POLITECNICO DI TORINO
Department of Control and Computer Engineering
Master's degree in Computer Engineering

Master Degree Thesis

Industrial data analytics from IoT sensors: an explorative study on coffee machines

Supervisors
prof. Elena Baralis
prof. Daniele Apiletti
prof. Frederic Le Mouel

Candidate
Alessandro Chiotti

December 2019

Summary

The ability of collecting data from any kind of device is becoming more and more important every day. Many companies have realised that data acquisition and analysis provides a way to find new and more efficient solutions to their problems and to open up new perspectives. In order to optimize costs and maximize results, an initial exploratory analysis is necessary: in this phase, the interaction and exchange between domain experts and data analysts is fundamental to guide the analysis towards the company's objectives and to correctly interpret the results obtained from the data.

For this reason, the thesis, conducted in collaboration with Lavazza, explores the data obtained from the telemetry sensors installed by the company inside their bar coffee machines. The study is focused on analysing the feasibility of predicting coffee quality, predictive maintenance and identifying the variables mostly related to the two previous objectives.

The analysis was developed in two phases: a first phase for studying the data provided by tests performed in the laboratory, in which both the data collected by the telemetry sensor and those collected from the cup were available. Afterwards, a second phase of analysis of data from coffee machines on the market has been performed, thus having including only data from telemetry sensors available.

Furthermore, the lLaboratory tests were structured in such a way that the overall external factors were reduced as much as possible: only double brewings (more stable than single brews) and always with the same type of blend were performed. In both phases, a preliminary analysis was necessary which showed that the three main parameters that characterise the coffee brewing process are: brewing time, flow-rate and quantity of coffee brewed in the cup. About these, Lavazza's domain experts provided initial quality thresholds, which were then questioned evaluated and updated as a result of the during the analyses of the thesis.

The results of the analyses were are promising.: data Data-driven results confirmed already known and hypothesized behaviours, and also revealed new possibilities. In fact, in some cases, the variation of the three external variables (dose, grinding and pressure) influenced flow, time and quantity of coffee in an evident way. In others, instead, the compensation effect covered up the consequences of the variations. Thanks to the identification of the correlation between the variables

measured from the cup and those extracted from the telemetry, it was possible the transition from the study was extended of data from the the laboratory to those extracted from the marketreal-world coffee machines in bars. The latterse revealed the limits of the data currently available in relation to the planned objectives.

However, the analyses performed so far are only a solid starting point. One possibility of progress, in fact, A promising approach is represented by the use of the time series: through a process of feature engineering, have been obtained four new variables have been designed, related to the single coffee brewing. From the first analyses conducted, the addition of this information, on one hand, is potentially useful to deepen the intuitions already identified, on the other hand it marginally increases the amount of data sent by the sensor sin a non-significant way.

Acknowledgements

Il mio primo ringraziamento va alla professoressa Elena Baralis e alla Lavazza, per avermi dato l'opportunità di lavorare a questo progetto di tesi.

Un grande ringraziamento anche al professor Daniele Apiletti, che mi ha aiutato e supportato durante tutto il percorso di tesi e da cui ho imparato molto. Ringrazio anche il mio compagno di viaggio Marco Lentini, che ha collaborato con me al progetto: un amico, sempre disponibile e propositivo.

Ringrazio ovviamente tutti i miei amici che sono stati al mio fianco sin dal primo giorno di Università o che ho conosciuto "in corsa": mi avete aiutato a "staccare" quando ne avevo bisogno.

Un ringraziamento speciale va anche alla mia ragazza Elena: nonostante la lontananza mi ha sopportato e supportato per diversi mesi, non facendomi mancare nulla.

Ringrazio anche i miei nonni, sempre curiosi di conoscere l'esito degli esami e l'andamento della tesi.

Infine un ringraziamento speciale va a alla mia famiglia: a mia sorella Giorgia che mi ha tenuto compagnia nei lunghi pomeriggi passati in biblioteca. E ai miei genitori che hanno cercato di facilitare il più possibile il mio percorso e mi hanno sempre spronato a dare il meglio: siete e sarete sempre un esempio per me.

Contents

List of Tables 7

List of Figures 9

1 Introduction 12

1.1 Content overview 13

2 State of the art 15

3 Context of Application 17

3.1 Professional coffee machines 17

3.2 Simplified structure 17

3.3 Calibration process 19

4 Dataset description 21

4.1 First experimental session dataset 22

4.1.1 Initial dataset 23

4.1.2 Modified dataset 26

4.2 Second experimental session dataset 26

4.3 Pc card dataset 28

4.4 Proposed experimental session dataset 29

4.5 Telemetry dataset 30

5 Data analysis techniques 33

5.1 Implementation details: Jupyter notebook and Python 33

5.2 Data Cleaning 34

5.2.1 Data cleaning applied to the individual test 35

5.2.2 Cleaning and preliminary alignment of telemetry data 35

5.2.3 Cleaning of the brews made by the machines on the market 36

5.3 Correlation and linear regression 38

5.4 Feature engineering 39

5.5 Statistical tests: ADF and KPSS 42

6 Experimental results 45
6.1 Experimental sessions in the laboratory 45
6.1.1 First experimental session 45
6.1.2 Second experimental session 46
6.1.3 Proposed experimental session 47
6.2 Coffee quality and thresholds analysis 50
6.3 Variables correlation 53
6.3.1 PC card - cup variables correlation 54
6.3.2 Telemetry - cup variables correlation 60
6.3.3 Telemetry - cup variables correlation on *Rancilio CL11* machine 65
6.3.4 Telemetry - cup variables correlation on *Faema E71* machine 71
6.4 Degradation 73
6.5 Time series analysis 81
6.6 Analysis of data from the market 92
6.6.1 Clients and machines modelling 93
6.6.2 Anomalous clients analysis 103

7 Conclusions 113
7.1 Analysis limits 114
7.2 Possible future improvements 114

Bibliography 117

List of Tables

6.1 Coffee quality in cup thresholds 51
6.2 Flow statistics from first experimental session data (test 1, group 2) 53
6.3 Proposed in cup thresholds from first experimental session data . . 53
6.4 PC card flow and in cup flow correlation and regression indexes . . 54
6.5 PC card quantity and in cup quantity statistics 58
6.6 PC card quantity and in cup quantity correlation and regression indexes 59
6.7 PC card flow and in cup flow correlation and regression indexes, second experimental session 59
6.8 PC card quantity and in cup quantity correlation and regression indexes, second experimental session 60
6.9 PC card quantity and in cup quantity statistics, second experimental session 61
6.10 telemetry flow and in cup flow correlation and regression indexes, second experimental session 62
6.11 telemetry quantity and in cup quantity correlation and regression indexes 63
6.12 Telemetry quantity and in cup quantity statistics 63
6.13 telemetry flows and in cup flows correlation and regression indexes . 67
6.14 telemetry flows and in cup flows correlation and regression indexes (*Faema E71*) 72
6.15 ADF test on group 2 time curve 76
6.16 KPSS test on group 2 time curve 76
6.17 ADF test on 2 time curve 77
6.18 KPSS test on group 2 time curve 77
6.19 Anomalous clients: Bar *RAL* data cleaned 91
6.20 Anomalous clients: Bar *RAL* data cleaned 92
6.21 In cup and telemetry thresholds 94
6.22 Analysed period 95
6.23 Threshold analysis *Nuvola* in the period 27/06/2019–11/07/2019. . 96
6.24 *Nuvola* telemetry variables statistics 98
6.25 *Nuvola* thresholds in the period 27/06/2019–11/07/2019. 98

6.26 Threshold analysis *Nuvola* in the period 27/06/2019–11/07/2019. . 100
6.27 *Nuvola* telemetry variables statistics, period 17/05/2019-30/05/2019. 100
6.28 *Nuvola* thresholds in the period 17/05/2019-30/05/2019. 101
6.29 Threshold analysis *Top Shop* in the period 08/07/2019 – 27/07/2019. 102
6.30 *Top Shop* tel. variables statistics, period 08/07/2019 – 27/07/2019. 103
6.31 Anomalous clients: Bar *RAL* data cleaned 105
6.32 Bar *RAL* telemetry variables statistics 107
6.33 Ristorante *il pepe* telemetry variables statistics 108
6.34 Anomalous clients: Ristorante *VITESSE SAS* data cleaned 108
6.35 Anomalous clients: Ristorante *GRUPPO SPES SCS ETIKO* data cleaned 108
6.36 Ristorante *VITESSE SAS* telemetry variables statistics 110
6.37 Ristorante *GRUPPO SPES SCS ETIKO* telemetry variables statistics110

List of Figures

3.1 Simplified coffee machine structure and components 18
4.1 Data flow diagram 21
5.1 Quantity distribution example 37
5.2 Regression line example 39
5.3 Time series of telemetry pulses (*Cimbali M100*) 40
6.1 Univariate analysis heatmap 51
6.2 PC card flow and in cup flow comparison, first experimental session (focus on 200 brews) 55
6.3 PC card flow and in cup flow regression line, first experimental session 55
6.4 PC card water quantity and in cup quantity comparison, first experimental session (focus on 200 brews) 56
6.5 In cup quantity univariate analysis 57
6.6 PC card water quantity and in cup quantity comparison after data cleaning 57
6.7 PC card quantity and in cup quantity regression line 58
6.8 PC card flow and in cup flow regression line, second experimental session 60
6.9 Telemetry flow and in cup flow comparison, second experimental session 61
6.10 Telemetry flow and in cup flow regression line, second experimental session 62
6.11 Telemetry water quantity and in cup quantity comparison, second experimental session 64
6.12 Telemetry water quantity and PC card water quantity trend, second experimental session 64
6.13 Telemetry flow and in cup flow trend (*Ranclio CL11*), proposed experimental session 65
6.14 Telemetry flow and in cup flow regression line (*Ranclio CL11*), proposed experimental session 66
6.15 Telemetry flow and in cup flow correlation and regression indexes (*Ranclio CL11*), proposed experimental session 66

6.16 Telemetry water quantity and in cup quantity trend (*Ranclio CL11*), proposed experimental session . . . 68
6.17 Telemetry quantity and in cup quantity regression line (*Ranclio CL11*) 69
6.18 Telemetry quantity and in cup quantity correlation and regression indexes (*Ranclio CL11*) . . . 69
6.19 Telemetry quantity and PC card quantity trend (*Ranclio CL11*), proposed experimental session . . . 70
6.20 Telemetry flow and PC card flow trend (*Ranclio CL11*), proposed experimental session . . . 70
6.21 Telemetry quantity and in cup quantity trend (*Faema E71*), proposed experimental session . . . 72
6.22 Telemetry quantity and in cup quantity regression line (*Faema E71*) 73
6.23 Telemetry quantity and in quantity flow correlation and regression indexes (*Faema E71*) . . . 73
6.24 Degradation group 2 Telemetry quantity trend (*Wega Urban*), proposed experimental session . . . 74
6.25 Degradation group 2 Telemetry time trend (*Wega Urban*), proposed experimental session . . . 75
6.26 Degradation group 2 tel. time stats comparation (*Wega Urban*), proposed experimental session . . . 75
6.27 Degradation Telemetry flow trend (*Wega Urban*), proposed experimental session . . . 77
6.28 Degradation group 2 tel. flow stats comparation (*Wega Urban*), proposed experimental session . . . 77
6.29 Degradation group 2 Telemetry quantity trend (*Wega Urban*), proposed experimental session . . . 78
6.30 Degradation group 2 Telemetry time trend (*Wega Urban*), proposed experimental session . . . 79
6.31 Degradation group 2 tel. time stats comparation (*Wega Urban*) . . 79
6.32 ADF test on group 2 time curve . . . 79
6.33 KPSS test on group 2 time curve . . . 79
6.34 Degradation Telemetry flow trend (*Wega Urban*), proposed experimental session . . . 80
6.35 Degradation group 2 tel. flow stats comparation (*Wega Urban*), proposed experimental session . . . 80
6.36 ADF test on time curve . . . 80
6.37 KPSS test on time curve . . . 80
6.38 Timeseries of telemetry pulses (*Cimbali M100*) . . . 82
6.39 Average flows scatter plot (group 1 *Cimbali M100*) . . . 83
6.40 Scatter plot with color formatting based on time (group 1, *Cimbali M100*) . . . 84

6.41 Scatter plot with color formatting based on in cup flow (group 1, *Cimbali M100*) . . . 85
6.42 Scatter plot with color formatting based on pressure (group 1, *Cimbali M100*) . . . 86
6.43 Scatter plot with color formatting based on grinding (group 1, *Cimbali M100*) . . . 87
6.44 Scatter plot with color formatting based on dose (group 1, *Cimbali M100*) . . . 88
6.45 Optimal and inside the thresholds brews view (group 1, *Cimbali M100*) 89
6.46 Optimal and inside the thresholds brews comparison (group 1, *Cimbali M100*) . . . 90
6.47 Optimal and inside the thresholds brews comparison (group 2, *Cimbali M100*) . . . 90
6.48 Cafeteria *Nuvola* selling data 27/06/2019–11/07/2019 . . . 95
6.49 Cafeteria *Nuvola* data cleaned . . . 96
6.50 Box plots *Nuvola* 27/06/2019 – 11/07/2019 . . . 97
6.51 Cafeteria *Nuvola* data cleaned, period 17/05/2019-30/05/2019 . . . 99
6.52 Cafeteria *Top Shop* data cleaned, period 08/07/2019 – 27/07/2019 . 102
6.53 Anomalous clients: Bar *RAL* data cleaned . . . 104
6.54 Anomalous clients: Bar *RAL* selling data . . . 104
6.55 Brewing time of Bar *RAL (left)* vs Ristorante *il pepe (right)* . . . 106
6.56 Brewing average flow of Bar *RAL (left)* vs Ristorante *il pepe (right)* 106
6.57 Brewing water quantity of Bar *RAL (left)* vs Ristorante *il pepe (right)*107
6.58 Brewing time of Ris. *VITESSE SAS (left)* vs Ris. *GRUPPO SPES SCS ETIKO (right)* . . . 109
6.59 Brewing average flow of Ristorante *VITESSE SAS (left)* vs Ristorante *GRUPPO SPES SCS ETIKO (right)* . . . 109
6.60 Brewing water quantity of Ristorante *VITESSE SAS (left)* vs Ristorante *GRUPPO SPES SCS ETIKO (right)* . . . 110

Chapter 1

Introduction

The new trend towards automation and data exchange, brought by the industry 4.0 era, leads companies operating in all sectors to adapt their systems to make them compatible with these new technologies. This is the case of Lavazza, where the objective is the insertion of an IoT sensor inside all their bar coffee machines. In fact, collecting, being able to send and analyse huge amount of data from any kind of device is becoming crucial for a company today: data acquisition and analysis provide a way to find new and more efficient solutions to problems and to open up new perspectives. If the information obtained turns out to be useful, considerable advantages can be obtained in both productivity and economic terms. However, the process towards an automated data collection and analysis system is not simple and straightforward. In order to optimize costs and maximize results, an initial exploratory analysis is necessary: in this phase, the interaction and exchange between domain experts and data analysts is fundamental to guide the analysis towards the company's objectives and to correctly interpret the results obtained from the data.

For this reason, the thesis, conducted in collaboration with Lavazza, explores the data obtained from the telemetry sensors installed by the company inside their bar coffee machines. The two primary objectives of the company are to increase product quality and to reduce maintenance and repair costs. Therefore, the study is focused on:

- analysing the feasibility of predicting coffee quality.
- predictive maintenance.
- identifying the variables mostly related to the two previous objectives.

It was developed in two phases: a first phase for studying the data provided by tests performed in the laboratory, in which both the data collected by the telemetry sensor and those collected from the cup were available. Afterwards, a second phase of analysis of data from coffee machines on the market has been performed, thus including only data from telemetry sensors. The company would like to able to identify clients brewing coffees below the standard and the causes involved. On the other side, with the purpose of reducing repair operations by the company, an analysis on the absence of maintenance has been conducted: the study is focused on the degradation of the variables characterizing the coffee due to the lack of washes.

1.1 Content overview

Chapter 1 provided an overview of the analysis conducted in this work of thesis, explaining reasons and objectives of the study. Chapter 2 offers a review of the literature related to the previous studies carried out about coffee and professional coffee machines. The aim is to underline the innovative approach of this work. Chapter 3 describes the coffee machines used for the tests and available on the market. The first part is dedicated to the their structure and their functioning: it is important to better understand the choices made and the results of the various tests. The second part describes the calibration process, a fundamental procedure to obtain a coffee above the standard. Chapter 4 presents a guide to understand the structure of all the datasets available for the analysis. Chapter 5 describes the various tool and techniques adopted during this work of thesis. After a focus on the procedures of data cleaning and regression technique applied, the process of feature engineering on the time series and the ADF and KPSS statistical tests are displayed. Chapter 6 is the core of the work of thesis. At the beginning there is the description of the experimental sessions in the laboratory, where it is highlighted the importance of structuring future tests according to the proposed sequence. Then, after the definition of the quality parameters and their thresholds, the two main analysis on laboratory data are described: the variables correlation and the degradation due to the lack of washes. Afterwards the study on the time series is presented: it demonstrates how powerful could be the addition of four new variables taken from the time series of the flowmeter pulses. Finally, there is the description of the analysis made on data from clients on the market: after a focus on trusted clients brews characteristic, sell in - sell out data are cross-referenced with telemetry variable analyses, to discover anomalous clients. Chapter 7 concludes the study, with the summary of the analyses conducted and the discussion about possible future works.

Chapter 2

State of the art

This chapter describes main studies made on coffee machines and coffee quality. Different type of sensors and techniques were used, but all the approaches were different from the one used in this work of thesis. Considering coffee quality, a lot of studies confirm that there are a lot of variables that contribute to the realization of a good coffee. For example, the study [9], tries to find a correlation between espresso coffee quality and water used: also the properties of water, can influence the the realization of the coffee in cup. They found that coffee foam depends on the water used to perform the brew: through an analysis of the foam volume, composition and persistency the difference is evident. The authors underline that in Italy the optimal time of brew is 25-30 *s* and the optimal quantity is around 25 - 30 *ml*. Even if the purpose is different because Lavazza is looking for coffee quality in the variables they can detect from telemetry, the approach used to make a coffee brew is similar: the coffee brew quantity and time assume values analogous to the ones considered and analysed in this work of thesis. Another study focuses on water in the coffee brew, but in this case the attention is in the influence of coffee/water ratio on the coffee quality [1]: the authors decided to analyse the psycho-chemical parameters of coffees, brewed with different amount of dose. They studied the results for three levels of dose (6.5 *g*, 7.5 *g* and 8.5 *g*) for three different qualities of coffee beans. The other parameters that were fixed for the brews were brewing time (18-24 *s*, a little low for the standards founded in this work of thesis) and the quantity (40 ± 2 *ml*, definitely high for Italian Espresso standards). The instrumental results show that the differences were not appreciable: the optimization of other parameters such as pressure, temperature and grinding, seemed to minimize the influence of coffee/water ratio on coffee quality. This conclusion supports the idea that it is difficult to find general rules because all the variables can compensate each other. The concept of analysing the composition of the coffee is adopted by other investigations, as in [11]: in this case, an electronic nose system was used to study the changes in the aromatic profile of espresso coffee. The study emphasizes the importance of the first 8 seconds of the coffee brew, because in this range the major amount of organic acids, solids and caffeine are extracted: it coincides in fact with the first part of the coffee brew, where the average flow is higher (as described in section 5.4 and 6.5). The analysis concludes that both time and grinding level significantly affect the aromatic profile of espresso coffee. More recent studies, instead, focus on introducing IoT sensors inside coffee machines: in paper [2], the aim is to automatize coffee buying process and to monitor machine status through an application. The part related to this work of thesis is the second one, where telemetry data are used to control machine condition: the predictive maintenance in this case is only related to deficiencies of the machine (low temperature, water level, lack of coffee powder) because the process of coffee brewing is automatized. There is no analysis on the degradation of coffee brew parameters, for example (see section 6.4). The analysis made in paper [7] are more related to predictive maintenance: the purpose is to install a telemetry sensor on an old coffee machine and, through machine learning techniques, to predict the last coffee before coffee bean depletion by collecting vibration data. While the application is questionable, the general purpose is definitely valuable for coffee companies: predictive maintenance techniques could make them save considerable repair and replacement costs.

In literature there is no a research like the one described in this work of thesis: the innovation is in the approach and the purpose of laboratory experimental sessions. In fact, the aim is to model the wrong behaviours of baristas, identified by Lavazza experts, to recognize them from telemetry data. In fact, those behaviours, lead to bad quality coffee brews and probably to degradation of machine components.

Chapter 3

Context of Application

3.1 Professional coffee machines

During the study several models of bar coffee machines on the market were taken into consideration:

1. La Cimbali M100
2. Wega MyConcept
3. Wega Urban
4. Faema E71
5. Rancilio CL5, CL7, CL11

All the machines listed above have almost the same structure and they are subjected to a calibration process based on the same principles. In this paragraph are described the general structure of a coffee machine and the procedure calibration.

3.2 Simplified structure

Almost all coffee machines have the following structure: the cold water is pumped into the circuit by a pump and with a precise pressure. Then, it passes through a boiler, so that it can be heated, and through a flowmeter, which regulates the flow: each flowmeter pulse 0.5 ml of water are emitted into the group. Through the group it reaches the panel of powdered coffee contained in the filter holder. The water that does not pass through the coffee panel remains inside the circuit [5]. The process is shown in the diagram in figure 3.1

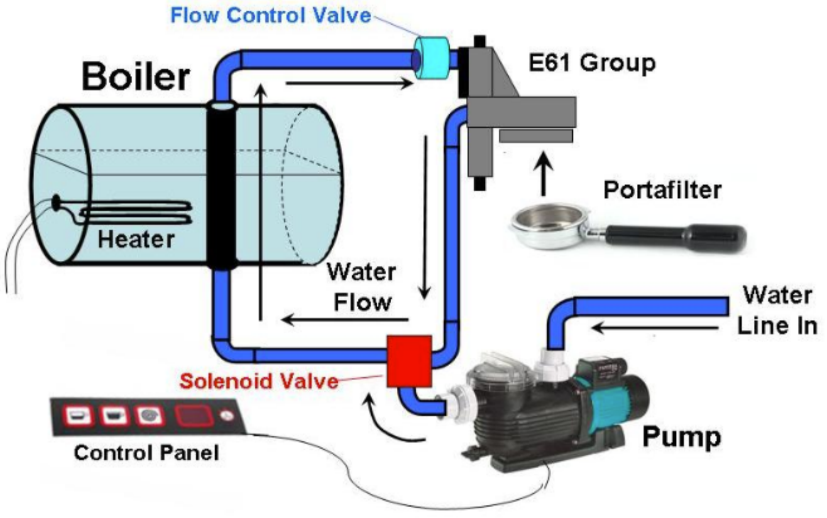

Figure 3.1. Simplified coffee machine structure and components

The telemetry device, which is able to send data to the company's portal, is located at the level of the flowmeter. The various models of sensors have slightly different functions: they prove different data or perform or not processing on site, depending on the machine on which they are installed. In chapter 4 the data supplied by the sensor will be described.

The flowmeter pulses are regulated during the calibration process: the machine is designed to brew a constant quantity of water for each coffee brewing. But the quantity of water passing through the coffee powder panel and turning into real coffee, depends on many other factors: the pressure and therefore the flow, but also the degree of grinding of the coffee beans, the quantity of dose inserted in the filter holder and the quality of the coffee itself.

To the coffee machine is therefore associated a coffee grinder. There are two types of coffee grinders:

- Dosing coffee grinders: the coffee is ground in advance and the bartender takes care of lowering the dose, using a manual lever, inside the filter holder. When the lever is operated, about 7 g of the dose are lowered, corresponding to the single dose (if operated twice, 14 g corresponds to the double dose). This is the type of coffee grinder that allows dishonest baristas to insert a smaller dose into the filter holder.

- On-Demand coffee grinders: the coffee is ground on the spot, so as to maintain the freshness of the beans. In addition, only the correct dose is brewed by the grinder-doser, thus preventing misuse.

3.3 Calibration process

One of the most important processes for the quality of coffee is certainly the calibration of the machine. It is a trial and error process, in which the final aim is to obtain, through a double brewing (chosen because it is more stable), 46 grams of coffee (with a double brew)in 25 seconds.

To achieve this, the grinding of the coffee is regulated by means of a special grid on the coffee grinder. The first step of the process is to press the self-learning button on the machine and then the button corresponding to the double dose: when the desired quantity of coffee (the cups are weighed with a scale) is reached, it is necessary to press the double dose button again to stop dispensing. At this point, the time of the coffee brewing is compared to the standard (25 seconds): if it was more than 25 seconds, it means that the coffee powder is too fine and vice versa. The process is then restarted, adjusting the grill on the coffee grinder: all the process is repeated until a brewing with a weight of around 46 grams is reached in about 25 seconds [3].

These are the indications given to the baristas. Domain experts, instead, are able to recognize an optimal coffee observing the flow of the coffee in the cup, the colour of the same coffee and other parameters.

After the calibration process, the obtain good quality coffees the following parameters must be maintained constant:

- the grinding must not be changed.
- the dose has to weight 14 gr for double coffee brewing and 7 gr for single coffee.

Moreover the machine has to be subjected to washes: after every coffee brew the purge program must be executed and at the end of every working day is necessary to perform the group cleaning.

Chapter 4

Dataset description

This diagram in figure 4.1 shows the flow of data that goes to compose the different datasets. In grey are the external variables that are taken into account by the study:

- Grinding, i.e. the degree of grinding of the coffee
- Dose, i.e. the quantity of coffee powder used for each brew
- Pressure, i.e. the pressure inside the coffee machine
- Cdv, the total pulses emitted by the flowmeter

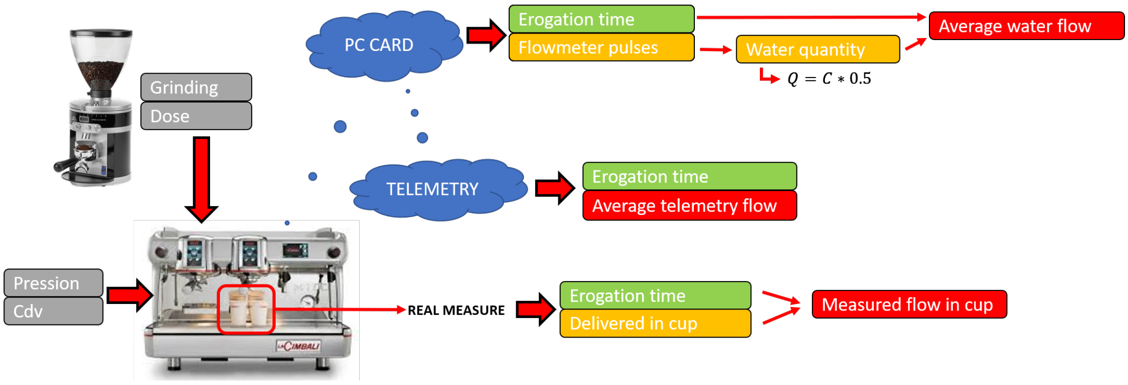

Figure 4.1. Data flow diagram

All these variables can be modified by the user: the first two depend on the setting of the grinder-doser, while the other two can be modified on the machine itself. The diagram also identifies the different data sources:

1. *PC Card data.* Data received by connecting the PC card directly to the machine: they represent the data obtained from the sensor placed at flowmeter level. These data are raw, not processed, but more difficult to extract and send. They were just detected for the first experimental session: only flowmeter pulses and coffee brew time are available, sampled every 200 milliseconds at each dispensing. In the future, pressure and temperature could be added with the same sampling.

