe. In both cases, two principal patterns leading to the technical intervention have been identified; the gasket patterns described 87% of the cases, whereas 71% of the core probe substitutions followed the defined patterns. Moreover, the model has been used for prediction in the second case study. It predicted correctly 87.5% of the cooking programs sequences.

Riassunto

Nell'era dell'Industria 4.0 i sistemi produttivi vengono dotati di un numero sempre maggiore di sensori e apparecchiature per la misura di differenti tipologie di variabili. Queste ultime rappresentano lo stato di uno specifico sistema e ne possono evidenziare eventuali malfunzionamenti o fasi non ottimizzate. La grande quantità di dati raccolti a frequenze elevate è pertanto una fonte preziosa di informazioni, la cui analisi può tuttavia risultare complessa e richiedere tempi molto lunghi. Attraverso l'utilizzo di tecniche statistiche multivariate si possono estrarre informazioni per l'ottimizzazione di processi e servizi, al fine di diminuire i costi di processo e manutenzione e migliorare la qualità dell'offerta in un mercato competitivo. L'obiettivo di questa Tesi consiste nel proporre per l'azienda Unox S.p.A. una procedura su misura finalizzata all'analisi dei dati attraverso tecniche statistiche avanzate. Unox è leader mondiale nel mercato dei forni professionali. Questi forni sono dotati di numerosi sensori per la misura delle variabili durante i processi di cottura e i dati raccolti vengono inviati attraverso Internet al *cloud* dell'azienda, dove vengono compressi e archiviati. In questa Tesi sono stati presi in considerazione dati relativi alla serie di forni MIND.Maps™ e, in particolare, alle categorie di forni Chef Top e Baker Top, appartenenti a tale serie. I risultati ottenuti dimostrano come la procedura di analisi dati proposta è in grado di fornire informazioni preziose per il monitoraggio e la predizione del comportamento delle singole apparecchiature, nell'ottica di un miglioramento continuo delle strategie di progettazione di prodotto e dei servizi offerti al cliente.

La Tesi è organizzata secondo il seguente schema concettuale.

Nel Capitolo 1 si presenta un'introduzione relativa al panorama industriale attuale e alle recenti innovazioni introdotte a livello di impresa con il concetto di Industria 4.0; successivamente si descrive brevemente la realtà aziendale di Unox S.p.A, azienda che ha fornito i dati oggetto del lavoro di Tesi, evidenziando le tipologie di forno prodotte e le opzioni di cottura/lavaggio implementate nel software installato nelle apparecchiature. Nella parte finale del capitolo si presentano gli obiettivi della Tesi.

Nel Capitolo 2 vengono descritti i fondamenti matematici e i campi di applicazione delle due principali tecniche statistiche multivariate (PCA e PLS) utilizzate per l'analisi dei dati.

Nel Capitolo 3 vengono analizzati i dati storici dei processi di cottura e di lavaggio registrati per tutte le apparecchiature connesse al *cloud*. Attraverso l'analisi dell'impiego dei forni sono

state individuate le funzioni e le modalità più utilizzate e quelle meno utilizzate. Per quelle meno utilizzate l'eliminazione potrebbe essere presa in considerazione, mentre quelle più usate possono essere ulteriormente sviluppate o modificate. Infatti, l'obiettivo principale è quello di semplificare l'interfaccia del pannello per migliorare l'esperienza dell'utente.

Per uno studio più approfondito dei dati vengono utilizzate le tecniche statistiche multivariate PCA e PLS. Questi metodi vengono implementati per il monitoraggio dei processi e la manutenzione predittiva. Il monitoraggio del processo (Capitolo 4) viene utilizzato per identificare anomalie nei programmi di cottura rispetto alle normali condizioni operative. Per ottenere un monitoraggio di alta precisione, è stato ottenuto un modello specifico per un singolo forno. Il modello di processo ottenuto riesce a catturate il 93% della varianza dei dati e viene implementato per il controllo del processo stesso al fine di ottenere le specifiche di prodotto desiderate. Infine, viene implementata la manutenzione predittiva allo scopo di includere la prevenzione nei servizi post vendita. Il Capitolo 5 illustra l'applicazione della manutenzione predittiva a due casi studio: la sostituzione della guarnizione e il malfunzionamento della sonda al cuore. Attraverso l'applicazione della tecnica PLS-DA sono stati ottenuti due modelli che individuano i pattern che seguono le variabili di processo durante le cotture che precedono l'intervento tecnico. In entrambi i casi, vi sono principalmente 2 pattern che descrivono 87% e 71% dei casi per l'intervento alla guarnizione e alla sonda al cuore rispettivamente. Inoltre, con il modello della sonda al cuore si è riusciti a predire correttamente l'87.5% delle sequenze di cotture. La previsione dei guasti delle apparecchiature viene implementata per evitare l'interruzione del funzionamento del forno programmando gli interventi tecnici. Inoltre essa consente di effettuare la manutenzione solo quando vi è effettivamente la necessità diminuendo i costi rispetto alla manutenzione preventiva.

Contents

Introduction

The fourth industrial revolution brings in advanced technologies that are meant to redefine the actual concepts of industry and society. Intelligent, autonomous and connected systems are going to affect everyday life and to create a costumer-centred industry. In a highly competitive and globalised market putting the costumers needs at the centre of each phase from design to after-sale services is the key for success. The new advanced technologies allow the companies to better understand the market through data collection and analysis and to design and optimise the products, the production processes and the services by diminishing the costs. Unox S.p.A. has already introduced some of the Industry 4.0 core technologies. Its professional ovens are equipped with smart sensors that measure the process variables during cooking and washing processes. Data are collected and sent through the Internet of Things structure to the cloud of the company where they are compressed and stored. The amount of data collected is enormous and its analysis is complex and time-consuming. However, data mining is necessary in order to extract useful information. Although some case-specific analysis have been already done by the company, there are no industrial procedures for data analysis. Therefore, the main objective of this Thesis is to propose an industrial procedure tailored to the company for data analysis through advanced statistical techniques. First, data are analysed for global cooking process understanding and data driven design. The aim of these analyses is the improvement of the user experience in the interaction with the machine. Then, advanced statistical techniques are implemented in product quality control and maintenance. Product quality control is gained through process monitoring; by ensuring the process variables to stay into the defined normal conditions the final desired quality is achieved. Maintenance can be optimised by the introduction of the predictive factor. Predictive maintenance allows carrying out maintenance activities at the first signs of imminent failures by avoiding both unnecessary repairs and catastrophic failure. This way failure prediction ensures cost reduction in the maintenance area, which is one of the largest expenses for a company.

The Thesis is composed of five chapters. The first Chapter gives a brief introduction to the Industry 4.0 concept and its key technological components. The problem is then contextualised in the company collaborating in the study (Unox S.p.A.), highlighting the types of oven produced and the variety of cooking/washing options implemented in the equipment software. In the final part of the Chapter the objectives of the Thesis are presented. The second Chapter

describes the used multivariate statistical techniques: Principal Components Analysis (PCA) and Partial Least Squares analysis (PLS). Their mathematical fundamentals and results interpretation for practical use are reported. These methods are implemented for process monitoring and predictive maintenance with the use of Matlab® software and the PLS Toolbox. The third Chapter reports the analysis of the historical data of cooking and washing processes both for process understanding and for data-driven design. Through the study of the oven utilisation, cooking modes are analysed in order to: *(i)* identify the least frequently used ones for future elimination or integration; *(ii)* propose further development and improvement of the most used ones. In fact, the objective is to simplify the panel interface in order to improve the user experience. In the fourth Chapter, process monitoring is used to identify anomalies in cooking programs compared to normal operating conditions. In order to obtain a high precision monitoring and thus the desired product specifications, an oven-specific model is implemented through the PCA analysis. Finally, predictive maintenance is implemented (in the fifth Chapter) in order to include prevention in the after-sales services and thus to avoid interruption of oven operation. Two case studies are presented: gasket and core probe substitutions. The prediction of specific equipment faults is carried out by applying the PLS-DA technique.

Chapter 1

Industry 4.0: a brief introduction

In this Chapter the background, the definition and the basic components of Industry 4.0 concept are described. Then, the key role of data-driven techniques (such as process monitoring and predictive maintenance) applied to manufacturing industry is highlighted. Since these data-driven techniques have been tested on a two-year dataset provided by Unox S.p.A, leader in professional ovens manufacturing, a brief description of the enterprise business and the provided dataset is given. Finally, the motivation and the objectives of the Thesis are described.

1.1 Industry 4.0: an overview

Industry 4.0 is a collective term used to define the current trend of integrating innovative production technologies and modern automation systems to improve working conditions, create new business models and increase the productivity and quality of production plants. The resulting race towards rapid implementation of technological innovation does not only impact on the manufacturing industry itself, but also on the service industry and society in general. In order to speed up the transfer from the old way of operating to the new one, a joint work between academy and industry is needed; on one hand, academic research would focus on defining, developing and sharing innovative models and methodologies; on the other hand, industry would focus on industrial machine update, intelligent products implementation and potential customers information.

1.1.1 Background and definition

The current industrial situation is the result of an historical path of development started in the 18th century. As represented in Figure 1.1, the first industrial revolution was characterised by the introduction of mechanical production powered by steam and water. Then, work division (Taylorism) and electrical energy lead to the second industrial revolution in the 19th century. The third industrial revolution started in the 1970s, when advanced electronics and information

technology further developed the automation of production processes. Nowadays the so-called fourth industrial revolution is taking place.

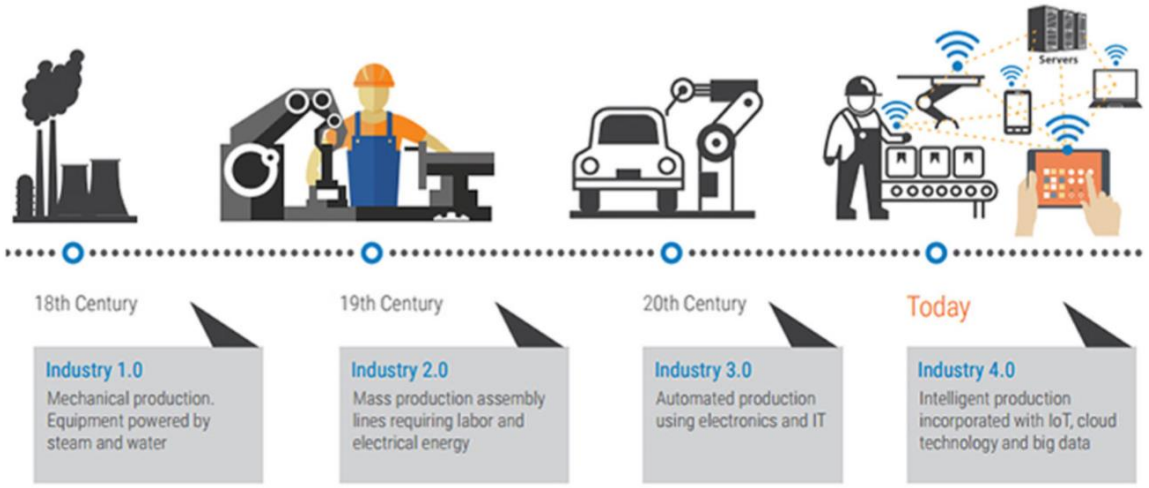


Figure 1.1. *Historical timeline of industrial revolutions [Oztemel and Gursev, 2018]*

The definition of Industry 4.0, firstly introduced as a strategic initiative of the German government in 2011 (Kagermann *et al.*, 2013), is not straightforward since continuous researches highlight different key aspects: from manufacturing digitalization (De Carolis *et al*., 2017) to the generation of the so-called smart factories (Kagermann *et al*., 2013); from communication technology improvement (Igor *et al*., 2016) to systems automation (Oztemel, 2010). Oztemel and Gursev (2018) in their review of Industry 4.0 recommend the following definition: "Industry 4.0 is a manufacturing philosophy that includes modern automation systems with a certain level autonomy, flexible and effective data exchanges encoring the implementation of next generation production technologies, innovation in design, and more personal and more agile in production as well as customized products." This definition clearly indicates the main elements of Industry 4.0: systems with automatized decision-making capability and data exchange platforms (IoT, Cloud) supporting the innovation and invention of future generation technologies as well as more profitable utilization of data (Big Data).

A similar concept, mainly used in the USA, is the so-called *Smart Manufacturing* (Kang *et al.,* 2016). Indeed, the agency of the Department of Commerce NIST (National Institute of Standards and Technology) defines *Smart Manufacturing* as "fully-integrated and collaborative manufacturing systems that respond in real time to meet the changing demands and conditions in the factory, supply network, and customer needs." Since both *Smart Manufacturing* and Industry 4.0 are based on the same principles, the two terms are considered equivalent. In the

following paragraph, the key components of Industry 4.0 and *Smart manufacturing* are described.

1.1.2 Key components

The goal of Industry 4.0 is to enhance and improve the efficiency of operations and the productivity of new business models, services and products. In order to achieve it, an integration of computing and physical processes is needed, the so-called Cyber Physical System (CPS) described by Lee *et al*. (2014). The CPS interacts with the physical system and expands its capabilities through computation, communication and control. The interconnection between different pieces of equipment and the development of human-machine interfaces allow for instant control of the processes and the services with an orientation to the costumers need. The Cyber Physical Systems are made of two main elements:

- a network of objects and systems communicating to each other through the Internet with a designated address;
- a virtual environment that is created by a computer simulation of objects and behaviours of the real world.

For the realisation of a CPS new technologies are needed: Internet of Things, smart sensors, advanced robotics, Big Data analytics, 3D printing, augmented reality, cloud computing and location detection (see Figure 1.2).

The connection between communicating physical devices (smart sensors, industrial robots and location detection technology) and the network connectivity that enables these elements to exchange and collect data constitute the so-called Internet of Things (IoT) (Aztori *et al*., 2016). One of the most important physical devices are the smart sensors. Smart sensors are placed in strategical points of the system and measure process variables during the entire process. The measured data are collected and stored in a cloud system. The cloud storage is only one of the resources of cloud computing. Indeed, the American National Institute of Standards and Technology (NIST) defines cloud computing as (Mell and Grance, 2011): “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.” This powerful tool has applications in different areas of a company; in particular, it makes the handling of Big Data easy and efficient. In fact, a large amount of data is continuously generated from different sources and their interpretation and analysis is complex and time-consuming

(Yan *et al*., 2017). Data mining consists of process analysis and information extraction from large datasets in order to obtain meaningful value. Data mining techniques are mainly used for classification and prediction in order to allow data-based management decisions. Other important elements of factories stepping in the fourth industrial revolution are advanced robotics and augmented reality. Advanced robotics consists in autonomous and cooperating industrial robots that enables flexibility in factories in order to tailor products and services to customers. Augmented reality is an enhanced version of physical real-world environments augmented with computer-generated images and provides benefits in designing products and production systems, training, maintenance and logistics. For design purposes, 3D printing is a fundamental technology. It allows the printing of prototypes with reduced costs, but also spare parts can be produced reducing inventories and transports. Finally, location detection allows tracking spare parts for maintenance purposes and products for both logistic and customer satisfaction purposes. In general, all these new and advanced technologies are meant to put the costumer, directly or indirectly, at the centre of attention through tailored products and services. Companies need to undergo a significant change in their common practices and related attitude. This yields to a complete re-developments and re-establishment of processes, products and services. However, a well-defined road map to Industry 4.0 is needed due to current absence of practical guidelines. For this reason, both researchers and companies are looking for defining the best way to obtain the maximum advantage from the fourth industrial revolution.

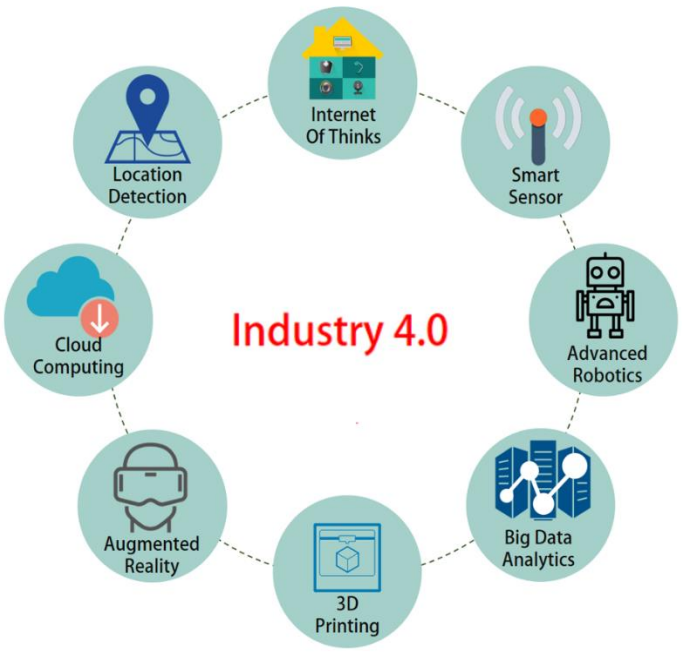


Figure 1.2. *Basic technologies of Industry 4.0*

1.1.3 Process monitoring and predictive maintenance

Data mining is a key element of the fourth industrial revolution. Data manipulation and analysis is fundamental to obtain valuable information. Moreover, data analysis through advanced statistical techniques allow the implementation of process monitoring and predictive maintenance. These two elements meet the goals of Industry 4.0: they allow efficient control of the process, production optimization, cost reduction and improved customer satisfaction. Process monitoring consists in real-time tracking of process condition through online measurements of process variables. It is important both for safety reasons and for product quality. Indeed, process monitoring allows taking corrective action when the process departs from normal operating conditions in order to avoid the final product going out of specification. The normal operating conditions are expressed by a model of the system that detects anomalous path of processes from early stages. The model of the system can be obtained through different approaches:

- State estimation approach: the online measurements are used to identify the theoretical state of the system on the basis of a fundamental model of the process;
- Knowledge-based approach: the model is built from the knowledge about the process;
- Multivariate statistical approach: the process model is empirically built through a multivariate statistic method based on historical data.

The last approach is the one used in this Thesis and consists in the creation of a process model from a multivariate statistical method of past successful processes. There are multiple examples of applications of the multivariate statistical approach in literature. One of this is described by Nomikos and MacGregor (1994). They use the Multi-way Principal Component Analysis (MPCA) for obtaining the process model of a semi-batch emulsion polymerisation of styrene-butadiene to make a latex rubber (SBR). The model is calculated from historical values of process variables of successful batches and the confidence limits for normal conditions are identified. The new batches are then projected into these limits in order to understand if they are normal or not while running. Another example of process monitoring using this approach is described by Largoni *et al*. (2015). They present the case of an industrial batch bioreactor used in avian vaccine manufacturing. Beside MPCA model, they use Multi-way Partial Least Square analysis (MPLS) to monitor the process with regard to the final product quality. They managed to predict not only whether the product would be or not on specification, but also the value of the quality parameter (in the case of on specification product).

The same techniques can be also used for predictive maintenance. Unlike the time-driven approach of preventive maintenance relying on industrial or in-plant average-life statistics (i.e., mean-time-to-failure) to schedule maintenance activities, predictive maintenance (Mobley, 2002) is a condition-driven approach. It uses direct monitoring of the mechanical condition, system efficiency, and other indicators to determine the actual mean-time-to-failure or loss of efficiency for each machine-train and system in the plant. It allows carrying out maintenance activities at the first signs of imminent failures by avoiding both unnecessary repairs and catastrophic failure. Predictive maintenance program is implemented through process variables monitoring and other non-destructive techniques (e.g., vibration monitoring, thermography, tribology). An example of application of predictive maintenance through process variables monitoring is presented by Marton *et al*. (2013). In this paper, a data driven approach based on PCA and PLS is used to detect abnormal patterns that lead an asynchronous generator to failure or malfunctioning and to predict these events in order to reduce maintenance costs.

1.2 Unox

This Thesis has been done in collaboration with Unox S.p.A., leader in professional cooking ovens market. The company produces different series of ovens and distributes them to more than 130 countries. Moreover, Unox offers after-sale services like cooking training, support and technical assistance.

1.2.1 Ovens description

Amongst all the ovens produced by Unox, the MIND.Maps™ series has been considered in this Thesis. It consists of two main categories: Chef Top MIND.Maps™ and Baker Top MIND.Maps™. As shown in Table 1.1, each category is split into two new sub-categories: ONE and PLUS; PLUS ovens are characterised by more advanced technical tools and a wider range of cooking options than ONE units. Each sub-category includes oven models with different hardware and software characteristics.

