Chemometrics and Intelligent Laboratory Systems 176 (2018) 134–148

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Functional extensions of Mandel's *h* and *k* statistics for outlier detection in interlaboratory studies

Miguel Flores [a], Javier Tarrío-Saavedra [b,*], Rubén Fernández-Casal [c], Salvador Naya [b]

[a] *Escuela Politécnica Nacional Quito, Ecuador*
[b] *Grupo MODES, CITIC, ITMATI, Departamento de Matemáticas, Escola Politécnica Superior, Universidade da Coruña. Ferrol, Spain*
[c] *Grupo MODES, CITIC, ITMATI, Facultade de Informática, Universidade da Coruña. A Coruña, Spain*

ARTICLE INFO

Keywords:
Interlaboratory studies
Functional data analysis
Outlier detection
Bootstrap
Data depth

ABSTRACT

Functional data analysis (FDA) alternatives, based on the classical Mandel *h* and *k* statistics, are proposed to identify the laboratories that supply inconsistent results in interlaboratory studies (ILS). ILS is the procedure performed by a number of laboratories to test the precision of an analytical method, to measure the proficiency of laboratories in implementing an analytical procedure, to certify reference materials, and to evaluate a new experimental standard. The use of outlier tests, such as *h* and *k* Mandel statistics proposed by the ASTM E691, is crucial to assess these aims, estimating inter- and intra-laboratory data position and variability from a univariate point of view. Considering that experimental results obtained in analytical sciences are often functional, the use of FDA techniques can prevent the loss of important data information. The FDA approaches of *h* and *k* statistics are presented and point-wise obtained to deal with functional experimental data. Both functional statistics are estimated for each laboratory, their functional critical limits are obtained by bootstrap resampling, and new FDA versions of *h* and *k* graphics are presented. Real and synthetic thermogravimetric data are utilized to assess the good performance of the proposed FDA *h* and *k* statistics and their advantages with respect to the univariate approach.

1. Introduction

Interlaboratory Studies (ILS) can be defined as the statistical quality control procedures implemented to evaluate the performance of an analytical method through collaborative trials, to develop bias tests of a standard measurement method, to measure the proficiency of laboratories that implement a specific analytical procedure, to certify reference materials, and to validate a new international standard [1–4]. In all cases, ILS statistical methods evaluate the precision and consistency of testing results obtained by different laboratories [1]. Two of the most common ILS are those applied in collaborative trials and bias tests. Collaborative trials provide estimates of precision, in terms of repeatability, reproducibility, and variability [2,5]. The development of ILS methodologies is absolutely necessary when a precision estimate of a new analytical method is required. On the other hand, bias tests are developed with a standard method. They aim to evaluate a standard measurement method bias or laboratory bias when they are used as a standard method [2]. The monographs of [2] and [5] provide more comprehensive information about experimental method precision, bias, and proficiency studies. The ISO standard regulates the implementation of bias tests and also defines collaborative trials [6]. This work will focus on outlier test applications in these types of ILS studies. In fact, new functional extensions for outlier tests are proposed and described.

Both in collaborative trials and bias tests, outlier detection procedures play a fundamental roll [7–9], where the aim is to detect the laboratories that provide results that are significantly different from the others and, thus, to discard the inconsistent data that they provide. Many outlier tests have been applied in ILS studies. All of them are developed from a scalar or univariate perspective. They can be classified into those that examine laboratory result variances and those based on mean differences. Standards usually propose the implementation of variance-based outlier tests (one-sided tests) over those based on laboratory mean differences [2]. The Cochran test is by far the most used variance based test in interlaboratory studies [2,10]. In addition, the F test is also employed by comparing intralaboratory variances with respect to repeatability variance [2]. It is important to stress that, unlike the Cochran test, it is necessary to discard those laboratories with outlying means before the application of an F test [2]. There are many more test focused on

[*] Corresponding author.
E-mail address: javier.tarrio@udc.es (J. Tarrío-Saavedra).

https://doi.org/10.1016/j.chemolab.2018.03.016
Received 11 July 2017; Received in revised form 30 March 2018; Accepted 31 March 2018
Available online 4 April 2018

detecting outlying means. Such tests include the Grubbs test (for single or double outliers) [11,12] and the Graf and Henning test [13]. In addition, some robust alternatives to classical outlier test approaches have been proposed. Namely, the median of absolute deviations from the median (MAD) [14], the robust mean and standard deviation calculation [15], and Tukey's biweight function [14], based on assigning less importance (weight) to less reliable data.

The use of graphical methods to interpret the results retrieved in ILS is closely linked to outlier detection. Thus, the use of diagrams such as boxplots [16], Youden plots [17], control charts, and bar plots, among others, is proposed in the different ILS protocols. Among the different existing graphical methods, Mandel's *h* and *k* statistics [18] are intensively used in ILS for detecting laboratories that provide inconsistent results, using graphical tools such as bar plots. The *h* statistic accounts for intralaboratory variability, i.e., the differences of laboratory means with respect to global mean, whereas the *k* statistic estimates the intralaboratory variability by comparing the repeatability variances corresponding to each laboratory. Thus, they are employed to detect outliers among the means (type 2) and among the standard deviations (type 3), but not among the replicates (type 1) [2]. Their use is proposed by different protocols corresponding to collaborative trials and bias tests [19], combined with other outlier tests such as Cochran, Grubbs, and F.