2. *Telemetry data.* They are the data sent directly to the portal of the company. These are obtained by processing sensor data. In addition to some diagnostic data of the machine and data related to the washing of the same, the brew time and the average telemetry flow are returned: unfortunately the data regarding the average flow is an elaborate data (for the majority of the coffee machines), obtained from the sensor data described above.

3. *Real Measurements data.* They are the data measured by hand after each single brew: after the tare, the cup is weighed to obtain the quantity of coffee brewed in cup, while the flow is obtained from the latter, divided by the brew time.

At the moment only telemetry data are available for machines in production. However, having the three data sources available for the laboratory sessions is useful for:

- derive data correlations and apply known principles on a data type derived in a different way.
- understand which variables are needed for the various studies, whether they can be obtained through regression techniques and which functionalities should be implemented in the future.

4.1 First experimental session dataset

The objective of data collection in the first experimental session is to observe the variation in the characteristics of the coffee (time of brewing, flow and quantity brewed in the cup) due to the variations of three external parameters (grinding, dose and pressure) compared to the optimal ones. In addition, only one of the two groups is washed (group 2), in order to observe the degradation of the dirty one. In particular, 27 tests are carried out per group, so as to cover all possible combinations of the variations of the external parameters, each one composed of 20 brews. The data and information available were proposed by the experts of the Lavazza: there are a lot of derived indices. Starting from this proposal, an analysis has been carried out to understand what variables are required and the dataset structure was redefined.

4.1.1 Initial dataset

This section describes the structure of the initial Excel file provided by Lavazza. The dataset contains 1080 rows and 40 columns. Each row corresponds to a double coffee brew while the columns describe its characteristics. The columns have been grouped according to the type of information contained. We proceed with the description of the meaning of the various columns, and the possible values they assume.

- *Informazioni generali*: provide the technical information of the coffee brew.
  - Test: identifier of the test to which the brew belongs. Each test corresponds to a certain combination of grinding, dose and pressure (e.g. "T_01", "T_02",...);
  - ID: unique identification number of the brew. Entire incremental from 0;
  - Giorno: date when brew was effected;
  - Orario: time of brew;
- *Dati della macchina*:provide information on the model of the machine used and the programming of the brew.
  - Macchina: indicates the model of machine on which the dispensing was performed. The only value assumed is "CIMBALI M100 Dosatron".
  - Gruppo usato: indicates the machine group from which the dispensing was carried out. It assumes only two values: "GROUP 1" and "GROUP 2";
  - Dose(singola o doppia): indicates the key pressed to make the dispensing. The only value present is "KEY 2", corresponding to double dose;
  - Dose programmata (colpi di ventolino) CdV=0,5ml: indicates the number of flowmeter pulses set by the machine. Each pulse corresponds to 0.5 *ml* of water brewed. This is the parameter related to the calibration of the machine. The only value assumed is 177.
- *Informazioni sulla miscela*: They provide information about the mixture used and how it was pressed.
  - Pressatura: indicates how the ground material was pressed. The only value assumed is "Tarata".
  - Miscela: indicates the mixture used to make the brews. The only value assumed is "TOP CLASS". A single mixture is then used.
- *Rilevazioni esterne*: are data measured externally by means of a scale or a reading from the machine's display.

  - Peso dose macinato(gr): weight of the ground material used to make the brew, measured with a scale;
  - Temperatura caldaia(Bar) da display: boiler temperature in bar, shown on the machine display;
  - Pressione(bar) a vista: boiler pressure, expressed in bar;
  - Tempo erogazione (sec): brew time measured by means of a chronometer;
  - Peso erogato in tazza(gr): weight of the coffee dispensed, referred to the double dose and measured with a scale;
  - Erogato in tazza singola(ml): millilitres of coffee dispensed for a single cup. Calculated from "Peso erogato in tazza (g)", divided by two and multiplied by 1.02, the specific gravity of the coffee.

- *Indici calcolati dalle rilevazioni manuali*: relate the various external measurements made, with the aim of providing more information.
  - Ratio(peso dose/peso erogato): indicates the relationship between the weight of the dose of ground and the weight of the quantity dispensed in the cup.
  - Flusso medio macchina (dose programmata ml / tempo erogazione sec): indicates the average flow calculated from the quantity of water supplied by the machine.
  - Velocità del flusso in tazza (peso erogato/tempo erogazione): average flow in the cup measured as g/ml in relation to the quantity of dispensed for the double dose.
  - Velocità flusso (ml/Sec): indicates the flow in the cup calculated in ml/s and referred to the single cup.

- *Confronto con le soglie*: comparison of the values of flow, brew time and coffee brewed in the cup with respect to the defined thresholds.
  - SOGLIA FLUSSO ml/sec (su dose singola): flow in the cup calculated as ""Erogato in tazza singola (ml)" divided by "Tempo erogazione";
  - SOGLIA FLUSSO (per calcolo): indicates the label assumed by the brew with respect to the defined flow thresholds. It can assume only three values: MINORE, OK e MAGGIORE.
  - TEMPO DI EROGAZIONE 20/27 sec: brew time calculated from the PC card of the machine, calculated in seconds. Corresponds to the attribute "Tempo di erogazione";
  - TEMPO DI EROGAZIONE (per calcolo): indicates the label assumed by the brew with respect to the defined brew time thresholds. It can assume only three values: MINORE, OK e MAGGIORE.

  - Erogato in tazza 18ml /30 ml: millilitres of brew in cup measured on the single dose. Corresponds to “Erogato in tazza singolo (*ml*)”;
  - EROGATO IN TAZZA (per calcolo): indicates the label assumed by the brew with respect to the defined brew time thresholds. As for the previous labels, it assumes only the following values MINORE, OK e MAGGIORE.

- *Dati Scheda PC*: measurements taken from the PC card, i.e. directly from the machine control unit.
  - Tempo erogazione: measurement of dispensing time in seconds;
  - Impulsi flussimetro: number of flowmeter pulses carried out during dispensing. Each pulse corresponds to 0.5 *ml* of water. It depends on the calibration of the machine and has a very small range of variation;
  - Quantità erogata(ml): quantity of water brewed referred to the single dose: it is calculated as “Impulsi flussimetro” multiplied by 0.5 *ml*, divided by 2, so as to have a figure relating to the single dose in the cup;
  - Flusso acqua (ml): flow referred to the water supplied by the machine. It is calculated as "Quantità erogata (*ml*)" divided by "Tempo di erogazione";
  - Calcolo variazione tempo estrazione (display Vs. macchina): difference between the time measured externally and that of the PC card. Useful for aligning the data of the two sources;
  - Calcolo della variazione su quantità erogata (ml/dose in tazza singola): ratio between the total quantity of water dispensed for the PC card and the quantity in a single cup measured externally.

- *Diagnostica macchina* information related to the washes.
  - Lavaggi: indicates whether or not the daily washing on the group has been carried out. It can only assume the values "Si a fine giornata" and "No";
  - Purge (prog. 4sec): indicates whether a short wash has been carried out after dispensing. It can only assume the values "Si" e "No";
  - Lavaggi Flusso: indicates a type of washing with a blind cover. Does not report any value.

- *Messaggi*: reports any error messages of the machine or device. They never assume valid values.

- *Indici calcolati dalla scheda PC*: indices calculated from PC card data trying to increase the available information.
  - Flusso acqua dai dati macchina: flow calculated from the PC card data as "Quantità erogata(*ml*)" divided by "Tempo erogazione". Repeat the field "Flusso acqua (*ml*)";

  - Soglia flusso dati macchina: application of the flow thresholds in the cup to the data of the PC card. It assumes values "MINORE", "OK" and "MAGGIORE".

4.1.2 Modified dataset

From the initial dataset, attributes considered unimportant were cleaned up: or they always have the same value or they were calculated from other data present in the file. In this way, both the subsequent analyses, able to process the data better, and the procedures for collecting data are optimized.

In particular, several columns have been deleted:

- "Tempo erogazione (sec)": since the data coming from the Pc card is consistent with that detected externally, it was decided to eliminate the recording of time by means of a chronometer;
- "Indici calcolati dalle rilevazioni manuali": eliminated because they are calculated from other data acquired;
- "Calcolo variazione tempo estrazione (display Vs. macchina)" and "Calcolo della variazione su quantità erogata (ml/dose in tazza singola)", because they don't carry any useful information;
- "Lavaggi flusso" and "Messaggi" because they don't indicate any value;
- "Indici calcolati dalla scheda PC", because they are repetitions of other attributes.

The information on the machine used and on the characteristics of the brew have been kept, even though they always assume the same value as they are useful in the case of experiments with different machines, calibrations and mixtures.

4.2 Second experimental session dataset

The second experimental session has the same approach as the first, with the difference that it aims to model minor variations in external factors. It again presents 27 tests that model the various combinations of external factors: each one is made up of 20 distributions, with the exception of the first test that presents 60 observations, as it is carried out with the optimal values of external factors and thus allows a more in-depth study of the optimal initial situation. In addition, the tests are carried out on both groups which, however, have different cleaning conditions. In fact, only one of the two groups is subjected to daily washing and purging. In this way it is possible to study a difference due to the different state of maintenance of the machine.

A further objective of this session is to study the correlation between telemetry data, PC card data and cup values. In fact, thanks to the *Cimbali M100* machine on which the experiment was carried out, it is possible to collect data from both sources that take data from sensors.

The structure of the dataset is very similar to that analysed for the previous experimental session, with some slight differences:

- Temperature and pressure from the display are not detected by the display as they are constant or changed voluntarily due to the test specifications;
- The telemetry detections are present, so the following columns have been designed:
  - tempo erogazione telemetria: brew time detected by the telemetry sensor, expressed in seconds;
  - calcolo variazione tempo estrazione: difference between the brew times detected by telemetry and PC card. It is very useful for alignment between PC card data and telemetry;
  - Impulsi flussimetro telemetria: number of flowmeter pulses calculated from the amount of water brewed per single dose. It is calculated as "Quantità erogata telemetria (ml)" multiplied by 4, because referred to double dose and each pulse of the flowmeter is equal to 0.5 ml;
  - Quantità erogata telemetria(ml): the amount of water in millilitres referred to the single dose, calculated from the telemetry flow for the double dose. It is calculated as "Flusso acqua telemetria(ml/s)" divided by 2 by "Tempo erogazione telemetria", to be referred to the single dose;
  - Calcolo della variazione sulla quantità erogata: quotient between "Quantità erogata telemetria (ml)" and "Quantità erogata (ml)" calculated by the PC card, expressed as a percentage;
  - Flusso acqua telemetria(ml/s): average flow rate measured by telemetry in relation to the double dose, expressed in millilitres per second;
  - Flusso acqua telemetria per singola dose(ml/s): average flow detected by telemetry in relation to the single dose. Calculated as ”Flusso acqua telemetria (ml/s)" divided by 2;
  - Calcolo della variazione su flusso acqua: "Flusso acqua telemetria per singola dose(ml/s)" divided by "Flusso acqua (ml/s)", expressed as a percentage.

The dataset therefore has 1160 rows and 30 columns. It should be noted, however, that there are 437 values not detected by telemetry. As confirmed by the analysis of data from the market in the 5.2.2 section, the lack of some data from the telemetry device is due to the cleaning carried out by the telemetry of the manufacturer Cimbali on the data of the brews.

4.3 Pc card dataset

For the first experimental session are also available the datasets containing the time series of the flowmeter pulses: each file is associated to a test (2 groups, 27 tests per group), so there are 54 excel file. Every file contains 20 coffee brews, but for each brew there are multiples row: a row represents a sample of a single brew, detected every 200 *ms*. As the coffee brew time is not constant, the number of rows associated to a brew is variable. The columns represent the characteristics detected by the PC card, the structure is the following:

- Date Time: date and time of detection of the brew by the PC card;
- SystemTime: system time of the PC card;
- time: time measured from the beginning of the detection of the brews;
- tempo: time in tenths of a second that measures the duration of each single brew: it is reset at the end of each brew;
- ngruppo: identification number of the group that is brewing;
- ntasto: identification number of the key pressed to make the brew ('2' = double brew);
- tempoerogazione: time in tenths of a second that measures the duration of each single brew: it is reset at the end of each brew;
- tempoinfusione: variable not detected;
- impulsiflussimetro: number of flowmeter pulses executed by the machine since the beginning of the brew (they zero after each brew);
- portata: variable not detected;
- temperatura: variable not detected;
- pressione: variable not detected;
- percentualeaperturavalvola: variable not detected;

The brews in which the tempoerogazione column increases while the impulsiflussimetro column remains at 0 are the measurements of the washes that occurred between two brews.

To analyse the result of the feature engineering process, it was created a script able to extract from the 27 excel files of each group a new file: each row represents a coffee brew. The new dataset contains the following columns:

- ID: unique identification number of the brew. Entire incremental from 0;
- Test: identifier of the test to which the brew belongs. Each test corresponds to a certain combination of grinding, dose and pressure (e.g. "T_01", "T_02",...);
- Tempo erogazione: ime in seconds that measures the duration of each single brew;
- Impulsi flussimetro: number of flowmeter pulses carried out during dispensing. Each pulse corresponds to 0.5 *ml* of water. It depends on the calibration of the machine and has a very small range of variation;
- Quantità acqua: quantity of water dispensed by the machine at each brew. It is calculated as "Impulsi flussimetro" multiplied by 0.5;
- Slope1: average flow of the first part of the brew, before the trend point (see 5.4 for further details);
- Slope2: average flow of the second part of the brew, after the trend point (see 5.4 for further details);
- Tempo al trend point: time in seconds in which the trend point is reached (see 5.4 for further details);
- Colpi di ventolino al trend point: number of the flowmeter pulses at the trend point (see 5.4 for further details);
- Quantità acqua al trend point: Quantity of water at the trend point. It is calculated as "Colpi di ventolino al trend point" multiplied by 0.5;

4.4 Proposed experimental session dataset

After the first two experimental sessions, some objectives to be modelled in the laboratory were defined. In particular, it was created a test session to be performed on different machines, in order to detect any differences that should be considered in the analysis of customers on the market. Five types of tests are defined, aimed at modelling different behaviours, whose composition and order of execution have been decided on the basis of the observations made on the first experiments.

The specifications and reasons for each test are described in the 6.1.3 section. Below are the cardinalities for each test session. The columns are the same as those of the second experimental session. However, please note that in some models of machine it is not possible to detect the data of the PC card and the telemetry at the same time. In this case the preference falls on the telemetry data, as it is closer to the real case of the machines on the market. The number of brews expected for each test, considering that the test is carried out on machines with two groups, is:

- intrinsic difference between the groups: 60 optimal brews for each group, for a total of 120 brews.
- Relationship between flowmeter pulses and cup brew: 10 tests of 10 brews per group, for a total of 200 observations. The total number of observations could be reduced if the number of pulses of the flowmeter moves too far away from the optimal number;
- Modelling of abnormal washes: 60 optimal brews without washes, plus 6 tests of 10 readings each, carried out for each group, for a total of 240 observations;
- Degradation due to lack of washes: 60 optimal brews to have the initial comparison, 500 brews to accumulate a degradation and as many to see the differences in the presence of washes. In total, the sum of the two groups' brews is 2120. In addition, the first 500 brews may not be sufficient to accumulate performance degradation, and brews without washing may be increased until performance degradation is observed. Moreover, only a small part of the totality of the brews is detected in the cup, so as to speed up the experiments;
- Variation of external parameters: 19 tests consisting of 20 brews to model combinations of variations of external parameters (dose and grinding), plus 60 initial optimal brews, for a total of 440 brews per group, or 880 total.

4.5 Telemetry dataset

In the second part of the study, data from the market, i.e. collected by telemetry, was analysed. They are downloaded from the Lavazza telemetry portal, which allows a maximum period of 41 days to be selected on one of the supervised machines.

Through the portal, you can select which information to download in an excel file. In our analysis, only the flow and time measurements for each brew have been taken into consideration. The downloaded file, however, is not in a convenient format to continue the analysis. In fact, it has a sheet for each of the selected characteristics, divided by brew group. In practice, for a machine with 2 groups of which you are interested in analysing brew time and flow, you will get a file with 4 sheets, two with the brew times of the two groups and two with the flows in the two cases.

In particular, each sheet has 3 columns of interest: the start date of the survey, the end date of the survey and the brew data, which varies depending on whether it is flow or time.

In theory, for each brew both the brew time and the average flow should be recorded. In some cases, however, only one of the two is detected: usually in cases where the brew is very short or very long. A more detailed description is given in the 5.2.2 section.

To proceed with the analysis it is very useful to have time and flow associated with the same time of detection. A program was then written that received the file downloaded from the portal, returns a new file with all the brews made on the machine, reported in the following format:

- Data: date and time of detection of the brew by the telemetry;
- Gruppo: group of the machine on which the brew was detected;
- Tempo erogazione telemetria: brew time measured by telemetry in tenths of a second;
- Flusso telemetria: average flow calculated from telemetry in ml/s.

In this way, all the information are immediately available, and the problem of manual alignment, which requires a lot of time and attention, is solved. In addition, those brews with only the detected time or flow are automatically discarded.

Chapter 5

Data analysis techniques

This chapter describes the techniques and programs used for data analysis, with a focus on some problems identified due to the lack and misalignment of some data. In particular, the chapter is divided into the following paragraphs:

- Jupyter notebook and Python
- Data cleaning process
- Correlation and inear regression
- Feature engineering
- Statistical tests: ADF and KPSS

5.1 Implementation details: Jupyter notebook and Python

For the analysis it was decided to use the web application Jupyter Notebook (formerly IPython Notebooks). It is a web-based computational and interactive environment to create Jupyter notebook documents (JSON documents). Each document contains an ordered list of input/output cells. Every cell can contain code, text (using Markdown), plots and mathematical formulas. A Jupyter Notebook document can be converted and downloaded as a different open standard format (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python). In this way, code and output results can be shared and easily red on different operative systems and platforms.

Jupyter Notebook provides a browser-based REPL built upon a number of popular open-source libraries:

- IPython

- ØMQ
- Tornado (web server)
- jQuery
- Bootstrap (front-end framework)
- MathJax

Jupyter Notebook can connect to different kernels, to allow programming in many languages. To carry out the analyses, the programming language chosen was Python: besides being perfectly integrated in Jupyter Notebook, it allows to easily manipulate and effectively analyse datasets in Excel format: thanks to the numerous libraries available, in fact, it is one of the most used languages concerning artificial intelligence, machine learning and data mining [13]. In particular, the libraries mainly used for analysis are the following:

- Pandas: it is an open source library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [10]. It is used for the conversion of excel files in data frame, a two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).
- Numpy: it is the fundamental package for scientific computing with Python. It is useful to deal with large, multi-dimensional arrays and matrices, and it contains a large set of high-level mathematical functions to operate on them.
- Matplotlib: it is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. [8] It is used to generate plots, histograms, bar charts, scatterplots, box plots and many other plots.
- Sklearn: it features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy [14].

5.2 Data Cleaning

Data cleaning is the process by which records of a dataset, that present values outside the domain or not acceptable, are recognized and corrected or deleted. In this way it is possible to eliminate the "noise", guaranteeing the correctness of the database.

This process turned out to be necessary for all the datasets before performing the analyses. In fact, some human errors have been found in the transcription of data, as well as some problems of the telemetry tool and other errors whose origin is unknown.

The three cleaning processes carried out are the following:

- Data cleaning applied to the individual test.
- Cleaning and preliminary alignment of telemetry data.
- Cleaning of the brews made by the machines on the market.

5.2.1 Data cleaning applied to the individual test

This is the cleaning of data from experiments on laboratory coffee machines. Each experiment, in fact, is associated with a dataset divided into tests: each test has a variation with respect to the previous one (for example, different external variables, presence or absence of washing, different calibration of the machine). Since the machine does not have an immediate response to the variations, it was necessary to carry out the cleaning on the single test, instead of on the complete dataset.

Since the first experimental session, it has been noted that the first coffee of each test had an unacceptable (too low) amount of coffee in the cup. For this reason it was decided to eliminate the data: for homogeneity, in addition to the elimination of the record corresponding to the coffee with minimum cup brew in the test, it was decided to eliminate also the one corresponding to the coffee with maximum cup brew.

5.2.2 Cleaning and preliminary alignment of telemetry data

The non-optimal accuracy of the telemetry sensor made it necessary to clean up the data obtained from the Lavazza online portal.

Concerning telemetry data, some alignment problems were found. In particular, three issues were highlighted:

1. Some brews are not detected by the telemetry sensor. In fact the number of total coffee brews, obtained from the analysis of sales, is higher than the data actually available. The sales analysis data are considered reliable, as they correspond to the number of times a key is pressed on the machine. Depending on the machine model, the number of missing data is greater or lower, or even zero.

2. Some of the missing coffee brews are due to what has been called " merging ": the flows and times of two successive coffees are added and displayed as a

single brew, thus reducing the number of total brews available. But, being the button pressed twice, both the brews are counted and appear in the sales analysis.

3. There are cases where, for a single brew, time is detected while the average flow is not and vice versa. The consequence is the loss of alignment of the data: for example, a missing value in the column of time causes all the subsequent times to scale up of one cell, thus losing the match with the cells in the column of average flows.

   In September 2019 an update of the telemetry software was introduced: the problems seem to have been partially solved. However, it was necessary to create a script that would automatically align the data. The average flows and brew times are in fact in two different sheets of the excel file downloaded from the portal. A matching is made based on the time of brew (present for both variables) and when a brew time is found with no corresponding flow (or vice versa), the record is deleted. In this way, the correspondence between brew times and average telemetry flows is restored.

5.2.3 Cleaning of the brews made by the machines on the market

Unfortunately, the telemetry data do not have the label relative to the key pressed by the customer to make a single brew. To overcome this lack and proceed with the analysis, it was necessary to implement a data cleaning process: in this way it was possible to compare data retrieved from machines on the market with the results obtained from the experimental session in laboratory, where all the brews were double coffees brews.

Telemetry data, especially when extracted over long periods of time, are very noisy. This is due to the variety of possible brews: normal, short, long (for each there is a single or double version) and continuous (or leva). Usually, among these seven types of brew, the most used are two: the normal single brew and the normal double brew.

The cleaning was carried out thanks to an analysis of the telemetry data: the average flow and the brew time. The first approach was to individually analyse their distributions and define thresholds for each. The thresholds are taken visually, so as to include the highest peak, which represents the most frequently brewed coffees: in this way the data, after cleaning, should correspond almost only to normal single or normal double brewing. There are therefore two possibilities:

1. Remove all brews with at least one of the two variables out of range.

2. Only remove those brews that have both variables out of range.

In the first case, too few brews are deleted, while in the second case the selection is too stringent. Therefore, it was decided to opt for a cleaning system that took into account the combined effect of the two variables: multiplying them, in fact, you get the amount of water brewed. Since the objective of the machine is to supply a constant quantity of water, set during calibration, the different types of brews (single, double, short, long, free) will have a different quantity of water. The data was then cleaned by eliminating the brews that have a quantity of telemetry water brewed outside the thresholds, selected visually on the graph of the distribution.

This method has been particularly effective because often a group is used to make the same type of brew: domain experts have confirmed that usually each group is associated with a single or double filter holder. The data confirmed this statement, as some groups had peaks in the amount of telemetry water around 14-17 ml (groups used for single coffee), others around 22-27 ml (groups used for double coffee). In this way it is possible to discriminate which groups are used to make the majority of single coffees and which groups are used to make the majority of double coffees. By comparing the sales data (in which it is possible to visualise the quantity of coffee produced for each type for a given period) with the number of brews remaining, there is a correspondence between the brews labelled as single and double after cleaning and the single and double brews actually brewed.

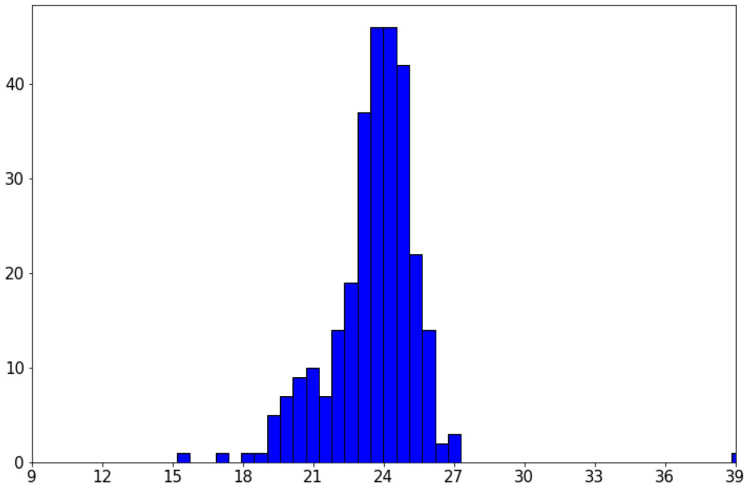

Figure 5.1. Quantity distribution example

Visual cleaning can be automated as the variation in the amount of telemetry water brewed around the peak is generally constant. In fact, once the peak has been identified, generally it is sufficient to include all brews 2 ml above and 2 ml below it. In the future this type of data cleaning won't be necessary because the information concerning the typology of coffee brewing will be available

5.3 Correlation and linear regression

A fundamental part of the study was focused on the search for the relationship between two data collections. Two indices frequently used to measure the statistical relationship between two variables are:

- *Covariance*:
$$Cov(X,Y) = \sum_x \sum_y (x - \mu_x) \cdot (y - \mu_y)$$
with X and Y variables and μ_x and μ_y mean of the two variables
- *Correlation*:
$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma y}$$
with σ_x and σ_y standard deviation of variables X and Y

The *covariance* measures if the two variables have a concordant trend: a positive value means the two variables have both an ascending or descending trend, while a negative one means the two characteristics shows an opposite behaviour. Finally, if the value is close to zero, the two variables have no relation. The *covariance* has two main problems: if the scale is small, it is always close to zero; and it has a different unit of measurement, making the comparison more difficult.

The *correlation* represents the solution to those problems because it is a pure and an dimensionless number, in the range [-1, 1]. For this reasons this is the index adopted in all the analysis performed.

