Statistical Quality Control with Functional Data. An application to Energy Efficiency

Salvador Naya*
Universidad de A Coruña, Ferrol, Spain – salva@udc.es

Javier Tarrío-Saavedra
Universidad de A Coruña, Ferrol, Spain – jtarrio@udc.es

Sonia Zaragoza
Universidad de A Coruña, Ferrol, Spain – szaragoza@udc.es

Miguel Alfonso Flores Sánchez
Universidad de Las Américas, Quito, Ecuador – mflores72000@gmail.com

Manuel Oviedo
Universidade de Santiago de Compostela and ITMATI, Santiago, España – manuel.oviedo@usc.es

Abstract

Functional, Complex or Big Data are a popular terms that is used to describe the large, diverse, or longitudinal datasets generated from a variety of instruments or sensors. This term refers not only to the size or volume of data, but also to the variety and the velocity or speed of data accrual. These new data pose new opportunities for researchers in statistical quality control and for innovative solutions in industry. In this work, we discuss several functional or big data application in statistical quality control. Our goal is to bring the research for complex and big data analysis in energy efficiency. Some suggestions to apply to data quality control in energy efficiency issues will be presented. In this type of control is common to have a large number of sensor data obtained every 5 minutes.

The air quality, thermal comfort and energy efficiency of buildings heating, ventilation, and air conditioning (HVAC) is checked and controlled. Since there are many variables critical to quality (temperatures, humidity, CO2 concentrations), multivariate control charts were applied to latent variables obtained by partial least squares (PLS) to solve this problem. In addition, new control charts in the framework of functional data analysis, are proposed to control functional variables such as the diary temperature (measured each 5 minutes or each hour) in a room.

Keywords: quality control; complex data; energy efficiency; data mining.

1. Introduction

The data produced from sensor measuring the thermal comfort of buildings represent an interesting example of what might be called complex data. To put these data every 5 minutes several variables such as temperature, CO2, and relative humidity (HR) is collected and facing the growing need to make sense of complex or big data. This notion is not new idea in statistical quality control, for example, W.E. Deming said “Information, no matter how complete and speedy, is not knowledge. Knowledge has temporal spread. Knowledge comes from theory. Without theory, there is no way to use the information that comes to us on the instant."

In this article we give an application of big data in statistical quality control can aid in making sense of complex or big data. We begin in Section 2 by trying to answer the question “What is Complex and Big Data?". In Section 3 we discuss the application in energy efficiency quality control. We provide concluding remarks in conclusion.

2. What is Functional and Big Data?

Functional Data and Big Data is a popular term that is used to describe the large, diverse, complex or longitudinal datasets generated from a variety of instruments, sensors. These challenges include the size, or volume, as well as the variety and velocity of the. Known as the 3V's, the volume, variety, and velocity of the data are the three main characteristics that distinguish big data from the data we have had in the past (Zikopoulos et al., 2013).

In the other hand, the term data mining generally refers to the process of mining through large data sets to search for useful information. Data mining is considered a field that merges computer science, artificial intelligence, and statistics. There is therefore a strong relationship between these emerging terms.

A major challenge of complex and big data analysis is how to automatically process and translate such data into new techniques. For example, in the paper of Jones-Farmer et al. (2013) gave an overview of data quality and discussed opportunities for research in statistical methods for evaluating and monitoring the quality of data.

A particular case of this type of data are functional data (FDA). Functional data originated within the field of chemometrics in the 1960s. The topic has received a great of attention from the statistics community ever since, because it covers a wide range of important statistical topics (Ferraty and Romain, 2011). The reader is referred to Ramsay and Silverman (2002, 2005) for an in-depth coverage of the topics involved in functional data analysis. Ferraty and Romain (2011) provided an excellent discussion in the fundamental mathematical aspects that are related to the big data aspect of FDA.

We view the use of images for process monitoring as a very promising area of statistical quality control with a wide range of applications. This is somewhat different from traditional Statistical Quality Control applications that focus on dimensional and discrete process data.

These challenges are important research areas to consider, especially requires Phase I in Qualtity analyses where the baseline is established and the parameters are estimated. For example, there is no discussion on the effect of estimation error on image-based control charts, and therefore it is not known what Phase I sample sizes are needed and whether they would be practical.