Table 1.1. *Categories, sub-categories, models and ovens analysed in the Thesis*

Categories	Sub-categories	Models	Ovens
Chef Top MIND.Maps™	PLUS	36	616
	ONE	6	25
Baker Top MIND.Maps™	PLUS	11	171
	ONE	4	14

The oven quantities reported in the last column of Table 1.1 do not reflect the actual amount of ovens sold by the enterprise, but they correspond to the number of ovens connected to the Internet cloud. Chef Top PLUS ovens represent 75% of the ovens registered in the database, whereas ONE ovens are the least connected as a choice of the users.

In general, ovens operate by implementing 8 different cooking programs that are identified by a code number (from 2 to 11). In five cooking programs, the user sets autonomously the desired cooking parameters by setting their values as piecewise constant/linear functions. The maximum autonomy and control of the process the user can exert is reached with the program that gives the name to the series (MIND.Maps™). It allows the user to design manually the curve of the values of the parameters that the process will follow in each moment. Finally, the other three programs are pre-set programs, i.e. the process parameters are already set by the company.

Code number 1 identifies a generic washing program set between two cooking phases. Washing programs quality is further specified by another key feature: the duration. Although washing duration slightly varies according to the oven model, five types of washing modes are identified:

1. Type 1: 6-10 min;
2. Type 2: 30-35 min;
3. Type 3: 41-43 min;
4. Type 4: 102-105 min;
5. Type 5: 143-153 min.

During a cooking program, some process variables are measured, recorded, collected and sent to the cloud through the Internet connection. The total number of variables is six: three of them are measured through the oven instrumentation, whereas the remaining three are manually or automatically set by the user whenever a pre-set program is chosen. The registered process variables are:

- the temperature measured in the oven chamber;
- the temperature measured by the core probe placed inside a food sample;
- the temperature set for the cooking program: it corresponds to the set-point(s) at which the oven works;
- the ventilation set-point: it is an integer value within the range [-4, 4], where negative integer values correspond to pulsed ventilation, positive values to continuous ventilation and zero means no ventilation in the oven;

- the humidity set-point: it is a percentage within the range [-100%, 100%], where negative values mean that the humidity is extracted from the oven;
- the humidity measured in the oven chamber.

1.2.2 Approaching Industry 4.0

Unox wants to define its own path throughout the innovation and the digitalization that Industry 4.0 philosophy brings in. Its goals is to improve the production process, but also to increase the quality of the after-sale services. As already mentioned, Unox collects the data from those ovens the clients agreed and/or had the possibility to connect to the Internet. Each oven is equipped with sensors that measure the variables values and transmits the information about the process state. Together with the measured ones, the other set variables are sent to the cloud through the Internet of Things structure. In the cloud, data are stored and compressed: the Big Data issue arise. The large amount of available data has to be studied and interpreted in order to get useful information. The dataset is actually used by the company to improve the technical services, but its potential is not fully developed. The information contained in the data can be further used to improve both products and services: some of the applications are shown in this Thesis as the following section describes.

1.3 Thesis objectives

The main objective of this Thesis is to propose an industrial procedure tailored to the company for data analysis through advanced statistical techniques. As previously stated, the large amount of available data can be studied and analysed in order to obtain useful information. First of all, the data of cooking and washing processes are analysed for both process understanding and data-driven design. Through the analysis of the measured and set variables, a global understanding of both cooking and washing processes is gained. Oven dataset is analysed to highlight the most and least used oven functions/modes in order to propose both further development of frequently used ones and adjustment/replacement of the least frequently used. The main goal of these actions is to simplify the panel interface in order to improve the user experience. PCA and PLS methodologies for data mining are implemented for process monitoring and predictive maintenance. Process monitoring is used to identify anomalies in the cooking programs with respect to the normal operating condition. In order to have high precision monitoring an oven-specific model can be obtained from a PCA model. Process

monitoring allows the user to have a tight control of the process in order to obtain the desired product. Predictive maintenance is implemented in order to include prevention into the after-sale offered services. The prediction of equipment failures is fundamental in order to avoid interruption of oven operation. Indeed, maintenance can be scheduled and action can be taken before the problems actually arise. Moreover, useless technical interventions dictated by preventive maintenance can be avoided with cost reduction. The prediction of specific equipment failures can be carried out through the application of the PLS-DA technique.

Chapter 2

Mathematical background

In this Chapter the theoretical background on the multivariate statistical techniques used in this Thesis (PCA and PLS) is presented, focusing both on mathematical fundamentals and results interpretation for practical use. The methodologies described in this Chapter have been implemented for process monitoring and predictive maintenance tasks in Matlab® software, with the help of PLS Toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA, 2015).

2.1 Principal Component Analysis

Principal Components Analysis (PCA) is a multivariate statistical method generally used to analyse datasets with a high number of variables. Since industrial processes are becoming ever more heavily instrumented and data are collected with ever higher frequency, PCA is one of the best technique to compress the huge amount of available data and to extract information from process variables. Its goal is to obtain a model that not only provides useful information about the studied system but that can also be used for process monitoring. The theoretical background of the method is explained in the following section.

2.1.1 PCA method

Mathematically, PCA is based on the eigenvector decomposition of the covariance of data matrix $\mathbf{X}$ ($I \times J$), where I is the number of observations and J is the number of collected variables. When the initial matrix is mean-centred (see pre-processing treatments §2.3), the covariance of $\mathbf{X}$ is calculated as:

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}'\mathbf{X}}{I-1}. \tag{2.1}$$

PCA method writes the matrix $\mathbf{X}$ of rank h as the sum of h matrices of rank 1, as follows:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \cdots + \mathbf{M}_h \,. \tag{2.2}$$

Each of these matrices can be written as the outer product of column vectors $\mathbf{t}_h$ and $\mathbf{p}_h$, respectively named as scores and loadings. Equation 2.2 can therefore be reformulated as:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \mathbf{t}_3\mathbf{p}_3' + \cdots + \mathbf{t}_h\mathbf{p}_h' \tag{2.3}$$

or, in a matricial form, as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}', \tag{2.4}$$

where $\mathbf{T}$ is the matrix that contains score vectors as columns, while $\mathbf{P}'$ is made of loading vectors as rows. The loading vectors are the eigenvectors of the covariance matrix $\mathrm{cov}(\mathbf{X})$; then, for each $\mathbf{p}_i$, the following equation is valid:

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_i\ , \quad i = 1,..., h \tag{2.5}$$

where λ_i is the eigenvalue associated with the eigenvector $\mathbf{p}_i$. The loading vectors contain information on how variables relate to each other. The score vectors contain information on how the samples relate to each other and can be written as the linear combination of the original data matrix and the corresponding loading vector, as follows:

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i, \quad i = 1,..., h. \tag{2.6}$$

Each $(\mathbf{t}_i, \mathbf{p}_i)$ pair is arranged in descending order according to the associate eigenvalue λ_i, that becomes a metric of the amount of variance described by each pair. The first pair capture the largest amount of information and each subsequent pair captures the greatest possible amount of variance remaining at that step. After an adequate truncation, the first K pairs can represent the initial system using less factors than the original variables.

The above defined vectors and eigenvalues have been calculated by applying the Singular Value Decomposition (SVD) algorithm implemented in Matlab® since it avoids numerical issues due to finite precision representation of real numbers. The algorithm decomposes the **X** matrix into the product of three matrices (see Figure 2.1), where the columns of **U** and **V** are orthonormal and **D** is diagonal matrix with positive real values. In particular, **V** is the loading matrix, **D** is a diagonal matrix whose terms are the eigenvalues of **X** and the score matrix is defined as the product **UD.**

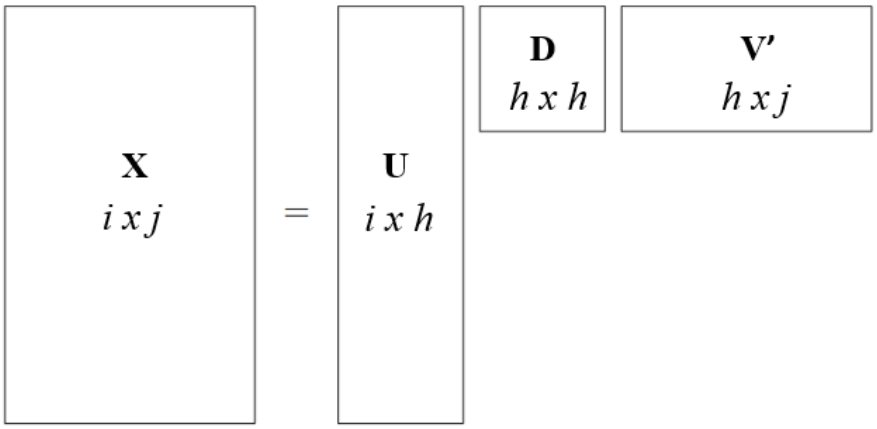


Figure 2.1. *Schematic view of Singular Value Decomposition of matrix* **X**

From a graphical point of view, principal components concept is represented in Figure 2.2 through an example for two variables. The first principal component direction is found by fitting the data with a line passing through the origin of the axes. The best fitting is found by minimizing the distance between each point and its projection on the fitting line or, equivalently, by maximising the distances between the projection of the points and the origin of the axes (SVD algorithm approach).

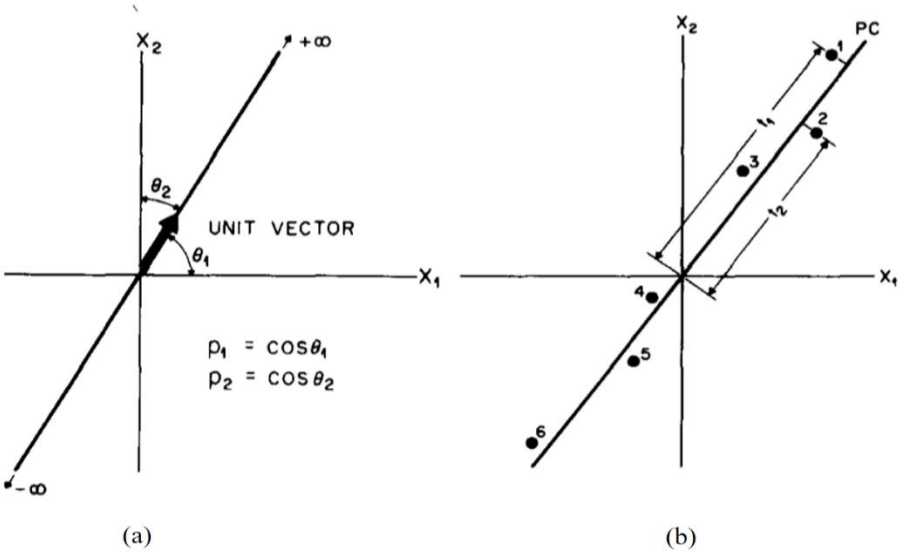


Figure 2.2. *Graphical representation of: (a) loadings and (b) scores of a two variables PCA analysis [Geladi and Kowalski, 1986]*

As it can be seen in Figure 2.2*b*, the distances of each point from the origin along the PC direction are the components of the score vector, the eigenvalue is the sum of the squared distances from the origin of each projection and the eigenvector is the versor that goes from the origin to the direction of the associated PC. Its projections on the axes of the plot (the direction

cosines) are the loadings, $\mathbf{p}_1$ and $\mathbf{p}_2$ (Figure 2.2*a*); in other words, they are the proportional contribution of each variable to the eigenvector. In general, when all the principal components are individuated, the variation around the origin for each PC is calculated as the sum of squared distances of each sample projection on PC from the origin (definition of eigenvalue) and divided by the sample size minus one ($I - 1$). Then, the variance captured by each PC is expressed as percentage of the total variance captured by all the PCs in order to highlight the importance of the principal components.

The original dataset can be generally described by extracting its first K principal components without significant information loss. The value of parameter K must be lower than the smaller dimension of the matrix $\mathbf{X}$ and it is determined by looking at the cumulative variance represented by the PCs. A limit value for the minimum cumulative variance captured by the model can be set to find out the number of PCs needed to achieve it. Otherwise, a more accurate way to define K can be used: the cross-validation technique. In this procedure, the dataset is generally divided into segments and a PCA model is calibrated on the matrix generated by extracting one of the segments. The model is then validated through the process data not used in the model calibration. The procedure is iteratively repeated and the prediction error is calculated in order to evaluate the model predictive power. This error is called Root Mean Square Error of Cross-Validation (RMSECV) and is calculated at each iteration j as follows:

$$RMSECV_j = \sqrt{\frac{PRESS_j}{I}}, \tag{2.7}$$

where PRESS is the Prediction Error of Sum of Squares. It is calculated as the sum of the squared difference of each sample value and its prediction ($\hat{x}_{ij}$) through Eq. 2.8:

$$PRESS_j = \sum_{i=1}^{I} \left(x_{ij} - \hat{x}_{ij}\right)^2. \tag{2.8}$$

The calculated error is then plotted as a function of the number of PCs used. The more PCs that describe large amounts of systematic variance are added to the model, the more the error decreases; alternatively, the more PCs describing only small noise variance are added, the more the error increases.

When the number of principal components K is chosen, the $\mathbf{X}$ matrix can be written as:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_K\mathbf{p}_K' + \mathbf{E}, \tag{2.9}$$

where $\mathbf{E}$ is the so-called residual matrix. The decomposition of the $\mathbf{X}$ matrix described by Eq. 2.9 is graphically represented in Figure 2.3: the K score vectors with ($I \times 1$) dimensions are

collected in the ($I \times K$) **T** matrix, the K loading vectors with ($1 \times J$) dimensions are collected in the ($K \times J$) **P** matrix and the residuals form a ($I \times J$) **E** matrix.

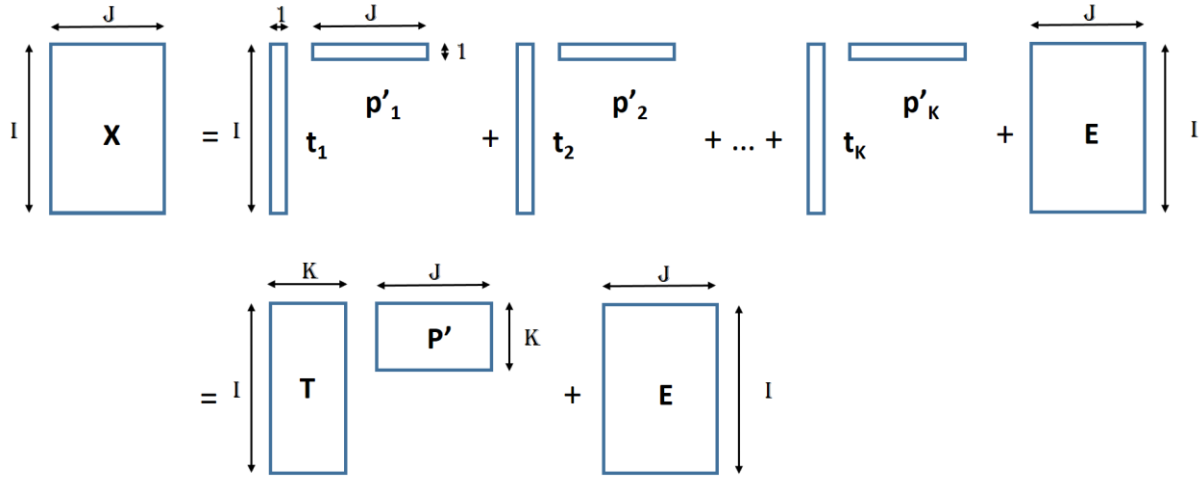


Figure 2.3. *Decomposition of matrix* **X** *in scores, loadings and residuals*

This way the PCA model is obtained and an example of its graphical representation can be seen in Figure 2.4: a three variables system represented by two principal components.

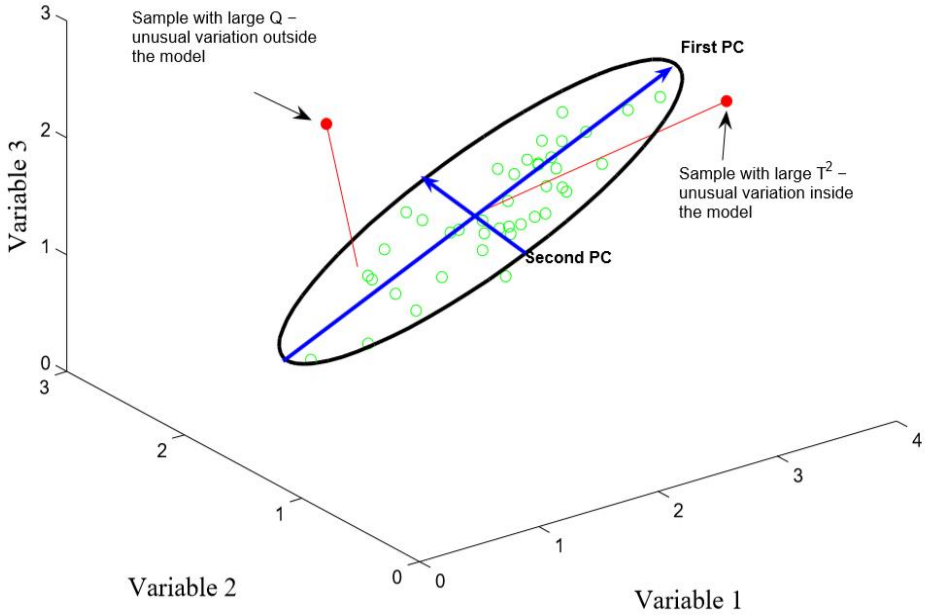


Figure 2.4. *Graphical representation of PCA results of a system with three variables and two PCA (blue axes): the green circles are the samples, the black ellipse represents the confidence limits and the red points are samples with high values of model evaluation statistics [Wise et al., 2006]*

The figure represents the samples as green circles in the three variables space that actually dispose themselves into a hyperplane identified by the directions of the two principal

components (blue axes). By considering the new coordination axes, the dimension of the system can be reduced. In this case the reduction is of one dimension only, but in larger system the reduction is higher and is essential for monitoring purposes. In order to achieve these goals the confidence limits can be identified (black ellipse) and the new samples can be projected on the hyperplane: through the location of the new samples projections the process can be controlled. This type of graphic is going to be used in Chapter 4 of this Thesis for process control.

2.1.2 Model evaluation statistics

There are some parameters that can be calculated in order to evaluate how the model represents the data: the Q and the T^2 Hotelling statistics.

Q statistic is the Euclidean distance of a point from the hyperplane generated by the retained PCs. It is used to evaluate the lack of model fit and it is calculated as the sum of squares of each row of $\boldsymbol{E}$, as follows:

$$Q_i = \mathbf{e}_i\mathbf{e}_i' = \mathbf{x}_i(\mathbf{I} - \mathbf{P}_K\mathbf{P}_K')\mathbf{x}_i', \quad i = 1,..., K \tag{2.10}$$

where $\mathbf{e}_i$ is the i^{th} row of $\mathbf{E}$, $\mathbf{x}_i$ is the i^{th} sample in $\mathbf{X}$, $\mathbf{P}_K$ is the matrix of the first K loading vectors retained in the PCA model and $\mathbf{I}$ is the identity matrix. Each row of the residual matrix $\mathbf{E}$ represents the Q contributions of a given sample and each component of $\mathbf{e}_i$ shows how much each variable contributes to the overall Q statistic. The Q statistic indicates how well the samples are represented by the model.

The Hotelling T^2 statistic is the sum of normalised squared scores defined by the following equation:

$$T_i^2 = \mathbf{t}_i\boldsymbol{\lambda}^{-1}\mathbf{t}_i' = \mathbf{x}_i\mathbf{P}_K\boldsymbol{\lambda}^{-1}\mathbf{P}_K'\mathbf{x}_i', \quad i = 1,..., K \tag{2.11}$$

where $\mathbf{t}_i$ is the i^{th} row of scores matrix $\mathbf{T}_K$ and $\boldsymbol{\lambda}$ is a diagonal matrix of the first K eigenvalues. The T^2 statistic is a measure of the distance from the multivariate mean to the projection of the point on the PCs hyperplane and it is a metric to quantify the variability within the model.