If the *correlation* index has an high value, it means that from the value of the independent variable can be approximately obtained the value of the dependent variable. If the *correlation* is high, then the next step to find a transfer function between the two variables is the graphical visualization: the scatter plot. In fact, in this plot, each data is represented as a point: the x-coordinate is the value of one characteristic and the y-coordinate the value of the other one. If the two variables have a linear relationship, the majority of the points tend to dispose in a straight line (example in figure 5.2) . Finally, if the *correlation* is promising (an absolute value close to one) and most of the points of the scatter plot are located in a straight line, the *linear regression* can be calculated.

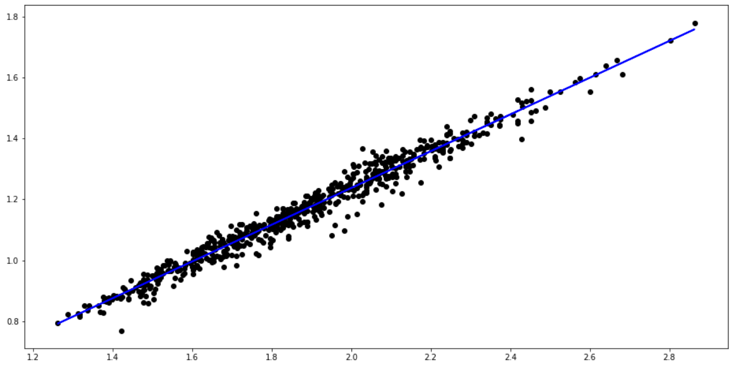

Figure 5.2. Regression line example

The *linear regression* method is able to compute the linear regression function between the two variables in exam, giving an equation in the form

$$y = \alpha + \beta \cdot x$$

where α is the intercept on the y axis and β is the slope of the straight line.

The coefficient used to evaluate how well observed outcomes are replicated by the model is called *coefficient of determination* (r^2 or *r-squared*). It is in range [0,1] and represents the proportion of the variance in the dependent variable that is predictable from the independent variable: an higher value implies a better *linear regression* [6].

The type of *linear regression* adopted for the analysis is the *OLS*(Ordinary Least Square) method: it finds the parameters α and β minimizing the squared errors. The error term is defined as the distance between a point of the scatter plot and the linear regression straight line. Therefore, minimizing the square of all the errors, there is no compensation between positive and negative errors.

5.4 Feature engineering

Feature engineering is the process of transforming given data to a less complex and easier to interpret form: exploiting domain knowledge, it can be used to improve machine learning algorithms. In this case, the company provided, in the first experimental session, a different and more detailed way of collecting data: by connecting the pc card to the coffee machine they were able to detect the time series of the flowmeter pulses (sampled every 300 *ms*). In fact, those are the data the sensor on the machine is able to detect, but after a data processing phase, it sends only the total time and an average flow. Therefore, the aim was to estimate the potential improvements given by the addition of those new information to the dataset. In order to develop a solution that would not significantly increase the amount of data sent (for cost-related reasons) and that can still be interpretable, a feature engineering solution was chosen.

In the figure 5.3 is shown a coffee brewing curve. The flowmeter pulses are converted to the quantity of water brewed with the formula 6.2 and displayed on y-axis. The graph seems divided in two nearly constant segments: after an initial increasing, there is a visible point of slope change. Being the two variables of the graph the time and the quantity of water brewed, the slopes of the two segments represent the average flow in the two phases. The domain experts state the shape is due to the two phases of the brew: in the first one the coffee panel inside the filter holder is wetted and an higher flow rate is necessary; then, the second phase represent the penetration of the water through the panel, where a lower flow rate is sufficient.

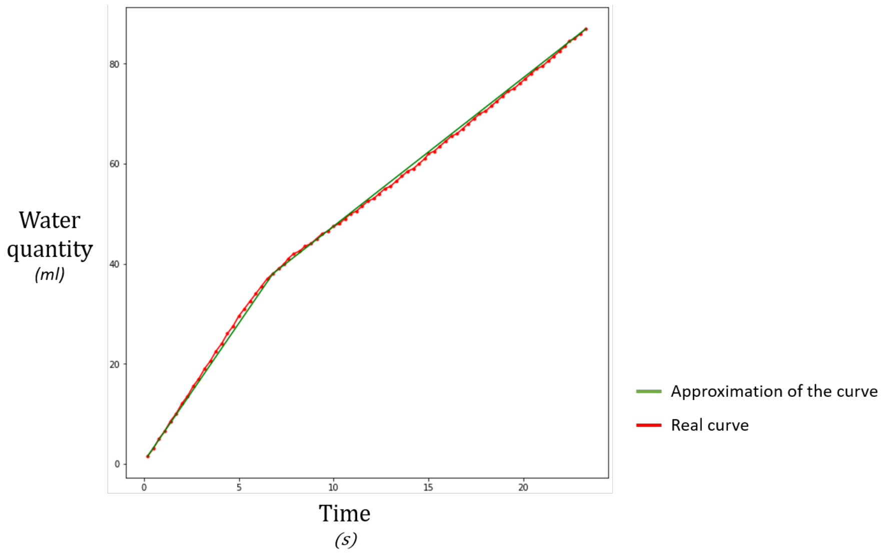

Figure 5.3. Time series of telemetry pulses (*Cimbali M100*)

The approximation of the curve in the graph is obtained finding the trend point: the point where the average flow become lower. Here is the algorithm implemented to find the index of the trend point:

```
def featureCalculator(t, f):
    x1, x2, y1, y2 = 0, 0, 0, 0
    N = 10
    diff = []
```

```
m1, m2 = np.zeros(N), np.zeros(N)

for i in range(0, len(f)):
    x1 = x2
    y1 = y2
    x2 = t[i]
    y2 = f[i]
    if(i != 0):
        m = (y2-y1)/(x2-x1)
        if(i < N):
            m1[i] = m
        elif(i < 2*N):
            m2[i%N] = m
        else:
            mean_m1 = np.mean(m1)
            mean_m2 = np.mean(m2)
            diff.append(mean_m1 - mean_m2)
            m1[i%N] = m2[i%N]
            m2[i%N] = m

max_diff = np.amax(diff)
index = diff.index(max_diff)+N
return index
```

Those are the steps:

1. the slopes between two consecutive points are computed. The first ten samples are inserted inside vector $m1$ and the second ten samples inside $m2$;
2. the average of each of the two is calculated and inserted in $mean_m1$ and $mean_m2$;
3. the difference between $mean_m1$ and $mean_m2$ is stored in vector $diff$;
4. vectors $m1$ and $m2$ are shifted
5. repeat from 2. until $m2$ contains in the last cell the last computable slope;
6. the maximum of $diff$ is computed and the index of the associated point is retrieved

The algorithm compute the index of the point where the difference between the mean of the previous 10 samples and the following 10 samples is maximum. Then, knowing this point, four new features are computed:

1. Q_{tr} = water quantity trend point (ml)
2. T_{tr} = time at trend point (s)

3. Φ_{slope1} = average flow before trend point (ml/s)

4. Φ_{slope2} = average flow after trend point(ml/s)

5.5 Statistical tests: ADF and KPSS

A stationary series is a series in which mean, variance and covariance are not time-dependent functions. To prove that a series is stationary, two tests frequently used in econometrics are the Augmented Dickey Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. They are called unit root tests because the presence of a unit root in a time series indicates that its properties varies with time. The demonstration is the following: having a time series defined as

$$y_t = \alpha \cdot y_{t-1} + \epsilon_t$$

where y_t represents is value assumed at time t and ϵ_t represents the error term at time t. The term y_{t-1} instead is computed as

$$y_{t-1} = \alpha \cdot y_{t-2} + \epsilon_{t-1}$$

Performing all the substitutions, the general formula to compute the n^{th} term is

$$y_t = \alpha_n \cdot y_{t-n} + \sum_{i=0}^{n} \epsilon_{t-i} \cdot \alpha_i$$

In the above equation, if the value of α is approximately 1 (unit), all the samples of the time series can be computed as y_{t-n} plus the sum of all errors from $t - n$ to t: it means that the variance increases with time, so the series is non-stationary. The ADF test and the KPSS tests operates checking if value of $\alpha = 1$.

ADF test

The ADF test is a unit root test that determine whether a time series is stationary or difference stationary (it can be made strict stationary by differencing). Those are the two hypothesis:

- **Null hypothesis**: the series has a unit root (non-stationary series)
- **Alternate hypothesis**: the series has no unit root (series is stationary)

Through the computation of the *p-value* and the *test statistic* indices, the null hypothesis can be accepted or rejected (and so the alternate hypothesis is accepted). In fact, if the *p-value* < 0.01 there is a very strong evidence against the null hypothesis ($0.01 <$ *p-value* < 0.05, strong evidence), so the *alternate hypothesis* is accepted (series is stationary). Otherwise if *p-value* > 0.05, the *test statistic* value is compared to the *critical values*(1%, 5% or 10%): if the *test statistic* is less than the *critical value* chosen (usually 5% critical value is selected), the null hypothesis is rejected and the series is stationary. Otherwise, the null hypothesis is accepted and the series can be linear or difference stationary [4].

KPSS test

The KPSS test is a unit root test that determine whether a time series is stationary or trend stationary (it can be made strict stationary by removing the trend). Those are the two hypothesis:

- **Null hypothesis**: the series has no unit root (series is stationary)
- **Alternate hypothesis**: the series has a unit root (non-stationary series)

Through the computation of the *p-value* and the *test statistic* indices, the null hypothesis can be accepted or rejected (and vice versa the alternate hypothesis). In this case, the null and alternate hypothesis are opposite of the ADF test: if the *p-value* > 0.1 there is a little (or no evidence) against the null hypothesis ($0.05 <$ *p-value* < 0.1, weak evidence), so the *null hypothesis* is accepted (series is stationary). Otherwise if *p-value* < 0.05, the *test statistic* value is compared to the *critical values*(1%, 5% or 10%): if the *test statistic* is greater than the *critical value* chosen (usually 5% critical value is selected), the null hypothesis is rejected and the series is trend stationary. Otherwise, the null hypothesis is accepted and the series is stationary [12].

Performing both the tests, there are four possible cases.

- **case 1**: if both the tests concludes that the time series is not stationary, it is not stationary
- **case 2**: if both the tests concludes that the time series is stationary, it is stationary
- **case 3**: if KPSS concludes that the time series is stationary, but ADF concludes that it is not stationary, the time series is trend stationary.
- **case 4**: if ADF concludes that the time series is stationary, but KPSS concludes that it is not stationary, the time series is difference stationary.

Chapter 6

Experimental results

6.1 Experimental sessions in the laboratory

This section will describe the structure and objectives of the experimental sessions in the laboratory. All coffees dispensed have the following characteristics:

- they are all regular double coffees. These type of coffees are in fact the most frequently brewed on the market and they are generally more stable.
- they have the *TOP CLASS* mixture, a good quality and very common one.

The three experimental sessions performed in the laboratory are:

1. **First experimental session**
2. **Second experimental session**
3. **Proposed experimental session**

6.1.1 First experimental session

It is related to the first series of tests available for the analysis. Performed on the machine *Cimbali M100*, the objective was to study the influence and the effects of the external variables (*dose, pressure, grinding*) on the variables measured in the cup (*time, flow, quantity*). It was decided to make significant variations with respect to the standard (optimal) values for all three parameters, to highlight the consequences. The external parameters have been changed as follows:

- the *pressure* cannot be changed from the outside, but only by removing some components of the machine (making more difficult for the customer to modify it). From the optimal value of 8.9 *bar*, the pressure was reduced by 3 *bar* (6 *bar*) and subsequently increased by 2 *bar* (11 *bar*).

- the *dose* has a reference value of 14 *g* (double dose). The two non-optimal values selected for the tests are 12 *g* (-2 *g*) and 16 *g* (+2 *g*). Lower dose tests are representative of a common behaviour in the market, which the company would like to be able to detect.
- the *grinding* grade does not have a precise reference value, but derives from the process of calibration of the machine (explained in section 3.3). Taking the optimal value as reference, the other two are obtained by rotating the ring nut of two markers clockwise (+2, finer grinding) and two markers counter clockwise (-2, coarser grinding).

The combinations of the three values of the three external parameters has therefore generated a matrix of 27 tests, 20 brews each. The whole has been executed for each group: group 1 has not been cleaned, while group 2 has been subjected to washes. Another objective was therefore to understand the influence of the lack of washes on the brews. Before analysing the data, it was necessary to select the relevant variables and to define the structure of the dataset, to be used as a template for subsequent sessions. In fact, all the variables derived from other already existing variables have been deleted within the data structure. Among these, only the flow has been maintained, useful for comparison with the quality thresholds (defined in section 6.2). The data of the first experimental session include the in-cup and PC card readings (for the structure of the datasets refer to section 4.1).

6.1.2 Second experimental session

For the second experimental session, a similar approach to the first one was chosen. The objectives were both to confirm and to deepen the results of the first session, through a different variation of the external parameters:

- the *pressure* from the optimal value of 8.9 *bar*, was decreased by 3 *bar* (6 *bar*) and increased by $2bar$ ($11bar$). There was no change from the first experimental session.
- the *dose*, compared to the reference value of 14 *g*, was altered to 13 *g* (-1 *g*) and 15 *g* (+2 *g*).
- the *grinding* compared to the reference one was increased by one mark (+1, finer grind) and decreased of a mark (-1, coarser grinding).

In particular, the aim was to verify whether it was possible to identify a variation in the variables in the cup again and to determine if it could be related to the value assumed by the external parameters. In this session, both the values obtained from the PC card and the telemetry variables were available. However, the experiment was not very effective because the values measured in the cup of the brews made with optimal external parameters were outside the thresholds. The different tests were then analysed with the aim of finding the relationship between the 3 groups of variables:

- in cup variables
- PC card variables
- telemetry variables

The results of the analysis are presented in sections 6.3.1 and 6.3.2.

6.1.3 Proposed experimental session

As a result of the first two experimental sessions, it was decided to modify the setting of the tests. First of all, the relevant objectives were defined together with Lavazza:

1. Intrinsic difference between groups;
2. Relationship between flowmeter pulses and amount of coffee in the cup;
3. Modelling of anomalous washes;
4. Detection of degradation due to lack of washes;
5. Influence of external parameters on coffee quality.

Therefore, a sequence of tests has been developed for each one of these objectives. The order of execution is important because it was designed to eliminate possible effects due to recalibration of the machine.

Intrinsic difference between groups

After selecting the machine model on which perform the tests, the first brews to be made are those aimed at identifying the possible **intrinsic difference between groups**: in fact, there may be differences between the mechanical components of the machine. Moreover, the calibration process is executed only on one group and the others are set according to the same parameters. In the first experimental sessions in the laboratory, differences in the flows and brew times of the two groups were highlighted: the objective of the test is therefore to understand if and in which measure these were due to mechanical differences. The sequence of operations to be carried out is the following:

1. bring the machine into ideal conditions by executing the washes;
2. calibrate the machine as indicated in the section 3.3;

3. perform 60 optimal brews for each group, making the purge after each brew.

For each brew it would be appropriate to obtain both the data of the telemetry device and detect the variables in the cup. This procedure should be repeated before each test, in order to have an ideal starting point.

Relationship between flowmeter pulses and amount of coffee in the cup;

From the comparison with the experts, it seemed that the calibration process did not introduce a significant variability, if done with all the precautions. Comparing, however, the second experimental session with the first one, a difference of 40 pulses of the flowmeter was highlighted. Being *Cdv* in the second experimental session higher, the amount of coffee dispensed in the cup was increased, exceeding the quality thresholds provided by the experts. For this reason it seems necessary to carry out a test to study the **relationship between flowmeter pulses and amount of coffee in the cup**. The test should be structured as follows:

- 10 optimal brews (after calibration);
- 4 tests of 10 brews each, keeping the optimal grinding and increasing each test by 10 *Cdv* (starting from the value of *Cdv* after calibration);
- 4 tests of 10 brews each, maintaining the optimal grinding and decreasing 10 *Cdv* per test (starting from the value of *Cdv* after calibration).

Tests have to be made on each group. The variation of 10 *Cvd* between tests has been chosen knowing that $1\ Cdv = 0.5\ ml$ of water: in reality this value could be different depending on the machine selected for the test.

Modelling of anomalous washes

Two main types of automatic washing available for a coffee machine are: long washes (group washes) and short washes (purges). The first ones are carried out daily, while the second ones are made at the end of each brew. They can be executed in three ways:

- without the group inserted, allowing water to flow through the pipes without any obstruction;
- with blind cup, limiting the water brewed inside the machine and carrying out a more effective;
- with blind cup and tablet, ensuring a deeper cleaning of the group.

Despite the presence of a special button for washing, many baristas select the double dose or the free dose for cleaning the machine. Some of them made also use of unofficial tablets, which may not provide the same performance and even damage the machine. The following sequence of tests is recommended for modelling **anomalous washes**:

- 60 optimal brews (after the calibration process) without executing washes;
- 3 test of 10 brews each with the wrong button, one with the blind filter and tablet, one with tablet and one without the group;
- 3 test of 10 brews each with the wright button, one with the blind filter and tablet, one with tablet and one without the group.

The aim is to find out if and how anomalous washes are detected by telemetry, so as to recognize them (to correct the wrong behaviour of the customer) and understand what effect they can have on the brews.

Detection of degradation due to lack of washes

One of the main objectives set by Lavazza is to evaluate the possibility of identifying customers who do not properly maintain the machine: the purges after each brew and the washing of the group at the end of the day in fact ensure good coffee quality and reduce the need for technical interventions. The objective of the test is therefore to simulate a customer who does not execute washes: it tries to highlight a **degradation due to lack of washes** through the study of the telemetry variables. The sequence to be followed should be:

- make 60 optimal brews (after cleaning and calibration of the machine) for each group;
- make 500 brews per group without purges;
- in a group make another 60 brews with purge (to see if these short washes lead to an improvement);
- in the other to carry out a washing of the group and then another 60 brews without washing (to measure the effect of the washing group).

The high number of brews required is due to the simulation of the day of a bar: it was estimated that 500 brews is the average number of coffees made in a day. But the degradation could not be visible: the brews should therefore continue until it is detected.

Influence of external parameters on coffee quality

As a consequence of the first two experimental sessions, it was observed that it was necessary to deepen the tests related to the **influence of external parameters on coffee quality**. Since pressure is a difficult parameter to modify by the customer, it was decided to perform the tests by varying only the grinding grade and the dose. In fact, it is common practice among customers to insert a lower dose into the filter holder (saving on the purchase of coffee) and sometimes to compensate it with a larger grinding grade: this leads to a loss of coffee quality, a real concern of the company. The variations selected for the external parameters are:

- dose, -2g, -1g, optimal e +1g (optimal corresponds to 14g)
- grinding grade, +2, +1, optimal, -1 (optimal corresponds to the grinding set at the time of calibration)

Following the same pattern of the first two experimental sessions, purges and washes are always executed on one group, while both are absent on the other. The test is divided into:

- 60 optimal brews (after calibration and washing) as a reference;
- 15 tests of 20 brews each, corresponding to all combinations of the 4 values chosen for dose and grinding (excluding the combination in which both are optimal);
- 20 optimal brews made each time before changing the grinding grade of the coffee beans, to have a comparison with the initial situation.

There are therefore a total of 20 tests, with 440 brews per group.

6.2 Coffee quality and thresholds analysis

A fundamental parameter for the analysis, discussed since the first experimental session, concerns the definition of quality coffee. Correct calibration, constant cleaning of the machine and brew according to optimal parameters are therefore the necessary behaviour for the brew of quality coffee (and to prevent any breakdowns). This is what Lavazza asks its customers to do in order to maintain a high standard. Through empirical studies, Lavazza experts have determined quality thresholds (one lower and one higher) for each of the three variables measured in the cup (listed in table 6.1).

Since tasting each coffee by an expert to determine its quality would have been too long, it was decided to use the thresholds. In order to study the influence of the modification of external variables on the quality of the coffee in the cup, a univariate analysis was performed, whose results are represented by the heatmap in figure 6.1.

	Min	Max
Flow (ml/s)	0.49	1.33
Time (s)	20	27
Quantity (ml)	18	30

Table 6.1. Coffee quality in cup thresholds

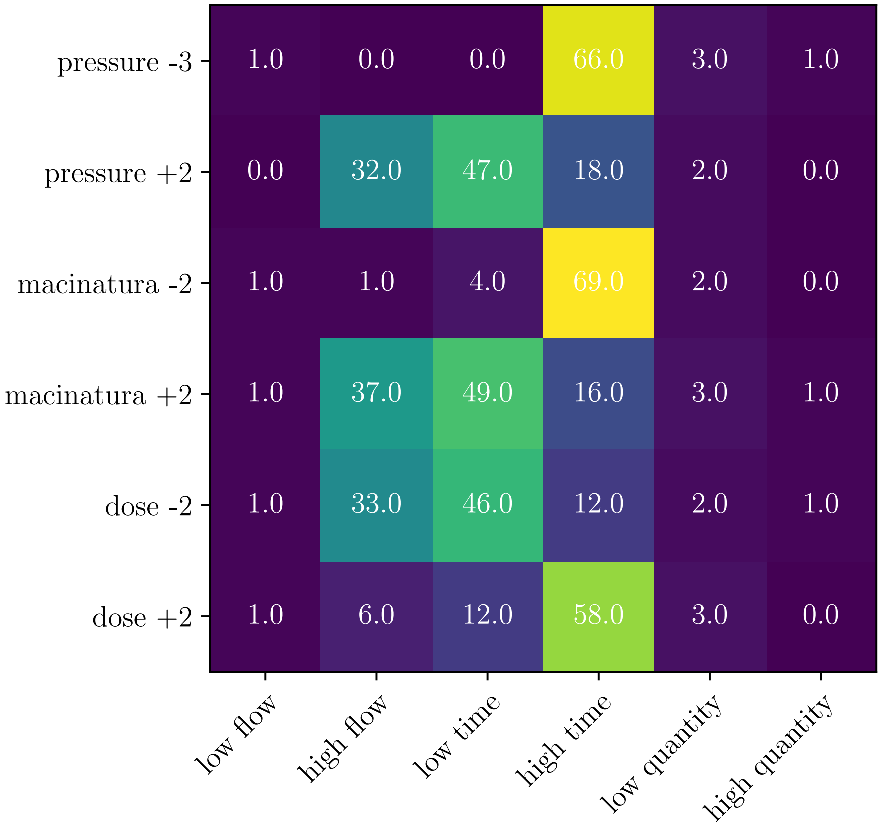

Figure 6.1. Univariate analysis heatmap

Each cell indicates the percentage of low quality brews according to the indicator in the column, caused by the external variable in the row. The results of the study are summarized as follows:

- the brew time is above the threshold in 60-70% of brews where the external variables are low pressure, a fine grinding or a high dose.
- the time is below the threshold (45-50% of brews) when the values of pressure, grinding and dose are complementary to the previous ones.

- the quantity is practically always inside the thresholds, despite the variations of the external parameters.
- he flow is above the threshold in 35% of brews with high pressure, coarse grinding or low dose.
- the flow is practically never below the threshold, despite variations in external parameters.

The behaviour of the time described above confirms the expectations of the domain experts. Concerning the quantity instead, the result is more relevant: despite the variation of the external parameters, it remains constant. It seems therefore only to derive from the process of calibration of the machine, i.e. from the *Cdv* programmed. What is most surprising is the behaviour of the flow: the causes that lead it to exceed the threshold are those expected (although with a relatively low percentage). The fact that it is practically never below the threshold, however, is suspicious. The cause of this behaviour was identified in the definition of thresholds: in fact, analysing the specifications used to carry out the calibration (46 g of coffee in 25 s), the flow should be centred on the value of 1 ml/s:

$$
\begin{aligned}
Q_{ml} &= Q_g \cdot \rho \\
\rho_{coffee} &= 0.98\ \frac{g}{ml} \\
Q_{coffee\_ml} &= 46 \cdot 1.02 = 46.92\ ml \\
\Phi_{double\_coffee} &= \frac{46.92}{24} = 1.95\ ml/s \\
\Phi_{single\_coffee} &= \frac{1.95}{2} = 0.98\ ml/s \approx 1 ml/s
\end{aligned}
$$

The thresholds of the flow were therefore discussed and redefined starting from the first experimental session data. Only the coffees brewed with optimal external parameters of group 2 (the one receiving the washes) were then analysed: the results of the study of the average and of the standard deviation of the flow are shown in the table 6.2.

The table 6.3 indicates the recalculated thresholds. The time and the quantity are unchanged: the time is more variable, while the quantity is very constant and therefore depends almost exclusively on the calibration. If correctly set, it should be around 24 ml, but the original thresholds have been kept to allow a wider range: the parameter is more subjective and in Italy it varies from region to region.

	Mean	Standard deviation
Flow (ml/s)	1	0.23

Table 6.2. Flow statistics from first experimental session data (test 1, group 2)

	Min	Max
Flow (ml/s)	0.77	1.23
Time (s)	20	27
Quantity (ml)	18	30

Table 6.3. Proposed in cup thresholds from first experimental session data

Analysing the second experimental session, it was discovered that the majority of the brews made with the optimal external parameters had a very high quantity of coffee brewed in the cup (outside the threshold): consequently, it was decided not to deepen the research on the quality thresholds, but to carry out the study of correlation of variables (section 6.3).

In conclusion, the aim of this study was to define new quality thresholds for coffee in the cup. Through the transfer functions (calculated in the section 6.3), they can be applied to telemetry data. In the section 6.6.1 a different approach was adopted: the quality thresholds on the flow were obtained from trusted customers on the market directly from the telemetry data. They were then compared with those obtained in this section.

6.3 Variables correlation

The main purpose of the laboratory sessions was from the very beginning to find a correlation between the three groups of variables available:

- in cup variables
- PC card variables
- telemetry variables

The correlation is studied because the final aim is to retrieve information from data of machines on the market: for these machines, only the variables detected by the telemetry sensor are available. If a correlation between those three sources of information is found, the analysis on coffee quality and thresholds (section 6.2), defined for the variables measured in the cup, can be transferred with similar results.

Telemetry variables represent an elaboration of PC card data, which identify the data as detected by the sensor: for this reason an initial analysis is made between PC card and in cup variables. If the correlation turned out to be satisfactory, then the study is redirected between telemetry variables and in cup variables.

6.3.1 PC card - cup variables correlation

In the first experimental session (performed on the *Cimbali M100*), the variables available are those measured in the cup and the one from the PC card. Since the brew time variable coincides, the objective was to find a correlation between the other two quality variables: flow and quantity. The flow measured in the cup and the one measured by the PC card are both derived quantities, according to the following formulas:

$$\Phi_{cup} = \frac{Q}{t} \quad where$$

- Q is the total weighted quantity of coffee in cup.
- t is the total time of brew.

$$\Phi_{PC} = \frac{Cdv \cdot 0.5 \cdot \frac{1}{2}}{t} \quad where \tag{6.1}$$

- Cdv is the number of pulses of the flowmeter.
- 0.5 are the ml of water brewed per pulse of the flowmeter.
- $\frac{1}{2}$ is the factor to switch from double to single brew.
- t is the total time of brew.