It is important to stress that all the outlier detection tests for ILS deal with scalar data. Nowadays, there are many experimental techniques related to applied chemistry, physics, and engineering where data are complexer rather than scalar. They are often high dimensional, even functional. In fact, spectra [20,21] (e.g. Mass Spectroscopy [22], Nuclear Magnetic Resonance spectroscopy [23], and Near-Infared Spectroscopy [24]), thermogravimetric [1,25–29], calorimetric [1,26], or dynamic mechanical curves [30] are special cases of infinite dimensional data, i.e., functional data. In fact, from a physical perspective, could be more informative to analyze a spectrum as a function rather than a vector of features due to the presence of high correlation among them, as pointed out by Saeys et al. [21]. We can find excellent examples of functional data in the domains of proteomics [22,31], where experimental techniques such as Mass Spectrometry are used for protein identification or quantification tasks, and metabolomics, where Nuclear Magnetic Resonance spectroscopy is usually applied [23]. Thus, the application of functional data analysis (FDA) techniques such as the proposed for ILS could be useful. FDA is a relatively new branch of statistics that deals with infinite dimensional data, i.e., those curves, surfaces, and volumes defined continually such as the time or frequency domain [20,32]. Taking into account recent advances in computing science and the increasing amount of functional data retrieved by experimental techniques and sensors, FDA has been a great development in recent years. In fact, a great deal of exploratory [33,34], regression [35,36], classification [37], analysis of variance [25,38], and time series [39] statistical methodologies have been developed and extended to a functional case. These techniques have been successfully applied in a wide range of scientific domains such as neuroscience [40], engineering [36], environmental sciences [41], material science [28,30], and chemistry [20]. The use of FDA statistical techniques is facilitated for practitioners by the development of various packages implemented in R software [42] such as fda [43] and fda. usc [44,45]. This fact has helped to increase the usability and generalization of these techniques.

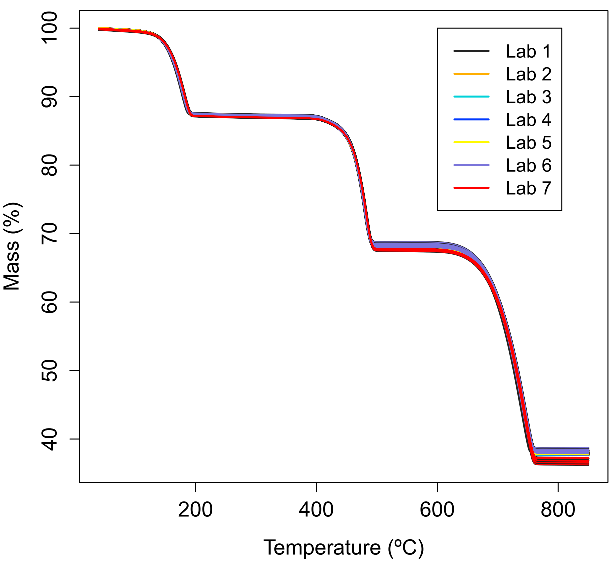

Fig. 1. TG data obtained by the different emulated laboratories.

Concerning ILS studies, FDA approaches for outlier detection based on functional data depth have been introduced in [1]. That work includes FDA exploratory analysis [33], functional ANOVA based on random projections (with false discovery rate correction) [38], and an FDA outlier detection method composed of functional data depth calculation (mode, Fraiman and Muniz, random projections depths) [33,45], and bootstrap resampling [34,45]. That approach identifies outliers among replicates (type 1 outliers), but it does not directly identify laboratories that provide inconsistent data (type 2 and 3). On the basis of scalar, Mandel's *h* and *k* statistics with their graphical tools are able to identify outliers of type 2 and 3. The development of functional extensions of *h* and *k* are justified when data obtained by laboratories are functional (each curve is a datum of infinite dimension). In fact, the application of a scalar test requires the previous extraction of one representative feature from curves or surfaces. Important information can be lost in this process, indeed, depending on the extracted feature, the test result could be different. In this work, the use of bootstrap resampling provides an alternative to develop functional extensions for these two statistics, first briefly introduced in [46].

2. Experimental data collection

Two different real datasets have been used to test the new FDA approximation for *h* and *k* statistics. The first one deals with thermal analysis analytical techniques such as thermogravimetry, whereas the second dataset accounts for temperature measurements obtained by three redundant sensors placed in the same room in a commercial area of a building.

2.1. Interlaboratory study based on thermogravimetric data

Thermogravimetric (TG) curves obtained from calcium oxalate monohydrate (99.0 + % purity) by Panreac, Ca(COO)2H2O have been used to evaluate the new FDA approach of *h* and *k* statistics. TG is a thermal analysis technique that provides information about material thermal stability by measuring the mass loss as a function of time or temperature. The goal is to assess the good performance of the proposed FDA extensions for Mandel's *h* and *k* by comparing their results with those corresponding to the *h* and *k* classical scalar approach. This database has been obtained and used in the authors' previous work where FDA outliers detection techniques based on data depth were introduced and compared with the classical univariate approach [1]. Then, its use is justified in order to properly assess the FDA *h* and *k* performance. As pointed out in the previous work, calcium oxalate is often used as reference material in calibration tasks due to its well-defined thermooxidative reactions, composed of three very well-defined mass loss steps.