Confidence limits are established in order to control the process and to define when a Q or T^2 value is considered statistically acceptable. They are based on the assumption of normal distribution of the scores. In fact, PCA models have the additional advantage that the scores produced which are linear combinations of the original variables, are more normally distributed than the original variables themselves. This is a consequence of the central limit theorem, which can be stated as follows: if the sample size is large, the theoretical sampling distribution of the

mean can be approximated closely with a normal distribution. Since typically the sampling size is large, we would expect the scores, which are a weighted sum like a mean, to be approximately normally distributed. The confidence limits on T^2 is then calculated as follows:

$$T^2_{K,I,\alpha} = \frac{K(I^2-1)}{I(I-K)} F_{K,I-1,\alpha} \tag{2.12}$$

where K is the number of principal components, I is the number of samples and $F_{K,I-1,\alpha}$ is the critical value of the F-distribution with K, (I -1) degrees of freedom and α significance level. The T^2 confidence interval defines the hyperspace in which the projection of the samples are located in normal conditions. In a two PCs model this limit is an ellipse, while in three PCs model it is an ellipsoid whose semi-axes **s** can be calculated as:

$$s_i = \sqrt{\lambda_i T^2_{K,I,\alpha}}, \qquad i = 1,..., 3. \tag{2.13}$$

The Q limit Q_α defines the distance off the plane or the space that is considered unusual for the point location. It is calculated as:

$$Q_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \tag{2.14}$$

where c_α is the standard normal deviate corresponding to the upper (1−α) percentile, h_0 is defined by Eq. 2.15 and θ_r by Eq. 2.16.

$$h_0 = 1 - \frac{2\theta_1\theta_2}{3\theta_2^2} \tag{2.15}$$

$$\theta_r = \sum_{i=K+1}^{h} \lambda_i^r, \qquad r = 1,..., 3 \tag{2.16}$$

where K is the number of principal components retained by the PCA model and h is the rank of **X**.

An example of graphical representation of confidence limits and samples with high values of Q and T^2 statistics is reported in Figure 2.4. The sample with a high value of Q is not represented by the model in a statistical satisfactory way; in fact, it is far distant from the hyperplane generated by the two PCs. The sample with high value of T^2 is located on the hyperplane defined by the principal components but it is out of the confidence limits represented by the ellipse.

2.1.3 Multi-way Principal Component Analysis method

In most cases, variables are measured with a certain frequency during the entire duration of a process (sample). Time becomes the third dimension of the initial data matrix as shown by Figure 2.5. Data have to be manipulated in order to obtain a two-dimensional matrix for PCA analysis.

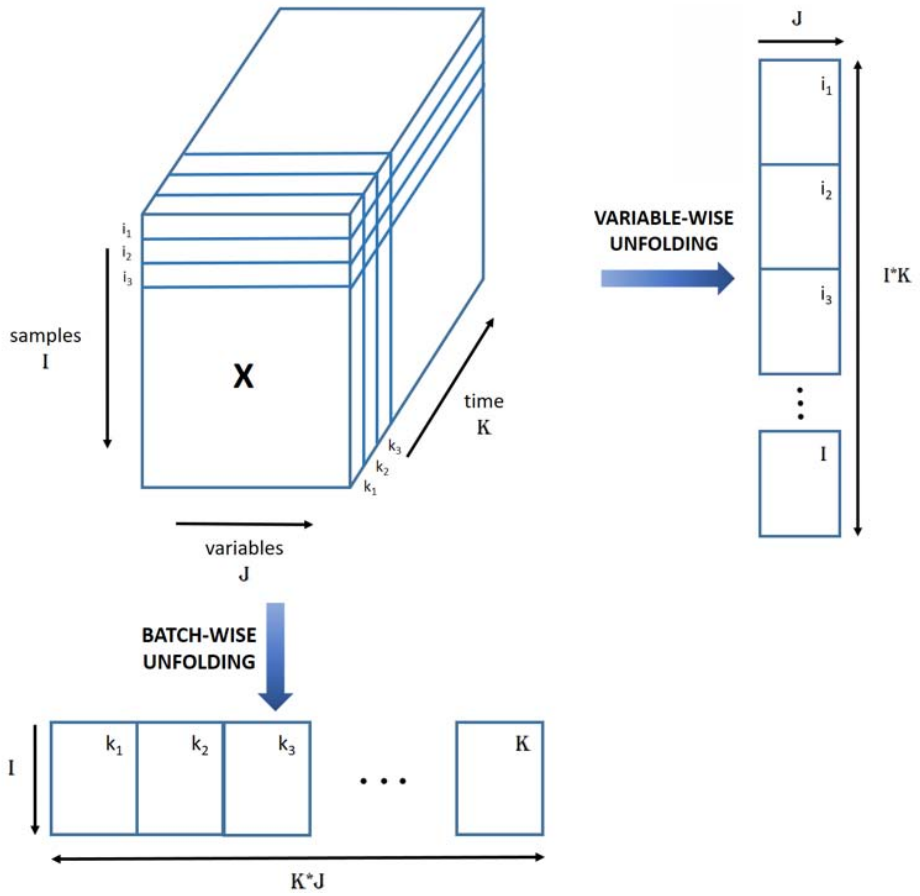


Figure 2.5. *3D data matrix and unfolding for MPCA*

There are several methods to convert a three-dimensional data matrix to a two-dimensional one, but Multi-way Principle Components Analysis (MPCA) is the most straightforward approach (Wise and Gallagher, 1996). It is equivalent to performing a PCA on a two-dimensional matrix obtained from the 3D matrix by the so-called *unfolding*. Depending on the aim of the analysis, the unfolding is mainly done in two ways (Figure 2.5):

- Variable-wise unfolding: the horizontal ($J \times K$) slices at fixed I are placed one below the other along the time axis. In this configuration the variables information is kept together and the evolution path of a single variable can be followed in time through the single sample.

- Batch-wise unfolding: the vertical ($I \times J$) slices at fixed time are placed side by side to the right along the variable axis. In this configuration the sample information is kept together and the evolution path of a single sample can be followed through the single variable in time.

2.2 Partial Least Squares

Similarly to PCA approach, Partial Least Squares (PLS) regression is a statistical method based on the concept of creating models that represent complex systems with a number of factors smaller than the number of original variables; however, the aim of PLS is to correlate two datasets (generally, process variables data with product quality outputs) with prediction purposes.

2.2.1 PLS method

PLS is a statistical technique that can be used to create models relating one or more product quality measurements (**Y**) to collected process variables (**X**). The main goal of PLS is to find factors that capture the greatest amount of variance in the predictor variables (**X**) and best correlate **X** with predicted variables (**Y**). These factors correspond to the principal components of PCA since they are the direction of maximum variability, but, unlike PCs, they are rotated in order to predict the dependent variables of **Y** matrix. The directions of maximum variability are called latent variables (LVs). The number of latent variables retained in a model depend on the captured variance and on the prediction power of the model. The predictor matrix **X** and the predicted matrix **Y** are then decomposed as product of a score matrix (**T** and **U** respectively) and loading matrix (**P** and **Q** respectively) plus an error matrix (**E** and **F** respectively):

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}'_a + \mathbf{E}, \tag{2.17}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}'_a + \mathbf{F}, \tag{2.18}$$

where A is the number of significant LVs retained in the model. Each score vector is expressed as linear combination of the original data matrix as:

$$\mathbf{t}_a = \mathbf{X}\mathbf{w}_a, \tag{2.19}$$

$$\mathbf{u}_a = \mathbf{Y}\mathbf{q}_a. \tag{2.20}$$

$\mathbf{t}_a$ and $\mathbf{u}_a$ are not independent, but they are linked by the "inner-relationship" $\mathbf{u}_a = b_a\mathbf{t}_a$. In Eq. 2.19 additional proportionality vectors $\mathbf{w}_a$ $(a = 1,2, \dots, A)$ are introduced: they are called weights and their aim is to maintain the orthogonality property of the scores. They have the same dimensionality of the loading vectors and express the contribution of each variable to the model definition. The PLS model is calculated by implementing Equations from 2.17 to 2.20 into an algorithm whose goal is to minimize the error matrices ||**E**|| and ||**F**||. The Multi-way Partial Least Squares (MPLS) analysis is led by performing a PLS on a two dimensional matrix obtained by unfolding the 3D matrix as previously seen in §2.1.3.

2.2.2 Partial Least Squares – Discriminant Analysis

Partial Least Squares – Discriminant Analysis (PLS-DA) is a specific PLS analysis used when the predicted variables **Y** are categorical. It means that the predicted values are restricted to a specific range and then divided into categories through thresholds. The predicted variables can be defined as:

1. a column vector of numbers indicating class assignment for each sample (row): for example $\mathbf{y} = [1\ 1\ 3\ 2]'$;
2. a matrix of one or more columns containing logical zeros (= not in class) or ones (= in class) for each sample (row).

   For example, $\mathbf{Y} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.

The first option is used when classes are liked by some type of interdependence, while the second option allows the use of classes without any relationship at all. In the second case, each column of the **Y** matrix corresponds to a different class and a one value appears only if the sample in the **X** matrix belongs to that class. This specific structure of **Y** ensures independence between classes. PLS-DA calculates the prediction probability and the classification threshold for each modelled class. The prediction probability is the probability of each sample to belong to a class. It is obtained by fitting the predicted *y*-values from the model to a normal distribution and calculating the probability of observing a given *y*-value. If there are two classes, the probability of a sample to belong to class 1 is calculated as:

$$\frac{P(y,1)}{(P(y,1)+P(y,0))}, \tag{2.21}$$

where y is the y-value predicted by PLS-DA model for a sample, $P(y,1)$ and $P(y,0)$ are the probabilities of measuring the given y-value for a class 1 sample and a class 0 sample respectively. These two probabilities are estimated by the y-values observed in the calibration data.

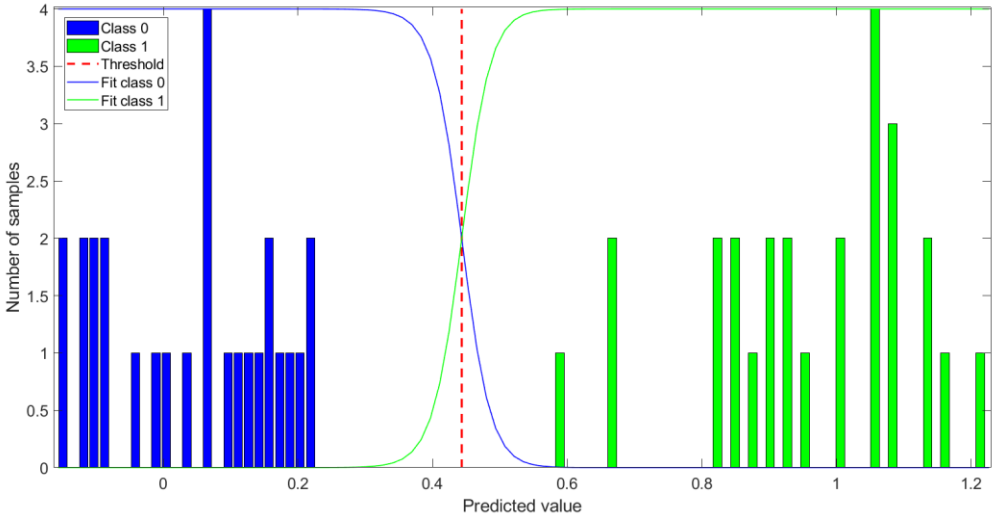


Figure 2.6. *Graphical representation of a 2 classes PLS-DA model predictions (blue and green bars) and fitting functions to normal distribution (blue and green lines): the red dashed line represents the threshold*

Figure 2.6 presents the model predictions of the calibration samples belonging to two different classes: the blue bars indicate the values predicted for samples of class 0 and the green ones the predictions for class 1. The distribution of each class is fitted to a normal one and the blue and green curves are obtained. As it can be seen, the curves cross only once and the corresponding x-value is the threshold between the classes (this happens because the number of samplings belonging to each class is balanced; otherwise, there would be more than one crossing points). It means that the probability of measuring a value of 0.44 for class 1 is equal to the probability of measuring the same value for a class 0 sample. Because of the previous normalization, the threshold represents the 50% probability of a sample to belong to class 1 (or 0). It divides the entire range of y-values into two areas: samples located in the left part of the graph have a higher probability of belonging to class 0 and vice versa. Finally, the validation samples are divided into the two classes by comparing their predicted y-value to the threshold and calculating their prediction probability.

2.3 Pre-processing treatments

PCA and PLS analysis is based on a data matrix with a defined structure. The collected samples are located on the rows of the matrix while the measured variables represent its columns. However, the variables are usually reported with different units and their own mean/variance could be different, as reported in Figure 2.7*a*. For this reason, data pre-processing treatments are often required.

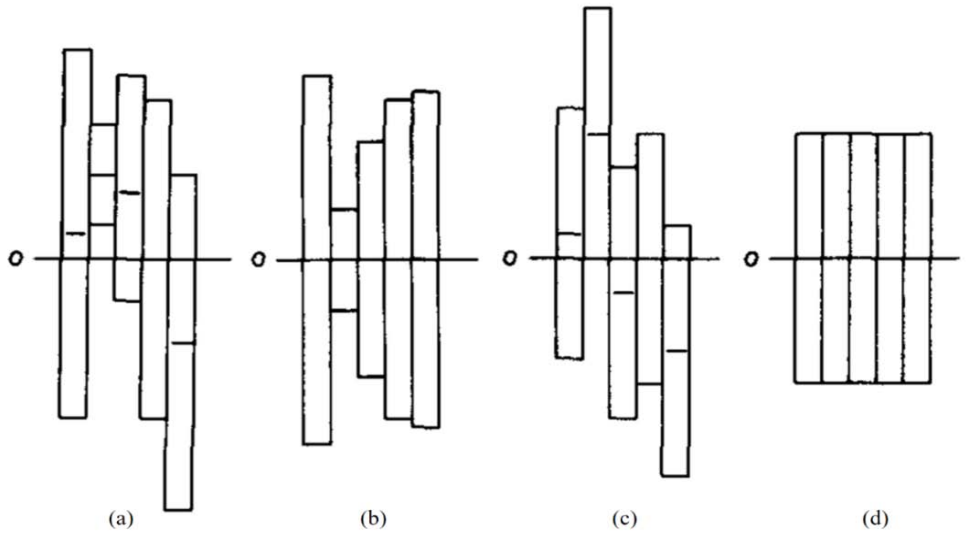


Figure 2.7. *Data preprocessing of: (a) raw data, (b) results of mean-centering, (c) results of variance-scaling and (d) results of auto-scaling [Geladi and Kowalski, 1986]*

Because of SVD algorithm requests (previously explained), data have to be mean-centred. The mean of each variable (column) is calculated by Eq. 2.22:

$$\bar{x}_j = \frac{\sum_{i=1}^{I} x_{ij}}{I} \tag{2.22}$$

where x_{ij} is the measured value of variable j for sample i in the matrix. Then, each value of the matrix is diminished by the mean of its own column, i.e. the mean profile is subtracted from the trajectory of the single variable (see Figure 2.7*b*). This way, a minor deviation from the mean profile is highlighted and its rejection is possible to ensure a tight control on the process.

When process variables have different measurement scales, variance-scaling is used in order to obtain unitary variance for all them. The procedure consists of dividing all the values by their standard deviations, calculated as follows:

$$\sigma_j = \frac{\sqrt{\sum_{i=1}^{I}(x_{ij}-\bar{x}_j)}}{I}. \tag{2.23}$$

Variance-scaling allows treating all the variables with the same weight and avoids prevailing of large measurement scale. The results of this pre-process treatment can be seen in Figure 2.7*c*. Finally, auto-scaling is a pre-processing treatment where data are both mean-centred and variance-scaled (see Figure 2.7*d*). The methods explained in this Chapter are going to be used in the following sections: PCA is used for process monitoring case study presented in Chapter 4, whereas PLS-DA is implemented in Chapter 5 for the predictive maintenance case studies.

Chapter 3

Preliminary analysis

In this Chapter, the available datasets are described: one Matlab table with cooking programs description, the .csv files with the process variables collected for each oven and one Excel file with information about all the technical interventions. Then, some preliminary analysis is run in order to get global understanding of the cooking and washing processes. The study of oven utilisation is fundamental for data-driven design: on the basis of the most and the least used functions and modes, the panel interface could be modified and simplified in order to improve user experience.

3.1 Datasets description

The datasets used for the analysis in this Thesis are extracted from Unox cloud and, after proper data manipulation, a Matlab table, a .csv file for each oven and an Excel table are obtained. The Matlab table contains information about all the cooking and washing programs of all the ovens.

	1 final_user	2 timezone	3 model	4 cookingStep	5 detergent	6 init_time	7 end_time	8 energy	9 flags	10 device	11 kindOfCooking
1	refectory	Europe/Prague	XEVC-2011-EPR	"1"	"0"	"1454998898300"	"1454998968420"	"619"	"1"	"73"	"2"
2	refectory	Europe/Prague	XEVC-2011-EPR	"1"	"0"	"1454998995150"	"1454999140430"	"942"	"1"	"73"	"2"
3	refectory	Europe/Prague	XEVC-2011-EPR	"1"	"0"	"1454999883070"	"1454999964260"	"102"	"3"	"73"	"11"
4	restaurant	Australia/Melbourne	XEVC-0711-EPR	"0"	"177"	"1455082195530"	"1455085835320"	"1632"	"1"	"76"	"1"
5	restaurant	Australia/Melbourne	XEVC-0711-EPR	"0"	"354"	"1455088023040"	"1455094256500"	"1940"	"1"	"76"	"1"
6	restaurant	Australia/Melbourne	XEVC-0711-EPR	"1"	"0"	"1455146081440"	"1455147885470"	"4988"	"1"	"76"	"2"
7	restaurant	Australia/Melbourne	XEVC-0711-EPR	"1"	"0"	"1455147985290"	"1455149818500"	"3708"	"1"	"76"	"2"

Figure 3.1. *Example of Matlab table structure*

Figure 3.1 shows the structure of the Matlab table: each row represents a cooking or washing program, while the columns contain the information about the programs. The data used in this Thesis are the following:

- final user/types of user: refectory, restaurant, deli (gastronomy) or pastry/bakery;
- timezone: geographical location of the user;
- model: code model of the oven;

- cooking step: number of steps made during the program ("0" is the washing step);
- detergent: amount of detergent used during the washing;
- initial and end time: starting and finishing time of each program;
- energy: amount of energy used during the program;
- flags: index of interruption of an infinite program (i.e. without ending time setting);
- device: ID of the specific oven;
- kind of cooking: typology of program ("1" is a washing program).

A .csv file is available for each oven. It contains the samplings of the process variables taken every 30 seconds during each cooking program. The samplings are chronologically ordered without separation between different programs. The collected process variables are 6: the chamber temperature, the probe temperature, the temperature set by the user, the fan velocity, the set humidity and the measured humidity. Finally, an Excel table with all the technical interventions is available. It reports the datetime, the type of oven and the description of all technical interventions made by the technical service.

3.2 Data conversion and selection

Time dataset is saved in Unix format; the counting is in milliseconds and starts from the 1st of January 1970. Firstly, the dataset is filtered to select the programs that were registered in the period 2016-2017.

Table 3.1. *Number of programs distributed by year*

Year	N° Programs	% Programs
2017	615902	72.62
2016	233733	27.18
other	10302	1.20

As it can be seen in Table 3.1, almost 73% of data belong to 2017, the 27% to 2016 and the remaining 1% (10302 programs) is related to different years due to synchronisation issues or time setting manipulation by final users. This last information is not implemented in the analysis due to its low contribution if compared to the data related to 2016/2017.

3.3 Reference time period selection

In order to calculate cooking and washing frequency for each oven, the actual operating time period has been identified as follows:

- the initial time is set equal to the time of the first cooking program registered for each oven (each oven is sold/connected to the Internet in different moments);
- the final time is chosen as the moment when the last cooking program was performed by the oven.

For the final time choice, another option has been initially considered: setting the final time as the moment of dataset download from the cloud. In order to take a decision, different time-profiles of the ovens have been studied. As it can be seen in Figure 3.2, the registration periods of different ovens start and stop at different moments (maybe because the ovens have been incidentally disconnected from the Internet, hence interrupting data transmission). If the moment of data download would have been chosen, some errors would have been introduced during the frequency calculation: for example, for device "213" this would mean dividing the number of cooking programs by 18 months instead of 6 months (it is the period of time in which the oven was operative). For this reason, the final time of utilization of each oven is set as the time of its last registered cooking program.