3. Functional Data and Big Data in Statistical Quality Control

In big data sets, the decomposition methods used with multivariate control charts can become very computationally expensive. Several authors have considered variable selection methods combined with control charts to quickly detect process changes in a variety of practical scenarios including fault detection, multistage processes and profile monitoring.

The least absolute shrinkage and selection operator (LASSO) test statistic of Tibshirani (1996) combined with a multivariate exponentially weighted moving average (EWMA) control chart to quickly identify process changes and to identify the shifted mean components. Capizzi and Masarotto (2011) developed a multivariate (EWMA) control chart using the Least Angle Regression (LAR) algorithm to detect changes in both the mean and variance of a high dimensional process. All of these methods based on variable selection techniques are based on the idea of monitoring subsets of potentially faulty variables.

4. Application in Energy Efficiency Data

4.1. Data collection

Real data in an office building corresponding to temperature (ºC), HR (%), and CO_2 concentration (ppm) in indoor air were obtained each 5 minutes and each 1 hour. In addition, the energy consumption of the heating, ventilation, and air conditioning (HVAC) installations is also measured to quantify de eficience energy in this systems. Overall, there are 10 variables critical to HVAC installations. Figure 1 shows the studied office and the sensor locations.

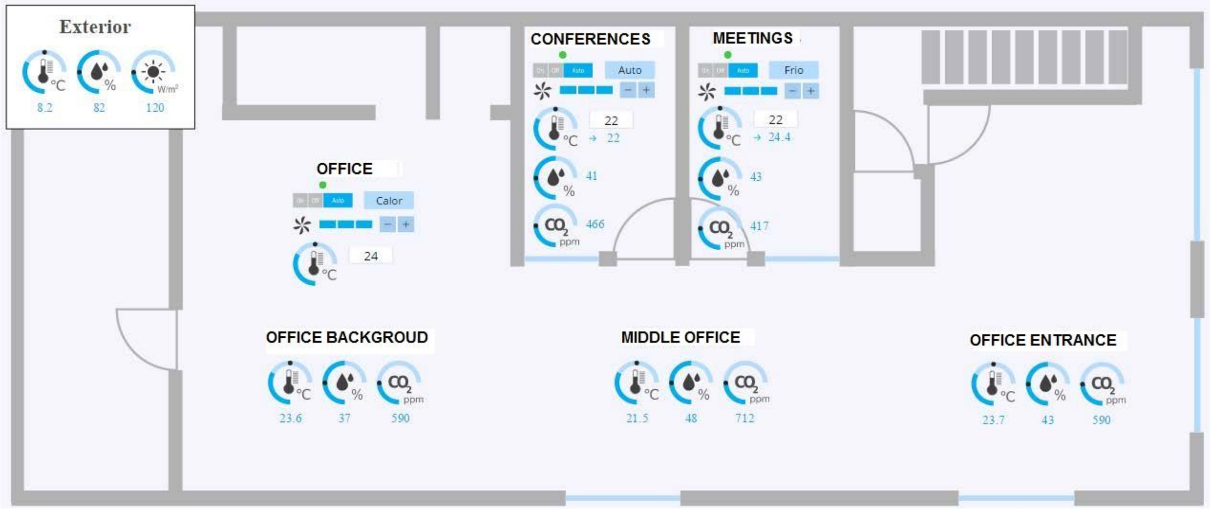


Figure 1. Squeme of the office where the data were obtained and sensor location.

Three different temperatures, 3 variables of HR and another 3 related to CO_2 content are measured, 9 sensors overall. Each variable is measured by one sensor. There are overall 3 sensors located in the office background, entrance and in the middle of the office. In addition, the HVAC energy consumption is measured (kW).

4.1. Results

4.1.1. Multivariate study

Since there are more than one variable critical to HVAC quality, this should be checked and controlled by the application of multivariate control charts such as T^2 Hotelling, MEWMA or MCUSUM (Montgomery, 2013). In the case of complex datasets where each individual is defined by a relatively high number of variables (such as the present case), the average run-length performance to detect a defined shift in the mean of these variables applying multivariate control charts increases, due to the shift with respect to the mean loses importance in the p-dimensional space of the process variables (Montgomery, 2013). Thus, we propose applying T^2 multivariate control charts to latent variables obtained by partial least squares (PLS) to solve this problem (Montgomery, 2013). Taking into account that the latent variables are sorted in descending order of explained variability of data, we can reduce the number of studied variables from 10 to 2 without losing relevant information.