In order to simulate a common ILS, 7 different laboratories were emulated by combining different testing instruments with different instrument calibrations. Each emulated laboratory tested 15 samples of calcium oxalate by thermogravimetric analysis, thus, overall 105 samples were used. Each sample was tested in a TA Instruments SDT 2960 or, alternatively, in a Rheometric STA 1500 simultaneous analyzer. TG curves were obtained heating each sample at a constant heating rate of 20 °C/min, between 20 and 900 °C under air atmosphere (50 mL/min).

Table 1
Label and description of each laboratory.

Laboratory 1	STA device with old calibration specifications is used
Laboratory 2	Core group of
Laboratory 3	laboratories that
Laboratory 4	provides consistent
Laboratory 5	data with an SDT device
Laboratory 6	Old calibration specifications are applied to an SDT instrument
laboratory 7	Biased calibration of 2 °C was applied

Fig. 1 shows the TG curves corresponding to the 7 studied laboratories. All the TG curves correspond to real data obtained in laboratory, but it is important to highlight that the ILS shown in this work is simulated. In fact, as pointed out above, each laboratory is simulated by slightly varying the instruments calibration, operators, and analyzer. Thus, there is a group of 4 simulated laboratories (corresponding to 4 different real operators) that provide the less unbiased and variable results. Moreover, a simulated laboratory that provides biased results due to a 2 °C biased temperature calibration with respect to the real value of zinc melting temperature is used. A Laboratory with higher variability has also been simulated using a 2-year old calibration. Finally, a seventh lab with higher bias and variability, using a different instrument, Rheometric STA 1500, with an old calibration, is used. Table 1 shows each laboratory label and summarizes its main features, indicating which of them should provide different results due to their differences in experimental testing procedure.

2.2. *Detection of anomalous performance in the case of temperature sensors in a heating ventilation and air conditioning installation*

In this section we present a case study of the study of redundant sensors [47]. That is, the installation of several sensors in the same area in order to ensure the correct measurement of a critical quality feature for the installation. In this case, the objective is to ensure the correct measurement of the temperature in the stores of a Galician textile company. This is a case study of a recently opened store located in a commercial center of Panama City. The aim of monitoring the temperature is to continually improve and control the energetic efficiency and thermal comfort of the store. Data are retrieved by Σqus, developer of web platforms [48], and Nerxus statistical consulting company for energy data. Fig. 2 presents the plan of the store and identifyies the three installed temperature sensors, one for each air conditioning machine (also called air handled unit, AHU). It is intended to identify if the sensors are measuring the same, that is, to identify if there is an erroneous operation or a change in the climatic characteristics in any of the areas of the store (which is undesirable and possibly related to a malfunction of the corresponding AHU). It is therefore a case study in the control of data monitoring systems through the analysis of sensors. It is intended to provide an alternative solution through the application of the functional approach to the Mandels h and k statistics. By means of these techniques, applied to the daily temperature curves, we intend to identify sensors that behave atypically. With the help of the maintenance professionals in the stores, whether the identified sensors have correct performance or not is verified. In particular, those responsible detected anomalous behavior in two of the sensors studied, due to technical reasons, during the time period that data was obtained. The objective of the application of the ILS FDA methodology will therefore be the detection of this situation automatically.

3. Functional methodology for inconsistent laboratories detection

In this section, the scalar h and k statistics, and their functional extensions, $H(t)$ and $K(t)$ statistics, are introduced, as well as the corresponding d^H and d^K test statistics. d^H and d^K, obtained from $H(t)$ and $K(t)$ statistics, permit the detection of laboratories that provide no consistent data in an ILS. In addition, some brief information about FDA, functional norm, and functional data depth is also included in order to present as self-contained a work as possible. The calculation of those measures is necessary for computing both the location and dispersion functional estimates and for outlier detection [33,34,44].

3.1. *FDA and functional data depth*

Taking into account the proposed methodology structure, functional data analysis and functional data depth are briefly described here.

Assume that the functional dataset $\{X_1(t), X_2(t), \ldots, X_n(t)\}$ was obtained as iid observations from a stochastic process $X(t)$, with continuous trajectories on the interval $[a, b]$, $\mu(t)$ being the functional mean and $\sigma^2(t) > 0$ the functional variance. We will consider the L_2-norm:

$$\|X\| = \left(\int_a^b X(t)^2 dt\right)^{\frac{1}{2}},$$

The data depth concept explains how a datum is centered with respect to a set of observations from a given population. Therefore, the deepest datum will be that surrounded by the highest number of neighbors. In the FDA context, deeper curves are identified as those closer to the center, which are usually estimated by the median [34]. Three of the most common approaches to calculating the functional depth are the depth of Fraiman and Muniz (or median depth) [33], the mode depth [34], and the depth based on random projections [49].

The functional data depth can be used for outlier detection. Febrero-Bande et al. [44] identify outliers in functional datasets, taking into account that depth and outlyingness are inverse notions (an outlier curve will have a significantly low depth). Therefore, a way to detect the presence of functional outliers is to look for curves with lower depths. In the present study, we use a procedure based on trimming for detecting outliers.