The two trends in the graph (6.2) show that the two variables are presumably correlated.

It is therefore decided to proceed with the computation of the correlation index and with the evaluation of the linear regression.

Pearson Corr.	R-squared	Function
0.98	0.90	y = 0.52x + 0.04

Table 6.4. PC card flow and in cup flow correlation and regression indexes

Both the graph and the values in the table show a strong correlation between the two flows: from the measurement of the flow through the PC card it is always

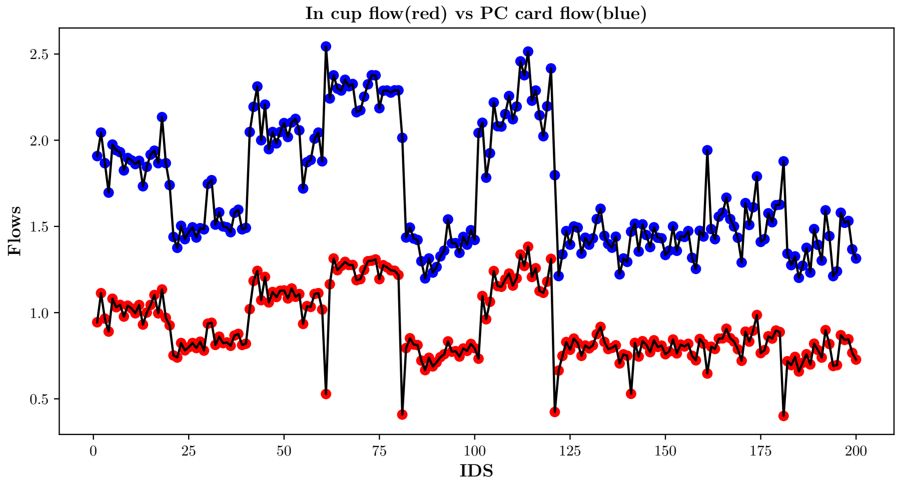

Figure 6.2. PC card flow and in cup flow comparison, first experimental session (focus on 200 brews)

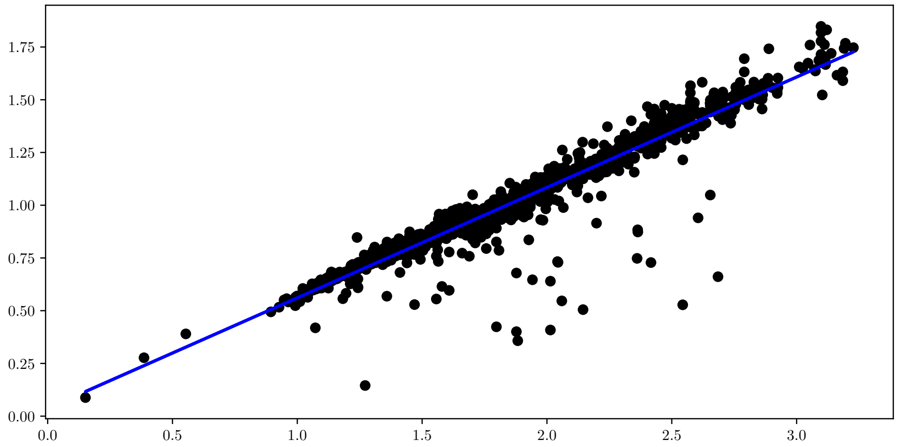

Figure 6.3. PC card flow and in cup flow regression line, first experimental session

possible to obtain the flow measured in the cup. However, it should be considered that the linear regression function is only valid in this phase for the machine used for the analysis (*Cimbali M100*) and for the current calibration (177 *Cdv*). Then, in order to proceed with the study of the variable quantity, the data related to the quantity of coffee weighed in the cup must be compared with the data related to the quantity of water measured by telemetry. The formula for obtaining the latter is the following:

$$Q_{PC} = Cdv \cdot 0.5 \cdot \frac{1}{2} \quad where \tag{6.2}$$

- Cdv is the number of pulses of the flowmeter.
- 0,5 are the ml of water brewed per pulse of the flowmeter.
- $\frac{1}{2}$ is the factor to switch from double to single brew.

Analysing the graph of the trends, we notice that both quantities are very constant, but the quantity measured in the cup presents some anomalies.

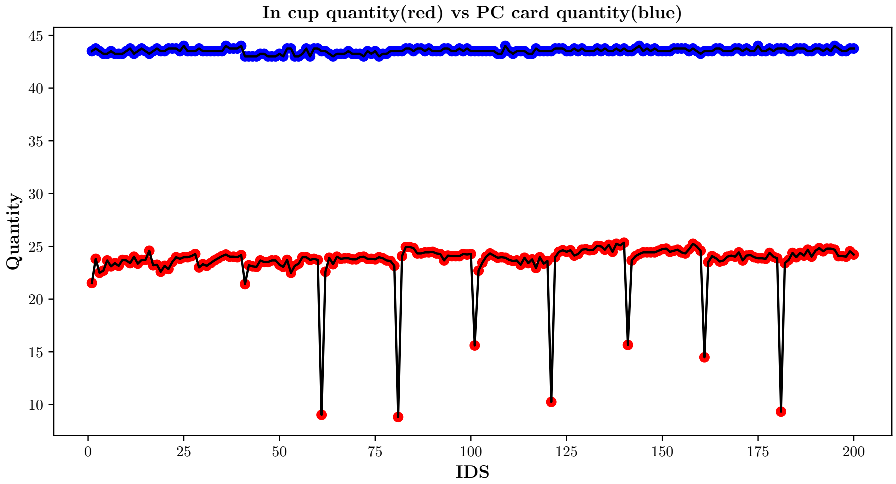

Figure 6.4. PC card water quantity and in cup quantity comparison, first experimental session (focus on 200 brews)

The behaviour is very unusual because the coffees are all made with the same calibration. It should be remembered, in fact, that the first experimental session was carried out by modifying the three external variables (*dose, pressure, grinding*) every 20 brews (tests). The quantity of coffee brewed in the cup, therefore, should have been constant within the same test (external variables fixed): instead, it can be observed that there are very marked lower peaks, in which the quantity of coffee brewed is much lower than the thresholds provided by the experts of the Training Center (6.1). For this reason, an analysis of the anomalies was conducted, aimed at highlighting a connection or a pattern between the excessively short brews. In graph 6.5 are shown the results of a univariate analysis: each histogram represents the number of anomalies related to the value of the external variable (not optimal) or of the variable in cup (out of threshold) indicated on the x axis.

The 25 anomalies (out of 1080 brews, 2%) have no recurrent pattern with respect to the external variables or to the variables in the cup. The only common scheme is that 17 outliers out of 25 (70%) correspond to the first brew of the test to which they belong: it seems to be due to a human error in the weighing of the coffee at the start of a new test. To eliminate these anomalies, the data cleaning process

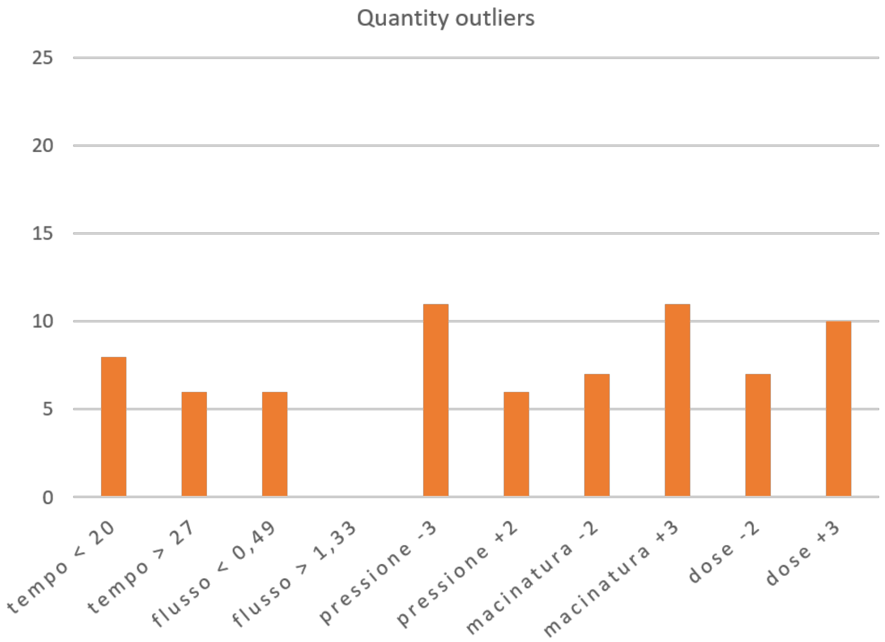

Figure 6.5. In cup quantity univariate analysis

described in paragraph 5.2.1 was applied. Graph 6.6 shows the trend of the brews after the data cleaning.

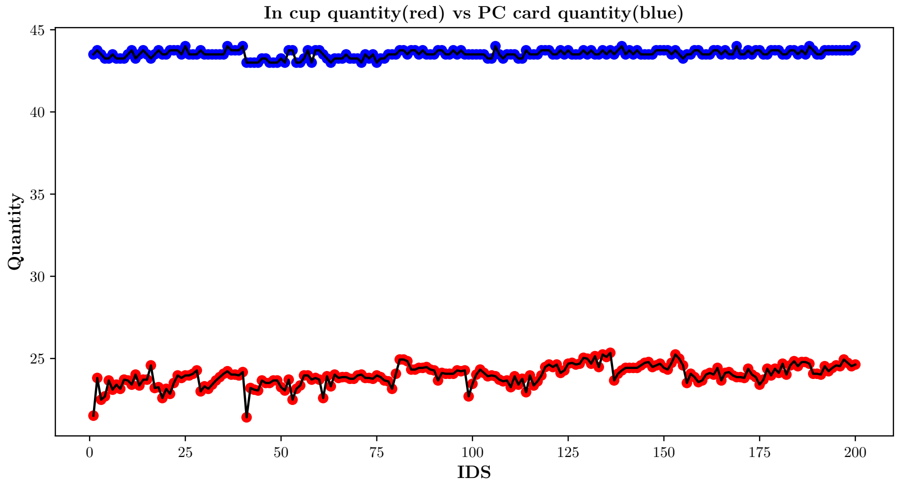

Figure 6.6. PC card water quantity and in cup quantity comparison after data cleaning

It is evident from graph 6.6 that the quantity of coffee brewed in the cup has a greater variance than the water brewed by the machine: the pulses produced by the flowmeter, in fact, are very precise and consequently the variance of the water brewed is very low (see table 6.5).

Quantity	Mean (ml)	Standard deviation (ml)
In cup	24.0	1.2
From PC card	43.5	0.3

Table 6.5. PC card quantity and in cup quantity statistics

It was decided to follow the same procedure as for the flow, calculating Pearson's correlation index and then proceeding with linear regression. The table 6.6 and the graph 6.7 show the results.

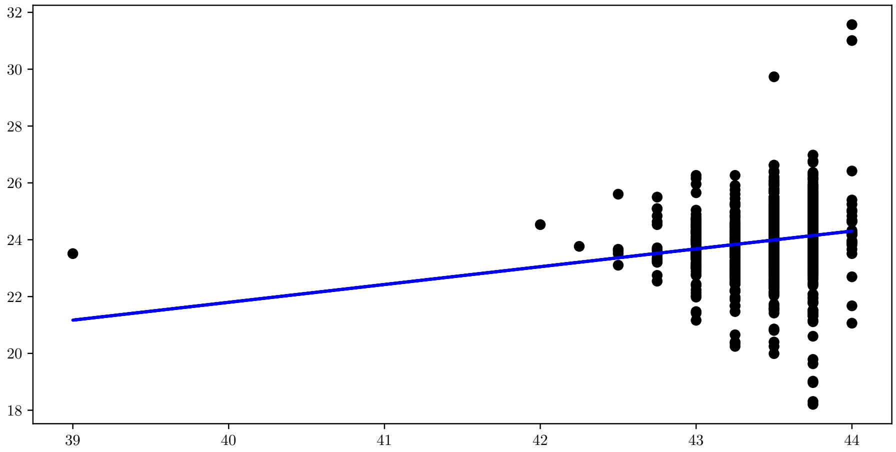

Figure 6.7. PC card quantity and in cup quantity regression line

The results in this case seem disappointing compared to those of the flow: both the correlation index of Person and the linear regression show negative results. In fact, the variations in the quantity of coffee in the cup between two brewing processes are not related to a variation in the quantity of water dispensed: the quantity of coffee brewed in the cup is affected by the variation in the external variables of this experiment. However, the sensor that measures the fan blows is not affected by external changes as it is upstream of the coffee block. In addition, in detecting the amount of coffee dispensed in the cup, there are two possibilities of human error that should not be overlooked:

1. the reading of the weight of the coffee dispensed is done by eye and the transcription of the data is done by hand.

Pearson Corr.	R-squared	Function
0.17	0.03	y = 0.63x + 3.28

Table 6.6. PC card quantity and in cup quantity correlation and regression indexes

2. After a few seconds the surface foam produced in the coffee brewing process evaporates, reducing the weight.

For the above-mentioned reasons and since the standard deviation of the brew in the cup is just over 1 *ml*, such deviation turns out to be not so relevant. The main consideration however is that there is a correspondence between the average amount of water dispensed (45.3 *ml*) and the average amount of coffee in the cup (24 *ml*). Summarizing, a *Cimbali M100* machine, calibrated at 177 *Cdv*, produces an average amount of water of 43.5 *ml*, which corresponds to 24 *ml* of coffee dispensed in the cup with a standard deviation of 1.2 *ml*.

To confirm and generalize the results, the same type of study was also performed in the second experimental session: the machine (always *Cimbali M100*) presents a different initial situation (calibration at 207 *Cdv*) and a different variation of external parameters (see section 6.1.2). The process followed is exactly the same as in the previous session: after evaluating Pearson's correlation index, the linear regression is calculated. The flow in the cup, as before, can be calculated accurately from the flow obtained from the PC board (see table 6.7 and figure 6.8). In this case the transfer function is different from the previous one (6.4): since the machine used and the type of mixture are the same (the variation of the external parameters is not relevant), this means that different transfer functions correspond to different calibrations.

Pearson Corr.	R-squared	Function
0.96	0.97	y = 0.60x + 0.03

Table 6.7. PC card flow and in cup flow correlation and regression indexes, second experimental session

The quantity weighed in the cup, as before, has some anomalies in relation to the first brew of the tests: with the cleaning process (5.2.1) the outliers have been eliminated. Therefore, proceeding with the calculation of correlation and linear regression, the same previous problems are encountered: it is not possible to accurately calculate the quantity in the cup from the flowmeter pulses (see table 6.8). In this case, however, the standard deviation of the quantity of coffee brewed in the cup is even lower (table 6.9). This is the result of a smaller variation in the

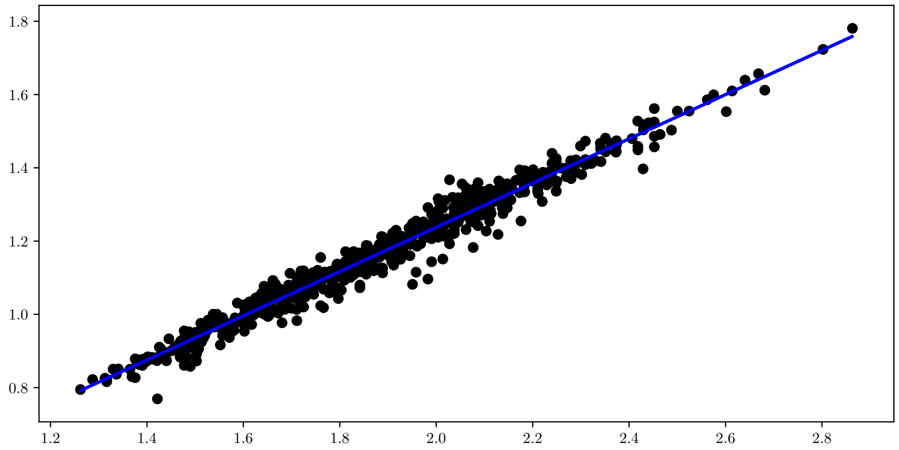

Figure 6.8. PC card flow and in cup flow regression line, second experimental session

external parameters (6.1.2).

Pearson Corr.	R-squared	Function
0.05	0.003	y = 0.21x + 21.42

Table 6.8. PC card quantity and in cup quantity correlation and regression indexes, second experimental session

The results obtained from the second experimental session, regarding the correlation between data measured in the cup and data obtained from the PC card, therefore confirm those of the first experimental session: the flow in the cup can be obtained from the flowmeter pulses (but the transfer function depends on calibration), while the quantity can only be approximated.

6.3.2 Telemetry - cup variables correlation

In the second experimental session, the variables obtained from the telemetry sensor were added: the average flow and the telemetry time. The first problems of lack and alignment of telemetry data (shown in section 5.2.1) have appeared: for this reason, the number of brews that can be analysed is 708 out of 1160 (39% of brews are missing). The correlation analysis carried out on the PC card data was intended to validate the correctness of the data acquired by the sensor and to verify the possibility of finding the transfer functions to obtain in cup variables values. The final objective, as mentioned before, is to find the correlation between the measurements in the cup and those of telemetry, the only ones available for the

Quantity	Mean (ml)	Standard deviation (ml)
In cup	32.0	0.9
From PC card	51.7	0.2

Table 6.9. PC card quantity and in cup quantity statistics, second experimental session

machines on the market.

In the same way as for the correlation between the PC card data and the data measured in the cup, the analysis started from the study of the flow: observing the trend of the two flows it is evident that also in this case the two can be correlated (graph 6.9). In fact, proceeding with the analysis of the statistical indices and with the linear regression, the correlation between the average telemetry flow and the flow measured in the cup is very strong (table 6.10 and figure 6.10). Since it has been shown that the transfer function between the flow in the cup and the flow of the PC card varies with the variation of the *Cdv*, we expect that the transfer function between the flow in the cup and the average flow of telemetry will behave in the same way: the comparison has been performed in the next two sessions (6.3.3, 6.3.4).

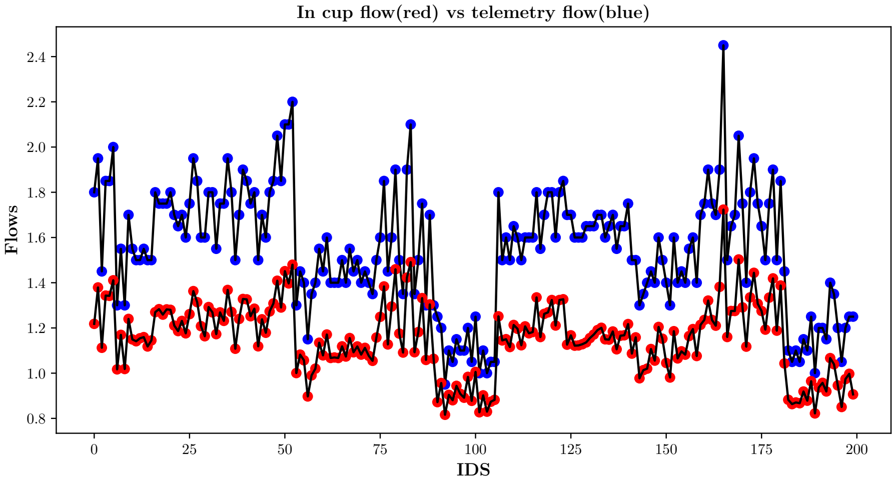

Figure 6.9. Telemetry flow and in cup flow comparison, second experimental session

Considering, instead, the quantity of water dispensed, this is not available as telemetry data. However, it can be derived from the average flow and brew time

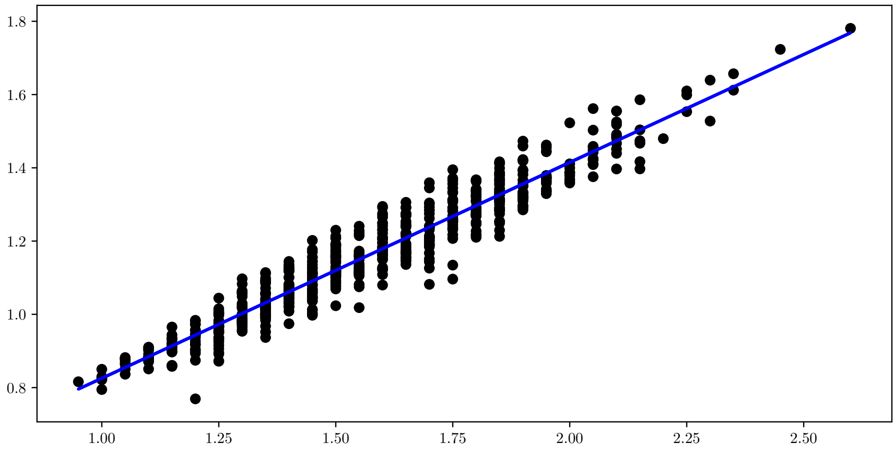

Figure 6.10. Telemetry flow and in cup flow regression line, second experimental session

Pearson Corr.	R-squared	Function
0.96	0.93	y = 0.59x + 0.24

Table 6.10. telemetry flow and in cup flow correlation and regression indexes, second experimental session

through the formula:

$$Q_{Telemetry} = \Phi_{average} \cdot t \quad where \tag{6.3}$$

- $\Phi_{average}$ is the telemetry average flow
- t is the telemetry brew time

Then, by calculating the correlation index and the linear regression between the quantity weighed in the cup and the quantity obtained from the telemetry, the results obtained are shown in table 6.11. Also in this case, it is not possible to find a precise regression function for the transition from the amount of telemetry water to the amount of coffee brewed in the cup. While previously the amount of water measured by the PC card was basically constant, now the amount of telemetry water has a more significant variance. This is demonstrated by the statistics shown in table 6.12. The case is very different from the previous one: while the quantity of coffee in cup is still constant, the standard deviation of telemetry water is no longer negligible. It is also evident from graph 6.11 the great fluctuations of the telemetry flow and the low correlation between the curves of the quantities.

In conclusion, the flow in cup can be effectively approximated: in this way the flow quality thresholds defined by the domain experts (and computed from data 6.2) can be applied. The two possible approaches that can be followed with production data are:

1. apply the transfer function to the cup thresholds, finding the corresponding telemetry thresholds to be applied to the average telemetry flow;
2. apply the transfer function to the telemetry flows and then compare the data with the thresholds in the cup;

Pearson Corr.	R-squared	Function
0.07	0.04	y = 0.07x + 28.97

Table 6.11. telemetry quantity and in cup quantity correlation and regression indexes

Quantity	Mean (ml)	Standard deviation (ml)
In cup	32.0	0.8
From Telemetry	43.9	2.4

Table 6.12. Telemetry quantity and in cup quantity statistics

The amount of coffee in cup, instead, cannot be accurately predicted from the telemetry water amount (formula (6.3)) in the *Cimbali M100* machine. Moreover, the value does not coincide and is much less constant than the quantity of water calculated from the PC card data (graph 6.12): this suggests that, since the telemetry time and the time calculated from the PC card coincide, the average telemetry flow is not calculated directly from the *Cdv*, but it is processed. This fact has been confirmed by Lavazza's domain experts, but unfortunately only the manufacturer of the machine is currently aware of the data processing algorithm.

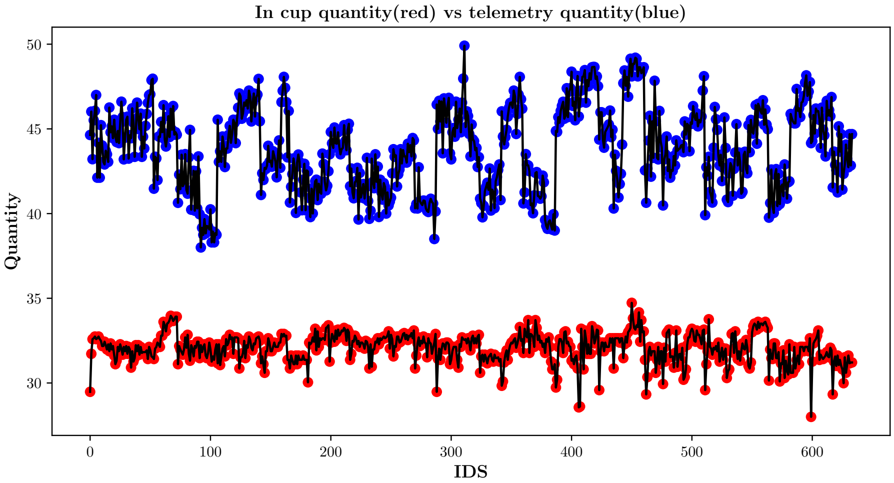

Figure 6.11. Telemetry water quantity and in cup quantity comparison, second experimental session

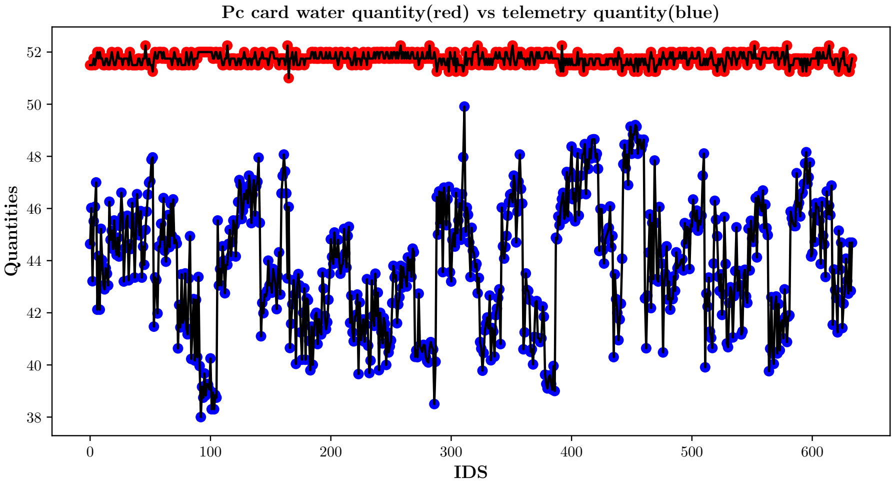

Figure 6.12. Telemetry water quantity and PC card water quantity trend, second experimental session

In order to deepen and verify the previous conclusions, the study was continued by analysing the data coming from the experiment based on the flowmeter pulses (6.1.3). In particular, the objectives are:

- compute the transfer functions of the telemetry flow for different calibrations (where only the *Cdv* changes) on the same machine.
- compute the transferring functions of the telemetry flow to different machines.