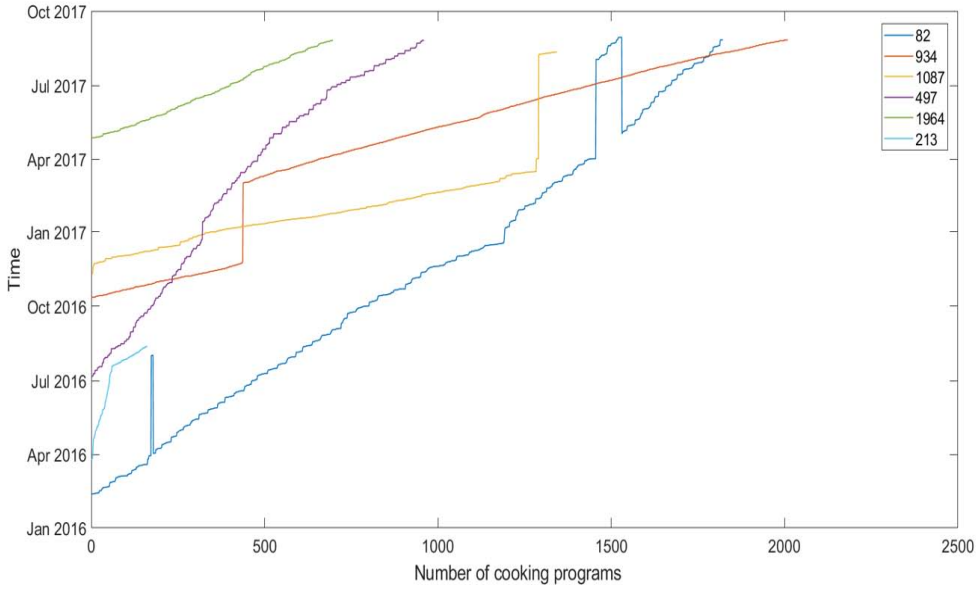


Figure 3.2. *Time profiles of cooking programs of 6 different devices*

Moreover, Figure 3.2 shows that some time-profiles present a rapid step up and down, like the ones of device "82" close to cooking programs 200 and 1500; this behaviour indicates that for a certain period of time the oven registered a wrong date and then it returned to the actual date time. These wrong steps have been eliminated: the counting for the operative time stops when a step starts, the duration between the cooking programs with the wrong date calculated and added to the total amount; then the counting starts again when the step comes back to acceptable

time values. The actual operating period is calculated with this procedure and it is used as reference for both cooking and washing programs. Finally, the number of cooking/washing programs is divided by the resulting operating period to obtain the weekly cooking/washing frequency of the specific oven.

3.4 Cooking programs analysis

The analysis of cooking programs is led in order to obtain a full insight of user utilisation and use the resulting information for design purposes: programs with low frequency of utilisation can be integrated into others or eliminated, while those with high frequency can be further improved and/or used to implement new specific oven functions.

First, the database is studied in an overall view by highlighting the most frequently used and the least frequently used ovens. Before implementing data analysis, some filters are introduced:

1. all the washing programs are excluded to focus only on cooking programs (kind of cooking ≠ "1");
2. all the cooking programs with duration smaller than 1/2/3/5 minutes are excluded because they are considered as technical maintenance events or errors;
3. all the ovens with a total number of cooking programs smaller than 5 are excluded because they are not statistically relevant;
4. the ovens with a total period of registration smaller than one week are excluded.

Table 3.2. Results of cooking duration based filtering for Chef Top, Baker Top and total ovens

	Chef Top		Baker Top		Total ovens	
Filter	**N° of eliminated programs**	**% of eliminated programs**	**N° of eliminated programs**	**% of eliminated programs**	**N° of eliminated programs**	**% of eliminated programs**
< 1 min	1102	0.2	4536	2	5641	0.7
< 2 min	13063	2.5	24599	10.7	37902	4.9
< 3 min	24793	4.7	35749	15.6	60993	7.9
< 5 min	52583	10.0	50292	21.9	103801	13.5

In order to choose the minimum duration time for the cooking programs retained in the analysis (point 2), different thresholds have been tested to avoid the elimination of useful data, as reported in Table 3.2. Note that the percentage of eliminated cooking programs is different between Chef Top and Baker Top ovens: the filters always eliminate a smaller percentage of

cooking programs for the first category than for the second one. This fact gives an indication about the general use of the ovens: program duration is smaller for Baker Top ovens than Chef Top ones. In fact, 22% of the cooking programs last less than 5 minutes (against the 10% of Chef Top typology). In order to avoid the elimination of almost 11% of Baker Top cooking programs with a 2-minutes filter, the 1-minute filter is chosen. In terms of total ovens, only 0.7% of programs are eliminated with this filter. After filtering, the total number of ovens considered decreases from 826 to 800. Then, the weekly frequency of cooking programs utilisation is calculated. The number of ovens for each frequency is represented in the histogram chart in Figure 3.3 (3 high frequency ovens are removed). Note that the distribution range is wide, since it goes from 0 to 180 cooking programs per week. 96.7% of the ovens perform less than 100 cooking programs per week: this is a high value considering that it means 14 cooking programs per day and that the average cooking program duration is 47 minutes. However, half of the ovens has a frequency smaller than 20 cooking programs per week (the red dot line in the chart), whereas the mean (the red dashed line in the chart) indicates a value of 29.2 cooking programs per week (about 4 a day). This means that, even if there are ovens with high frequency, the major part have a low cooking frequency.

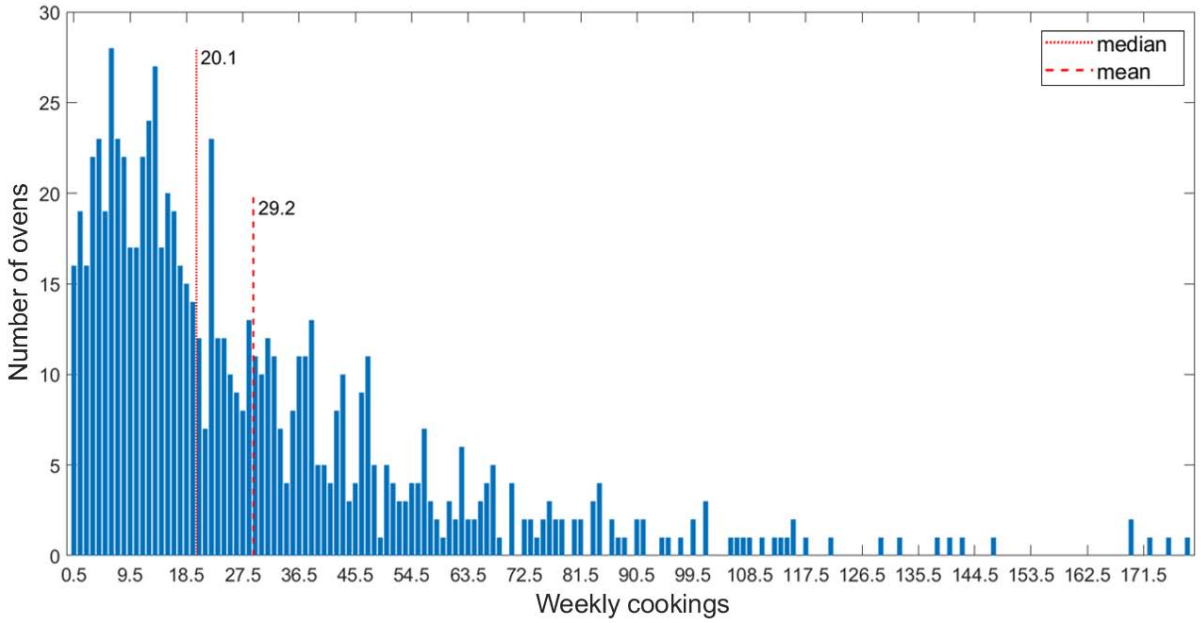


Figure 3.3. *Distribution of weekly frequency of cooking programs utilization of all ovens. The vertical red dot line indicates the median value, the vertical red dashed line indicates the mean*

The frequencies are then studied in terms of comparison between Chef Top and Baker Top ovens: their frequencies distributions are represented in Figure 3.4. Because of the difference in the total number of ovens between the two categories, the amount of weekly cookings for

each category is related to the percentage of the total number of ovens registered for each category.

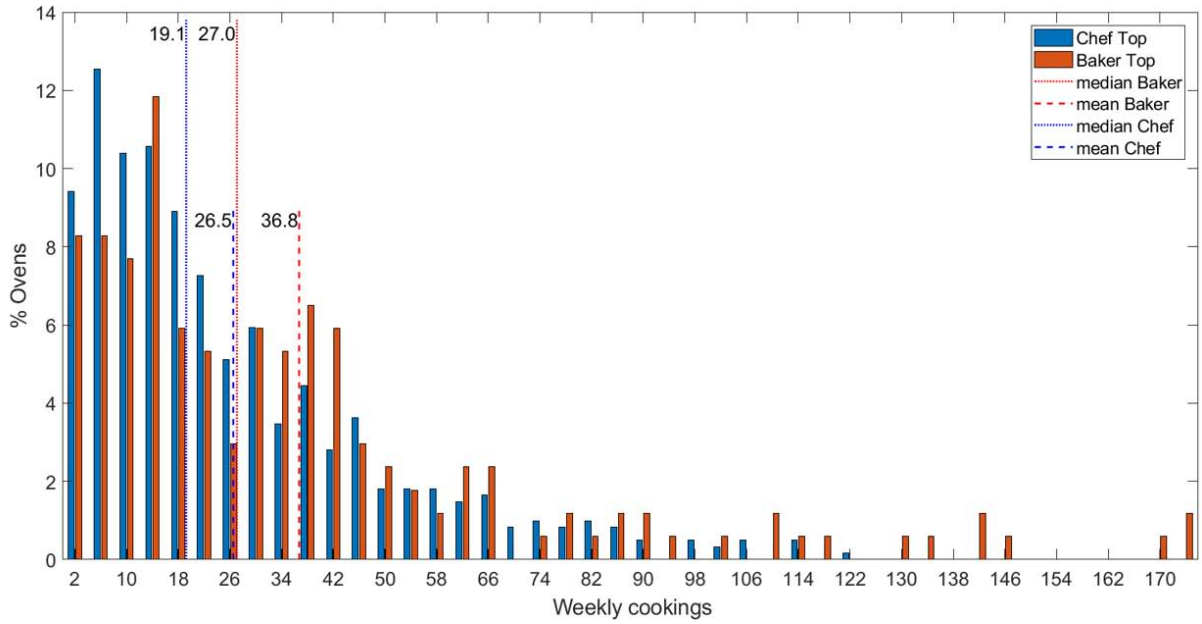


Figure 3.4. *Comparison between distributions of weekly frequency of cooking programs utilization of Chef Top (blue) and Baker Top (red) ovens. The vertical red dot line indicates the median value, the vertical red dashed line indicates the mean*

As in the previous analysis, both distributions have high percentage of ovens in the left part of the chart, the low frequency area. However, the maximum for Baker ovens is 174 whereas for Chef ovens is 122. Moreover, 92.3% of Chef ovens remain under the limit of 100 cooking programs per week, whereas the percentage of Baker ovens with less than 100 cooking programs per week is 98.5%. The highest peaks of the two distributions have different locations: 13% of the Chef Top ovens show 8 cooking programs per week, whereas 12% of the Baker Top ovens cook 14 times a week. This trend is confirmed by both mean and median: they present higher values for Baker Top ovens than for Chef Top ones. All these differences indicate that Baker Top ovens users cook more often, but the mean duration of a cooking program is significantly lower than the one of Chef Top cooking programs. In particular, the mean duration for Baker Top ovens is 24 minutes, while Chef Top cooking programs last 58 minutes on average (this tendency was anticipated by Table 3.2). Similar results are obtained if Chef Top PLUS and Baker Top PLUS ovens are compared.

Before studying the typologies of cooking programs used by high and low frequency ovens, a general view of cooking program utilisation is presented.

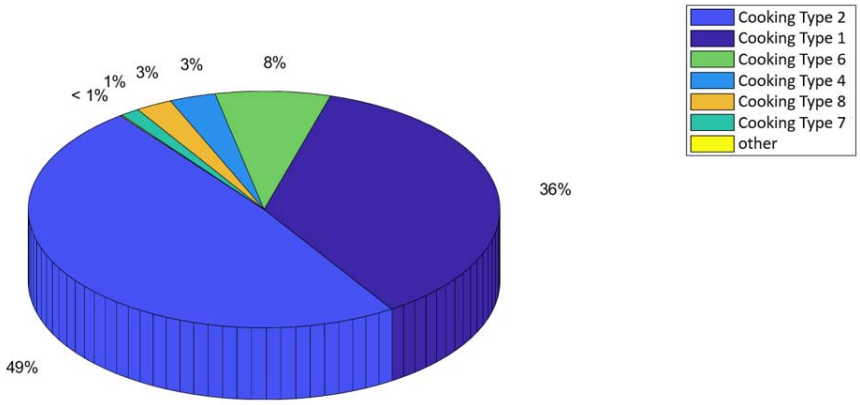


Figure 3.6. *Cooking programs analysis for all the ovens*

As shown in Figure 3.6, the most used cooking programs are types 1 and 2; in other words, the users prefer setting the cooking parameters autonomously in 85% of the cases (both by setting them at the beginning of the process or saving them for future cooking). Unox pre-set programs are selected in 12% of the cases: the most used pre-set program is cooking type 6 (8%). Finally, percentages smaller than 1% are obtained for cooking types 3 and 5. Although these two last programs could not be eliminated/integrated in other programs due to their significance in terms of marketing strategy, the other programs with small percentage could be integrated or improved through final user feedback. Finally, since the first two types are the most used programs some simplifications or improvements can be implemented for a better user experience.

The typologies of high and low frequency ovens are now studied and Figure 3.7 shows the cooking programs of the 10 most used (left) and the 10 least used ovens (right). It can be immediately noticed that, for both cases, cooking types 1 and 2 prevail. However, they occupy different pie portions in the two charts. For the high-frequency ovens the cooking type 2 dominates in the user preference with a percentage of 71%, whereas 28% of the cooking is run in the first cooking mode. The two modes are more balanced for the ovens used less frequently: the first program is selected in 48% of cases, whereas cooking type 2 has a 45% of preference. Furthermore, note that (excluding the two main programs) the other kinds of cooking programs represent a total percentage of 6% for the least used ovens, whereas they contribute for the 1% for the most frequently used ovens. Once more, the importance of the first two cooking programs is highlighted, especially for the most frequently used ovens. Since among them there are ovens with up to 170 cooking programs per week (see Figure 3.3), the implementation of new features for these two programs appears fundamental for the improvement of user experience.

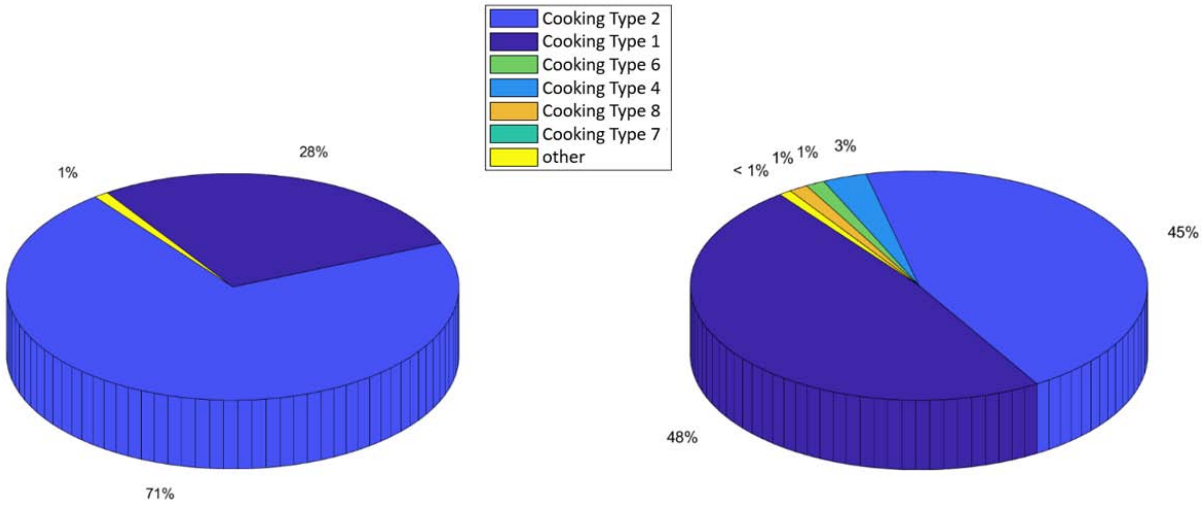


Figure 3.7. *Comparison of the 10 most used (left) and 10 least used (right) ovens by type of cooking programs*

At this point, it is of great interest to study if the same behaviour is shown by the ovens of the two categories Chef Top and Baker Top classes; the comparison is shown in Figure 3.8. Knowing that, in general, Baker Top ovens have a higher cooking frequency than Chef Top ones, the two categories follow the same trend described in the comparison between most and least used ovens. Although the first two cooking types are the most used for both categories, Baker Top ovens show 78% of cooking type 2 against 15% of type 1, whereas Chef Top models use cooking type 1 in 45% of cases against 36% of type 2. Moreover, Chef Top ovens use different cooking programs, whereas only cooking type 6 has a percentage higher or equal to 1% in Baker Top ovens. The results obtained are compatible with the main use of the ovens of the two categories; as the name says, Baker Top ovens are mainly used in bakery, where the variety of plates to cook is not as large as in restaurants/canteens where Chef Top ovens are mainly employed. Moreover, the fermentation processes in bakery are very sensible to changes in process parameters; the user thus could prefer to save the cooking settings used in a successful cooking and use them rather than change them (cooking type 2). Nonetheless, it is possible to use Chef Top ovens in bakery and vice versa, so the exact individuation of a typical behaviour is not straightforward. Similar analyses have been carried out for all the other sub-categories but they are not discussed here due to the low number of registered ovens/cooking programs.

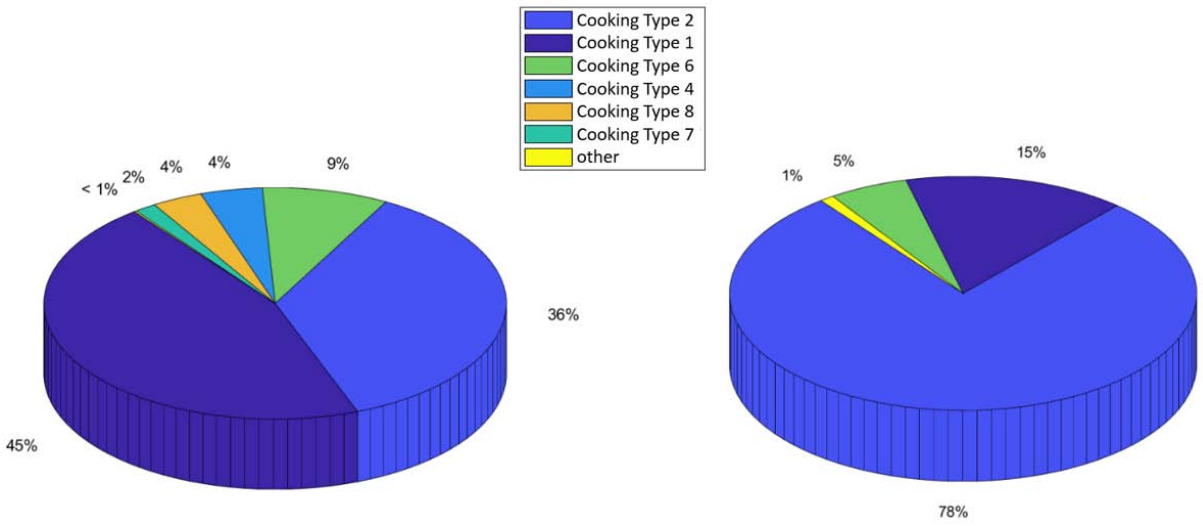


Figure 3.8. *Comparison between Chef Top (left) and Baker Top ovens (right)*

3.5 Washing programs analysis

The aim of this analysis is to study the washing programs (kind of cooking = "1") used by different types of ovens. It would allow the company to understand the user habits in order to implement useful improvements to the washing system.

The different types of washing programs are tagged by a code that varies locally, depending on the specific language of final user; this issue caused some problems in washing type identification due to foreign language characters recognition issues during data extraction phase. Moreover, washing categories could not be split into groups according to their duration because the programs could have been interrupted by the user before the usual washing protocol end. The problem was solved by: (*i*) grouping all the washing programs with the same untranslated code; (*ii*) assigning them to the most similar coded category in terms of mean duration time. The following cases have been removed from the analysis:

- washings that last more than 3 hours: when the detergent finishes during the program, the oven waits for a new one and the registration period does not stop at the end of the time program;
- washing programs that are only used by Unox commercials to show the washing capability of the ovens during product presentation;
- washing programs without name or enough information for the analysis;
- washing programs associated to ovens with a higher number of washings than cooking programs: these ovens are commonly used by Unox commercials in product presentation.