3.2. *Scalar and functional extensions of Mandel's h and k statistics*

In the ILS, a set of observations $\{X_1^l(t), \ldots, X_n^l(t)\}$ are obtained for each lab $l, l = 1, \ldots, p$. Each laboratory experimentally tests n samples,

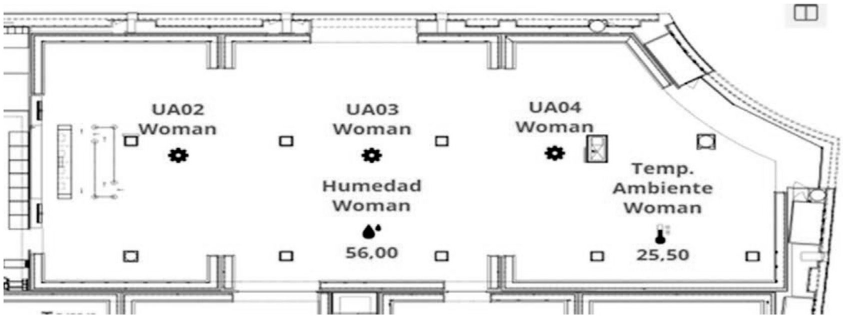

Fig. 2. Plan of the store, situation of the three temperature sensors associated with the three AHUs.

obtaining n different curves. The functional $H_l(t)$ and $K_l(t)$ statistics are estimated for each laboratory and the null hypothesis that there is no statistical difference between laboratory measurements is considered. The null hypotheses for R & R studies are described below:

The null hypothesis of reproducibility states that

$$H_0 : \mu_1(t) = \mu_2(t) = \ldots = \mu_p(t), \tag{1}$$

where $\mu_l(t)$, $l = 1 \ldots p$ are the populational functional mean for each laboratory l.

To test reproducibility of the laboratory results, the previous calculation of the $H(t)$ statistic is necessary. It is defined as

$$H_l(t) = \frac{X_i^l(t) - \overline{X}(t)}{S_l(t)}; l = 1, \ldots, p,$$

where $\overline{X}_l(t)$ y $S_l(t)$ are the mean and functional variance pointwise calculated for the l laboratory.

The null hypothesis of repeatability states that there are no differences in the laboratory variability:

$$H_0 : \sigma_1^2(t) = \sigma_2^2(t) = \ldots = \sigma_p^2(t), \tag{2}$$

where $\sigma_l(t)$, $l = 1 \ldots p$ are the theoretical functional variances corresponding to each laboratory l.

The repeatability test is based on the $K(t)$ statistic, defined as

$$K_l(t) = \frac{S_l(t)}{\sqrt{\overline{S^2}(t)}}; l = 1, \ldots, p,$$

where, $\overline{S^2}(t) = \frac{1}{p}\sum_{l=1}^{p} S_l^2(t)$.

On the one hand, in order to test the reproducibility hypothesis, we define the d^H test statistic as

$$d_l^H = \|H_l(t)\| = \left(\int_a^b H_l(t)^2 dt\right)^{\frac{1}{2}},$$

considering that d^H values corresponding to inhomogeneous laboratories will tend to be high. On the other hand, to test the repeatability hypothesis, we also define $d_l^K = \|K_l(t)\|$, taking into account that higher values of d^K correspond to non-consistent laboratories.

3.3. Bootstrap algorithm for critical values estimation

A bootstrap algorithm to test if the d_l^H and d_l^K are significantly high is proposed. The proposed bootstrap procedure pretends to reproduce the distribution of these statistics under the corresponding null hypothesis, (1) and (2) respectively. Assuming that a significance level α was fixed (typically $\alpha = 0.01$), the algorithm consists of the following steps:

1. Remove atypical observations (in this case, curves), grouping all the curves in a single set (null hypothesis), and applying the procedure based on trimming for detecting outliers.
2. Using the smoothed bootstrap proposed in [34] to obtain bootstrap samples of size $p{\cdot}n$ from the overall dataset once outliers are previously discarded. The bootstrap observations are randomly assigned to the laboratories in each bootstrap sample.
3. For each bootstrap sample, the $H_l^*(t)$ and $K_l^*(t)$ functional statistics, and the corresponding d_l^{H*} and d_l^{K*} test statistics, are computed for each laboratory $l = 1, \ldots, p$.
4. Approximate the critical values c_H and c_K of the test statistic (d_l^{H*} and d_l^{K*}) from the empirical $100(1-\alpha)/p$ percentile of the distribution of the corresponding $p{\cdot}B$ bootstrap replicates. In ILS, critical values for outlier tests are usually calculated for a signification level of $\alpha = 0.01$.
5. Finally, the confidence bands for the $H(t)$ and $K(t)$ statistics are computed. They are determined by the envelope of bootstrap samples defined by a norm that is less than the corresponding critical value.

It is important to note that, for each laboratory, once d_l^{H*}, d_l^{K*}, and their respective critical values, c_H and c_K, are calculated, the null hypotheses of reproducibility (1) (or repeatability (2)) will be rejected if $d_l^H = \|H(t)\| > c_H$ (alternatively, $d_l^K > c_K$ when repeatability is tested).