- deepen the correlation between the amount of water dispensed in the cup and the calculated amount of water of telemetry.

6.3.3 Telemetry - cup variables correlation on *Rancilio CL11* machine

The first machine considered is the *Rancilio CL11*. The test sequence of tests on this machine is complete:

- the first test of 20 brews with machine calibrated to 148 *Cdv*;
- 4 tests with 10 brews, each with 10 more *Cdv* than the previous one (from 158 to 188 *Cdv*);
- 4 tests with 10 brews, each with 10 less *Cdv* than the previous one (from 138 to 108 *Cdv*).

After cleaning the dataset (according to the procedure shown in section 5.2.1), the first analysis was performed on the flow. Like in the previous section, it was analysed the trend of the telemetry flow compared to that of the flow in the cup. As you can see in graph 6.13 (all the brews are present), the trends are similar, but the relative distance between the points is variable: for this reason, it is probable that there is more than one transfer function for the single machine. The correlation analysis confirms this hypothesis: the correlation and regression indices in table 6.15 are low and many points are far from the regression line shown in graph 6.14.

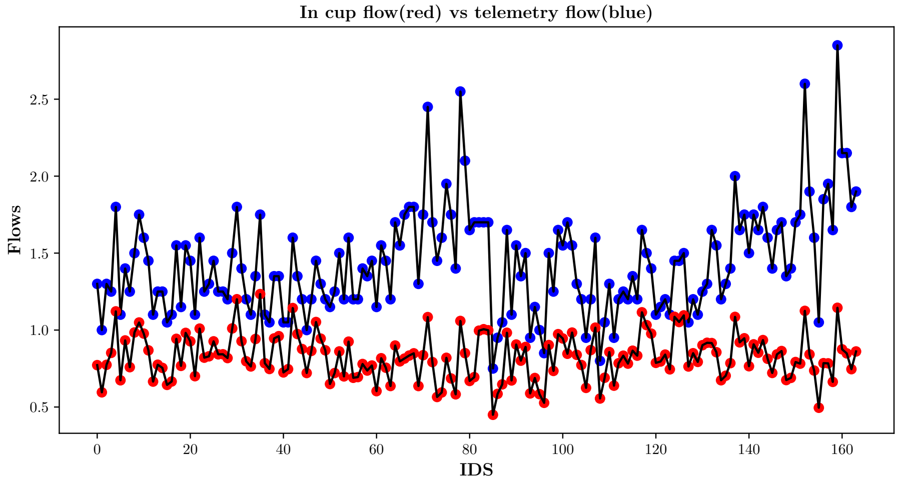

Figure 6.13. Telemetry flow and in cup flow trend (*Ranclio CL11*), proposed experimental session

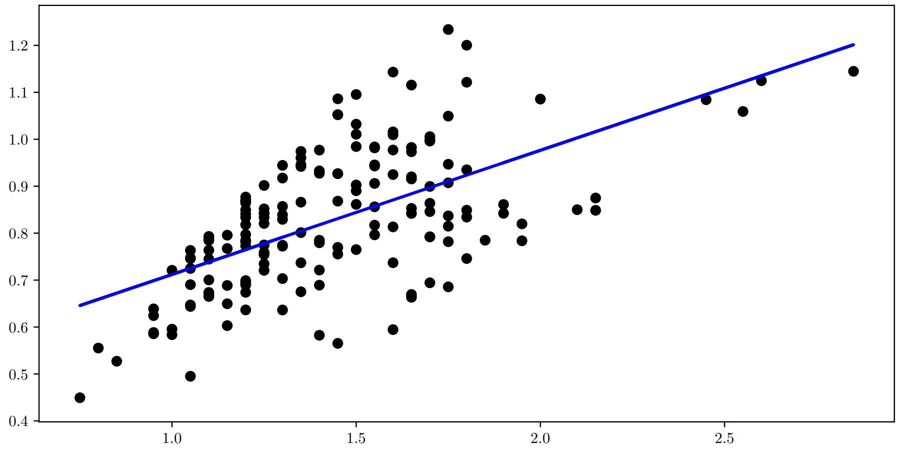

Figure 6.14. Telemetry flow and in cup flow regression line (*Ranclio CL11*), proposed experimental session

Pearson Corr.	R-squared	Function
0.61	0.37	y = 0.27x + 0.44

Figure 6.15. Telemetry flow and in cup flow correlation and regression indexes (*Ranclio CL11*), proposed experimental session

Previous analyses have shown that for the same machine, with two different calibrations, two different linear regression functions were required. It was therefore decided to study the correlation between the average flow of telemetry and the one measured in the cup separately for each test. Table 6.13 shows correlation and regression indices for all 18 tests: both Pearson Correlation and r-squared index confirm a high correlation between the average telemetry flows and those measured in the cup, if calculated per test. It is therefore evident that each transfer function is related to a different number of flowmeter pulses, as expected.

		Pearson Corr.	R-squared	Function
Group 1	test1 (148 *Cdv*)	0.97	0.95	y = -0.02 + 0.63x
	test2 (158 *Cdv*)	0.99	0.99	y = 0.08 + 0.58x
	test3 (168 *Cdv*)	0.99	0.99	y = 0.02 + 0.65x
	test4 (178 *Cdv*)	0.99	0.99	y = 0.70x
	test5 (188 *Cdv*)	0.99	0.99	y = 0.02 + 0.71x
	test6 (138 *Cdv*)	0.99	0.99	y = 0.57x
	test7 (128 *Cdv*)	0.99	0.98	y = 0.02 + 0.51x
	test8 (118 *Cdv*)	0.95	0.91	y = 0.03 + 0.44x
	test9 (108 *Cdv*)	0.98	0.97	y = -0.06 + 0.44x
Group 2	test1 (148 *Cdv*)	0.99	0.99	y = 0.04 + 0.57x
	test2 (158 *Cdv*)	0.97	0.94	y = 0.10 + 0.58x
	test3 (168 *Cdv*)	0.99	0.99	y = 0.08 + 0.59x
	test4 (178 *Cdv*)	0.99	0.99	y = 0.06 + 0.65x
	test5 (188 *Cdv*)	0.99	0.99	y = -0.03 + 0.74x
	test6 (138 *Cdv*)	0.99	0.99	y = 0.04 + 0.52x
	test7 (128 *Cdv*)	0.99	0.99	y = -0.04 + 0.54x
	test8 (118 *Cdv*)	0.99	0.98	y = 0.13 + 0.38x
	test9 (108 *Cdv*)	0.97	0.94	y = 0.08 + 0.37x

Table 6.13. telemetry flows and in cup flows correlation and regression indexes

The comparison between the quantity measured in the cup and the telemetry quantity started by the graphical analysis: in figure 6.16 are shown the trends of the two quantities compared. Analysing them in a visual way, they seem correlated, as the trend is very similar, and the relative distance between the points of the two curves appears constant for all the brews.

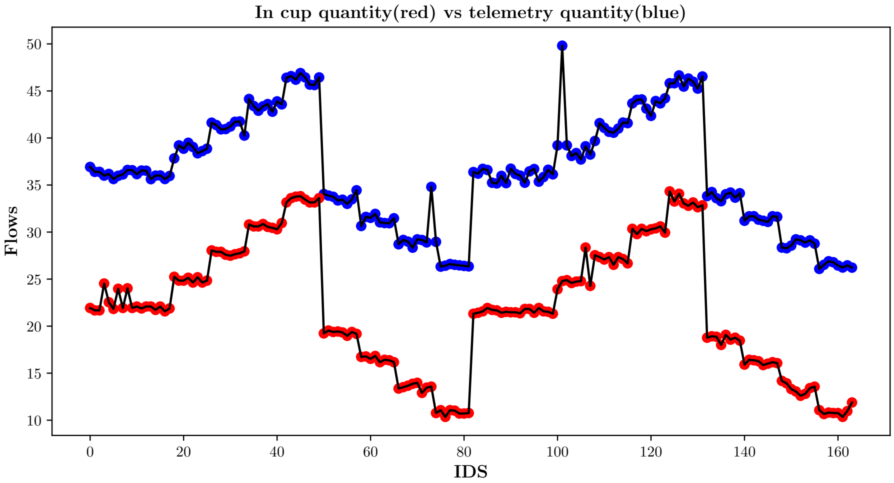

Figure 6.16. Telemetry water quantity and in cup quantity trend (*Ranclio CL11*), proposed experimental session

Continuing with the analysis of the correlation indices, the visual intuition is confirmed: the points are all located in proximity of the regression line (graph 6.17) and both the correlation index of Pearson and the r-squared index show a high correlation between the two types of data (table 6.18). On the *Rancilio CL11* machine it is therefore possible to predict the amount of coffee brewed in the cup from the data of telemetry (unlike the machine *Cimbali M100*). As in the test performed on the *Cimbali M100*, for each brew there is the number of *Cdv*. This allows to make a comparison between the quantity of water computed starting from them (according to the formula (6.2)) and the one calculated from average flow and telemetry time.

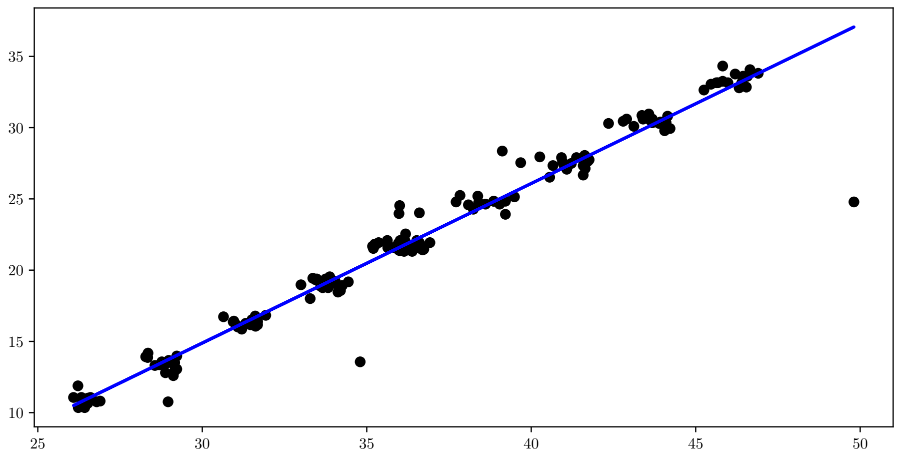

Figure 6.17. Telemetry quantity and in cup quantity regression line (*Ranclio CL11*)

Pearson Corr.	R-squared	Function
0.99	0.98	y = 1.15x + 19.6

Figure 6.18. Telemetry quantity and in cup quantity correlation and regression indexes (*Ranclio CL11*)

The graph in figure 6.19 shows that the quantity values for each individual brew correspond almost perfectly. The next step is to check the flow: if the average telemetry flow coincides with the flow obtained from time and *Cdv* (according to the formula (6.2)), it means that the machine does not process any data detected by the sensor. The analysis on the flow in theory would be superfluous: verified that the quantities correspond (the brew time detected by the PC card should coincide with that reported in telemetry), the flows obtained from them can only coincide. In fact, the two flows correspond, as shown in the figure 6.20. There are cases in which the telemetry data is slightly different from the one obtained from the *Cdv*: this happens because the data relative to the number of flowmeter pulses is obtained from the reading of the data set on the display (constant) and not from the PC card. The latter in fact is not stable because either the machine is not precise in the emission of pulses, or the sensor is not precise in the counting.

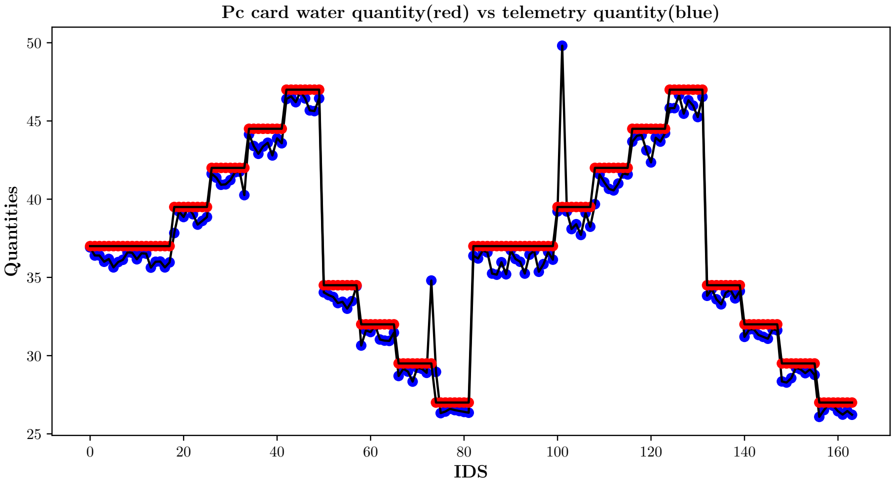

Figure 6.19. Telemetry quantity and PC card quantity trend (*Ranclio CL11*), proposed experimental session

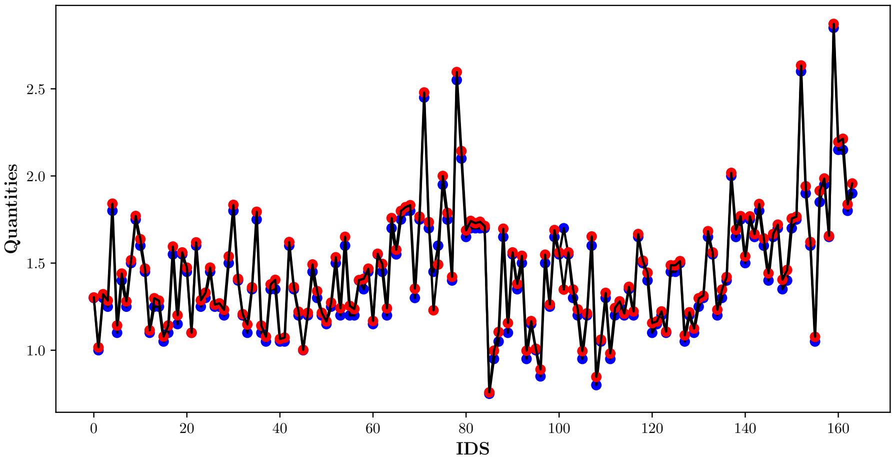

Figure 6.20. Telemetry flow and PC card flow trend (*Ranclio CL11*), proposed experimental session

In conclusion, this analysis showed that the *Rancilio CL11* machine does not process the telemetry data at all, compared to the *Cimbali M100* machine. The average flow is then calculated directly from the pulses of the flowmeter using the formula $\Phi_{Tel} = Cdv \cdot 0{,}5 \cdot \frac{1}{2}$ which corresponds to the formula (6.1).

6.3.4 Telemetry - cup variables correlation on *Faema E71* machine

The second machine examined is the *Faema E71*. In this case the test pattern is not complete, as there are a total of 6 tests available for each of the two groups:

- first test of 60 brews with optimal calibration (180 *Cdv*);
- 2 tests with 20 brews with higher calibration (190-200 *Cdv*);
- 3 tests with 20 brews with lower calibration (170-160-150 *Cdv*).

Although the laboratory experiment is not complete, the same analyses have been applied in the same way as for the machine *Rancilio CL11*. Before analysing the data, the dataset cleaning was performed, using the process explained in section 5.2.1. So here are the results:

- The flow has several transfer functions, depending on the calibration (see table 6.14). Apart from test 5 of the first group and tests 3 and 6 of the second group, as before, the linear regression approximates with very good precision the flow brewed in the cup, starting from the telemetry flow data.
- The analysis of the correlation between the telemetry quantity and the quantity in the cup has given negative results: from graph 6.21, it can be seen that the two trends are only partially aligned. Calculating Pearson's correlation and linear regression function, the indices are very low (see table 6.23).

		Pearson Corr.	R-squared	Function
Group 1	test1 (180 *Cdv*)	0.94	0.89	y = -0.08 + 0.83x
	test2 (190 *Cdv*)	0.94	0.88	y = 0.08 + 0.58x
	test3 (200 *Cdv*)	0.98	0.96	y = 0.02 + 0.65x
	test4 (170 *Cdv*)	0.99	0.98	y = 0.70x
	test5 (160 *Cdv*)	0.70	0.50	y = 0.02 + 0.71x
	test6 (150 *Cdv*)	0.97	0.93	y = 0.57x
Group 2	test1 (180 *Cdv*)	0.99	0.98	y = 0.04 + 0.57x
	test2 (190 *Cdv*)	0.97	0.94	y = 0.10 + 0.58x
	test3 (200 *Cdv*)	0.73	0.53	y = 0.08 + 0.59x
	test4 (170 *Cdv*)	0.95	0.90	y = 0.06 + 0.65x
	test5 (160 *Cdv*)	0.99	0.99	y = -0.03 + 0.74x
	test6 (150 *Cdv*)	0.76	0.58	y = 0.04 + 0.52x

Table 6.14. telemetry flows and in cup flows correlation and regression indexes (*Faema E71*)

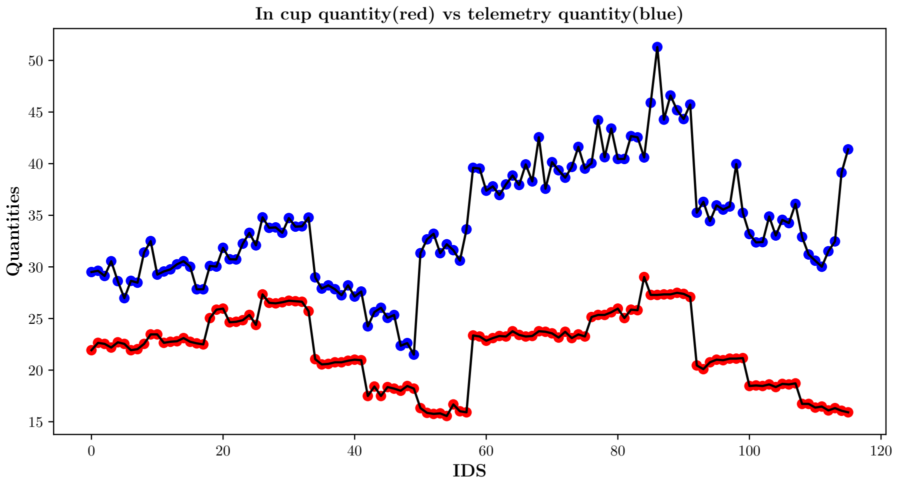

Figure 6.21. Telemetry quantity and in cup quantity trend (*Faema E71*), proposed experimental session

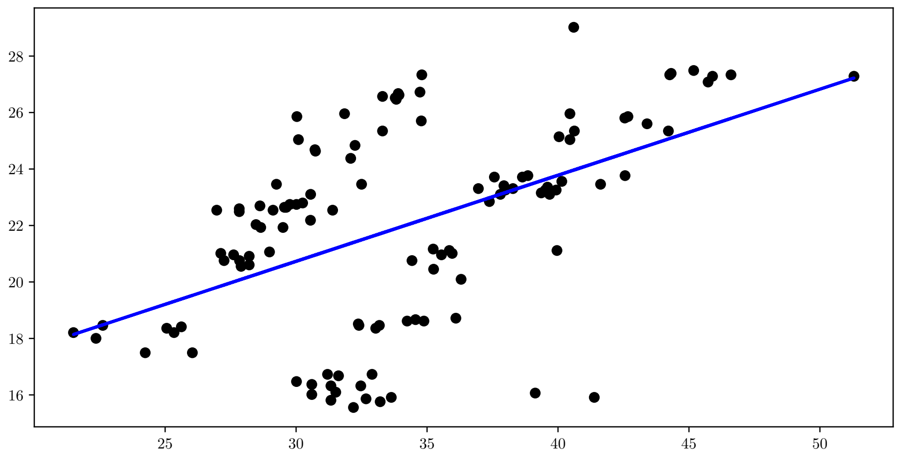

Figure 6.22. Telemetry quantity and in cup quantity regression line (*Faema E71*)

Pearson Corr.	R-squared	Function
0.50	0.25	y = 0.30x + 11.59

Figure 6.23. Telemetry quantity and in quantity flow correlation and regression indexes (*Faema E71*)

These results lead to the conclusion that the machine *Faema E71* (such as the *Cimbali M100*) apply an algorithm for processing the data detected by the sensor in calculating the average telemetry flow. The flow in the cup can therefore be predicted with precision, but the transfer function varies depending on the calibration. The quantity, however, due to data processing, cannot be estimated from the telemetry data.

6.4 Degradation

This section describes the analysis of degradation due to lack of washes (a test of the proposed experimental session, section 6.1.3), the first step towards predictive maintenance. The machine used for dispensing is the *Wega Urban*. The available data are:

- 60 initial brews with washes;
- 209 supplies for group 1, no washes;
- 527 group 2 supplies, no washes;

- 50 brews for each group executed with washes (purge after each brew and washing of the group every 20 brews).

The group considered in this work of thesis is group 2, for the greater number of coffee extractions. It was decided to analyse graphically the trend of the of the three telemetry variables (*flow, time, quantity*), i.e. the only ones available for machines on the market. Intuitively, as the machine's ducts become more and more dirty, brew after brew, the expected behaviours were the following:

1. The amount of water brewed remains constant, since calibration does not change throughout the experimental session.
2. The brew time increases.
3. The average flow decreases.
4. After group washes, the times and flows of the last brews are then comparable with those of the first 50 brews.

Analysing the water brewed (calculated with the formula (6.3)), the first unexpected behaviours are revealed. In fact, the dispensed water is not constant for all the 319 brews as forecasted: although flowmeter pulses remain constant, the quantity of water dispensed seems to depend on the day in which the brews are made (see figure 6.24). It appears, therefore, that the external factors (humidity, temperature and change of the coffee pack, for example) have an important effect on the brewing process.

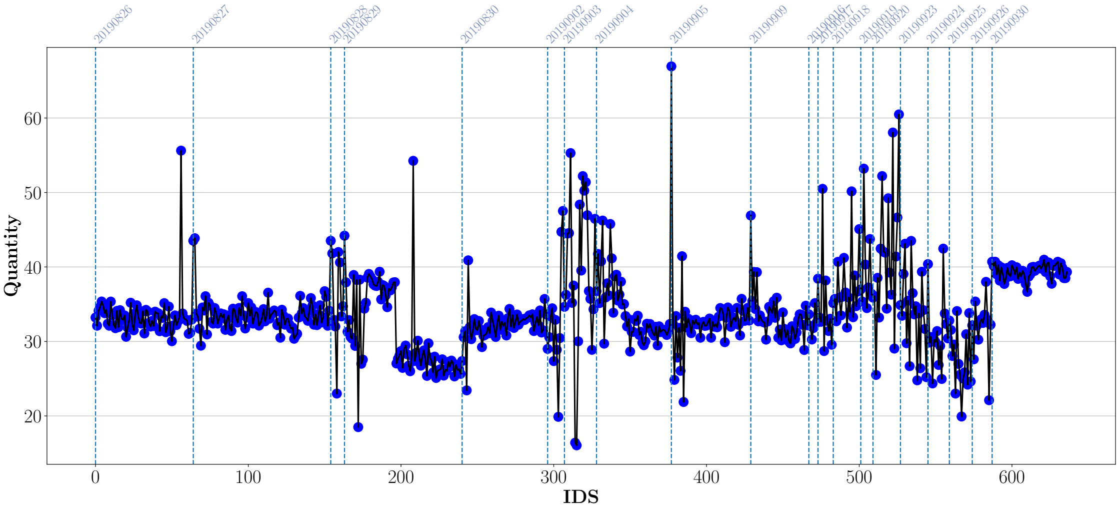

Figure 6.24. Degradation group 2 Telemetry quantity trend (*Wega Urban*), proposed experimental session

Proceeding with the analysis of the evolution of the brew time (shown in figure 6.25), it is not possible to identify visually any trend. The two boxes show the comparison between the first 60 brews (purges) and the last 50 brews (washes and purges): the brew times decrease, thanks to washes, but they are not comparable with those of the first brews. In fact, as shown in table 6.26, the average brew time between the two brew blocks differs by 10 seconds.

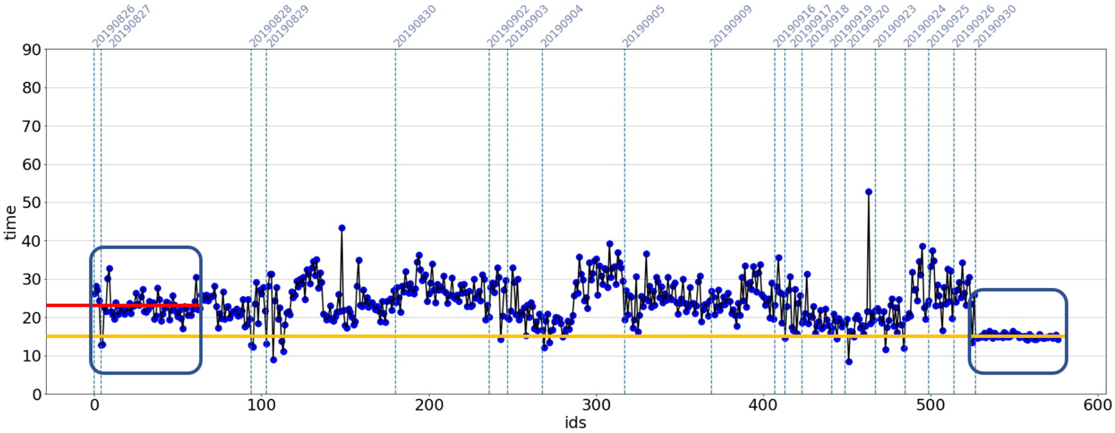

Figure 6.25. Degradation group 2 Telemetry time trend (*Wega Urban*), proposed experimental session

brews	average time (s)	standard dev. (s)
First 60	26.92	2.93
Last 50	16.14	0.82

Figure 6.26. Degradation group 2 telemetry time statistics comparation (*Wega Urban*), proposed experimental session

To deepen the analysis, two statistical tests have been used: the Dickey Fuller Test and the KPSS test (described in section 5.5). For the tests, the first 60 brews and the following 209 without washing were considered. The results of the two tests are given in tables 6.15 and 6.16. In the Dickey Fuller test, which measures if a curve is difference-stationary, the p-value is very low, close to zero (not all decimal places are given): this means that the null hypothesis can be rejected and it is assumed that the trend of time is stationary. In the KPSS test (measure if a curve is trend-stationary), the p-value is higher than 0.1: the value shows that the null hypothesis can not be rejected in this case and it is therefore assumed that the trend of the time is stationary (in this test null hypothesis and alternative hypothesis are reversed).