In addition to these filters, the same filters previously described for cooking programs selection are maintained and the same operating period is considered to calculate washing frequencies, in order to get a common basis for the following analysis. The total number of ovens after filtering decreases from 800 to 771 because not all the ovens have a registered washing activity. Following the same procedure as for cooking programs, the weekly frequencies of washing programs utilisation are obtained and split into different categories. The distribution of washing frequencies for all the ovens is represented in Figure 3.9. Washing frequencies go from 0.05 to 19.05 washings per week, but they are mainly concentrated in the left part of the chart (low frequencies). In fact, 91.5% of the ovens are washed less than once a day. Moreover, ovens are washed 3 times a week on average, while the median indicates a value of 2.1: half of the ovens are washed less than twice a week. Another relevant result is that 5.1% of the ovens are washed less than once a month. The fact that ovens are washed so rarely has a negative impact on their performances (beside the smell of burning dirt). Dirt accumulates all over the oven, also on the

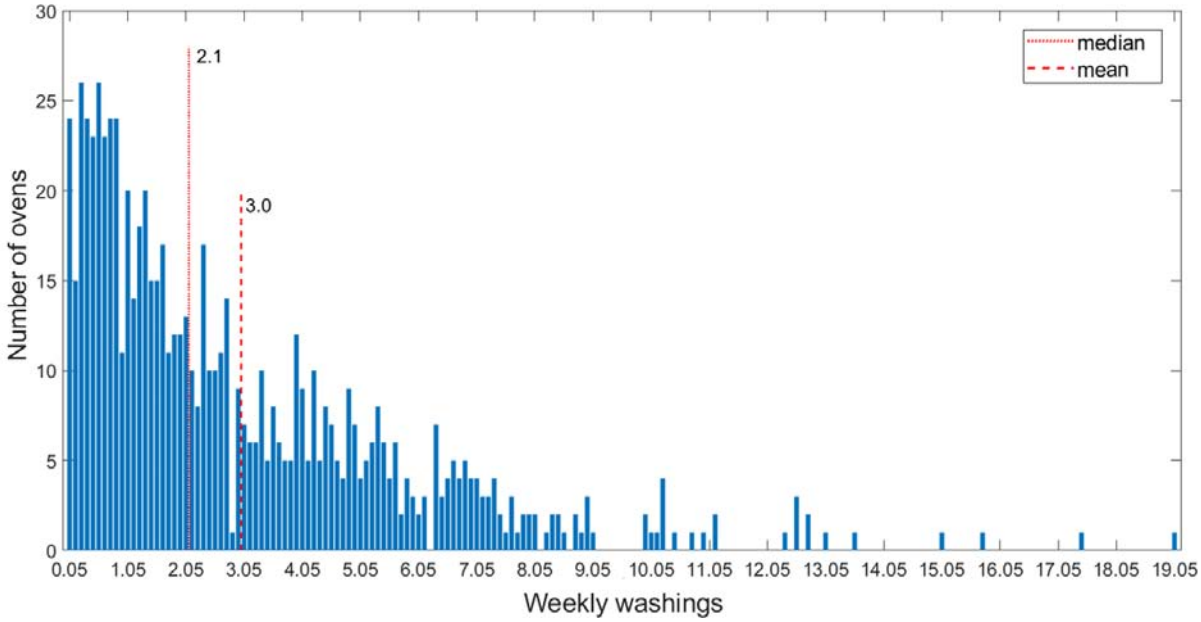


Figure 3.9. *Distribution of weekly frequency of washing programs utilization for all the ovens. The vertical red dot line indicates the median value, the vertical red dashed line indicates the mean*

resistances, hence slowing down the energy transfer; the resulting temperature field is not uniform even though the fans partially compensate for this effect. Moreover, lack of washing can have consequences on safety; the thick fat layer accumulated on the resistances can create hotspots and even start to burn. For this reason, the company has introduced an automatic washing program after a certain amount of chicken cooking programs. However, fat accumulates not only during chicken cooking programs and the introduction of an automatic

washing program after a pre-set amount of cooking programs should be taken into account; the choice has to be based mainly on the type of cooking programs and their duration. From expertise analysis, it has to be noticed that cooking programs of Baker Top ovens are usually cleaner than the Chef Top. For example, cooking bread (Baker Top) does not dirty the oven as much as cooking chicken (Chef Top).

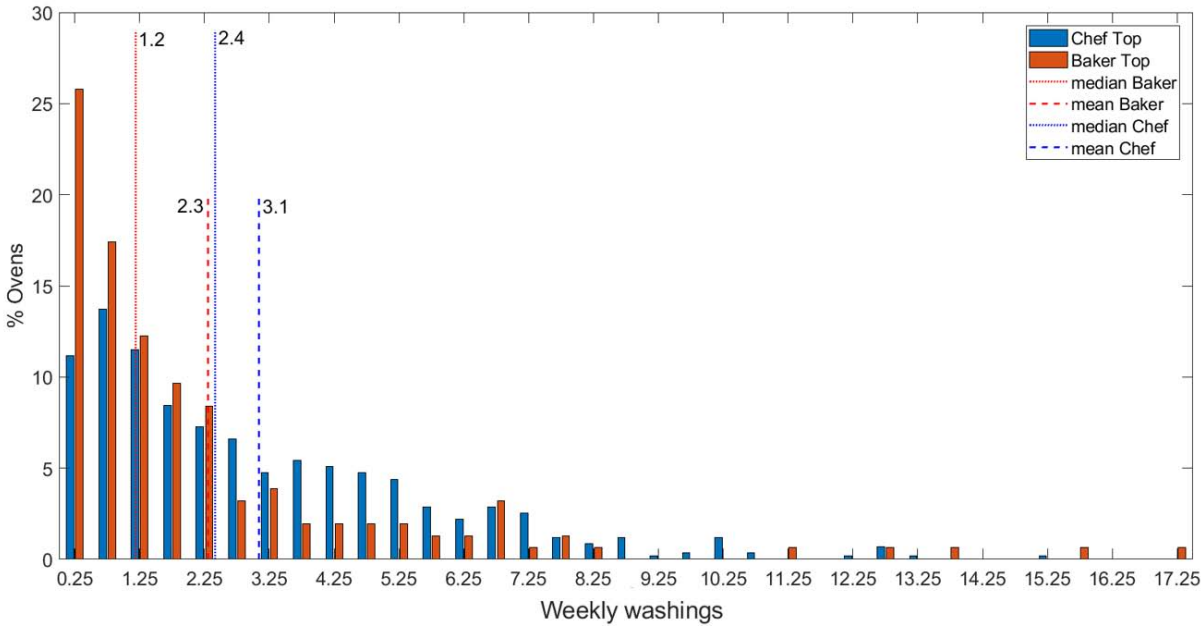


Figure 3.10. *Comparison between distributions of weekly frequency of washing programs utilization of Chef Top (blue) and Baker Top (red) ovens. The vertical red dot line indicates the median value, the vertical red dashed line indicates the mean*

This hypothesis is partially confirmed by Figure 3.10 that represents the comparison between distributions of weekly frequency of washing programs utilization of Chef Top and Baker Top ovens. Even though both distributions are located in the low-frequency area (left part), the Baker Top distribution shows an initial pick of 26% of the ovens that are washed about once a month or less. Chef Top distribution is smoother and reaches a maximum value of 14% of the ovens in correspondence with a washing frequency of 3 times a month. Moreover, half of the Baker Top ovens are washed less than 1.2 times a week, while 50% of Chef Top ones are washed less than 2.4 times a week. However, the range of washing frequency values for Baker Top ovens is wider than the one associated to Chef Top units (some Baker Top ovens are washed more than 17 times a week). This difference in the ranges is also highlighted by the differences in the mean values that are smaller than the median values. Baker Top ovens are washed on average 2.3 times a week whereas in Chef Top ones the washing program is activated 3.1 times a week. The same analysis has been done for other sub-categories, but the results are similar to those previously explained and they are not discussed in this Thesis.

Finally, in order to have a complete insight of the oven utilisation, the study of the different washing programs is carried out.

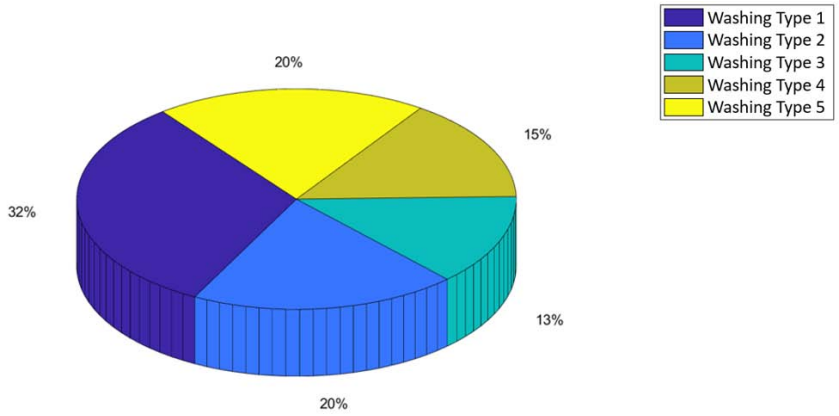


Figure 3.11. *Washing programs analysis for all the ovens*

The results of the global analysis on washing programs reported in Figure 3.11 show that there is no predominance of any of the 5 washing typologies, although there is a slight tendency to use the fastest ones: 65% of washings use the first three programs. Moreover, note that 20% of washings use the washing type 5. This means that users prefer to run short or long programs, while intermediate programs are less used. These results can be linked to the washing frequency through Figure 3.12: it shows the programs of the 30 most washed and the 30 least washed programs.

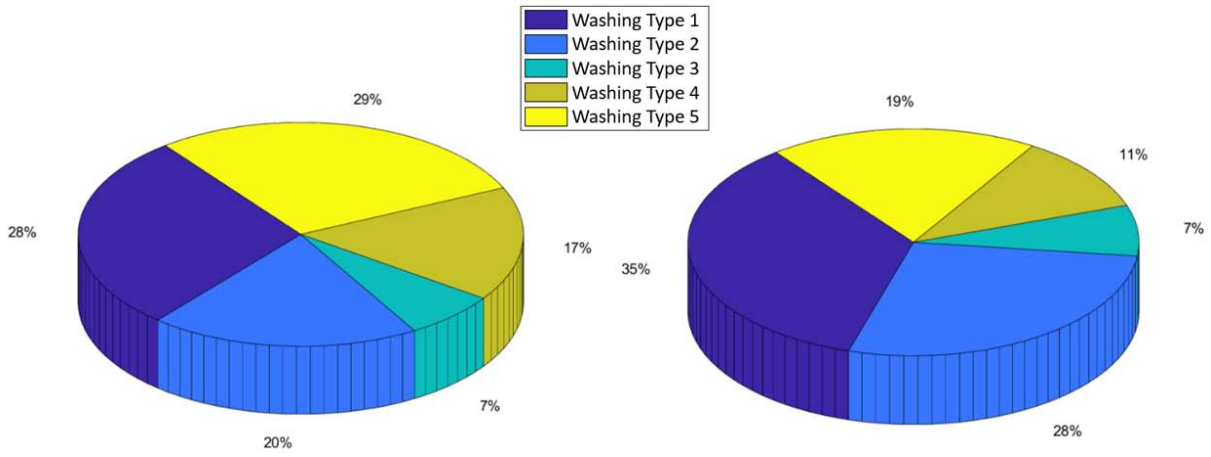


Figure 3.12. *Comparison of 30 most used (left) and 30 least used (right) ovens by type of washing programs*

All the programs are used both from very frequently used and infrequently used ovens, even though for the most used ovens a preference for longer programs is shown. In fact, there is a prevalence of washing types 4 and 5 for the most frequently used ovens, while the least frequently used use mainly shorter programs: in 70% of cases one of the first three washing modes is used. Moreover, it can be noticed that washing type 3 is always the least used for all

the type of ovens. For this type of program the integration with the washing type 2 can be taken into consideration since there is a difference of only 10 minutes of duration. However, the most important aspect highlighted by the results is that the ovens that are washed more frequently are also the ones with longer programs and vice versa. This result is contradictory to what one may expect; in fact, if the oven is frequently washed, less dirt accumulates and shorter programs are enough to clean it up. However, it can be explained by considering that ovens that are frequently washed with long programs are the one that perform dirtier cooking.

Since the cleanliness and dirtiness have been linked to the category to which the oven belongs, the washing programs typology analysis is run over both Chef Top and Baker Top ovens. The results of the comparison is reported in Figure 3.13. There is no significant differences between the two categories: all the washing programs show very similar percentages of usage. Washing types 1 and 2 have the same percentage in both cases while there is only a slight increment of type 5 and a subsequent decrease of type 3 for the Baker Top ovens (and vice versa). This result contradicts the hypothesis that Baker Top ovens perform cleaner cooking, but confirms the fact that ovens from both categories can be used by different users to cook different type of food, so there is not a net distinction of user type.

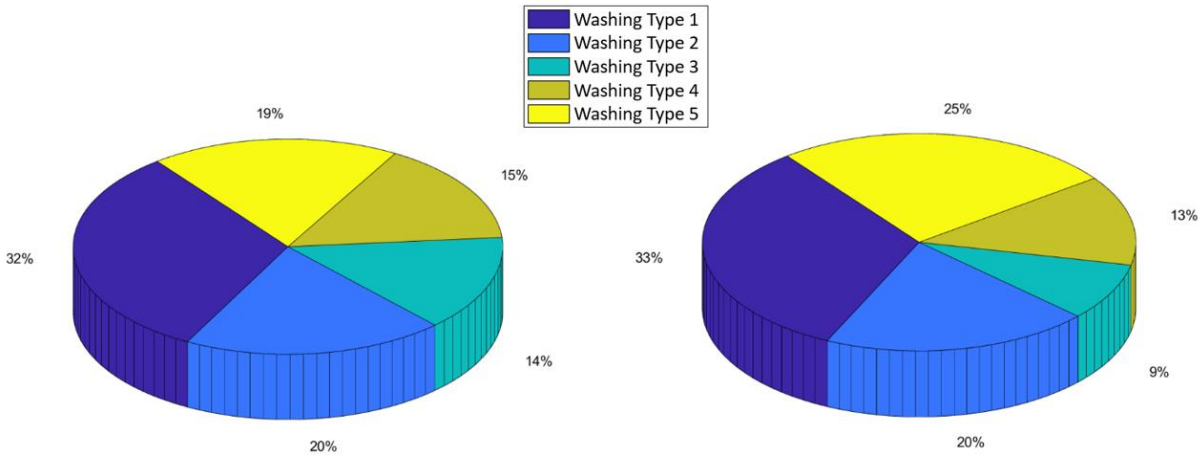


Figure 3.13. *Comparison between Chef Top (left) and Baker Top ovens (right)*

The analysis of different sub-categories produces results similar to those presented above and are not reported in this Thesis.

Chapter 4

Case study 1: Process monitoring

In this Chapter an exemplificative case study of process monitoring is shown. Process monitoring is obtained through a PCA model based on historical data collected for the Baker oven "1816". Although the model is oven-specific, the proposed procedure can be used for each oven in the database to detect (during early stages of the program) the cooking programs that do not follow the normal operating conditions.

4.1 Oven and data selection

In this Thesis, the process monitoring procedure is described for a specific oven. Theoretically, a model could be obtained for all the ovens or for a specific category of ovens, but increasing the number of ovens the model would include different behaviours, hence becoming less sensitive to small variations and changes.

The oven presented in this Thesis as an illustrative case study is a Baker one with a low weekly washing frequency. The oven identified by the ID number "1816" has been chosen because it belongs to the 26% of the Baker ovens that are washed less than once per month (the major category represented in Figure 2.10). In fact, it has registered only one washing program during the 5 months of operating period. Moreover, the oven presents a low weekly frequency of cooking programs: it registered 2.5 cooking programs per week on average.

As already mentioned in Chapter 3, the samplings of the process variables are stored in a .csv file for each oven. Since the different cooking programs are not discernible in these files, they are identified through the Matlab table that contains starting and finishing time of all the programs (see §2.1). Crossing the information given by these two files allows identifying the process variables associated to a specific cooking program (as shown in Figure 4.1). Moreover, since a batch-wise unfolding approach is not directly feasible due to the different duration of each cooking program, a preliminary downsampling procedure is conducted by selecting the same number of equidistant time samples for each cooking program. In this case, the total number of cooking programs identified for this oven is 82. For each program, 20 evenly spaced

rows are selected; the programs with less than 20 rows are eliminated. After this step, the number of cooking programs is reduced to 52: this means that 30 cooking programs last less than 10 minutes. It is important to notice that all the cooking processes of the oven are run using the cooking type 1. Finally, the 3D data matrix is created as shown in Figure 4.1: the 52 cooking programs are arranged along the vertical side (samples), the 6 process variables along the horizontal side and the 20 sampling times recede into the figure.

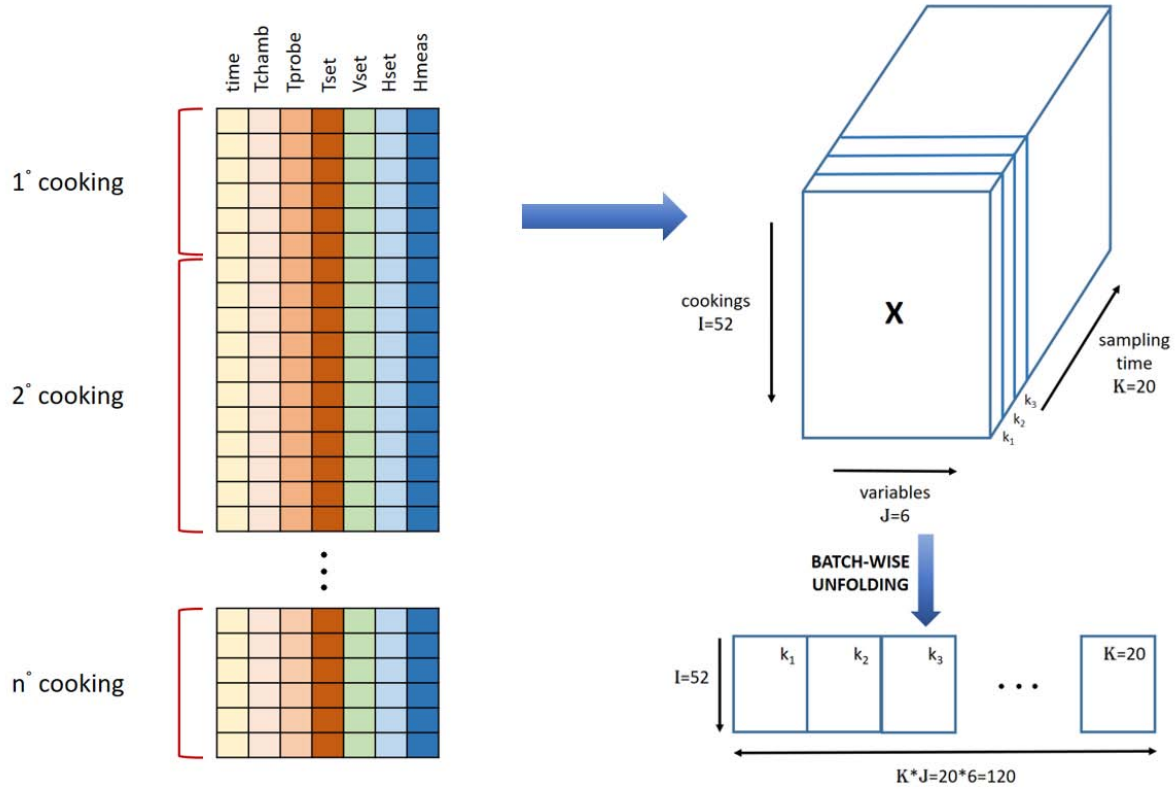


Figure 4.1. *Graphical representation of 3D matrix creation and unfolding for PCA analysis*

4.2 Pre-processing and unfolding

In order to implement PCA analysis, batch-wise unfolding is carried out on the 3D matrix. The vertical slices *(I×J)* at constant time sampling are placed side by side along the variable axis. The resulting 2D matrix (represented in Figure 4.1) has 52 rows and, as columns, the 6 variables set/measured at the first sampling time, then the 6 variables for the second one and so on to the last one (sample 20). The final matrix dimension is *52×120*.

Before PCA implementation, pre-processing treatments are done on the matrix. Since the analysis is run with the SVD algorithm and the process variables have different measurement scales, data matrix auto-scaling is performed.