3.4. Methodology implementation in the ILS R package

In order to provide direct and easy access to the developed FDA methodologies, a special emphasis has been placed on programming computational tools using the statistical software R. The aim is that practitioners of academia and industry could apply these techniques to different types of functional data. In particular, the interlaboratory study package (ILS) [50] has been developed and designed for the detection of laboratories that provide non-consistent (atypical) data in the field of interlaboratory studies. The ILS package allows for the application of the univariate statistical tools recommended by the ASTM standard [19], among which are the h and k, Grubbs, and Chocran scalar statistics. But, what is more important in the present case, the ILS package allows for the estimation of the functional statistics, $H(t)$ and $K(t)$, as well as the test statistics d_H and d_K. They are used for testing the hypothesis of repeatability and reproducibility based on the critical values c_h and c_k and estimated using the bootstrap algorithm described in [46].

4. A simulation study

Two scenarios are considered to observe the performance of the new FDA approximations for Mandel's h and k statistics. They account for the mean and variability deviations from the group formed by the consistent laboratories. Accordingly, the first scenario consists of varying the Gaussian process mean of one lab with respect to the data simulated for the consistent laboratories, whereas the second one is defined by fixing a different result variance for one laboratory. The above-mentioned scenarios allow for the evaluation of the d^H and d^K statistics power and also to develop new functional control charts for $H(t)$ and $K(t)$ functional statistics. The use of the latter provides additional information about the time/temperature range where curves become outliers, and consequently about the chemical or physical process and substances involved (depending on the experimental data analyzed).

Each scenario is composed by p laboratories (each one has tested n samples). The TG curves, results of each lab, are simulated from a Gaussian process $Y(t) = \mu(t) + \sigma(t)\varepsilon(t)$, where $t \in [0,1]$ is the simulated time/temperature interval measured in arbitrary units (u.a.). The $\mu(t) = \frac{c}{(1+\exp(b(t-m)))^{\frac{1}{\tau}}}$ is the trend function and it corresponds with the generalized logistic model, whereas $\sigma(t)^2 = c_0\left(5 + \nu\left(1 - \left(\frac{t}{0.5} - 1\right)^2\right)^3\right)$ is the deterministic variance with $c_0 = 10^{-6}$. In addition, ε is a second order stationary process defined by 0 mean and $\exp(-|s-t|/0.3)$ covariance. The generalized logistic model has been used to model TG data in many works dealing with kinetic and classification studies [29,51]. These two scenarios are defined taking into account the meaning of the generalized logistic parameters. In fact, the m parameter accounts for the time or temperature corresponding to the maximum curve slope or inflexion point, c is the initial sample mass (in this case it is fixed to 1, in u. a), b is related to the rate of change (in the case of TG curves, rate of thermal degradation), and τ accounts for the asymmetry degree of curves. In the interest of simplicity, and accounting for previous studies, one symmetric ($\tau = 1$) degradation process defined by an initial mass of 1 u. a. ($c = 1$), and rate of degradation characterized by $b = 10$ is assumed in the present simulation study. Summarizing, $\mu(t) = \frac{1}{1+\exp(10(t-m))}$.