Dickey Fuller Test	
ADF Statistic	-12.593
p-value	0.000
Critical value (1%)	-3.442
Critical value (5%)	-2.867
Critical value (10%)	-2.569

Table 6.15. ADF test on group 2 time curve

KPSS Test	
KPSS Statistic	0.259
p-value	0.100
Critical value (10%)	0.347
Critical value (5%)	0.463
Critical value (1%)	0.739

Table 6.16. KPSS test on group 2 time curve

Therefore the visual test and the two statistical tests have concluded that the evolution of time does not present a trend, making it impossible to detect the degradation of the machine. Finally, in the last part of the analysis, the flow trend was studied.

The same techniques used for the study of the time evolution have been adopted. The results are specular:

- visually, no trend can be detected (figure 6.27).
- The last 60 brews are not comparable with the first 50 brews again: the average flow after the washes increases, but is not comparable with the one of the first brews. The difference is about 1 ml/s (see graph 6.27 and table 6.28).

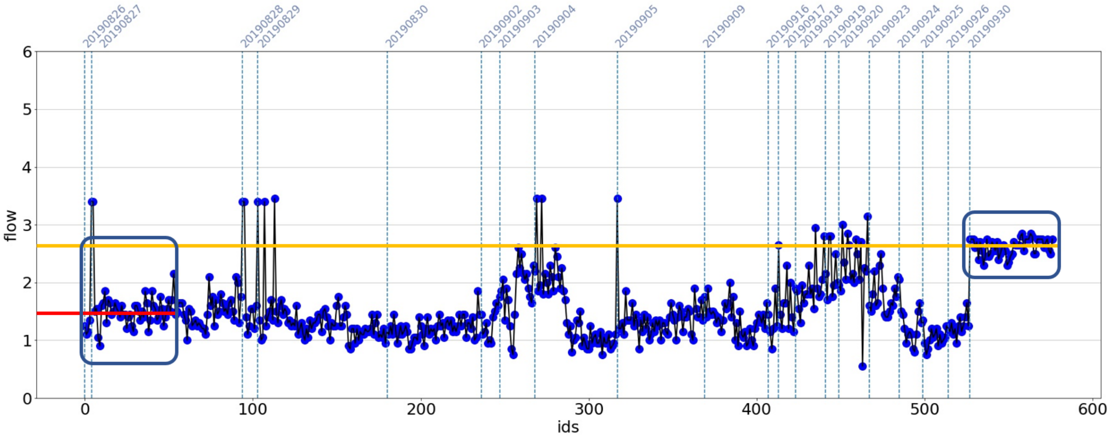

Figure 6.27. Degradation Telemetry flow trend (*Wega Urban*), proposed experimental session

brews	average flow (ml/s)	standard dev. (ml/s)
First 60	1.37	0.17
Last 50	2.37	0.17

Figure 6.28. Degradation group 2 telemetry flow statistics comparation (*Wega Urban*), proposed experimental session

Finally, the two statistical tests (ADF and KPSS) give a negative result: both show that the flow evolution is stationary (see tables 6.17, 6.18).

Dickey Fuller Test	
ADF Statistic	-12.593
p-value	0.000
Critical value (1%)	-3.442
Critical value (5%)	-2.867
Critical value (10%)	-2.569

Table 6.17. ADF test on 2 time curve

KPSS Test	
KPSS Statistic	0.259
p-value	0.100
Critical value (10%)	0.347
Critical value (5%)	0.463
Critical value (1%)	0.739

Table 6.18. KPSS test on group 2 time curve

The same analysis was carried out on group two: although the number of brews was higher (637 in total), the results obtained were the same as for group 1. The results are shown in the graphs.

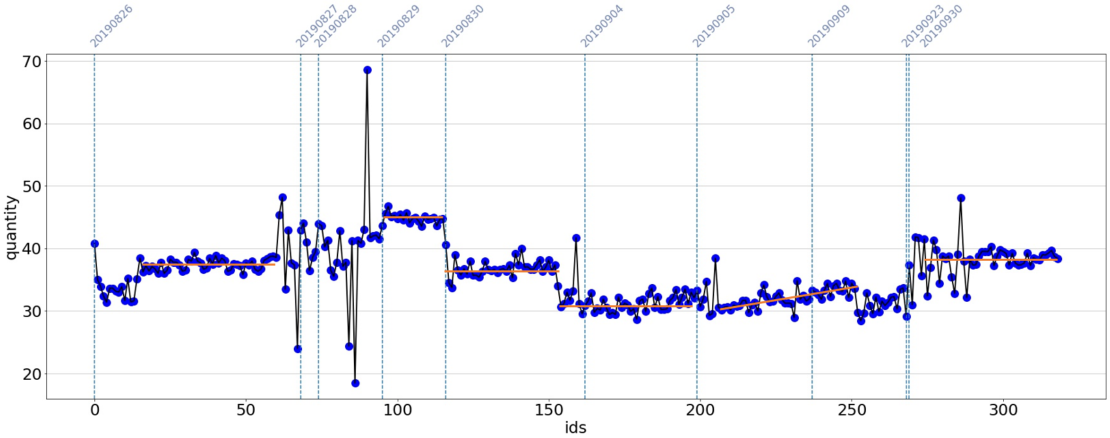

Figure 6.29. Degradation group 2 Telemetry quantity trend (*Wega Urban*), proposed experimental session

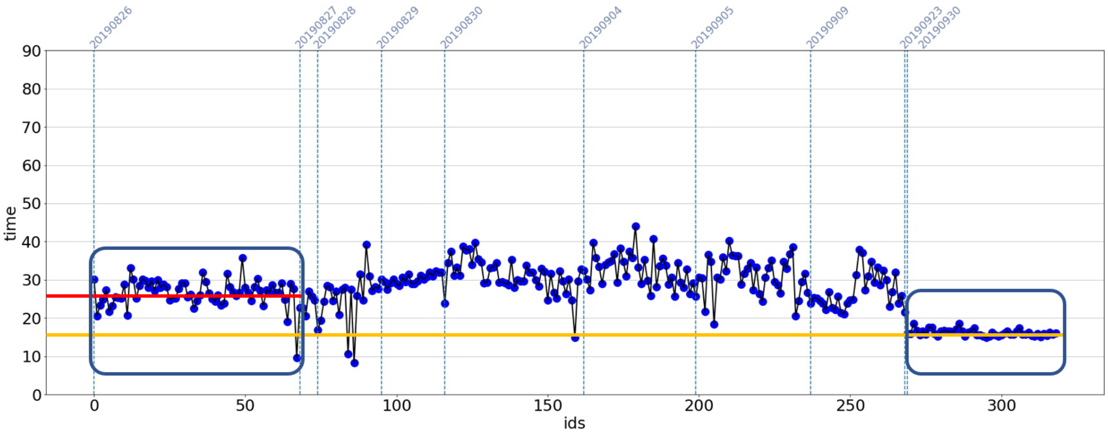

Figure 6.30. Degradation group 2 Telemetry time trend (*Wega Urban*), proposed experimental session

brews	average time (s)	standard dev. (s)
First 60	26.92	2.93
Last 50	16.14	0.82

Figure 6.31. Degradation group 2 telemetry time statistics comparation (*Wega Urban*), proposed experimental session

Dickey Fuller Test	
ADF Statistic	-12.593
p-value	0.000
Critical value (1%)	-3.442
Critical value (5%)	-2.867
Critical value (10%)	-2.569

Figure 6.32. ADF test on group 2 time curve

KPSS Test	
KPSS Statistic	0.259
p-value	0.100
Critical value (10%)	0.347
Critical value (5%)	0.463
Critical value (1%)	0.739

Figure 6.33. KPSS test on group 2 time curve

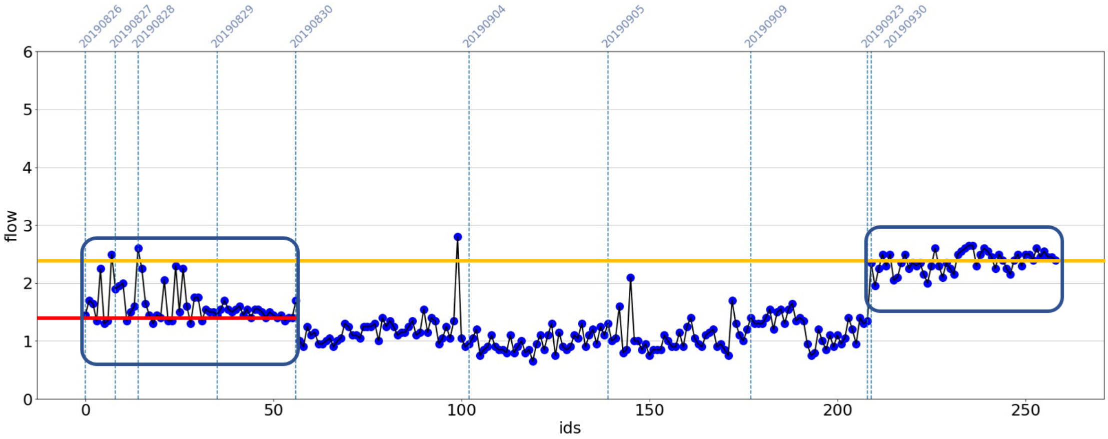

Figure 6.34. Degradation Telemetry flow trend (*Wega Urban*), proposed experimental session

brews	average flow (ml/s)	standard dev. (ml/s)
First 60	1.37	0.17
Last 50	2.37	0.17

Figure 6.35. Degradation group 2 telemetry flow statistics comparation (*Wega Urban*), proposed experimental session

Dickey Fuller Test	
ADF Statistic	-12.593
p-value	0.000
Critical value (1%)	-3.442
Critical value (5%)	-2.867
Critical value (10%)	-2.569

Figure 6.36. ADF test on time curve

KPSS Test	
KPSS Statistic	0.259
p-value	0.100
Critical value (10%)	0.347
Critical value (5%)	0.463
Critical value (1%)	0.739

Figure 6.37. KPSS test on time curve

In conclusion, despite the negative outcome of the study (no degradation due to the absence of washes was detected), new observations and considerations emerged:

- 527 brews without purge and washes were not sufficient to highlight the degradation of the machine.

- the flow and the time of brew (and consequently also the quantity), were very variable, despite the fact that and *Cdv* were constant. External factors, due to tests carried out on different days, with packages of coffee necessarily different (because of the many brews), have prevailed as an effect over the degradation.
- The 60 final brews show that just one washing of the unit is sufficient to bring the system into a different state (even if different from the initial one). The additional two washes (after and 20 and 40 brews), do not bring any change in flow patterns and brew time pattern.

6.5 Time series analysis

For the first experimental session, the time series related to the single brew are also available. For each coffee, in fact, the flowmeter pulses are sampled every 300 ms: it is therefore possible to draw a curve for each brew, in which each point corresponds to the total of the flowmeter pulses generated up to that instant of time. Moreover, the system already seems to be designed for sampling pressure and temperature, which would bring even more information. The solution that would certainly give more accurate results would be to transfer, for each brew, the entire time series of the flowmeter pulses. The latter, however, would significantly increase the costs of data management and transfer. Since the analysis is still in an exploratory phase and the company wants to proceed in a conservative manner, a feature engineering solution was chosen. The telemetry sensor system, in fact, obtains the brew time and the average flow from the time series: the brew time corresponds to the last time sampled of each brew, while the average flow is calculated starting from time and flowmeter pulses. The study of the curve generated by the flowmeter pulses over time allows to identify a trend point. The coffee brew process is divided in two phases: a first phase with a higher flow and a second phase (after the trend point) with a lower flow. It is assumed that the higher flow corresponds to the first phase of infusion of the coffee panel, while the lower one to the actual brewing. Through the algorithm described in section 5.4, four new features are obtained:

1. Q_{tr} = water quantity trend point (ml)
2. T_{tr} = time at trend point (s)
3. Φ_{slope1} = average flow before trend point (ml/s)
4. Φ_{slope2} = average flow after trend point(ml/s)

In graph 6.38 are represented the real curve, the approximate curve and the 4 features extracted thanks to the detection of the trend point.

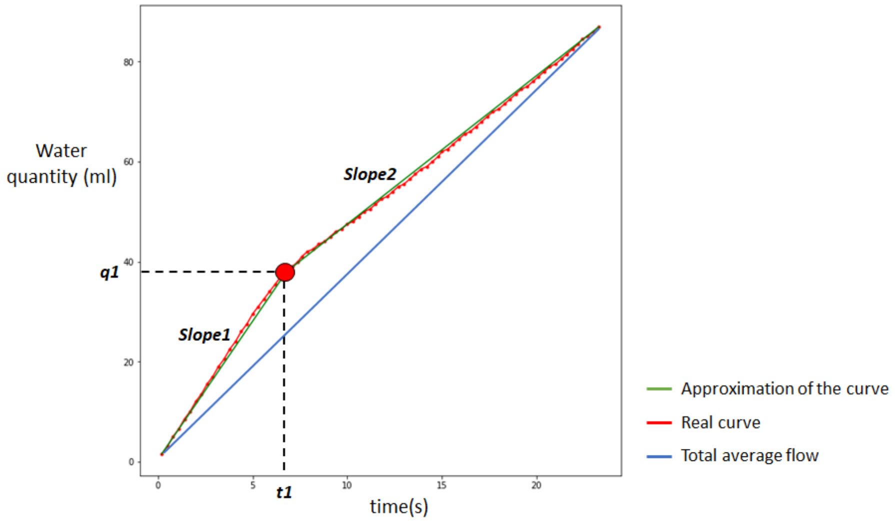

Figure 6.38. Timeseries of telemetry pulses (*Cimbali M100*)

The four features are automatically extracted from the files containing the timeseries and inserted into a single dataset, so that it is possible to perform the preliminary analysis. In order to simplify the study and the graphic visualization, it was decided to consider only the two average flows (Φ_{slope1} and Φ_{slope2}) for the analysis.

In the dispersion graph 6.39, each point represents a brew: the x-cord corresponds to the average flow before the trend point, while the y-coordinate corresponds to the average flow after it.

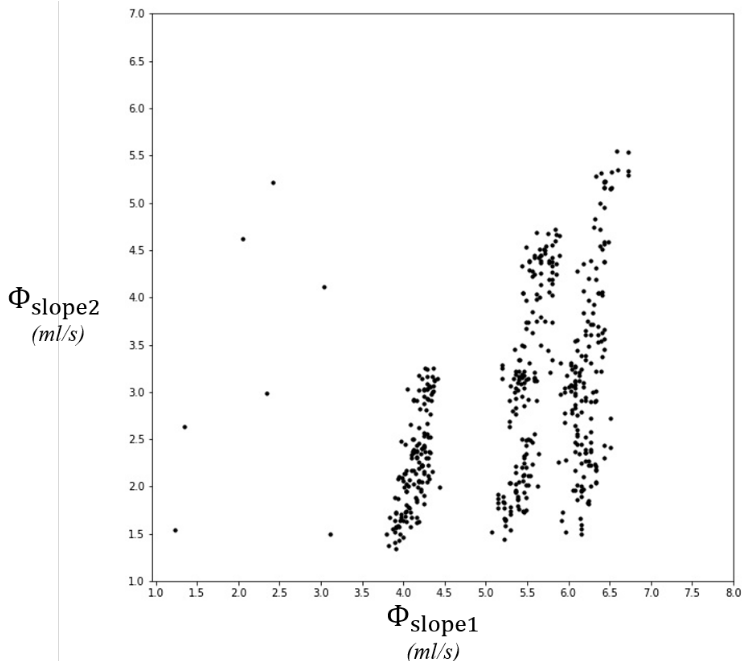

Figure 6.39. Average flows scatter plot (group 1 *Cimbali M100*)

The first part of the analysis was focused on the identification of the link between the three labels (inside the threshold, above the threshold, below the threshold), assigned to external variables (flow in the cup and brew time), and the value of average flows (Φ_{slope1} and Φ_{slope2}). The points of the scatter plot (each represents a brew), were then coloured according to the value assumed by the variable examined (refer to the legends of the graphs). The graph 6.40 shows how the brew time depends only on the average flow after the trend point: if it is too low, the brews exceed the upper threshold of 27 s, while if it is too high, the brews will be too fast (less than 20 s).

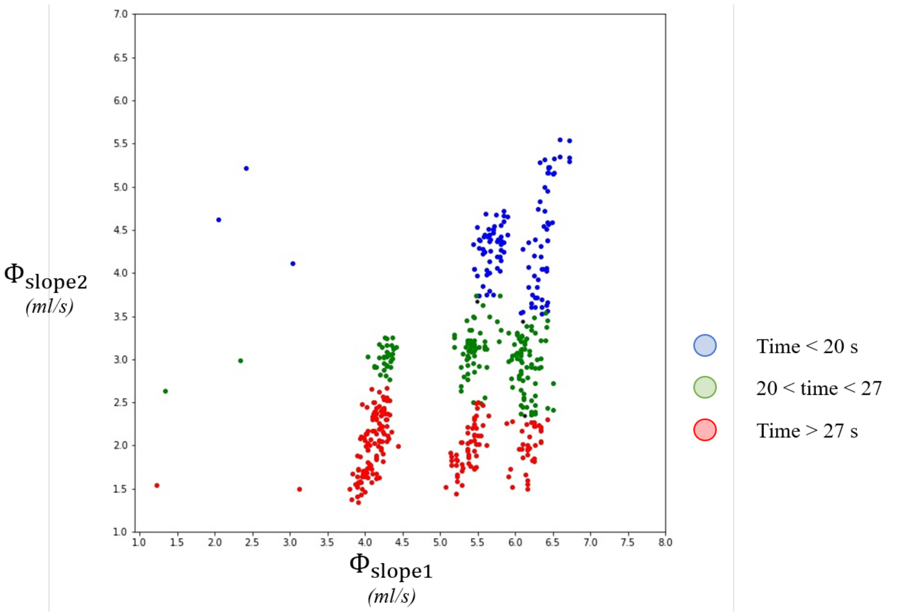

Figure 6.40. Scatter plot with color formatting based on time (group 1, *Cimbali M100*)

The flow in the cup has a dual behaviour with respect to time and specular with respect to the Φ_{slope2}. Both considerations have mathematical basis: the flow in the cup is the consequence of the Φ_{slope2} and is inversely related to the brew time (graph 6.41).

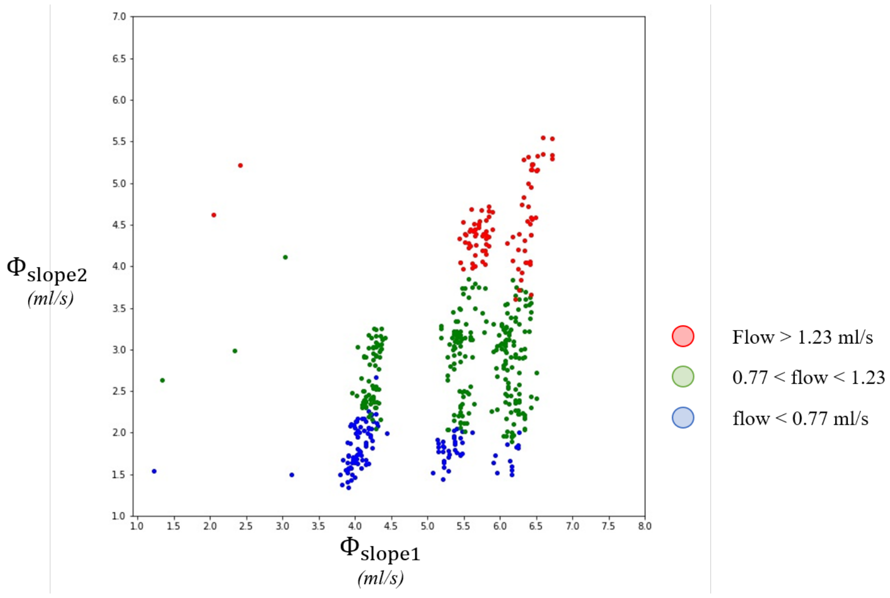

Figure 6.41. Scatter plot with color formatting based on in cup flow (group 1, *Cimbali M100*)

The second part of the analysis was focused on the research of a link between the external parameters and the two average flows (Φ_{slope1} and Φ_{slope2}). In a similar way to the previous study, the distributions, represented by the points in the scatter plots, were coloured with respect to the value of the external parameter under examination. The first variable taken into consideration is the pressure: the graph 6.42 shows how the division of the brews in three macro areas is determined by the pressure. The latter acts by modifying the initial average flow (Φ_{slope1}): by increasing the pressure, the value of the initial average flow increases. The central area represents the optimal pressure brews.

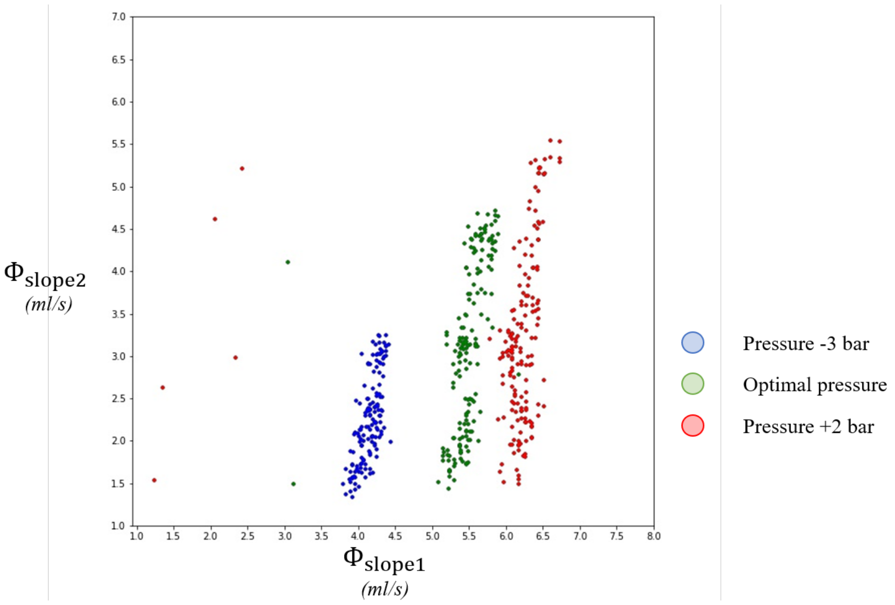

Figure 6.42. Scatter plot with color formatting based on pressure (group 1, *Cimbali M100*)

The analyses that gave the most difficult-to-interpret results were those related to dose and grinding grade. Looking at the grinding graph 6.43, there is no clear division of brews compared to the previous cases. It can be said, however, that brews with finer grind are more associated with a lower (Φ_{slope2}) (positioned at the bottom of the graph). The brews made with a coarser grinding instead are related to a higher (Φ_{slope2}), while those made with an optimal grinding are mostly positioned in the central part. This behaviour was expected: a finer grinding makes the coffee panel less easily penetrable by the water, causing the machine to reduce the flow (and vice versa with the coarser grinding). The brews that instead seem to be in the "wrong" range of Φ_{slope2} , are the result of the compensation of the variables dose and grinding: the brews with grinding +2 and -2 *g* for example, appear in the central area (or near this), close to the brews made with optimal parameters.

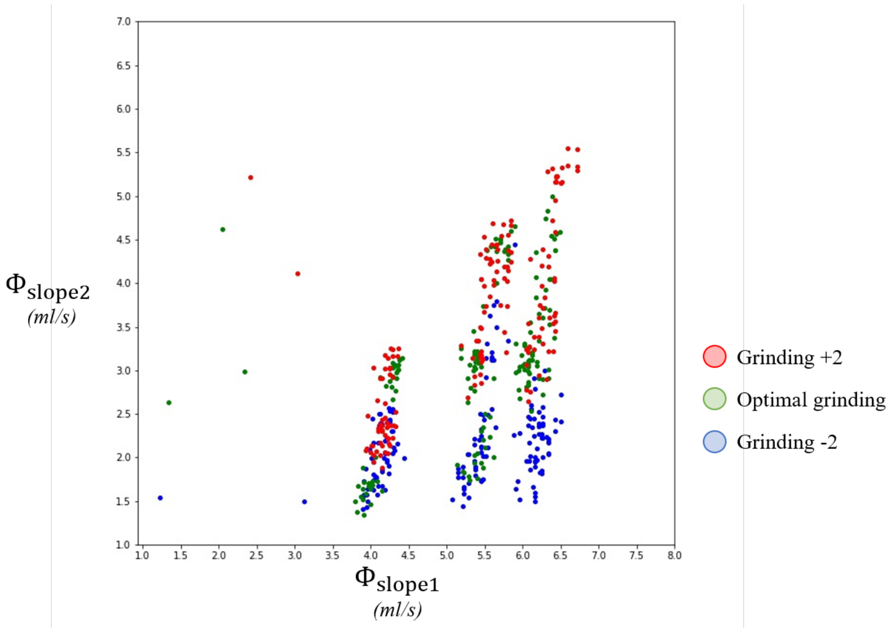

Figure 6.43. Scatter plot with color formatting based on grinding (group 1, *Cimbali M100*)

Analysing the dose graph (6.44), you can see a dual effect with respect to the grinding effect: a higher dose is associated with a lower (Φ_{slope2}), while a lower dose is associated with a higher (Φ_{slope2}). Again, the behaviour was predictable: a higher dose would be a more difficult obstacle to overcome, generating a lower (Φ_{slope2}) and vice versa. Of course, there are no three distinct bands due to compensation.

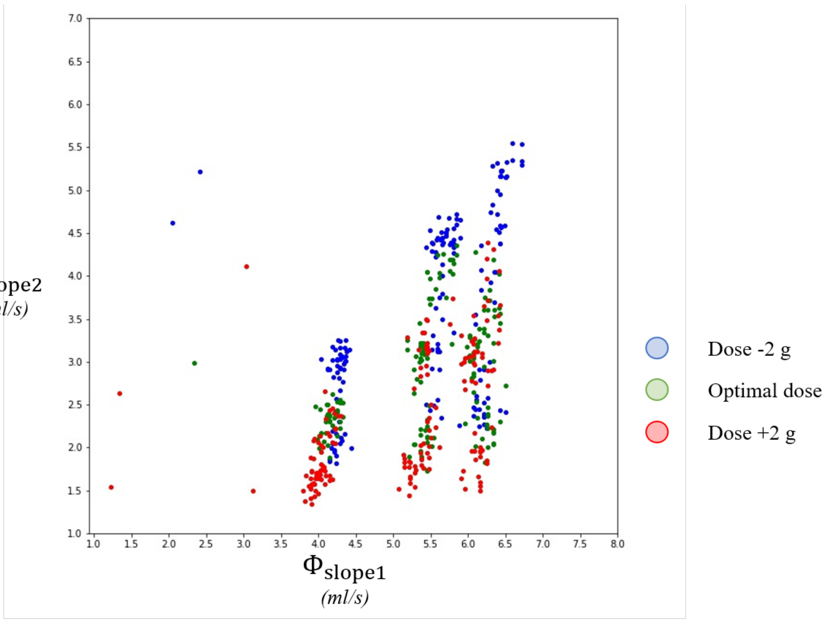

Figure 6.44. Scatter plot with color formatting based on dose (group 1, *Cimbali M100*)

It can therefore be said that the pressure variation can be effectively detected by the values assumed by the average flow before the trend point (Φ_{slope1}). The average flows are probably less effective for the creation of a prediction model because of the compensation phenomena. In the last part of the study, the possible contribution that the insertion of the two variables related to the average flows (Φ_{slope1} and Φ_{slope2}) is analysed. Graph 6.45 represents in green the brews made with all the optimal external parameters: they are in the middle zone with respect to the Φ_{slope1} (optimal pressure) and also with respect to the Φ_{slope2} (optimal dose and grinding), summarizing the previous analyses. In red are represented the brews with not optimal external parameters (only the pressure is optimal, because of less interest), but quality variables (flow and time) inside the threshold.