4.3 Process model

In order to select the number of principal components to retain in the model, the PCA analysis is run with all the possible principal components (i.e., the minor size of the matrix) and the graphical representation in Figure 4.2 is obtained.

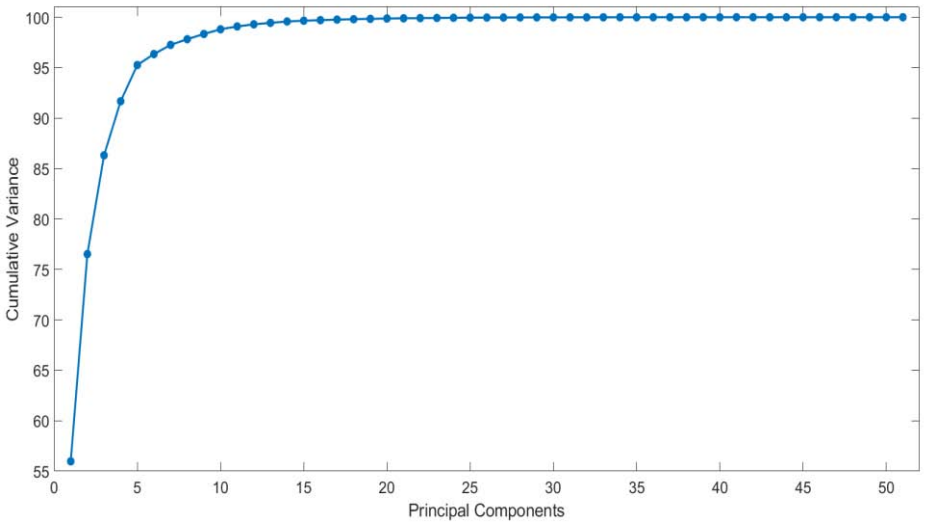


Figure 4.2. *Graphical representation of the cumulative variance captured by the principal components*

Figure 4.2 represents the monotonically increasing profile of the cumulative variance with respect to the number of principal components. It allows the evaluation of the contribution of each subsequent PC to the total cumulative variance. The curve reaches 99% with the 10th principal component, but a value of cumulated variance higher than 95% can be achieved with 5 components. As pointed out in Table 4.1, the first component captures the largest amount of variance (56% of the variance in data is represented by the first PC). The second one captures 21% of variance and the third one another 10% of data variability. In summary, 86% of data variability is explained with only three principal components. From the fourth principal component the captured variance is less than 5%.

Table 4.1. *Variance captured by the first three principal components*

PC	% Variance Captured	% Cumulative Variance Captured
1	55.98	55.98
2	20.54	76.52
3	9.79	86.31

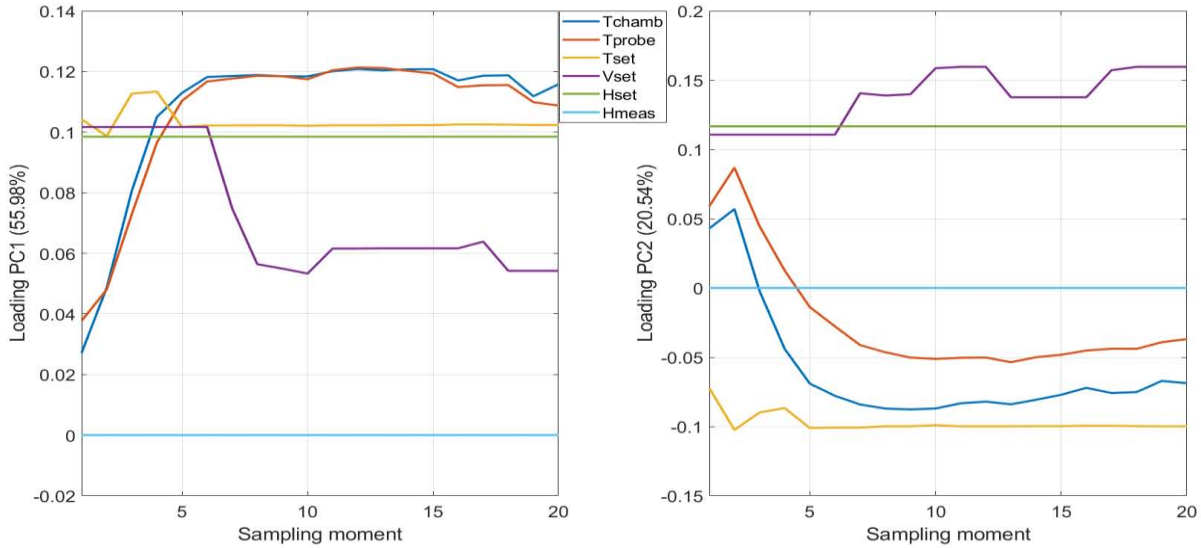


Figure 4.3. *Time profiles of variable contributions to PC1 (left) and to PC2 (right)*

As already mentioned, the number of retained principal components is chosen on the basis of the percentage of the cumulative variance described. For this first explorative PCA analysis, 86% of the variance is considered enough to represent the system since information loss is low; for this reason, a three principal components PCA model is built.

In Figure 4.3 the time profiles of the contribution of the original variables to the definition of the principal components are shown: the left chart represents the contributions of each original variable to the definition of PC1, whereas the right one shows the contributions to PC2. Only these two PCs are shown because they capture the highest amount of data variance. As it can be seen, the measured humidity has no effect on both PCs, whereas the set humidity has a positive effect on both of them (similar values). The set temperature has symmetrical effects on the PCs. For both PCs, it has a constant profile after a small initial variation: it assumes positive values for the first PC, whereas negative values for the second. The set ventilation has a positive effect on both PCs. The profiles start from (almost) the same positive value and at the 6th sampling moment decrease for the first PC and increase for the second one. Then, they both stabilise around a fixed value with slight variations. As expected, the chamber temperature and the probe temperature vary in the same way for both PCs, even though there is a small difference between the two profiles for PC2. Both profiles start from positive values, vary till the 6th sampling time and then stabilise to a constant value. However, while for the first PC they increase and stabilise to a positive value, for the second one they decrease and stabilise to a negative value. Set ventilation, chamber and probe temperatures present correlated patterns: in fact, the two temperatures vary at the beginning when set ventilation is constant, whereas set

ventilation starts to vary when the two temperatures are almost constant. This fact represents the actual behaviour of the system. During the preheating phase, when ventilation is set by default to 4 (constant), the goal is to heat up the oven in minimum time. This default choice can be explained by considering the heat transfer mechanism. Heat is first transferred from the hot serpentine to the adjacent air by conduction and then it is transferred to the whole oven by convection. In order to reach the goal and since forced convention (created through external ventilation) has a higher heat transfer rate than natural convection (without external ventilation), the maximum ventilation is used. Then, when the oven reaches the set temperature, the ventilation varies according to the user settings. Finally, the slight variation towards zero of the measured temperature profiles at the end of the cooking can be ascribed to oven door opening for final food quality check.

4.4 Model evaluation statistics

Once PCA model is built, T^2 Hotelling and Q statistics are evaluated as explained in Chapter 2. The cooking programs that present high values of these statistics usually present anomalies and should be examined one by one to assess if they must actually be neglected in model calibration.

4.4.1 T^2 Hotelling statistic

The scores plot is reported in Figure 4.4; the cooking programs are represented with different colours and numbers in chronological order as the colorbar shows. Moreover, the dashed line ellipsoid represents the 95% confidence limits of the Hotelling T^2 statistics. It can be seen that cooking programs 50 and 52 are located out of the limits. This means that they present some values in the process variables that are far away from the multivariate mean of oven "1861" cooking programs. However, while cooking 50 departs from the ellipsoid mainly along PC3 direction, cooking 52 is located out of the confidence limits along the first two principal components directions. Since PC1 and PC2 represent the highest percentage of data variability, the samples located far from the confidence limits in these directions are the ones that present the largest distances from the multivariate mean. The variable profiles of the 52^{th} cooking present anomalies if compared to the other cooking programs included in the model. This statement is confirmed by Figure 4.5 where the T^2 statistic values of each cooking program are represented by a bar. In fact, cooking 52 assumes the largest value and is represented by the highest bar in the plot. Cooking programs 50 and 26 also (slightly) overcome the red dashed

line that represents the 95% confidence limits, but their T^2 values are smaller than the one related to the 52th cooking.

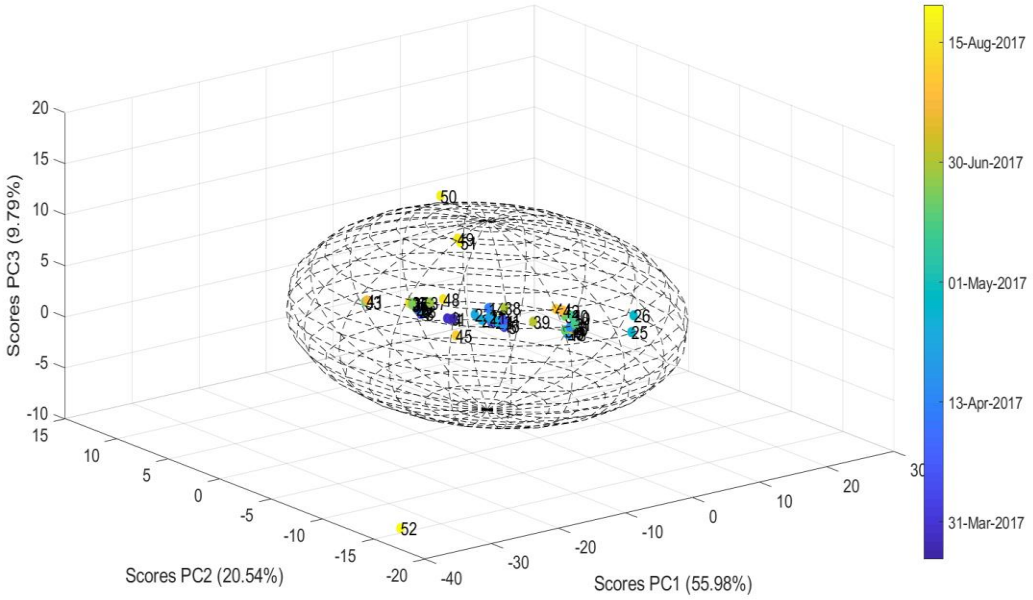


Figure 4.4. *3D scores plot of the cooking programs of oven "1816": the colorbar on the right represents the chronological sequence of the programs and the ellipsoid defines the confidence limits*

As already noted, the T^2 statistic points out samples that do not follow the normal operating conditions, but are still in the subspace defined by the principal components retained in the model. In order to understand the reason of the high values, the time profiles of the 6 process variables of programs with high T^2 values are studied and compared to the time profiles of the cooking with the lowest T^2 value. First, the cooking program that is more similar to the mean profile is studied; in this case, it is cooking 14, because it is associated to the minimum value of T^2. The variable profiles of cooking 14, shown by Figure 4.6, represent normal operating conditions for the considered oven. In particular, the set temperature and ventilation are set equal to 120°C and 4, respectively. The measured and set humidity assume zero value. The only variables that change are the temperature in the chamber and the temperature of the probe placed inside the food. They both start at 37°C, increase for about three minutes, then overcome the set-point. However, since the set-point on the temperature is used to control the chamber temperature, this last stabilizes slightly above the set-point (120°C), whereas the probe temperature increases to a maximum of 137°C and then gradually decreases. Note that temperature drops for both the two measured temperature profiles. The fast rate of temperature decrease and rise around 13.5 min can be attributed to door opening and closing for quality food

checking, respectively. This behaviour has not been considered as an anomaly since it belongs to normal operating conditions.

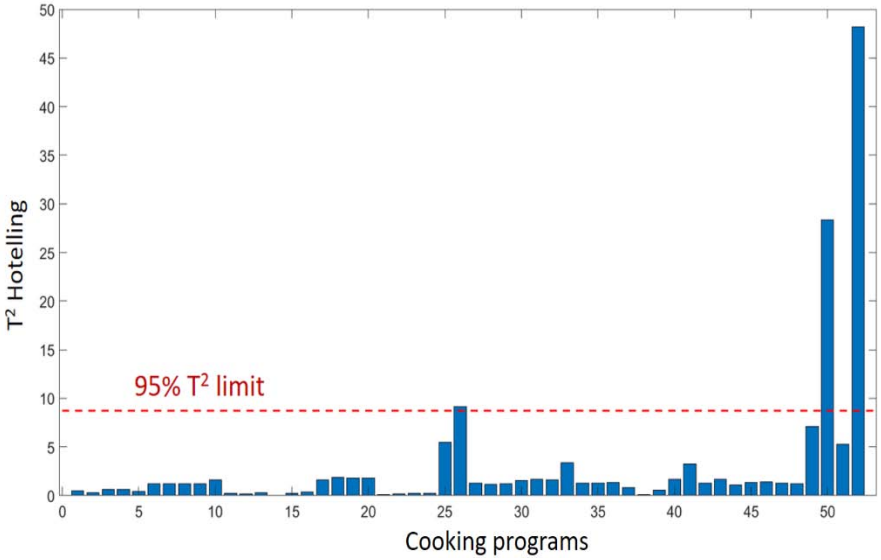


Figure 4.5. *Hotelling T^2 statistic for each considered cooking program of oven "1861": the red dashed line represents the confidence limit of 95%*

Cooking program 14 has been considered as the reference state for the cooking programs in oven "1861" in order to compare it with programs with high values of T^2 statistic. As an exemplificative case, the cooking program with the highest T^2 statistic (number 52, as shown in Figure 4.5) is now considered. Its process variables profiles are reported in Figure 4.7. The most important difference between this program and the reference case is the temperature set value (30°C, the minimum possible for the oven). Another difference consists in the variation of the temperature profiles for both the probe and the oven chamber. Unlike program number 14, the two temperatures do not increase concurrently, but, apart from an initial transient, the chamber temperature decreases as the probe temperature increases. This behaviour is due to the fact that the food placed in the oven is at a low temperature (8.8 °C). When heating is switched on, the temperature in the chamber rapidly increases and reaches the set-point, whereas the core probe does not measure this trend. In fact, the temperature rise has an immediate effect only on the external layer of the food, which is heated up and presents preliminary superficial water evaporation. The evaporation causes a small decrease in the food internal temperature (after about 2.4 min) because the latent heat is absorbed from both the chamber and the food core. Although this effect is usually negligible and is not measured, here it is intensified by the humidity absorption set by the user. Then the probe temperature increases to a final value of 22.5°C, while the chamber temperature slowly decreases under the set-point.

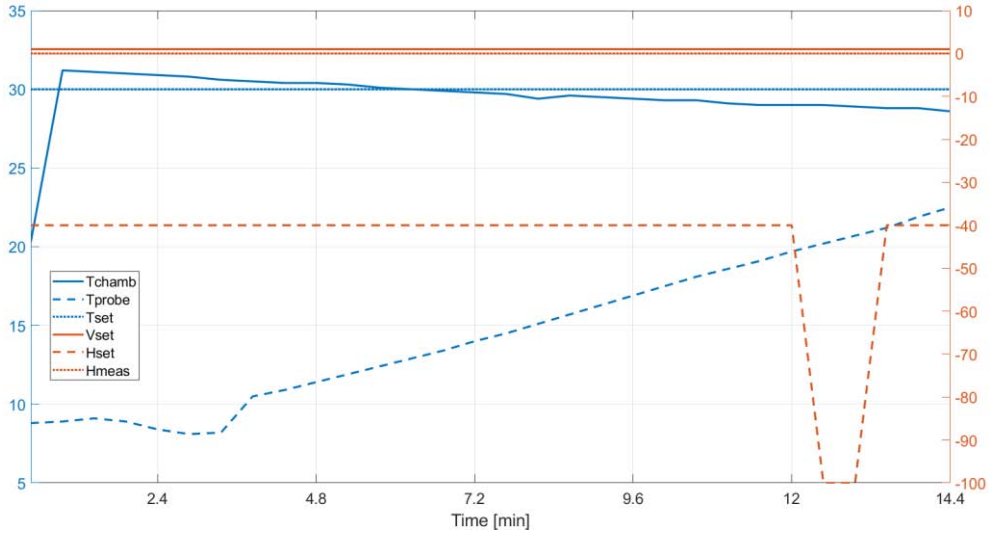


Figure 4.7. *Variables time profiles of cooking program 52*

Moreover, a difference in ventilation setting can be noticed: in cooking program 52 it is set equal to 2 against the maximum value (4) of program 14. This partially explains the difference in the temperature rate of change. Finally, the reason why cooking 52 is considered an "abnormal" cooking is that the oven was used to warm up some cold food at low temperature while all the other are programs are used to cook the food.

As reported in Figure 4.5, the cooking program 50 presents a high T^2 value. It has been studied in the same way as presented before. The discussion is not reported here; yet, program 50 has been eliminated together with cooking program 52 since similar differences in the process variables have been detected.

4.4.2 Q statistic

Q statistic is evaluated to detect the cooking processes that are not represented by the model. Unlike the T^2, the Q statistic underlines the points that are not included in the subspace defined by the first three principal components. As reported in Figure 4.8, the programs that overcome the confidence limit of 95% are number 25, 26, 48, 49 and 51. Cooking program 26 has the highest bar in the plot, meaning that its behaviour is the worst represented by the model. Its variables profiles are reported in Figure 4.9. If compared to the profiles of program 14, significant differences are not detected. In cooking program 26 the chamber and the probe temperatures increase at first while the ventilation is set to 4 by default: this is the preheating range.

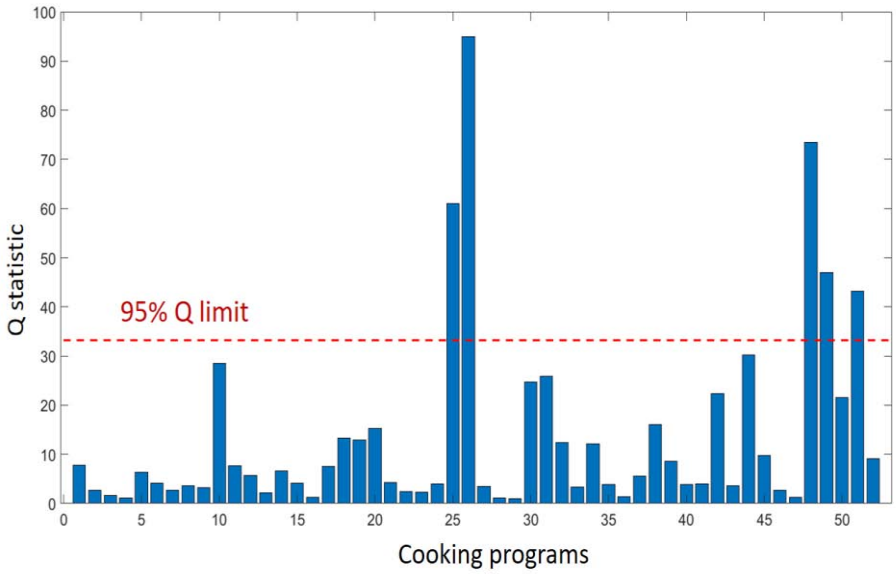


Figure 4.8. *Q statistic for each considered cooking program of oven "1861": the red dashed line represents the confidence limit of 95%*

Then, when the preheating finishes, the ventilation assumes the value set by the user and a negative pick is registered. It can be interpreted as the oven opening (that corresponds to a temperature decrease) in order to place the food and then door closing that allows the temperature to increase again. Since this behaviour is also present after 4 min in program 14 (it is not so evident due to the effect of downsampling), it is not considered as an anomaly. The same downsampling effect can be seen on the measured temperatures of cooking program 26; there is not an evident pick just before the process end that indicates the door opening/closing to check the food, like in program 14. It is barely visible around 14 min. In general, the Q statistic high value of cooking program 26 is not justified from a physical point of view; the time profiles of the process variables do not present anomalies and process behaviour do not differ from a normal cooking process. For this reason, sample 26 is not excluded from the database.

On the contrary, for other cooking programs the high values of Q statistic are justified by looking at their process variables time profiles. As shown in Figure 4.10, temperature profiles of chamber and core probe of cooking program 49 are very irregular. They start from about 80°C and they never reach the set-point. The profiles decrease many times during the cooking duration because of the oven door opening and no preheating phase is registered (users can choose to skip it). In fact, the ventilation is initially set to "3" (not to the default value "4") and then switched to "4". It can be concluded that cooking program 49 does not represent the normal

conditions of a general cooking for the studied oven. For this reason, it is eliminated from the database as an outlier.