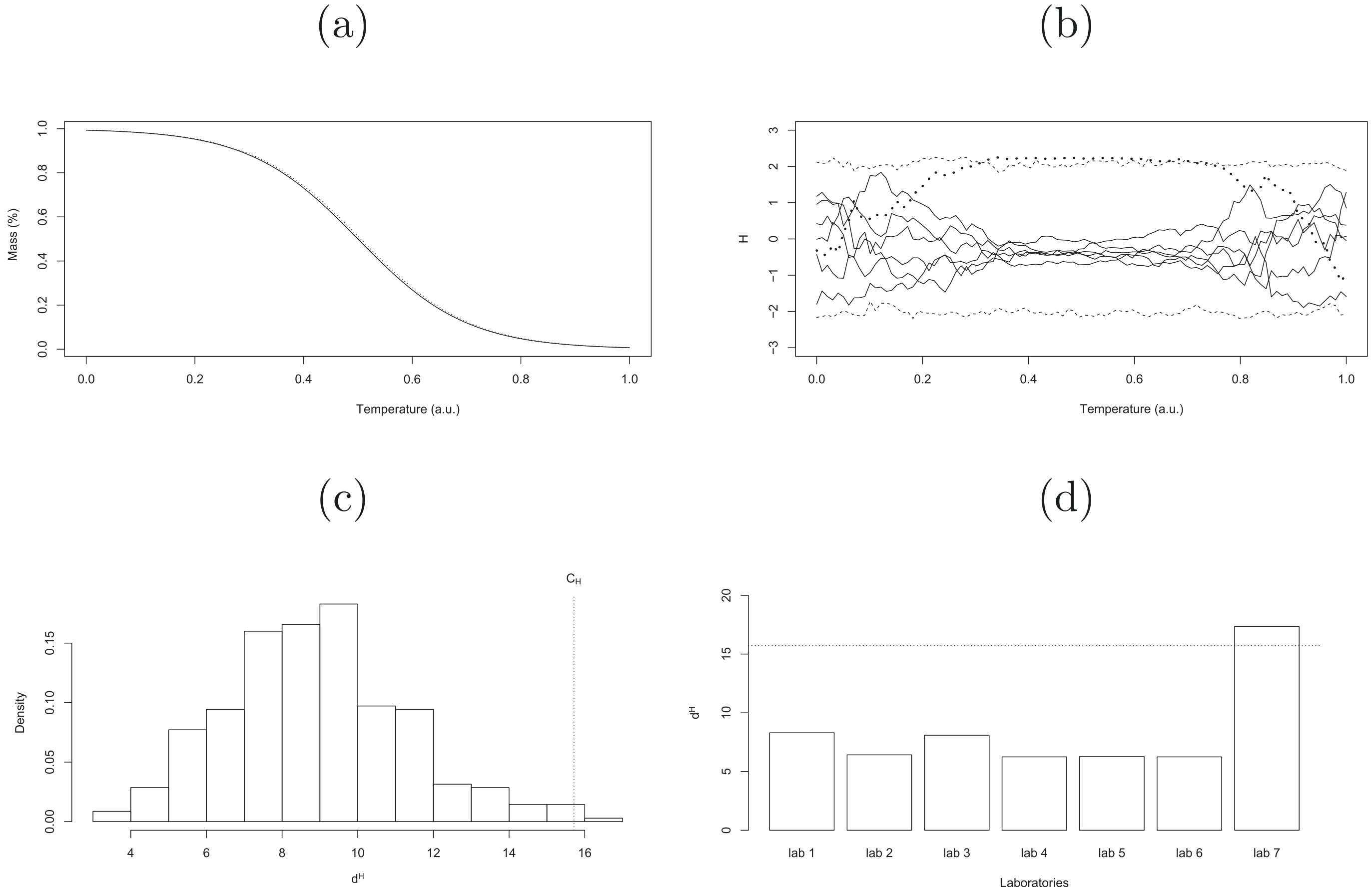

Fig. 3. (a) Theoretical means for TG curves simulated under the null hypothesis $H_0 : m_0 = 0.5$ and the alternative, $H_1 : m_1 = m_0(1 + \delta_H)$ ($\delta_h = 0.005$), (b) $H(t)$ functional statistic realizations for each of the 7 simulated laboratories, (c) d^h statistical distribution estimated by bootstrap resampling, and (d) d^H statistic realization for each laboratory sample.

4.1. A simulation study to test reproducibility

In the first simulated scenario, the null hypothesis is formulated in terms of m parameter. It is defined by $H_0 : m_0 = 0.5$, i.e., the results of laboratories are consistent when $\mu(t) = \frac{1}{1+exp(10(t-m_0))}$. On the other hand, the alternative hypothesis is defined by $H_1 : m_1 = m_0(1 + \delta_H)$, where $-0.005 \leq \delta_H \leq 0.005$. The theoretical mean for $m_0 = 0.5$ and $m_1 = m_0(1 + 0.005)$ are presented in Fig. 3 a. As can be observed, there are very slight differences between the two scenarios.

Accordingly, and taking into account previous studies with real data, TG curves corresponding to $p = 7$ laboratories, each obtaining one $n = 10$ replicate, are simulated. The data corresponding to a core of 6 laboratories are simulated under the H_0, whereas the results of 7th lab are obtained under the H_1. Considering that interlaboratory differences are induced, the null hypothesis of reproducibility has to be tested using the proposed d^H and $H(t)$ statistics. In this regard, 1000 simulations were done, each one composed of 500 bootstrap resamples for the purpose of estimating the d^H statistic distribution (Fig. 3c). Once the distribution of d^H statistic is estimated by the bootstrap procedure, the critical value test, c_H, corresponding to $\alpha = 0.01$, is also calculated and shown in Fig. 3 c (dotted line).

Note that one of the advantages of Mandel's h and k has currently been that both statistics provide very intuitive graphical tools for identifying the laboratories characterized by non-consistent results. Thus, the present FDA proposal intends to reproduce and even complete the graphical outputs corresponding to the univariate approach. Namely, in Fig. 3 b, the $H_l(t)$ (with $p = 1, 2, \ldots, 7$) sample realizations of functional statistic are shown. The corresponding d_l^H are shown in the bar plot of Fig. 3 d, following the same style as the univariate approach [1]. It provides a graphical way to perform the hypothesis test, where c_H is a dotted horizontal line that determines the edge above which a laboratory could be considered an outlier. In fact, the 7th laboratory (of which the data have been simulated under H_1) is successfully identified as an outlier, $d_7^H > c_H$ (Fig. 3d). The same result can be achieved by observing Fig. 3 b, where the central section of the $H_7(t)$ curve, shown as a segmented line, is out of the region defined by the functional quantiles (shown in dashed lines) obtained from c_H. It is important to note that the functional approach, shown in Fig. 3 b, also provides information about the time/temperature interval, where the results of lab 7 are different from the others. In this case, this interval corresponds to the TG step region, where the slope of the curve varies (related to a simulated degradation process). This information could be useful not only for thermal analysis curves but to analyze other analytical technique results.