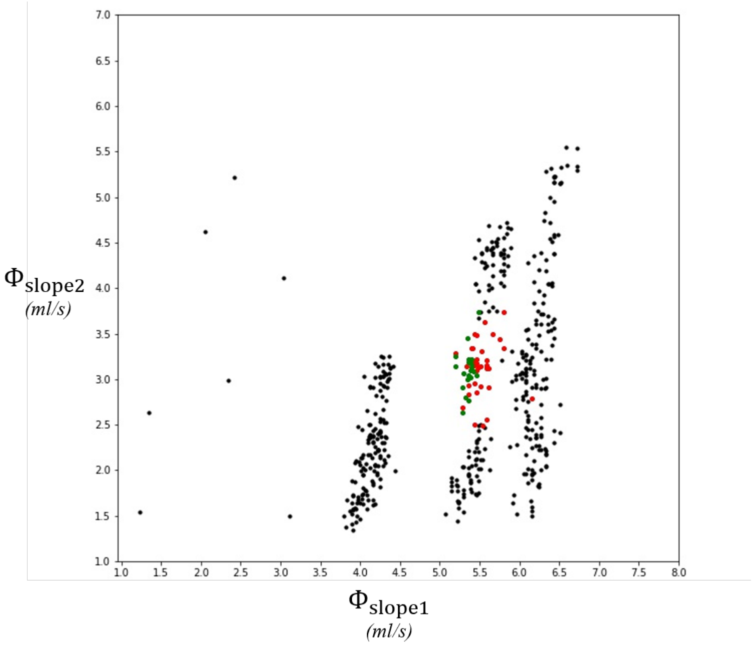

Figure 6.45. Optimal and inside the thresholds brews view (group 1, *Cimbali M100*)

The graphs 6.46 6.47 and the table 6.19 show the minimum and maximum data thresholds for the two brew sets.

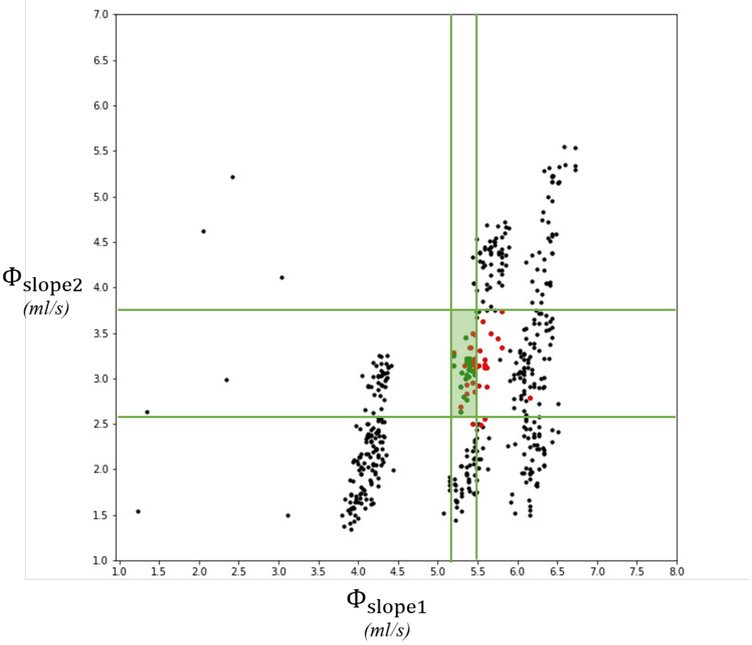

Figure 6.46. Optimal and inside the thresholds brews comparison (group 1, *Cimbali M100*)

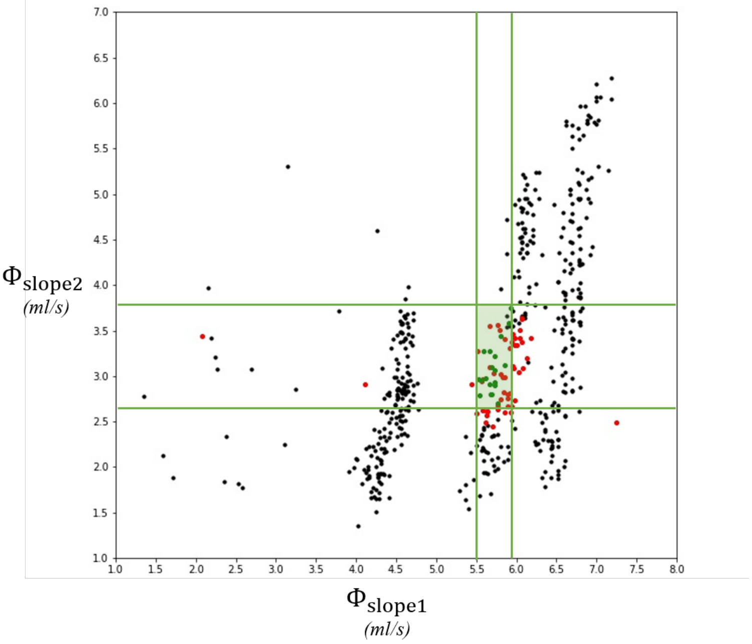

Figure 6.47. Optimal and inside the thresholds brews comparison (group 2, *Cimbali M100*)

	Φ_{slope1}	(ml/s)	Φ_{slope2}	(ml/s)
	min	max	min	max
Group 1	5.19	5.48	2.64	3.73
Group 2	5.53	5.93	2.70	3.75

Table 6.19. Anomalous clients: Bar *RAL* data cleaned

Using the flow thresholds Φ_{slope1} and Φ_{slope2}, it is possible to identify those coffees that have flow and dispensing time inside the thresholds, but not optimal external parameters. The two tables show the results: in group 1 there is about a 51 % improvement (18 out of 35 false optimal coffees are recognized), in group 2 there is a 67 % improvement (35 out of 52 false optimal coffees are recognized).

The advantage of introducing these new variables, obtained through feature engineering, is highlighted by the test 25 performed on group 2 (in the first experimental session): coffee is dispensed at optimal pressure, but with grinding -2 and dose - 2 g. This type of brew could represent those of a customer who tries to save on the dose, acting on the grinding to compensate the effects. Analysing the brews (see table 6.20) through the quality thresholds on flow and time, 65 % of them result inside the threshold: in the remaining 35 %, only one presents both the flow and the time outside the threshold, while the others have only the time slightly below the threshold. The customer would therefore seem to be making coffee of acceptable quality.

Test	ID	Pressure (bar)	Dose (g)	Grinding	Flow (ml/s)	Time (s)	Φ_{slope1} (ml/s)	Φ_{slope2} (ml/s)
25	1021	Optimal	-2	-2	0.86	25.4	7.24	2.49
25	1022	Optimal	-2	-2	1.06	22.4	5.81	3.02
25	1023	Optimal	-2	-2	1.20	19.7	6.03	3.58
25	1024	Optimal	-2	-2	1.15	20.6	6.19	3.41
25	1025	Optimal	-2	-2	1.16	20.4	6.05	3.42
25	1026	Optimal	-2	-2	1.05	22.4	6.08	3.09
25	1027	Optimal	-2	-2	1.02	21.6	6.03	3.04
25	1028	Optimal	-2	-2	1.14	21.0	6.00	3.35
25	1029	Optimal	-2	-2	1.16	20.4	5.98	3.42
25	1030	Optimal	-2	-2	1.13	22.4	6.13	3.20
25	1031	Optimal	-2	-2	1.04	19.5	5.92	3.55
25	1032	Optimal	-2	-2	1.12	21.1	6.07	3.38
25	1033	Optimal	-2	-2	1.04	21.7	5.96	3.11
25	1034	Optimal	-2	-2	1.33	17.4	6.08	4.30
25	1035	Optimal	-2	-2	1.03	22.2	5.97	3.09
25	1036	Optimal	-2	-2	1.16	19.9	6.10	3.64
25	1037	Optimal	-2	-2	1.17	19.7	6.10	3.68
25	1038	Optimal	-2	-2	1.25	18.7	6.05	3.81
25	1039	Optimal	-2	-2	0.88	25.8	5.70	2.45
25	1040	Optimal	-2	-2	0.79	28.0	5.92	2.16

Table 6.20. Anomalous clients: Bar *RAL* data cleaned

However, by analysing the values of the flows Φ_{slope1} and Φ_{slope2} and comparing them with the thresholds, only 2 coffees (10 %) would be within the thresholds. The additional information would then allow to identify customers with apparently "correct" behaviour. With a more in-depth analysis and including the other two variables ($Q_{tr} =$, $T_{tr} =$), the accuracy would probably be even better.

6.6 Analysis of data from the market

After the analyses performed on the laboratory data, it was decided to carry out studies on the data collected by the telemetry device from the machines on the market. The telemetry sensor is able to detect different information for each group: for example total water brewed, pump pressure and boiler temperature. The only variables that describe the brew are average flow and brew time. From these, the total water brewed is also calculated, using formula 6.3. To select only the two variables useful for the analysis, a script has been created. It extracts, for each group, the average flow and time of brew: the matching is done on the basis of the time of detection. Through the operation described in detail in section 5.2.2, the variables of time and flow relative to the single supply and associated to the number of the group to which they belong are aligned. The dataset is the one described in the section 4.5. The information obtained from the telemetry device is therefore reduced compared to the laboratory data and further limited by the lack of homogeneity of the brews made by customers:

- the mixture used is probably different from the *TOP CLASS* used in the laboratory and could also be different between the various brews.
- the status of the machine's calibration is unknown.
- The type of brew (single, double or continuous) and the dose (normal, long, short) are not identified.

The last problem would invalidate the analysis, since it would be impossible to compare different types of coffee: not discriminating between double coffees, it is not possible to halve the flow. For this reason, for the first analysis on trusted customers, it was decided to select periods in which most of the coffee produced would be normal double. The dataset was then cleaned using the cleaning process described in section 5.2.3: in this way, most of the remaining brews should consist of double brews, i.e. those analysed in the laboratory.

For the analysis of anomalous customers, it was selected for all the period of September: the brews are therefore more heterogeneous, however without invalidating the analysis (see section 6.6.2).

6.6.1 Clients and machines modelling

The first customers to be analysed were suggested by the company because they were considered trusted: in fact, they follow all the good practices necessary to obtain an excellent coffee quality (calibration, cleaning). The selected customers are the cafeteria *Nuvola*, based in Turin, and the *Top Shop*, based in Milan.

Both coffee shops are equipped with the Faema E71 machine model: thanks to tests on the variation of the flowmeter pulses, the transfer function between the telemetry flow and the flow measured in the cup is known (see section 6.3.4). Applying it to the thresholds defined in the cup, the telemetry thresholds are obtained. Thanks to these, it was therefore possible to compare data from two different periods, collected by the same machine, and data from different machines (and customers), of the same model.

Table 6.21 summarises the thresholds provided by the experts (*original thresholds*) and those derived from the data of the first experimental session (*proposed thresholds*, see section 6.2), and shows the corresponding telemetry thresholds.

	In cup	Telemetry
time	20 - 27 s	20 - 27 s
original thresholds	0.49 - 1.33 ml/s	0.65 - 2.04 ml/s
proposed thresholds	0.80 - 1.20 ml/s	1.17 - 1.86 ml/s

Table 6.21. In cup and telemetry thresholds

The analysis of each period is done through a series of steps:

- the comparison between the number of telemetry detections and the sales count, in order to verify their correct correspondence and the type of brews made in that period.
- Data cleaning, with the aim of eliminating merged brews and short and long brews, as described in the 5.2.3 section. During the analysis of the water brewed it is also possible to recognize the type of brew made on each group, whether single or double.
- Assignment of quality labels to brews on according to the thresholds indicated in table 6.21, and general analysis of them.
- Computation of the statistics (mean and standard deviation) of the variables brew time, flow and quantity of water dispensed of each group in the selected period.
- Starting from the statistics calculated, quality thresholds referring to telemetry data are obtained: they are calculated in particular for trusted customers (*Nuvola, Top Shop*), to be used as a term of comparison with the brews of other customers.
- Transformation of the thresholds just obtained in telemetry in thresholds referred to the data in the cup, in order to make a comparison with the thresholds provided by experts.

In the table 6.22, the details of each period are listed: remember that the choice of the period was made on based on the percentage of normal double brews.

Cliente	Selected period	Double brews percentage
Nuvola Lavazza	27/06/2019 – 11/07/2019	91%
Nuvola Lavazza	17/05/2019 – 30/05/2019	85%
Top Shop	08/07/2019 – 27/07/2019	88%

Table 6.22. Analysed period

The analysis carried out does not take into account the different mixtures used and probable differences in the grinding of the coffee.

The first data analysed are those of the cafeteria *Nuvola* of Lavazza. It is the most trusted customer: purges are executed after each brew, multiple daily washes are made and calibration is constantly monitored. In addition, higher doses of coffee powder are used to make the coffee more intense. It will be used as a reference for the entire analysis.

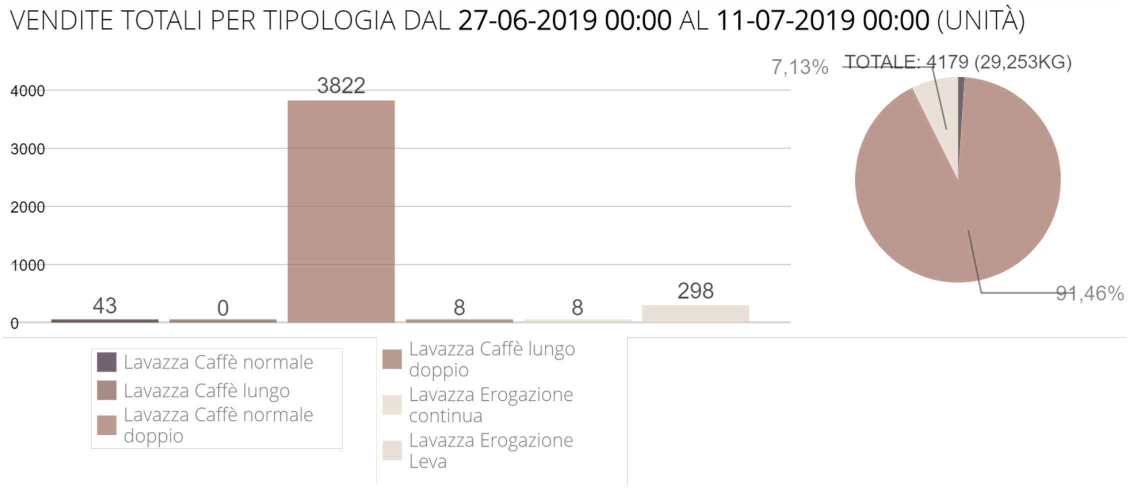

Figure 6.48. Cafeteria *Nuvola* selling data 27/06/2019–11/07/2019

Thanks to the selling data, shown in figure 6.48, it is discovered that, during the period analysed, were made:

- 3830 double coffees, corresponding to 1915 brews, as each double brew corresponds to two coffees sold;
- 349 other coffees (single, continuous and leva);
- **Total number of brews**: 2264.

The telemetry instead detects a total of 1471 brews: number of data do not match, only about 2/3 of the expected brews are detected. The cause of this discrepancy is probably due to the fact that Faema E71 has a telemetry device that processes the data collected, discarding the brews that do not meet certain thresholds (set during the first calibration of the machine). However, the analysis is performed, conscious of the absence of a significant number of brews.

The table 6.49 summarizes the results of the cleaning process, indicating both the thresholds manually chosen to filter the brews, and the number of detections before and after it. In addition, an interpretation of the use of the group is also given.

	quantity thresholds (ml)	total brews	remaining brews	brews guessed typology
Group 1	17 - 30	297	290	double coffee
Group 2	8 - 27	519	498	single coffee
Group 3	18 - 30	659	621	double coffee

Figure 6.49. Cafeteria *Nuvola* data cleaned

Group 2 has been identified as a group brewing mostly single coffees because it is characterized by a lower water quantity than the other groups (approximately halved compared to group 1). In addition, group 2 has a more variable flow rate, and consequently larger cleaning thresholds: it is probably used to make also other type of coffees (with short and long doses).

After having multiplied by two the flow of all the detections of group 2 (to make them comparable with the double brews), the brews are compared to the thresholds indicated in the table 6.21. The table 6.23 shows the analysis performed dividing the data in three different ways: all the brews together, groups 1 and 3 (double coffees), group 2 (single coffees)

Thresholds	General			Groups 1 and 3			Group 2		
	<	OK	>	<	OK	>	<	OK	>
Time	113	882	405	60	643	204	53	239	201
Original flow threshold	44	1348	8	42	865	0	2	483	8
Proposed flow threshold	851	524	25	744	163	0	107	361	25

Table 6.23. Threshold analysis *Nuvola* in the period 27/06/2019–11/07/2019.

It is immediately clear that with the original flow thresholds, the 95% of the brews are inside the thresholds. It could be justified by the fact that the cafeteria

Nuvola produces excellent coffees. But probably the thresholds proposed by the experts are very wide, accepting also coffees with a very slow flow. The proposed flow thresholds, on the other hand, show a large presence of flows below the threshold, particularly in the groups 1 and 3 (double brews), where only 18% of the coffees are inside the thresholds. Analysing the brew time, most of the coffees are inside the thresholds, but around 30% of them have a brew time higher than the threshold: the high presence of coffee with high brew times and very low flows could be due to the fact that the cafeteria uses an higher dose of coffee powder. This is made in order to increase the flavour of the coffee. It should also be remembered that the thresholds proposed refer to coffees made with *TOP CLASS* mixture.

The distribution of data is then analysed using the boxplots of figure 6.50 and calculating average and standard deviation (table 6.24) in order to better understand the trend of the three main variables.

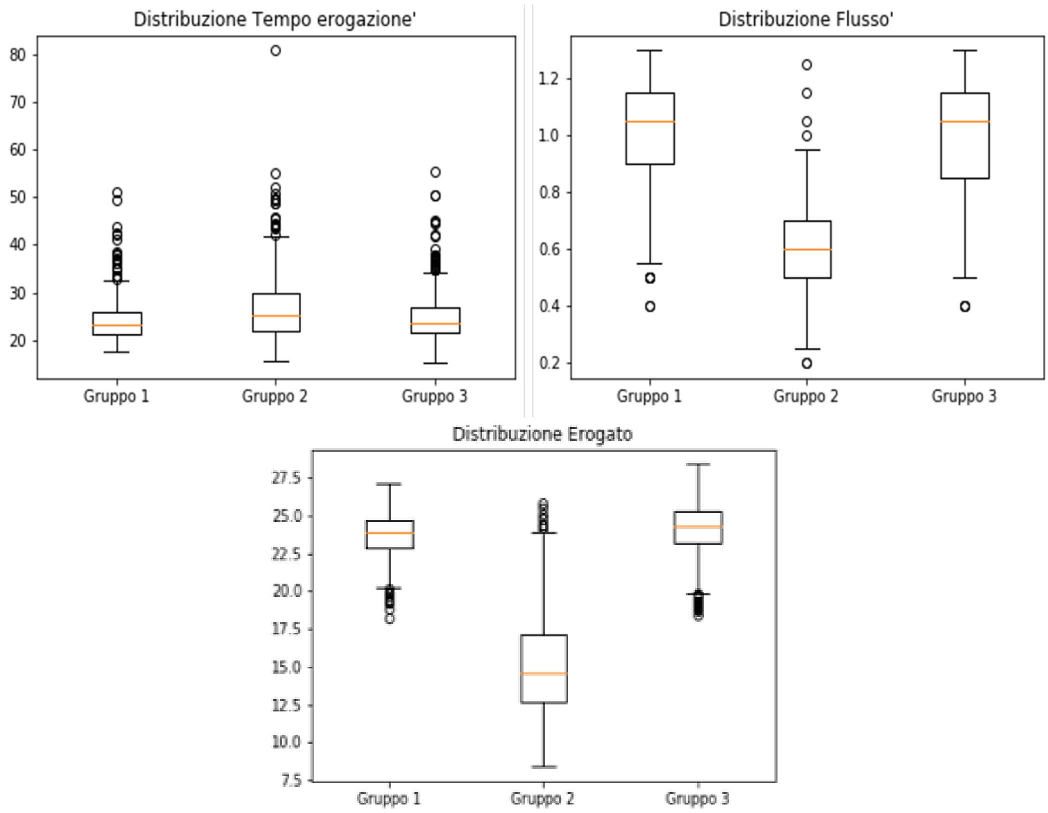

Figure 6.50. Box plots *Nuvola* 27/06/2019 – 11/07/2019

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	1.00	0.20	24.58	4.98	23.56	1.63
Group 2	1.16	0.30	26.83	7.01	30.04	7.20
Group 3	1.00	0.19	24.95	4.81	24.07	1.77

Table 6.24. *Nuvola* telemetry variables statistics

It is clear that group 2, which makes single doses, behaves differently from the other two groups. In fact, the water brewed has a much wider variation, probably representative of the fact that the single group is used to perform free brews. Therefore, the brew times are also slightly longer and more variable. Regarding the flow, the average is slightly greater than the one of the double groups.

Analysing in particular the table 6.24, we can see that the brew time of the double groups perfectly agrees with the rules defined by the experts for calibration: 24 - 25 seconds.

Since the cafeteria *Nuvola* is considered the most trusted customer, the quality thresholds in telemetry are calculated: $mean \pm dev.std.$, using group 1 as a reference. In addition, using the transfer functions for Faema E71 (6.14, 6.23)), the corresponding values in cup are also shown in table 6.25

	flow (ml/s)		time (s)		quantity (ml)	
	min	max	min	max	min	max
Telemetry	0.8	1.2	20	30	22	25
In cup	0.57	0.80	20	30	15	17

Table 6.25. *Nuvola* thresholds in the period 27/06/2019–11/07/2019.

By making a comparison with the thresholds in the cup obtained from the first experimental session, it emerges that, although the thresholds of the brew time are comparable, the flow is much lower (and consequently the amount of coffee). This is certainly due to the process calibration: a consultation with the cafeteria *Nuvola* bartender showed that the dose used was higher (18g for the double dose, instead of 14g), while the amount of coffee dispensed is lower than the average. These two elements guarantee, in fact, an higher quality: if the dose of coffee powder is higher, while the grinding remains unchanged, the obstacle created by by the coffee panel to the water is major. The result is a lower flow, but a more intense coffee flavour. The standard of the *Nuvola* cafeteria is not replicable by customers in the market, but it seemed appropriate to carry out the analyses anyway as well as having a benchmark.

Comparison of different detection periods on the same machine (cafeteria *Nuvola, Faema E71*)

The analysis continues with the comparison of two different periods on the same machine to check if the customer's behaviour varies. The analysis was carried out on the period from 17/05/2019 to 30/05/2019, in which, according to the sales data, they were supplied:

- 3534 double coffees, corresponding to 1767 brews detected by telemetry.
- 702 coffees of different types (single, continuous and leva brews).
- **total number of brews**: 2470.

Also in this case, a great difference was found between the number of coffees sold and the number of brews detected: in fact, there are only 1384 detections, just beyond half of those expected. The table 6.51 contains the results after the data cleaning operation based on the amount of water dispensed. Also in this case, group 2 has an average flow rate that is half of the other two groups, confirming itself as the group that makes the majority of single brews.

	quantity thresholds (*ml*)	total brews	remaining brews	brews guessed typology
Gruppo 1	17 - 30	379	354	double coffee
Gruppo 2	8 - 27	483	467	single coffee
Gruppo 3	18 - 30	522	483	double coffee

Figure 6.51. Cafeteria *Nuvola* data cleaned, period 17/05/2019-30/05/2019

The thresholds used for cleaning the data are the same as for the first period analysed: this is a demonstration of considerable constancy and attention in dispensing and maintaining the machine.

After having multiplied again by two the flow of all the detections of group 2 (to make them comparable with the double brews), the brews are compared to the thresholds indicated in the table 6.21. The results are shown in table 6.26

Thresholds	General			Groups 1 and 3			Group 2		
	<	OK	>	<	OK	>	<	OK	>
Time	104	812	391	55	564	338	69	248	153
Original flow thresholds	7	1292	8	7	831	0	1	461	8
Proposed flow thresholds	839	447	21	724	113	0	115	334	21

Table 6.26. Threshold analysis *Nuvola* in the period 27/06/2019–11/07/2019.

Comparing the results of the period under analysis with the values contained in the table 6.23, it can be observed a practically identical trend: many brews with slow flows and high brew times, due to the high quantity of coffee powder used, while the single group produces the majority of the coffees inside the thresholds. In addition, comparing the statistics calculated for the two periods (shown in tables 6.24 and 6.27), it can be appreciated that the average values of the three characteristics practically coincide, while the standard deviations of the second period are slightly higher.

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	0.98	0.16	24.90	3.5	23.90	1.20
Group 2	1.14	0.26	25.50	5.6	28.20	3.80
Group 3	0.99	0.16	25.50	4.20	24.70	1.60

Table 6.27. *Nuvola* telemetry variables statistics, period 17/05/2019-30/05/2019.

Starting from the computed statistics, it is possible to obtain again quality thresholds on the data coming from the market, always using group 1 as reference.

	flow (ml/s)		time (s)		quantity (ml)	
	min	max	min	max	min	max
Telemetry	0.82	1.14	21	28	23	25
In cup	0.59	0.77	21	28	16	17

Table 6.28. *Nuvola* thresholds in the period 17/05/2019-30/05/2019.

From the comparison of the thresholds obtained from the two periods, contained in the tables 6.25(June) and 6.28(May), it can be concluded that the data obtained from the same machine are consistent.

Through the analysis of these two periods of the Lavazza *Nuvola*, the flow thresholds and the reference brew time for the telemetry data were defined. In particular, the following thresholds were chosen:

- For the flow it was decided to use the less restrictive thresholds calculated with the data for the month of June: 0.8 - 1.2 ml/s.
- For the brew time, on the other hand, the stricter thresholds obtained from the data for the month of May were chosen, as they are more consistent with those of the experts: 21 - 28 seconds.

As stated above, the thresholds calculated from the data of the *Nuvola* cafeteria represent a reference that can hardly be reached by customers in the market.