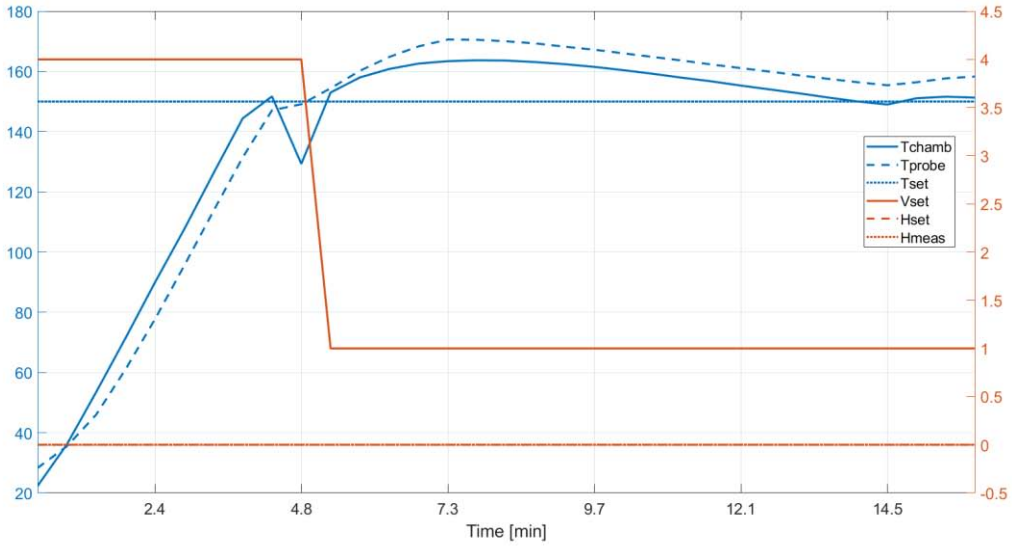


Figure 4.9. *Variables time profiles of cooking program 26*

This analysis is done for all the other programs that are not well represented by the model and, when physical explanation for high values of statistics are found, the program is eliminated. In the studied model, cooking programs number 48, 49, 50, 51 and 52 are eliminated and the new model can be developed.

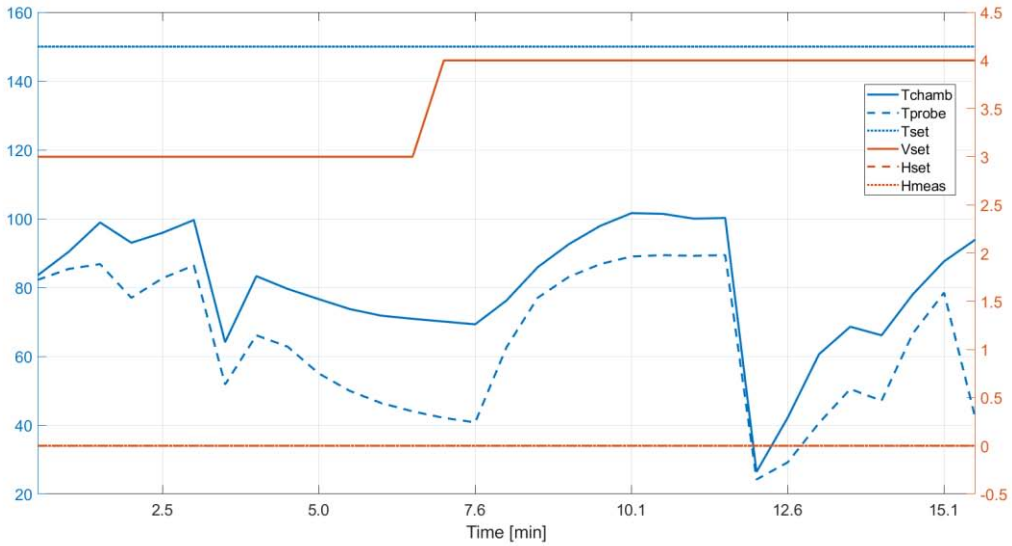


Figure 4.10. *Variables time profiles for cooking program 49*

4.5 Final process monitoring model

After outliers elimination, the final PCA model is recalculated by following the same procedure described in the previous sections. The retained number of principal components is maintained equal to 3 and the model captures 93% of the total variance; with an elimination of only 5 samples the model gains a 7% in the captured variance.

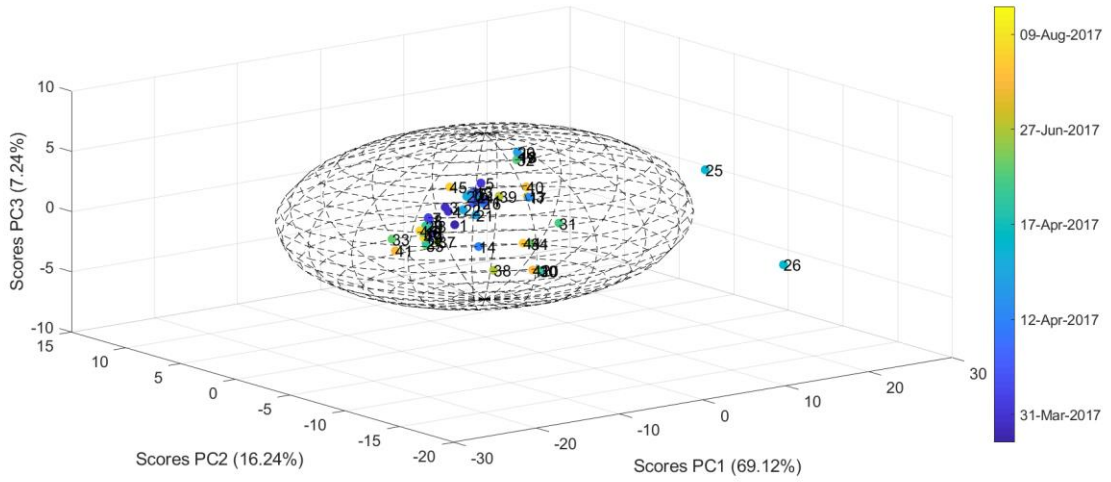


Figure 4.11. *3D scores plot of the cooking programs of oven "1816" final model*

Figure 4.11 represents the score plot of the model obtained. Samples 25 and 26 are placed out of the confidence limits. The common characteristic of the two cooking programs is the variation in the ventilation set by the user after the preheating phase (as it can be seen in Figure 4.9 for cooking 26). Even though this condition is not represented by the model, the captured variance reaches a high value without eliminating these two samples and the model is retained. However, if higher accuracy is needed, the presented procedure can be repeated iteratively and the definitive model can be found.

The obtained model can be used for future monitoring of cooking programs of the selected oven. Then, the programs running in conditions different from the normal ones can be detected. The user can identify these cooking programs and modify the operating conditions in order to match the product specifics. Moreover, the intervention of the technical service can be requested in order to detect malfunctioning through further analysis.

Chapter 5

Case study 2: Predictive maintenance

Prediction of equipment damage or rupture would allow the client service to take action before problems actually arise during normal operation. In this context, PLS-DA methodology is a suitable tool to predict abnormal process conditions leading to oven malfunctioning. In this Chapter the application of PLS-DA technique to two case studies dealing respectively with gasket and core probe damage prevention is discussed.

5.1 Gasket substitution

Oven gasket is used to guarantee both perfect insulation and uniform temperature profiles inside the oven. However, process operation at critical conditions such as high temperatures and high humidity may reduce gaskets lifetime, hence lowering oven performances. Although the company has already implemented technical solutions to decrease gasket damage occurrence, no data-driven procedure is currently available to predict damage/rupture events. Although 800 ovens are linked to the company cloud storage, the total number of gasket substitutions registered in 2016/2017 is equal to 15. On the one hand, it means that the technical solutions implemented by the company are already efficient; on the other hand, the low number of registered damages limits the activity related to this case study for PLS-DA model calibration. Further data should be collected to validate the model in a statistically sound way.

The available data have been rearranged as shown in Figure 5.1. Each row of **X** matrix contains the values of the 6 process variables (chamber, core probe and set temperatures, ventilation and measured and set humidity) sampled at 20 evenly spaced time points for 20 consecutive cooking programs. In total, the number of columns of matrix **X** is 2400 (6 variables × 20 sampling times per cooking program × 20 cooking programs). This matrix structure has been chosen to consider gasket damage dependence on previous oven history. The response matrix **Y** is reduced to a column vector **y**, whose i^{th} component assumes value "0" if no technical intervention is registered for the final cooking program of the sequence represented in the i^{th} row of **X** matrix, or "1" in case of gasket substitution after the 20^{th} program. The total number of rows of **X** and

y is 30 in order to have the same proportion of sequences of cooking programs with gasket substitution (15) and sequences of cooking programs without technical intervention (15). Although this choice reduces the samples to a small number, it avoids the misbalance between classes and allows maintaining the predictive capability of the model. If class "0" would contain a larger number of samples than class "1", PLS-DA analysis would not detect the differences between the two. In fact, it would create a model that interprets the cooking sequences with a technical intervention as abnormal as other cooking sequences without intervention, only due to their distance from the multivariate mean. This way the prediction would not be accurate even with a high number of latent variables.

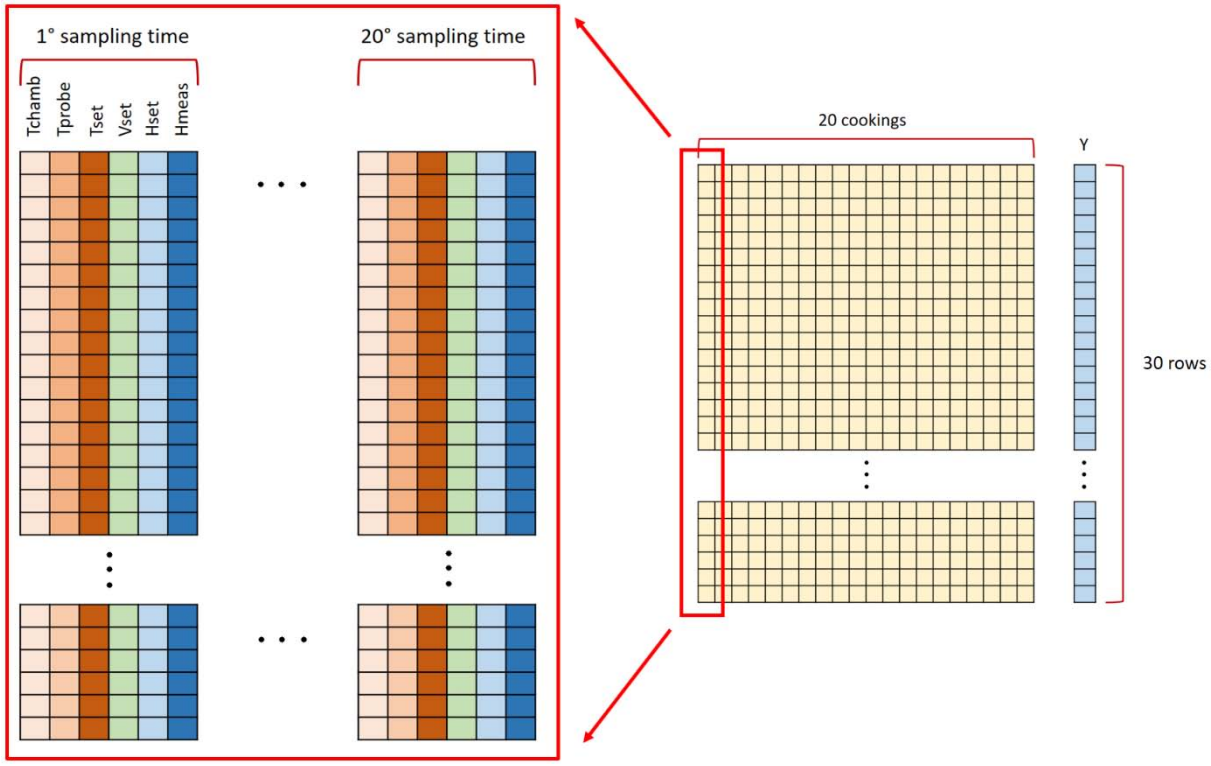


Figure 5.1. *Graphical representation of data matrix **X** and vector **y** structure used for PLS DA analysis*

In order to select the sequences of cooking programs belonging to class "0", 50 simulations are run. For each simulation, 15 sequences are randomly chosen and included in **X** matrix. The following results are related to the case with the best prediction for **y** vector.

The auto-scaled **X** matrix and the response vector **y** are used to calibrate a model with a PLS-DA analysis. As previously mentioned, this type of analysis is based on the rotation of the principal components in order to explain the variability in **y** (i.e. to highlight the differences between the cooking sequences followed by a technical intervention and those that are not). As

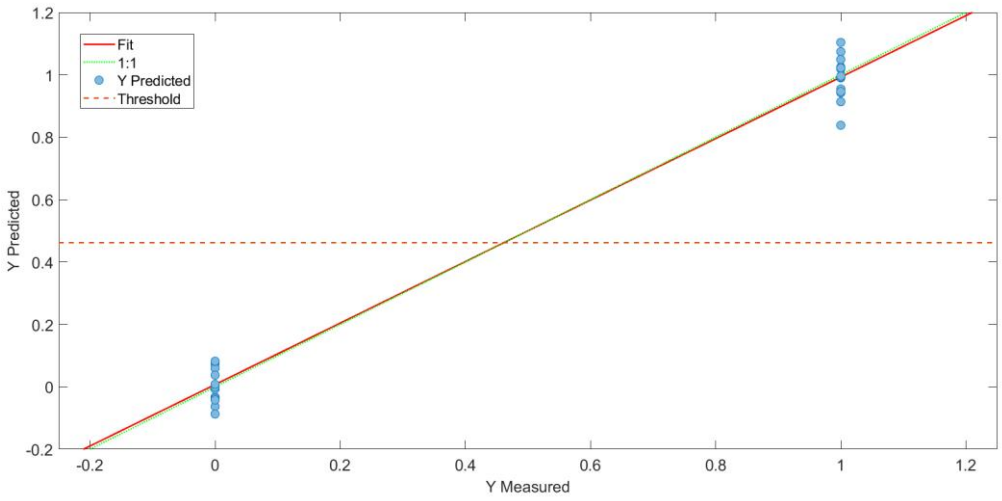
shown in Table 5.1, 86% of the **y** variability (and 22% of the variability of the **X** matrix) is explained with only 2 latent variables.

Table 5.1. *Variance captured by the first four latent variables for both **X** and **y**-block*

	X-Block		y-Block	
LV	**% Variance Captured**	**% Cumulative Variance Captured**	**% Variance Captured**	**% Cumulative Variance Captured**
1	17.27	17.27	34.51	34.51
2	4.74	22.01	51.62	86.13
3	5.76	27.77	9.79	95.92
4	5.68	33.44	2.63	98.55

The prediction power of a model is evaluated through its ability of fitting the calibration data. Even though the 2 LVs model is able to assign all the samples to their original class, its prediction power is not sufficient: the model has a low coefficient of determination R^2 (0.86). For this reason a 4 LVs model is calculated; the model reaches a R^2 of 0.99 with only two more latent variables. The fitting of the model is shown in Figure 5.2. The blue spots are the samples predictions made by the model versus their original values (measured). The **y** vector contains zero and one values (because of its construction), while the predictions show distributions that spread in a wider range. If these distributions were fitted to a normal one, they would cross each other in correspondence to the threshold. This means that a predicted y-value of 0.46 has a 50% chance of being in class "1" (or "0"). By setting this threshold of 50% of probability, the two area of the classes are delimitated: samples with higher values than 0.46 have a higher probability of belonging to class "1" and vice versa.

Figure 5.2. *Capacity of 4 LVs model of fitting the measured data: the blue spots are the samples predictions, the horizontal dashed line represents the threshold and the red and the green lines represent the fitting and the identity line respectively*



In Figure 5.2 there are no spots crossing the threshold meaning that all the predictions are accurate. Moreover, we obtained high goodness-of-fit since the red line (fitting) and the 1:1 green line are almost overlapped.

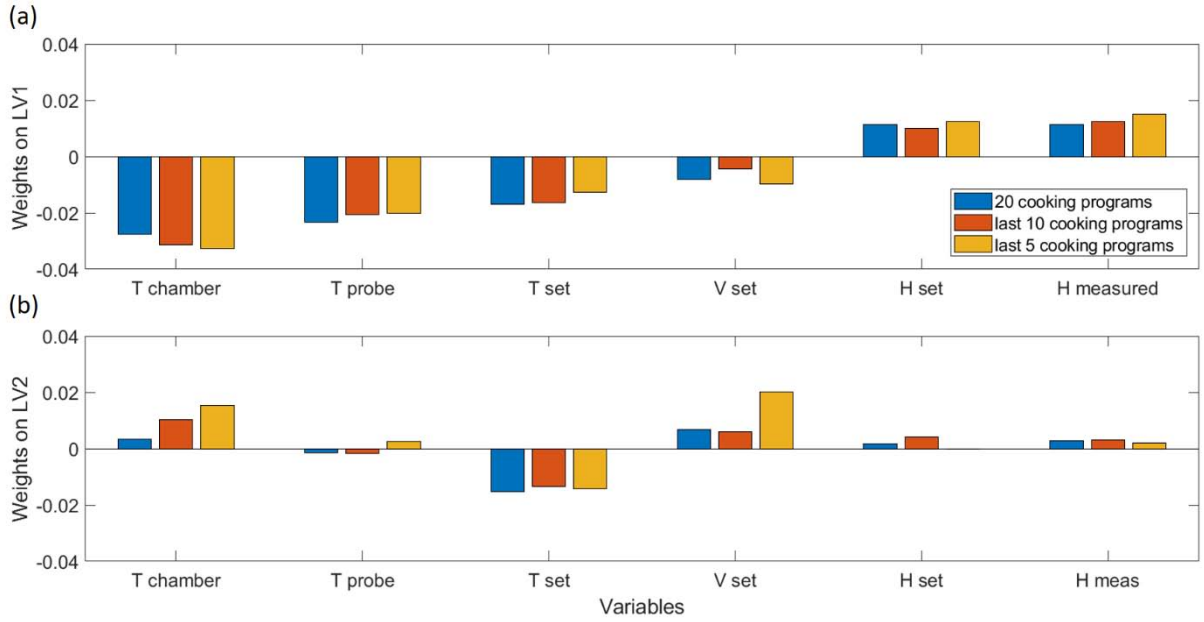


Figure 5.3. *Mean weights of the 6 variables on the 1st (a) and 2nd LV (b): the blue bar represents the mean weights of all the cooking programs and the red and the yellow bars represent the mean of the last 10 and 5 programs before the intervention respectively*

The effect of each of the 6 measured/set variables and their importance is studied through their weights on the principal LVs. The variables mean weights on the first and the second latent variable are shown in Figure 5.3: the blue bar represents the mean weights of all the cooking programs and the red and the yellow bars represent the mean of the last 10 and 5 programs before the intervention, respectively. The first latent variable is characterised by low temperatures (Figure 5.3*a*); with respect to the ones of user set and core probe, the chamber temperature is lower and it diminishes especially in the last 5 cooking programs. This result is justified by the gradual sealing loss of the gasket that does not provide a perfect insulation, thus preventing the achievement of the temperature set-point (the temperature in the chamber is lower than the one set). The humidity has a positive contribution; the cooking programs before the technical intervention have high values of humidity. As it could be expected, high humidity conditions can ruin the gasket. Finally, the ventilation gives a small negative contribution to the first LV; the cooking programs are characterised by pulsed ventilation.

The second latent variable (Figure 5.3*b*) is characterised by positive values of chamber temperature and ventilation that increase in the last 5 cooking programs. This means that 52%

of the cooking programs have high temperatures in the chamber and continuous ventilation. Negative contribution to the second LV is given by the set temperature. At first glance, the anti-correlation between the chamber temperature and the set one seems contradictory, especially if compared to the previous results. However, it points out the existence of a new behaviour that is analysed through the variables time profiles of the programs sequence that occupies the 29th row of the **X** matrix (Figure 5.4).