4.2. A simulation study to test repeatability

The second simulation scenario has been defined in order to test the behavior of the new methodology in detecting outlier laboratories taking into account intralaboratory variability. Thus, the null hypothesis is formulated in terms of variability, modifying the v parameter defined in section 3. It is defined by $H_0 : v_0 = 5$, i.e., the results of laboratories are consistent when $\sigma(t)^2 = c_0\left(5 + 5\left(1 - \left(\frac{t}{0.5} - 1\right)^2\right)^3\right)$. On the other hand, the alternative hypothesis is defined by $H_1 : v_1 = v_0(1 + \delta_K)$, where $0 \leq \delta_K \leq 2$. The theoretical variance for $v_0 = 5$ and $v_1 = v_0(1 + 2)$ are presented in Fig. 4 a. The variances have been chosen accounting for previous studies dealing with real TG data [1]. As in the case of the first

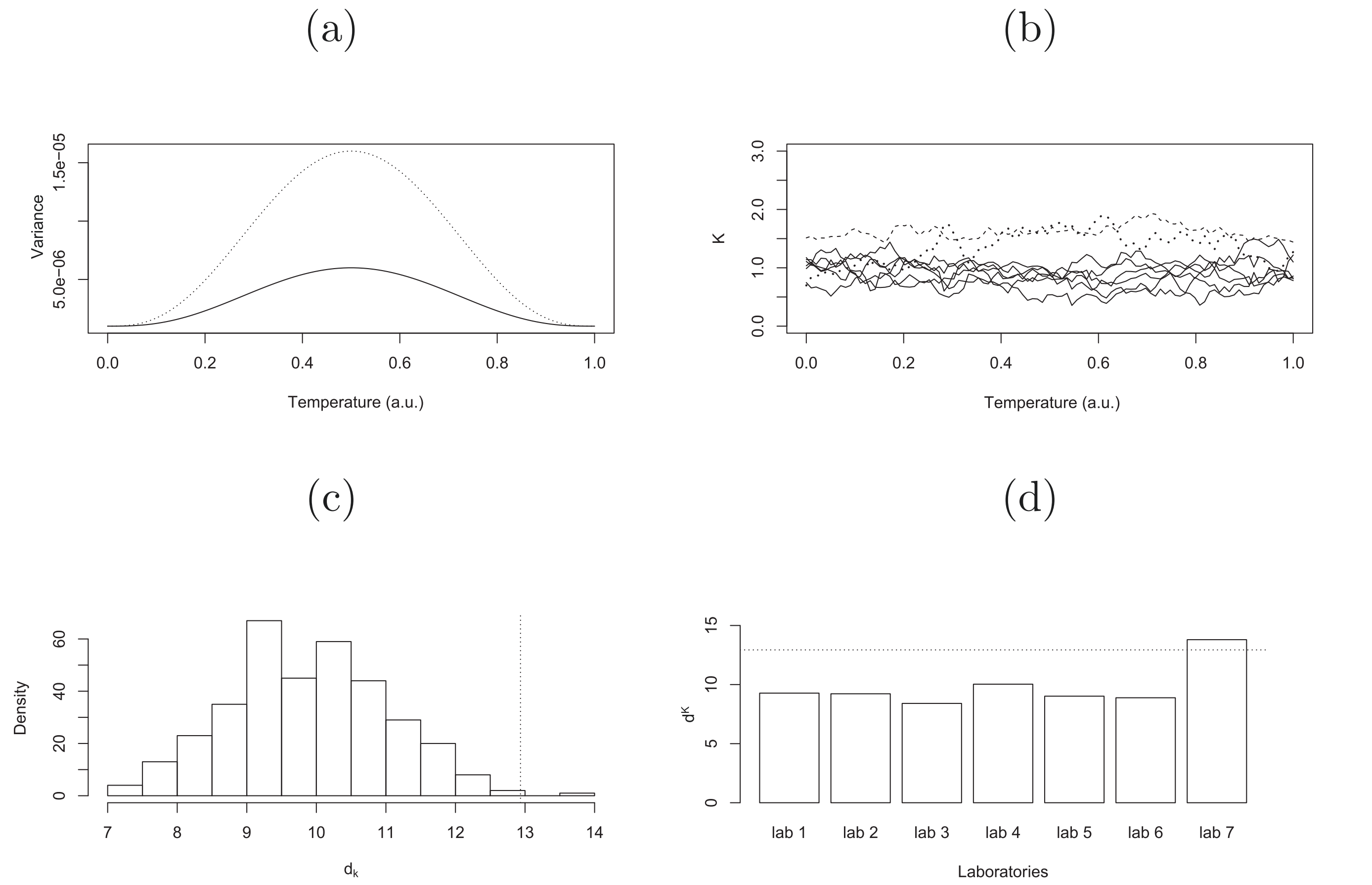

Fig. 4. (a) Theoretical means for TG curves simulated under the null hypothesis $H_0 : \nu_0 = 5$ and the alternative, $H_1 : \nu_1 = \nu_0(1 + \delta_K)$ ($\delta_k = 2$), (b) $H(t)$ functional statistic realizations for each of the 7 simulated laboratories, (c) d^K statistical distribution estimated by bootstrap resampling, and (d) d^K statistic realization for each laboratory sample.