Since in the future it will be possible to distinguish the various types of brew from telemetry, through a larger and more accurate study on customers, more specific thresholds can be defined: it would be appropriate to differentiate them with respect to the type of brew (long, short) and based on the dose (single, double). At the end of the tests concerning the variation of flowmeter pulses on each machine model, then, the relationship between telemetry data and data measured in the cup should be clear, making the analysis even more accurate.

Comparison of machines of the same model (*Faema E71*) but from different customers

The analysis continues with the comparison between two machines of the same model, but belonging to two different customers. The analysis takes into account the period 08/07/2019 - 27/07/2019 of the customer *Top Shop* in Milan. The machine model is therefore the same as that of the *Nuvola*, with the only difference that it has only two groups instead of three. Moreover, the *Top Shop*, like the *Nuvola*, is also a high quality customer.

The sales data show that during the selected period were sold:

- 1432 double coffees, corresponding to 716 brews detected by telemetry.

- 48 other coffees (single, continuous and leva).
- **Total number of brews**: 764.

Telemetry revealed 597 brews, or approximately 3/4 of those expected. After cleaning the data, we obtain a number of brews as described in the table 6.52.

	quantity thresholds (*ml*)	total brews	remaining brews	brews guessed typology
Group 1	17 - 30	291	258	double coffee
Group 2	21 - 27	306	272	single coffee

Figure 6.52. Cafeteria *Top Shop* data cleaned, period 08/07/2019 – 27/07/2019

It can be observed that in this case, since there are only two groups, there is not a group dedicated only to single brews. The brews are then directly compared with the thresholds indicated in the table 6.21.

Thresholds	General			Groups 1 and 3			Group 2		
	<	OK	>	<	OK	>	<	OK	>
Time	63	417	63	45	186	31	18	231	32
Original flow thresholds	5	538	0	4	258	0	1	280	0
Proposed flow thresholds	472	71	0	236	26	0	236	45	0

Table 6.29. Threshold analysis *Top Shop* in the period 08/07/2019 – 27/07/2019.

Data in table 6.29 show that, although most brews are inside the thresholds in terms of time, most of them have a flow below the thresholds (compared to the proposed thresholds). This could be caused by the use of a higher dose of coffee powder, as is the case with the Lavazza *Nuvola*, as a good quality customer.

The statistics are then calculated for the various characteristics during the period under analysis.

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	1.00	0.15	23.00	3.40	22.00	1.15
Group 2	1.00	0.13	23.60	2.60	23.80	1.20

Table 6.30. *Top Shop* telemetry variables statistics, period 08/07/2019 – 27/07/2019.

From the table 6.30 it can be observed that the behaviour of the two groups is coherent, practically coincident. A comparison with table 6.24 shows that the two periods have very similar statistics: the difference that can be seen is that the *Top Shop* tends to dispense coffee in a shorter average time, thus dispersing even less water (obtained as a flow per time). However, the difference is minimal and can be attributed to a different calibration or to the use of different mixtures.

It can therefore be concluded that the behaviour of the two machines of the same model is similar: in particular, this is possible because the two customers analysed have similar characteristics and it is certain that they make good quality coffees.

6.6.2 Anomalous clients analysis

In this section it is analysed whether it is possible to find out which customers are using less coffee powder than the recommended one (7 g for single brew, 14 g for double brew). The clients to be analysed were selected by comparing sell in and sell out data: calculating $\Delta_{sell_{out}-sell_{in}} = sell_{out} - sell_{in}$, if $\Delta_{sell_{out}-sell_{in}} > 0$, the customer sold less coffee than he bought from Lavazza. The only possible reasons are the following two:

1. The customer buys the coffee from a different supplier than Lavazza.
2. The customer inserts less dose into the filter holders to save on the coffee powder.

For the analysis, the $\Delta_{sell_{out}-sell_{in}}$ of August 2019 was calculated: two "anomalous" customers were selected having two different machines (*Rancilio CL11* and *WegaMyConcept*) and each one was compared with a "regular" customer having the same type of machine. The selected customers are:

- Ristorante *il pepe* (*Rancilio*, "anomalous")
- Bar *RAL* (*Rancilio*, "regular")
- Ristorante *GRUPPO SPES SCS ETIKO* (*WegaMyConcept*, "anomalous")
- Ristorante *VITESSE SAS* (*WegaMyConcept*, "regular")

The period selected for comparison is September 2019: the month is immediately following the $\Delta_{sell_{out}-sell_{in}}$ information available and the telemetry data is more accurate, as at the end of August the online platform firmware was updated. In addition, the lack of data about the key pressed to make the brew requires the cleaning process described in the paragraph 5.2.3. The goal is to keep only the single and double brews brewed most frequently. The first two customers with the *Rancilio* machine are then compared: the Ristorante *il pepe* and the *RAL* Bar. The table 6.53 summarises the results of the cleaning process: by comparing the number and quantity of water dispensed from the remaining coffees with the sales data (figure 6.54), it seems likely that group 1 was used for long double brews, while group 2 for single long ones.

	quantity thresholds (ml)	total brews	remaining brews	brews guessed typology
Group 1	38.5 - 41	867	542	double long coffee
Group 2	22 - 24	1700	1261	single long coffee

Figure 6.53. Anomalous clients: Bar *RAL* data cleaned

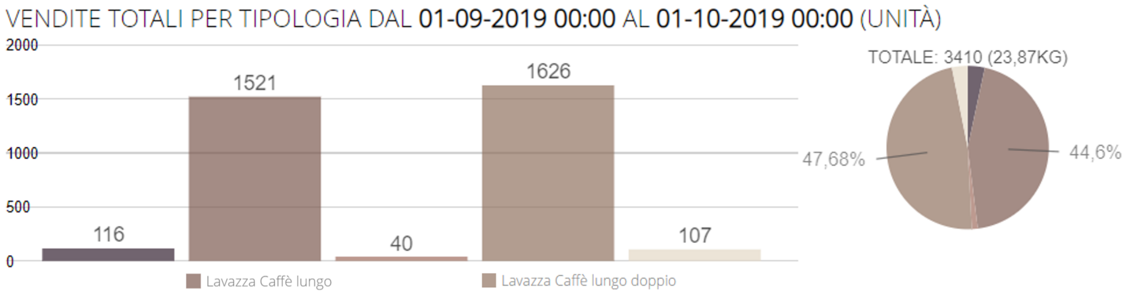

Figure 6.54. Anomalous clients: Bar *RAL* selling data

It is important to highlight the following considerations (already explained in the chapter 5.2.3), which are valid for all the analyses carried out on the machines on the market:

- the quantity thresholds are chosen visually on the distribution of the quantity of water supplied by telemetry.
- the sales data count the number of coffees produced, while the telemetry data count the number of brews made. The 542 group 1 coffee brews should therefore correspond to double brews since they have high quantity thresholds (38.5-42 ml), amounting to 1084 coffees. The 1261 group 2 coffee brews should

correspond to single coffees (quantity thresholds 22-24 *ml*). Comparing the numbers with the sales data, the assumptions are confirmed.

- The average flow rate for double coffees should be halved (thus halving the quantity), so that it can be compared with the quality thresholds (defined in the chapter 6.2). Since double brews are performed more frequently, the flows of all telemetry groups have been halved. In this way, the double coffee groups already have the correct data. The flows of the groups with the majority of single brews, on the other hand, are brought back to their original value before comparison.

The telemetry data of the Restaurant *il pepe* underwent the same cleaning process and subsequent comparison with the sales data: the results are shown in the table 6.31.

	quantity thresholds (*ml*)	total brews	remaining brews	brews guessed tipology
Group 1	33 - 35	2761	258	double long coffee
Group 2	21 - 23	3217	402	single long coffee
Group 3	33 - 35	3578	352	double long coffee

Table 6.31. Anomalous clients: Bar *RAL* data cleaned

The two machines are then compared and the data is divided by group. The boxplots in the figure 6.55, 6.56 and 6.57 show the characteristics of time distributions, average flows and quantities of water brewed. In the table 6.32 and 6.33 the statistical data of the two customers are reported, relating to the 3 telemetry variables and divided by group. The average brew time of the Restaurant *il pepe* is lower than that of the Bar *RAL* and also compared to the thresholds telemetry from the *Nuvola* analysis (21 - 27s, see chapters 6.2 and 6.6.1). Analysing the average telemetry flows of the 3 groups of the Restaurant *il pepe*, they are all very high and above the threshold of the *Nuvola*: low brew times, associated with high average flows and a $\Delta_{sell_{out}-sell_{in}} > 0$ suggest that the "anomalous" customer is probably introducing fewer doses of coffee powder to make the brews.

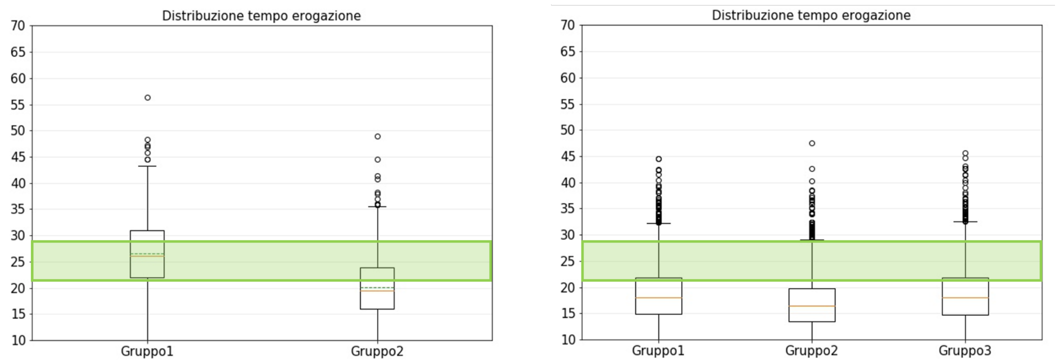

Figure 6.55. Brewing time of Bar *RAL (left)* vs Ristorante *il pepe (right)*

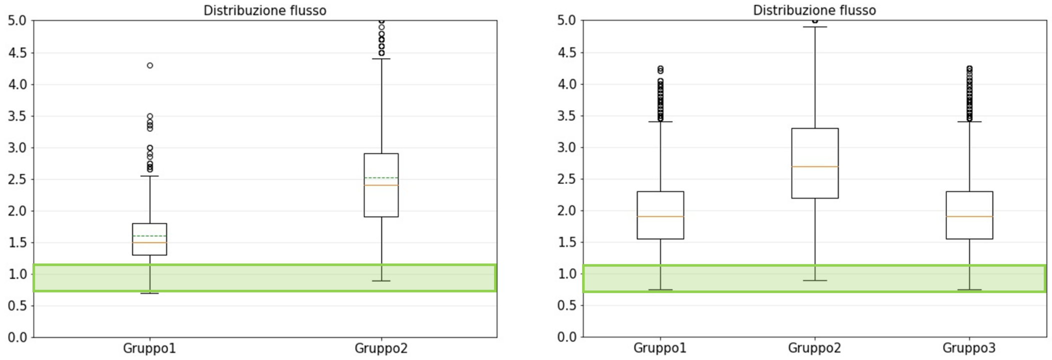

Figure 6.56. Brewing average flow of Bar *RAL (left)* vs Ristorante *il pepe (right)*

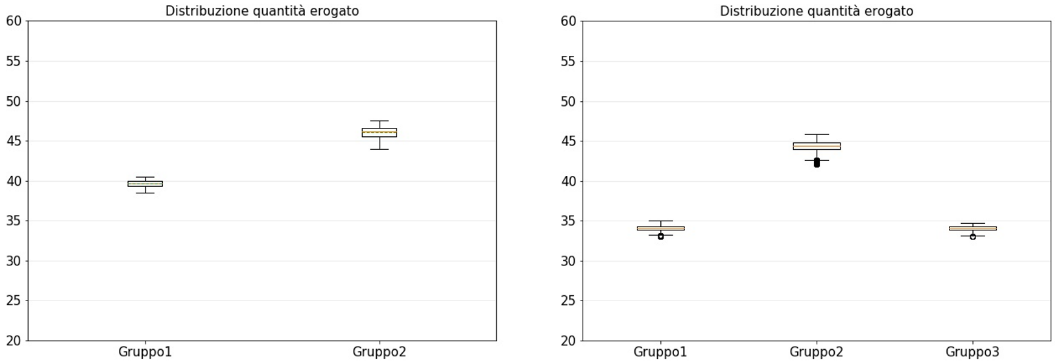

Figure 6.57. Brewing water quantity of Bar *RAL (left)* vs Ristorante *il pepe (right)*

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	1.60	0.46	26.62	6.26	39.66	0.44
Group 2	2.52	0.85	20.14	6.05	46.02	0.77

Table 6.32. Bar *RAL* telemetry variables statistics

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	1.99	0.60	18.70	5.45	34.10	0.30
Group 2	2.80	0.85	17.00	5.00	44.30	0.60
Group 3	1.98	0.60	18.70	5.50	34.10	0.30

Table 6.33. Ristorante *il pepe* telemetry variables statistics

Group 1 of the bar *RAL* is more in line with expectations: the average brew time is within the thresholds, while the average flow is slightly higher.

The next step was to study and compare the two customers with the *Wega MyConcept* machine: the *VITESSE SAS* restaurant, identified as the reference customer, and the *GRUPPO SPES SCS ETIKO* restaurant, identified as a customer with possible anomalous behaviour. After cleaning the data, groups were identified with a majority of single brews and those with a majority of double brews. The cleaning results are shown in the tables 6.34 and 6.35.

	quantity thresholds (ml)	total brews	remaining brews	brews guessed typology
Group 1	28 - 33	2761	207	double coffee
Group 2	28 - 33	3217	1341	double coffee
Group 3	12 - 17	3578	1611	single coffee

Table 6.34. Anomalous clients: Ristorante *VITESSE SAS* data cleaned

	quantity thresholds (ml)	total brews	remaining brews	brews guessed typology
Group 1	11 - 15	3070	1496	single coffee
Group 2	24 - 29	2302	1533	double coffee
Group 3	23 - 28	1672	1159	double coffee

Table 6.35. Anomalous clients: Ristorante *GRUPPO SPES SCS ETIKO* data cleaned

The two machines are then compared and the data is subdivided by group. Again, the boxplots (in the figure 6.58, 6.59 and 6.57) and the tables 6.32 and 6.33 show the statistical data of the two customers, relating to the 3 telemetry variables and divided by group. The average brew time of the Restaurant *VITESSE SAS* is lower than that of the Restaurant *GRUPPO SPES SCS ETIKO* and very low compared to the telemetry thresholds (21 - 27 s, chapters 6.2 and 6.6.1). Analysing the average telemetry flows of the 3 groups of the *VITESSE SAS* Restaurant, they are all very high and above the threshold of the *Nuvola*. The flows and times of the Restaurant *GRUPPO SPES SCS ETIKO* are in line with the thresholds obtained from the study on the machine of the *Nuvola*: the "anomalous" customer therefore seems to produce good quality coffees, while the "correct" customer probably has an incorrect calibration of the machine.

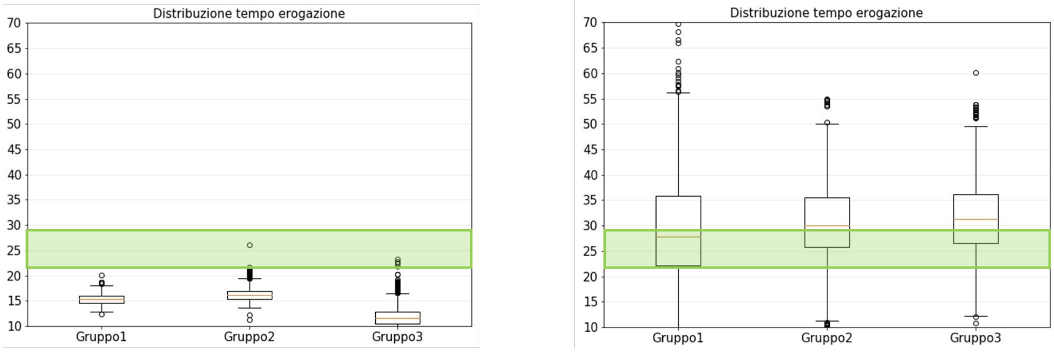

Figure 6.58. Brewing time of Ristorante *VITESSE SAS (left)* vs Ristorante *GRUPPO SPES SCS ETIKO (right)*

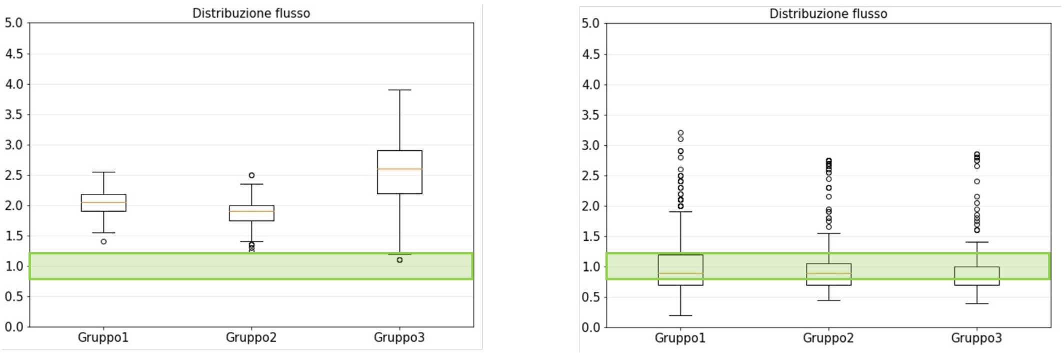

Figure 6.59. Brewing average flow of Ristorante *VITESSE SAS (left)* vs Ristorante *GRUPPO SPES SCS ETIKO (right)*

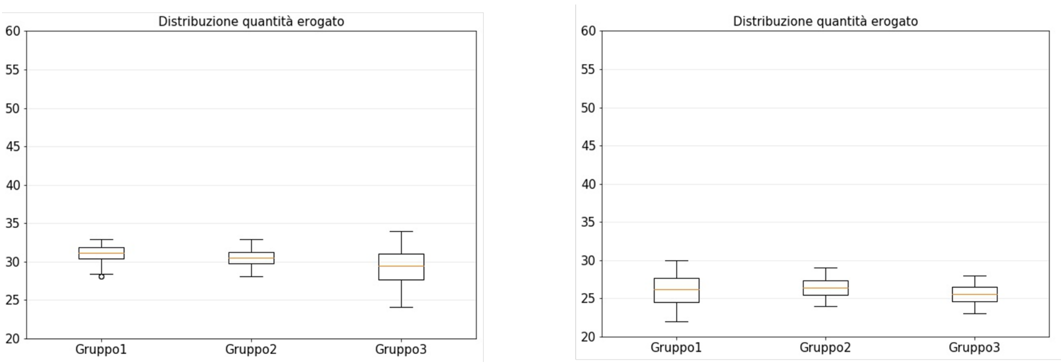

Figure 6.60. Brewing water quantity of Ristorante *VITESSE SAS (left)* vs Ristorante *GRUPPO SPES SCS ETIKO (right)*

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	2.03	0.20	15.42	1.16	31.04	1.07
Group 2	1.87	0.19	16.32	1.34	30.48	1.09
Group 3	2.54	0.53	11.90	1.95	29.33	2.30

Table 6.36. Ristorante *VITESSE SAS* telemetry variables statistics

	flow (ml/s)		time (s)		quantity (ml)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
Group 1	1.01	0.40	29.39	10.33	26.17	2.04
Group 2	0.92	0.31	30.88	7.17	26.44	1.24
Group 3	0.86	0.28	31.79	7.25	25.57	1.23

Table 6.37. Ristorante *GRUPPO SPES SCS ETIKO* telemetry variables statistics

In conclusion, there are two possibilities among customers who insert less dose:

1. calibration was carried out with the correct dose of coffee powder.

2. calibration was carried out with a lower dose.

In the first case, if the customer decides to brew coffee with a lower dose, the expected effects should be an increase in flow and a shorter brewing time. If these conditions occur with a customer with $\Delta_{sell_{out}-sell_{in}}$ >0, that customer is probably using a smaller dose to brew the coffee.

In the second case, if the grinding is "adapted" (finer yield) to the lower dose during the calibration of the machine, the variables time, flow and amount of telemetry are probably in line with those of a coffee brewed in optimal conditions. Even though the customer has a $\Delta_{sell_{out}-sell_{in}} > 0$, there is no evidence of using a lower dose.

Chapter 7

Conclusions

This thesis presents an exploratory analysis of data concerning the characteristics of coffee. In particular, the aim was to study the correlation of data collected in the cup with data obtained thanks to a telemetry device installed inside professional coffee machines: knowing the relationship between the two, the quality thresholds defined for in cup variables, can be applied to telemetry data. To achieve this objective, laboratory tests have been carried out, in order to collect data from different sources (direct measurements, electronic card and telemetry) and try to find their relationship: the results show that in cup measurements can always be obtained through PC card data, but the transfer function of the quantity depends only on the coffee machine model, while the flow transfer function depends on the machine and on the number of flowmeter pulses (derived from the calibration process). The flowmeter data are derived from the Pc card data: if the coffee machine manufacturer decided to not elaborate the average flow data, telemetry data are equivalent to Pc card data. In the other case, the transfer function from telemetry quantity to in cup coffee quantity is not accurate.

After achieving the first results, an experimental session structure was proposed: the aim was to modelling different behaviours (lack of washes, anomalous washes, *Cdv* - coffee quantity relation, influence of external parameters and intrinsic difference between groups). One of the test already conducted in laboratory was the analysis of the degradation due to the lack of washes: the aim was to study if the telemetry variables presented a trend, attributable to a poor maintenance of the coffee machine. The results shows that after 500 brews (standard number of dispensing in a working day) is not appreciable any trend: the external factors, due to tests carried out on different days, with packages of coffee necessarily different (because of the many brews), have prevailed over the degradation. The test must therefore be repeated, simulating better a classical working day or it must be performed on a machine on the market.

Finally the analysis was extended to on the exploration of data from the market. First of all an analyses on the behaviour of some trusted customers was conducted, with the aim of finding reference thresholds for telemetry data. After that, two "anomalous" clients (having two different machine models) were selected: according to selling data, in fact, they buy less coffee than the one necessary to make all the brews (monitored by the telemetry sensor). The telemetry data of each "anomalous" client were compared with those of a "correct" client (with the same machine model): a lower brew time in conjunction with an higher flow in the "anomalous" client data, indicates that probably he is inserting less dose inside the filter holder. If the telemetry variables are standard, then probably the client compensates the lower dose with a finer grinding.

The last part of the analysis conducted focused on the study of the time series: through a process of feature engineering, four new variables, related to the single coffee brewing, have been designed. The preliminary analysis conducted considers only two of the four variables extracted: the average flow before the trend point and the one after the trend point. In fact, the trend of the brewed water time series is composed by two straight line separated by a trend point. From the first analyses conducted, the addition of this information, could significantly improve the accuracy of coffee quality detection.

7.1 Analysis limits

The analysis carried out presented some technical difficulties. First of all, the long time it takes to carry out the experiments in the laboratory because of the large number of disbursements to be made. In addition, the alignment of data from different sources took a long time: at the beginning it was done by hand, later it was partly speeded creating some scripts. Finally, some analyses are limited due to the possibility of selecting only limited periods of time from the telemetry web portal. Moreover, there is a lack of the information on the type of brew made. In fact, it is not possible to distinguish between short, long, single, double and continuous brews. We have tried to propose a cleaning method based on the distribution of the quantity of water brewed, which makes the analyses more reliable. In the future this data will be available.

7.2 Possible future improvements

The studies carried out are very promising. In particular, information on the type of brews linked to the detection will significantly increase the accuracy of the analyses. In addition, the ability to retrieve telemetry data over longer periods of time (through direct access to the database), will allow to monitor customers on the long term and to calculate in a more accurate way. In addition, with the monitoring of long periods, the theme of predictive maintenance can be deepened, which allows to predict when an intervention is necessary by observing anomalous variations in the data. The tests of the proposed experimental sessions will provide a lot of useful information on the behaviour of the machine and on how different machines react differently. In particular, it will be possible to analyse the influence of less extreme variations of external parameters on coffee quality. Moreover, the time series analysis can be deepened: in fact, the four new variables would only marginally increase the amount of data sent by the sensor. But the introduction of those new parameter will certainly improve the results of the previous analysis.

Bibliography

[1] Susana Andueza et al. «Influence of coffee/water ratio on the final quality of espresso coffee». In: *Journal of the Science of Food and Agriculture* 87.4 (2007), pp. 586–592.

[2] Rajat Vaidya Ashish Kumar Singh Aditya Sinha and Hrishikesh Vijay Kulkarni. *A Review Paper on IoT Based Coffee Vending Machine*. 2007.

[3] *Come regolare la macinatura del caffè per un espresso perfetto*. `https://www.ilcaffeespressoitaliano.com/2018/come-regolare-la-macinatura-del-caffe-per-un-espresso-perfetto/`. [Online].

[4] W.A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics. Wiley, 2009. URL: `https://books.google.it/books?id=tI6j47m4tVwC`.

[5] *Funzionamento della macchina espresso*. Italian. `https://www.carioka.it/funzionamento-della-macchina-espresso/`. [Online].

[6] L.A. Kirkpatrick and B.C. Feeney. *A Simple Guide to IBM SPSS: For Version 20.0*. Cengage Learning, 2012. URL: `https://books.google.it/books?id=HR\_1CAAAQBAJ`.

[7] Benjamin Koke et al. «Sensor retrofit for a coffee machine as condition monitoring and predictive maintenance use case». In: Feb. 2019.

[8] *matplotlib*. `https://matplotlib.org/`. [Online].

[9] L. Navarini and D. Rivetti. «Water quality for Espresso coffee». In: *Food Chemistry* 122.2 (2010). 5th Conference on Water in Food, pp. 424 –428.

[10] *Python Data Analysis Library*. `https://pandas.pydata.org/`. [Online].

[11] C. Severini et al. «Changes in the Aromatic Profile of Espresso Coffee as a Function of the Grinding Grade and Extraction Time: A Study by the Electronic Nose System». In: *Journal of Agricultural and Food Chemistry* 63.8 (2015), pp. 2321–2327.

[12] Ewa M. Syczewska. *Empirical power of the Kwiatkowski-Phillips-Schmidt-Shin test*. Working Papers. Department of Applied Econometrics, Warsaw School of Economics, 2010. URL: `https://ideas.repec.org/p/wse/wpaper/45.html`.

[13] Wikipedia contributors. *Project Jupyter — Wikipedia, The Free Encyclopedia.* `https://en.wikipedia.org/wiki/Project_Jupyter`. [Online].

[14] Wikipedia contributors. *scikit-learn — Wikipedia, The Free Encyclopedia.* `https://en.wikipedia.org/wiki/Scikit-learn`. [Online].