It is important to note that there are at least two possible data dimension reduction alternatives: principal component analysis (PCA) and PLS. The goal of PLS is to deal with the relationship between X variables or process parameters and one process feature Y. This technique creates a set of weighted averages of the X variables and Y in order to predict the y values taking into account the x ones. In fact, in the present case the Y feature critical to energy efficiency quality is the energy consumption and the process parameters or inputs are the temperatures, HR and CO_2 content. This makes PLS an adequate technique to deal with energy efficiency data. More information about PLS technique can be obtained in Montgomery (2013) and Frank and Friedman (1993).

Figure 2 shows the T^2 control chart applied to the first two PLS components. The calibration sample corresponds to 10 first days of July while the monitored sample is composed by the remaining 20 days of the same month.

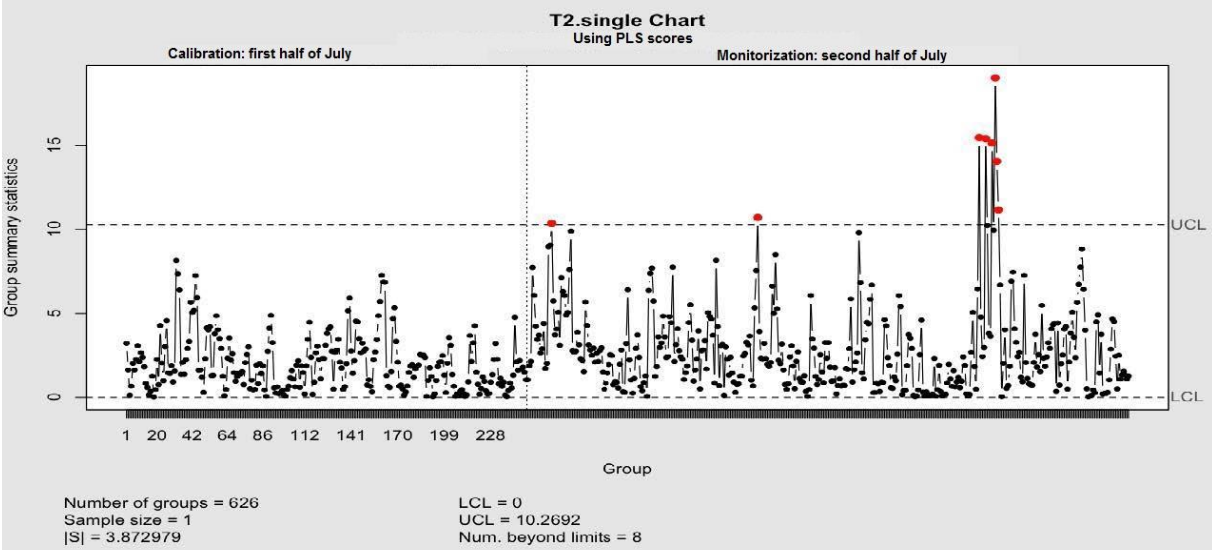
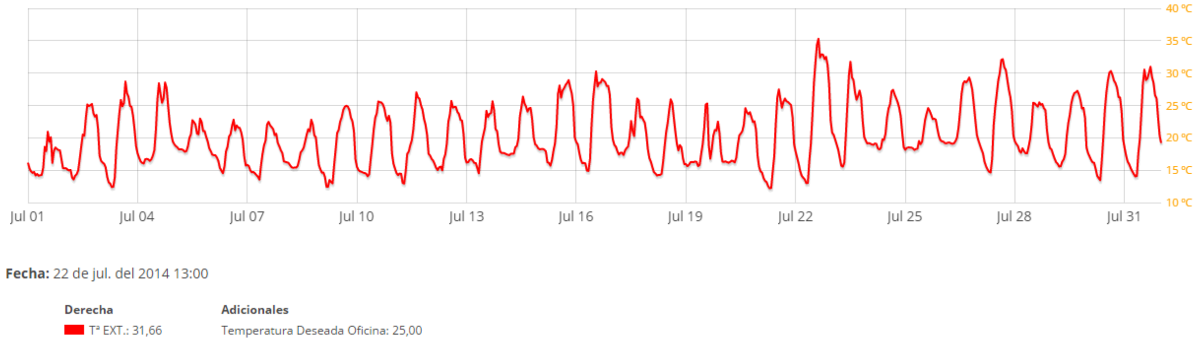


Figure 2. T^2 Hotelling control chart using 2 first PLS components of the 10 variables critical to HVAC quality.