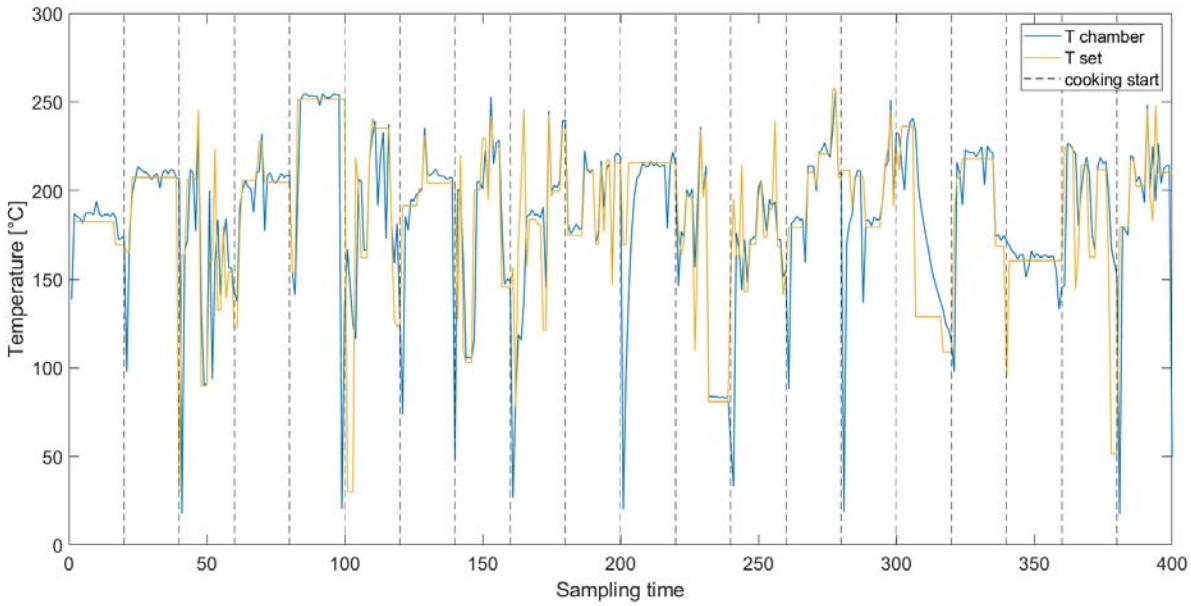


Figure 5.4. *Time profiles of chamber temperature (blue) and set temperature (yellow) of the programs sequence occupying the 29th row of **X**: the dashed lines represent the starting point of a new cooking program*

In general, the profiles are irregular with both positive and negative peaks. The peaks of the chamber temperature can be explained by:

- the oven opening that causes significant heat loss;
- the changes in the set-point.

The peaks in the set temperature profile are due to the user action that changes the set-point temperature during the cooking process. In particular, the negative deflections are very pronounced. The chamber temperature profile cannot follow properly the set-point profile when it goes down to low values because of the absence of a cooling system; the dispersion of the excess heat takes place only through transfer to the food and to the outside of the oven. However, these heat transfer processes are slow and during the time intervals needed to reach the low set-point, the chamber temperature turns out to be higher than the set one. This observation explains the anti-correlation between the two variables. Finally, it can be concluded

that these sudden changes in temperature act on the gasket deteriorating its elastic properties and thus its insulation capacity.

Summing up, a technical intervention to the gasket is preceded by:

- in 35% of cases, cooking programs with high humidity, low temperatures and pulsed ventilation;
- in 52% of cases, cooking programs with frequent changes in set temperature which determine a gap between the chamber temperature (which is high) and the set temperature (which is low) and continuous ventilation.

5.2 Core probe intervention

The following analysis is run on technical interventions on core probe with the aim of predicting malfunctioning and failure. This would avoid the interval between the damage and the substitution during which the client is not able to exploit all the oven potentiality and he/she is forced to change cooking program with negative consequences on the final product. If this may not seem a big problem for restaurants or canteens, the repeatability of the result is fundamental for large distributions where all the recipes are fixed and strictly followed.

The core probe is placed inside the food needed to be cooked and it is usually exposed to a wide range of temperatures that goes from some grades below zero to almost 300°C. The food temperature and the cooking conditions can damage the sensors or the core probe in general. The total number of this type of intervention is 21 during 2016/2017.

Table 5.2. *Variance captured by the first six latent variables for both **X** and **y** block*

	X-Block		y-Block	
LV	**% Variance Captured**	**% Cumulative Variance Captured**	**% Variance Captured**	**% Cumulative Variance Captured**
1	10.19	10.19	41.81	41.81
2	8.53	18.72	28.56	70.38
3	16.72	35.44	7.93	78.30
4	7.09	42.53	10.49	88.79
5	4.67	47.21	6.54	95.34
6	2.60	49.81	3.39	98.73

The **X** matrix and the **y** vector are built as shown in Figure 5.1 for the previous case study. **X**-block has the dimensions of 2400 columns per 42 rows, while the **y** vector is composed of 21 cooking programs sequences followed by technical interventions (class "1") and 21 sequences

without intervention (class "0"). As in the previous case, the 21 cooking sequences without technical intervention are chosen randomly from the ovens that do not present technical interventions at all during the studied period. The randomness component of the dataset makes necessary a large number of simulations: here the best case is presented. Before the PLS-DA analysis is started, both **X** and **y** are split into calibration and validation dataset with the proportion 80/20. After auto-scaling, the model is calibrated on a (34 × 2400) **X**-block and (34 × 1) **y**-block. As it can be seen in Table 5.2, the first 3 latent variables capture 78% of the **y** variance with only 35% of the **X** variance. Like the previous case study, the model is able to perfectly predict all the samples, but fitting is not close enough to the real data. Since goodness-of-fit is important to obtain a reliable predictive model, the number of latent variables is increased to 6 obtaining a model that captures 99% of the **y** variance and 50% of the **X** one.

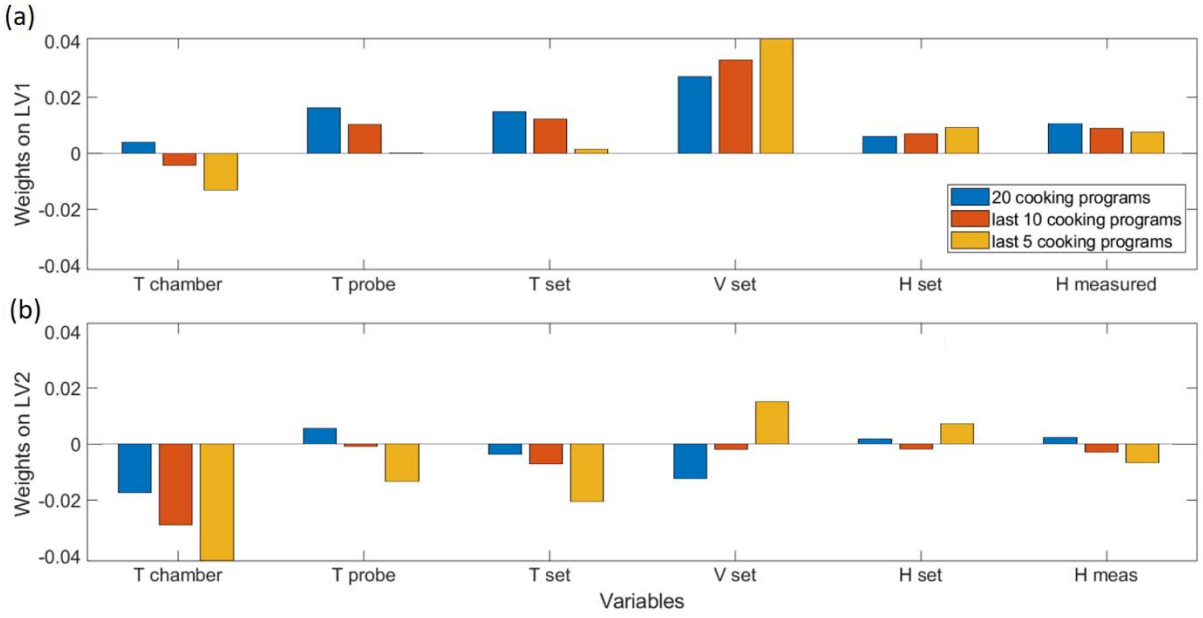


Figure 5.5. *Mean weights of the 6 variables on the 1st (a) and 2nd LV (b): the blue bar represents the mean weights of all the cooking programs and the red and the yellow bars represent the mean of the last 10 and 5 programs before the intervention, respectively*

Figure 5.5 shows the mean weights of the measured/set variables on the first two LVs: the blue bar represents the mean weights of all the cooking programs, while the red and the yellow bars represent the mean of the last 10 and 5 programs respectively. The major contribution to the first latent variable is the ventilation (Figure 5.5*a*). It assumes positive values (i.e. continuous ventilation) that increase for the last 5 programs. Ventilation is the key element in heat exchange through forced convection (like in this case); the higher the air speed, the higher the heat exchange and the temperature variation. Besides low set-point temperature and heat dispersion,

this aspect contributes to the negative trend of the chamber temperature. The high rate of heat exchange between the chamber and the food – that has a lower temperature in the last 5 cooking programs – produces a decrease in the chamber temperature even below the set-point.

42% of cooking programs represented by the first LV are characterized by chamber temperatures that are constantly lower than the set-point. Probe and set temperatures also have a decreasing trend: they assume high values during the first cooking programs and then their contributions is almost null during the last 5 programs. This could be explained by the fact that the initial high temperatures measured by the core probe could have damaged the core probe itself. Finally, the first LV is characterised by positive and almost constant values of humidity both set and measured. The predominant contribution to the second latent variable is given by the chamber temperature; the cooking process run before the technical intervention presents low chamber temperature values for the 5 immediately preceding cooking programs. Low chamber temperature means low heat exchange rate and thus high time-contact between the core probe and the cold food. Low temperature could have damaged the core probe. In fact, core probe temperature assume negative values during the last cooking programs. Notice also the inversion of the core probe temperature contribution: the first cooking programs are characterised by positive values, while the last 5 programs assume negative values. In almost 30% of cases, the core probe damage is preceded by cooking programs that register firstly high core probe temperature and then low. This inversion can cause mechanical problems to the core probe itself. A similar but opposite inversion can be observed for the ventilation. It goes from negative values to positive ones in the last 5 programs, i.e. the first cooking programs are run with pulsed ventilation while the last 5 present continuous ventilation. Finally, the humidity gives a small contribution to the second LV that in absolute value increases for the last programs.

Even though core probe temperature gives some contributions to both latent variables, it is not the predominant variable that indicates a core probe malfunctioning or damage as one could imagine. However, in both LVs it assumes a decreasing trend that could be associated to sudden variations in the measured temperature.

Summing up, the core probe substitution is preceded by:

- in 42% of cases, cooking programs with high continuous ventilation, chamber temperature constantly lower than the set-point, decreasing temperatures with high initial values for the probe core (that could have damaged it) and relatively high humidity;

- in 29% of cases, cooking programs with low chamber temperature, inversion in the core probe temperature contribution from positive to negative values and change in the type of ventilation from pulsed to continuous.

Finally, the model can be validated. The data used for validation are collected in a (8 × 2400) matrix and processed in the calibrated PLS-DA model for core probe damage prediction.

Table 5.3. *Confusion matrix of PLS-DA prediction*

		Actual class	
		No intervention	**Fault**
Predicted class	**No intervention**	37.5%	0
	Fault	12.5%	50%

The results of the prediction are shown in the confusion matrix of Table 5.3. The model is able to correctly predict the outcome of 87.5% of the cooking sequences; furthermore, the only sample that has not been properly classified corresponds to a normal cooking sequence misclassified as a set of cooking programs with final technical intervention. This means that a cooking sequence that leads to a technical intervention is always identified; this fact guarantees the reliability of the obtained model. However, the fact that technical interventions can be done on ovens that works properly means money loss. Further data should be collected to improve the performance of the current model.

In general, it can be stated that the proposed methodology is useful for the identification of the patterns that precede a technical intervention. The process variables express the state of the system allowing the prediction of failures and malfunctioning. The strength of this method consists of its ability of including in the model multiple patterns and process variable covariation to predict the appropriate moment to perform maintenance.

Conclusions

In this Thesis an industrial procedure for data analysis through advanced statistical techniques is proposed. Some specific analyses are run in order to extract valuable information from the large amount of available data. Cooking and washing processes are studied for data-driven design purposes. Through user preferences study, most and least frequent used modes are identified and their improvement through simplification or new features introduction would positively affect the user experience.

Multivariate statistical techniques are then used for process monitoring and predictive maintenance. Process monitoring is implemented through PCA analysis and a model is obtained to detect cooking processes that do not follow normal operating conditions. The aim is to take action on the process in order to reject disturbances that can affect the product quality thus keeping the process under control. In order to obtain a highly sensitive model an oven-specific one has been created. The analysis of the cooking processes with high values of Q and T^2 statistics was used to define oven normal operating conditions and delete outliers. In order to evaluate the outliers exclusion their effect on the captured variance of the model have been considered, too. The final model was able to capture 93% of the data variance with only 3 PCs. The same procedure can be applied to other ovens or to categories or sub-categories to obtain a model that can be automatically implemented as the controlling law for the processes.

Predictive maintenance is implemented through PLS-DA technique to two different technical interventions: gasket and core probe substitutions. The aim is to predict a failure before it occurs in order to avoid production interruptions and money loss. In both case studies, the amount of data used for model calibration and validation was limited because of the short time period taken into consideration. In fact, for the gasket substitution the validation of the model was not even possible. In the core probe case study the validation was done on a small database (8 samples). Moreover, because of the lack of data, the approach tried for model creation was the one of predicting the events just before their occurrence, i.e. the prediction of the last cooking program before the technical intervention was implemented. Despite all these inconveniences, models describing the data variance in a satisfactory way were obtained. Moreover, the main patterns that lead to technical interventions were characterized. In both cases, the main patterns in the variables behaviour of the 20 cooking processes before the intervention were described through two latent variables. The variables profiles were then linked to physical effects on the

damaged equipment. Finally, the prediction power of the model was tested for the core probe intervention and 87.5% of the cooking sequences were correctly predicted. Only one cooking programs sequence was misclassified as followed by a technical intervention. This means that the model was able to predict all the failures; this is an important result that guarantees the reliability of the obtained model. However, the fact that a needless intervention can be performed on the oven because of the wrong prediction means money loss. For this reason, the model has to be further improved through the use of a larger dataset for both calibration and validation. The presented procedure can be implemented for different failures detection with great benefits for both clients and company through costs reduction.

Nomenclature

a, i, j = generic subscripts

A = number of latent variables

c_α = standard normal deviate

cov(**X**) = covariance of matrix **X**

D = diagonal matrix of SVD decomposition containing the eigenvalues

E = residual matrix

||**E**|| = norm of matrix **E**

$\mathbf{e}_i$ = i^{th} row vector of residual matrix **E**

$\mathbf{e}_i'$ = transpose of residual vector $\mathbf{e}_i$

F = residual matrix of matrix **Y**

||**F**|| = norm of matrix **F**

$F_{K,I-1,\alpha}$ = F statistical distribution

h = rank of matrix **X**

I = number of samplings

I = identity matrix

J = number of collected variables

K = number of principal components

M = matrix of rank 1

P = loading matrix

P' = transpose of loading matrix **P**

$\mathbf{P}_K$ = matrix of the first K retained loading vectors

$\mathbf{P}_K'$ = transpose of loading matrix $\mathbf{P}_K$

$P(y, i)$ = probability of measuring the given y value for a class 'i' sample

$\mathbf{p}_i$ = i^{th} column vector of the loading matrix **P**

$\mathbf{p}_i'$ = transpose of loading vector $\mathbf{p}_i$

Q = loading matrix

Q' = transpose of matrix **Q**

Q_α = Q statistic limit

$\mathbf{q}_a$ = a^{th} column vector of loading matrix **Q**

$\mathbf{q}_a'$ = transpose of loading vector $\mathbf{q}_a$

s_i = i^{th} ellipsoid semi-axes

T = score matrix

$\mathbf{T}_K$ = matrix of the first K retained score vectors

$T^2_{K,I,\alpha}$ = confidence limits of T^2 Hotelling statistic

$\mathbf{t}_i$ = i^{th} column vector of score matrix **T**

$\mathbf{t}_i'$ = transpose of score vector $\mathbf{t}_i$

U = matrix of SVD decomposition/score matrix of matrix **Y**

$\mathbf{u}_a$ = a^{th} column vector of the score matrix **U**

V = matrix of SVD decomposition corresponding to loading matrix

$\mathbf{w}_a$ = a^{th} weights vector

X = process data matrix

X' = transpose of matrix **X**

$\mathbf{x}_i$ = i^{th} row vector of matrix **X**

$\mathbf{x}_i'$ = transpose of vector $\mathbf{x}_i$

x_{ij} = ij^{th} value of matrix **X**

$\hat{x}_{ij}$ = ij^{th} predicted value of matrix **X**

$\bar{x}_j$ = mean value of the j^{th} column elements of matrix **X**

Y = quality measurements matrix

y_{ij} = ij^{th} value of matrix **Y**

Greek letters

α = confidence level

λ = eigenvalue

$\boldsymbol{\lambda}$ = diagonal matrix

$\boldsymbol{\lambda}^{-1}$ = inverse of diagonal matrix $\boldsymbol{\lambda}$

σ = standard deviation

Acronyms

CPS = Cyber Physical System

IoT = Internet of Things

MPCA = Multi-way Principal Component Analysis

MPLS = Multi-way Partial Least Squares

PC= Principal component

PCA = Principal Components Analysis

PLS = Partial Least Squares

PLS-DA = Partial Least squares – Discriminant Analysis

PRESS = Predicted Residual Error Sum of Squares

RMSECV = Root Mean Square Error of Cross-Validation

SVD = Singular Value Decomposition

References

Atzori, L., A. Iera, G. Morabito (2017). Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Netw.*, **56**, 122-140.

De Carolis, A., M. Macchi, E. Negri, S. Terzi (2017). Guiding manufacturing companies towards digitalization. A methodology for supporting manufacturing companies in defining their digitalization roadmap. In: *2017 International Conference on Engineering Technology and Innovation (ICE/ITMC)*, 487-495.

Geladi, P., B. R. Kowalski (1986). Partial Least-Squares Regression: a tutorial. *Anal. Chim. Acta*, **185**, 1-17.

Igor, H., J. Bohuslava, J. Martin (2016). Proposal of communication standardization of industrial networks in Industry 4.0. In: *2016 - 20th Jubilee IEEE International Conference on Intelligent Engineering Systems (INES)*, 119-124.

Kagermann, H., W. Wahlster, J. Helbig (2013). *Recommendations for implementing the strategic initiative Industrie 4.0.* Final report of the Industrie 4.0 Working Group.

Kang, H.S., J.Y. Lee, S. Choi, H. Kim, J.H. Park, J.Y. Son, B.H. Kim, S.D. Noh (2016). Smart manufacturing: Past research, present findings, and future directions. *Int. J. Pr. Eng. Man-GT.*, **3**, 111-128.

Largoni, M., P. Facco, D. Bernini, F. Bezzo, M. Barolo (2015). Quality-by-Design approach to monitor the operation of a batch bioreactor in an industrial avian vaccine manufacturing process. *J. Biotechnol.*, **211**, 87-96.

Lee, J., B. Bagheri, H.-A. Kao (2014). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manuf. Lett.*, **3**, 18-23.

Marton, I., A. I. Sánchez, S. Carlos, S. Martorell (2013). Application of data driven methods for condition monitoring maintenance. *Chem. Eng. Trans.*, **33**, 301-306.

Mell, P., T. Grance (2011). The NIST Definition of Cloud Computing. *NIST Special Publication*, **800(145)**, 1-7.

Mobley, R. K. (2002). *An introduction to Predictive Maintenance* (2nd ed.). Butterworth-Heinemann, Massachusetts (USA).

Nomikos, P., J.F. MacGregor (1994). Monitoring batch processes using multiway principal component analysis. *AIChE J.*, **40**, 1361-1375.

Oztemel, E. (2010). Intelligent Manufacturing Systems. In: *Artificial Intelligence Techniques for Networked Manufacturing Enterprises Management* (L. Benyoucef and B. Grabot, Ed.), Springer Series in Advanced Manufacturing. Springer, London (UK).

Oztemel, E., S. Gursev (2018). Literature review of Industry 4.0 and related technologies. *J. Intell. Manuf.*, 1-56.

Wise, B. M., N. B. Gallagher (1996). The process chemometrics approach to process monitoring and fault detection. *J. Process Contr.*, **6**, 329-348.

Wise, B. M., N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, R. S. Koch (2006). *Chemometrics Tutorial for PLS_Toolbox and Solo*. Eigenvector Research, Wenatchee, WA (USA).

Yan, J., Y. Meng, L. Lu, and L. Li (2017). Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. *IEEE Access*, **5**, 23484 - 23491.

Websites

http://www.eigenvector.com/