Observations out of the control limits are observed in the monitored sample. The assignable cause is the wide daily variation external temperature in the second half of July. This is supported by the results of the EWMA control chart applied to the ARIMA model residuals of the temperature measured in the middle of the office (Figure 3). The time series model is applied in order to deal with the autocorrelation in temperature variable while EWMA chart is proposed due the model residuals are not Gaussian (Montgomery, 2013). This result revealed that the installed power of HVAC systems is not enough



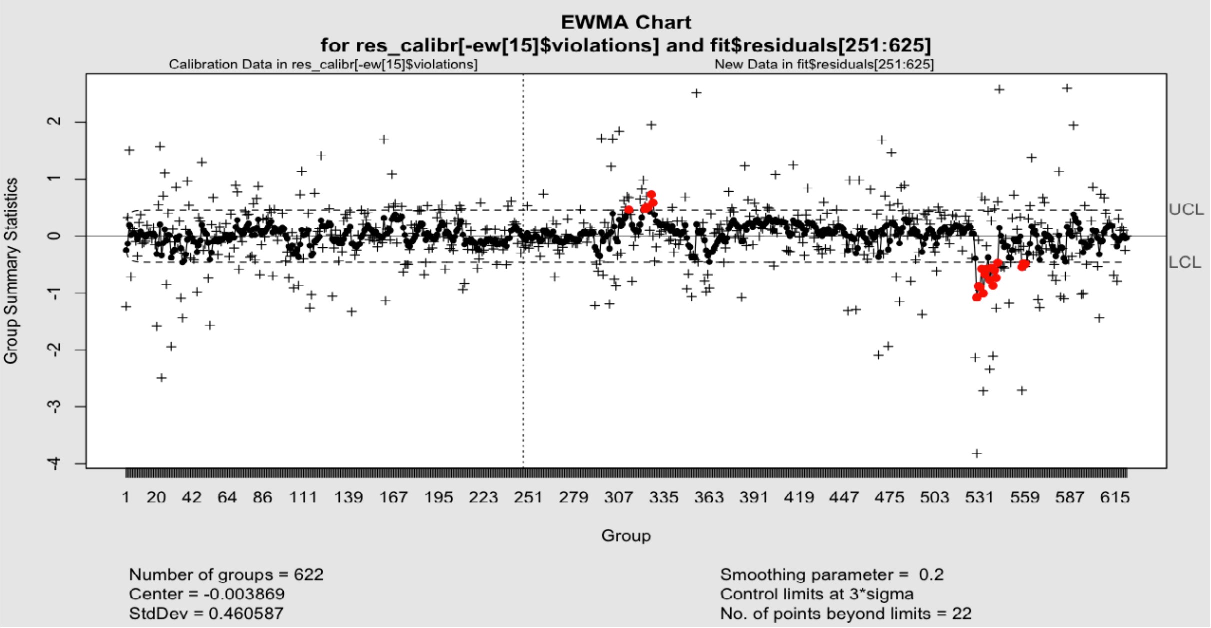


Figure 3. EWMA chart corresponding to the ARIMA model residuals applied to the temperature in the middle of the office.

4.1.1. FDA study

An alternative to control the energy efficiency variables is the application of FDA techniques. Each datum is a daily curve composed of hourly (or taken each 5 minute) measurements. Figure 4 shows the smoothed daily temperature curves plotted with the functional median and trimmed median. We can observe that the temperature in the middle of the office is more constant during the day than the other temperatures measured in the back and entrance.

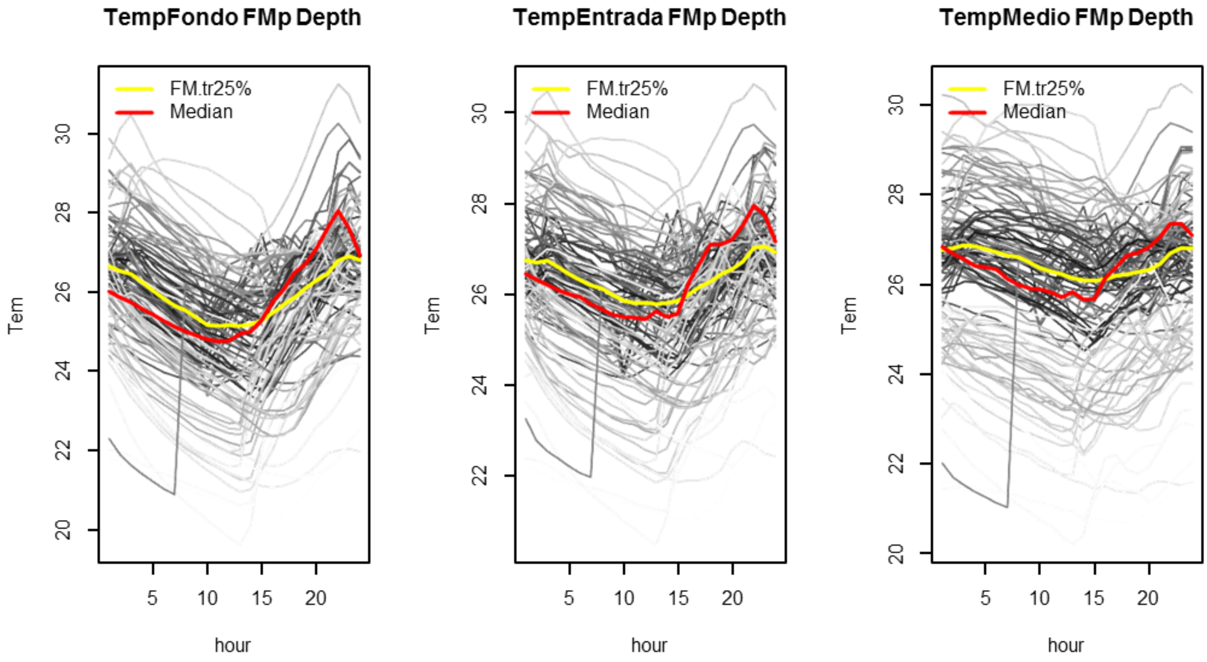


Figure 4. Back (TempFondo), entrance (TempEntrada) and middle (TempMedio) of the office daily temperatures with functional medians.

We deal with the problem of variables quality control and monitoring using alternative FDA classification techniques: DD-plot. Information about the DD-plot obtaining and characteristics is available in Li et al. (2012) and Cuesta-Albertos et al. (2015). DD-plot was applied using fda.usc R package (Febrero-Bande and Oviedo de la Fuente, 2012). The aim is to estimate if the aggregated daily energy consumption (scalar variable) is high or low depending on the 3 curves of daily temperatures or depending on the 3 curves of daily HR or the 3 curves of daily CO_2 content. For this purpose two groups are defined: one group corresponding to the days where the energy consumption was higher than the median and another composed of the days where the energy consumption was lower than the consumption median. The depth of each functional datum (curve) is calculated using the mode depth with respect to the two subjacent distributions (high or low consumption). Logistic classifier is applied to find the boundary that divides the plane in two regions: in red the high consumption region and in blue the low consumption region. The results are shown in Figure 5. The daily temperature or CO_2 curves do not differentiate between high and low energy consumption. The daily HR perfectly discerns between the two groups. The plane is built plotting the two mode depths of each three temperature, HR of CO_2 curves. Each point in the plane represents the one day. If the point is placed in the red region, we can infer that the energy consumption corresponding to the HR measured in that day was higher than the median. The two groups can be recalculated defining one corresponding to the consumption outliers and other corresponding to the normal consumption.

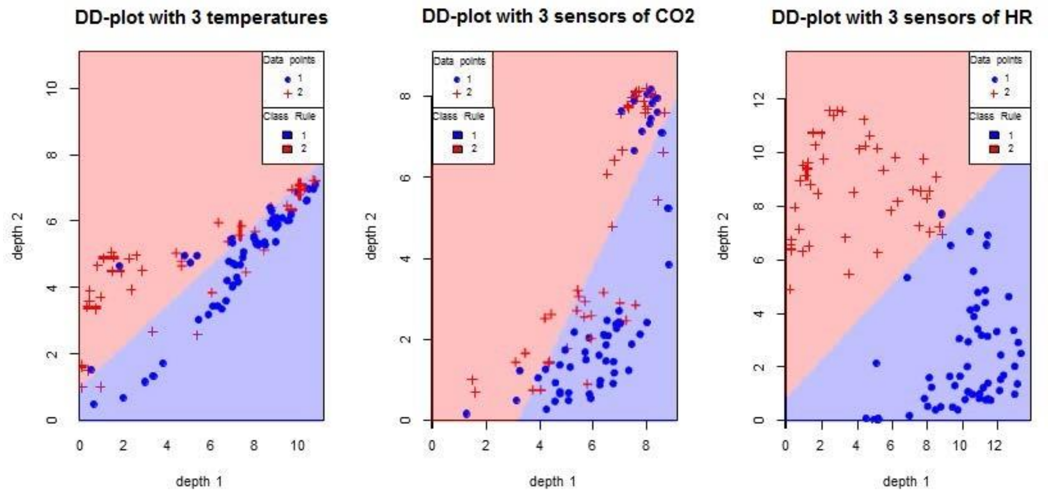


Figure 5. Left panel: DD-plot corresponding to the mode depths of the three diary curves of temperatures. Center panel: DD-plot corresponding to the mode depths of the three curves of CO_2. Right panel: DD-plot corresponding to the mode depths of the three curves of HR.

5. Conclusions

In this paper, we provided brief overview of functional data analysis and the role of quality engineers dealing with advancing big data. It is clear that FDA is an evolving field with numerous applications, some of which can present solutions to quality control problems. Standard Statistical Qualtiy Control procedures are not applicable to energy efficiency big data due they have not fulfill the standard assumptions, e.g. these type of data are heavily autocorrelated and they are not Gaussian distributed. The application of complexer (PLS) or even new (FDA DD-plot) control and monitoring techniques may play a role to properly deal with the problem. We have focused on the application of FDA techniques and also multivariate statistical control charts to solve control problems and alarms detection in energy efficiency. We have presented the problem related to the control of the air quality, thermal comfort and energy efficiency of HVAC installations.

Since the studied data is composed of 10 variables (temperatures, humidity, CO_2 concentrations, energy consumption), and energy consumption is the critical to quality variable that depends on the other 9 process parameters, T^2 multivariate control chart has been successfully applied to the two first PLS components to solve the control and alarm detection problem.

The DD-plot, an FDA alternative to multivariate control charts has been proposed to estimate if each analyzed day corresponds to high or low energy consumption from the daily HR curves. The strong relationship between these two variables (consumption is scalar and HR is functional) has been proved attending to the almost perfect classification obtained. The relationship between temperature and consumption is not as strong due high energy consumption could be due very low temperatures (heating consumption) and also high temperatures (cooling energy consumption).

These studies open a new way in the big data era, for the automation and control in HAVC installation thanks to the functional alarms.

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation, grant MTM2011-22393 (ERDF included), and by the Ministry of Economy and Competitiveness, grant MTM2013-41383P (ERDF included). We also express our gratitude to FRIDAMA company.

References

Capizzi G., Masarotto G. (2011). A Least Angle Regression Control Chart for Multidimensional Data. Technometrics, 53 (3):285-296.

Carter P. (2011). Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO. International Data Corporation (IDC),

Cuesta-Albertos JA, Febrero M, Oviedo de la Fuente M. The DDG-classifier in the functional setting. Submited to Test. 2015. Retrieved in http://arxiv.org/abs/1501.00372

Deming W. E. (2000). The new economics: for industry, government, education. 2nd ed. The MIT Press, Cambridge, MA. USA.

Febrero-Bande M, Oviedo de la Fuente M. Statistical Computing in Functional Data Analysis: The R Package fda.usc. J Stat Softw. 2012; 51, 1–28.

Ferraty F, Romain Y. (2011). The Oxford handbook of functional data analysis. Oxford handbooks. Oxford University Press, Oxford ; New York.

Jones-Farmer LA, Ezell JD, Hazen BT. (2013). Applying control chart methods to enhance data quality. To appear in Technometrics.

Francisco-Fernández M, Tarrío-Saavedra J., Mallik A., Naya S. A comprehensive classification of wood from thermogravimetric curves. Chemom Intell Lab Syst. 2012; 118:159–72.

Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993; 35: 109–135.

Li J, Cuesta-Albertos JA, Liu RY. DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. JASA. 2012; 107: 737-753.

Montgomery DC. (2013). Introduction to statistical quality control. 7th edn. Wiley, Hoboken, NJ.

Ramsay JO, Silverman BW. (2002). Applied functional data analysis: methods and case studies. Springer series in statistics. Springer, New York.

Ramsay JO, Silverman BW. (2005). Functional data analysis. Springer series in statistics, 2nd edn. Springer, New York.

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B, 58 (1): 267-288.