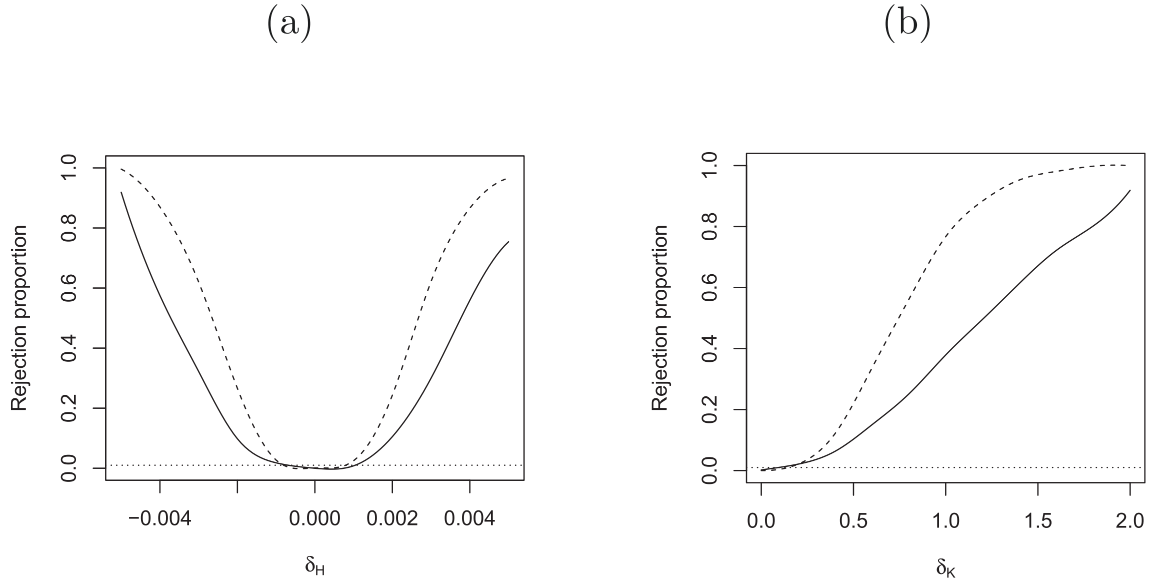

Fig. 5. (a) Power curves (rejection proportion) for the d^H statistic corresponding to $n = 10$ and $n = 20$ samples per lab, (b) power curves for the d^K statistic corresponding to $n = 10$ and $n = 20$ samples per lab. Segmented lines correspond to $n = 20$, whereas solid lines correspond to $n = 10$.

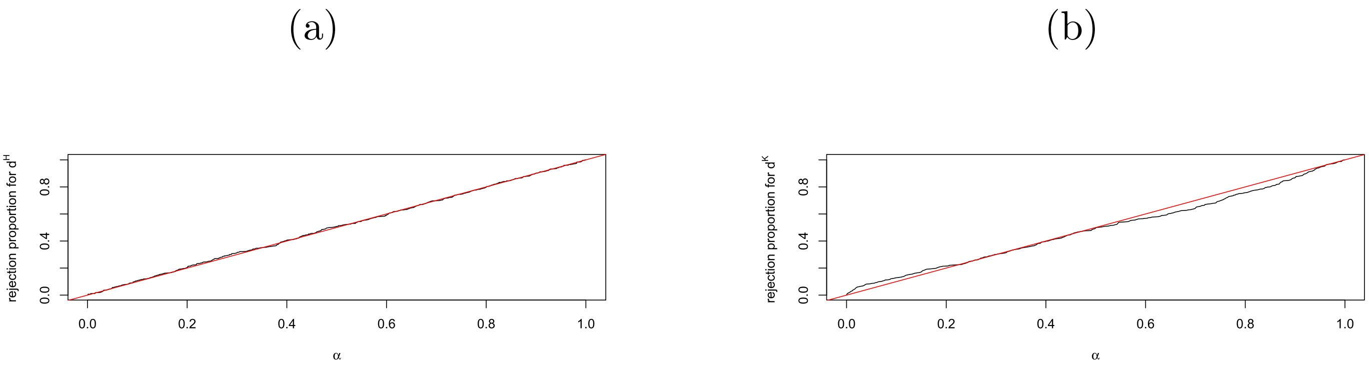

Fig. 6. (a) p-values distribution for the d^H statistic under the null hypothesis, (b) p-values distribution for the d^K statistic under the null hypothesis.

scenario, the data corresponding to the first 6 laboratories are simulated under the $H_0 : \nu_0 = 5$, whereas the synthetic curves of the 7th lab are obtained assuming the alternative hypothesis, $H_1 : \nu_1 = \nu_0(1 + 2)$.

Taking into account that intralaboratory differences are introduced, the null hypothesis of repeatability is tested using the proposed d^K and $K(t)$ statistics. A thousand simulations were performed, each one composed of 500 bootstrap resamples in order to estimate the d^K statistic distribution (Fig. 4c). The test of critical value, c_K, defined by, is also estimated for $\alpha = 0.01$ (Fig. 3 c, defined by a dotted line). Moreover, the $K_l(t)$ (with $p = 1, 2, \ldots, 7$) sample realizations of functional statistics are shown in Fig. 4 b. The corresponding d_l^K are plotted in Fig. 4 d, providing a graphical tool to perform the hypothesis test, where c_K is a dotted horizontal line above which a laboratory could be considered an outlier, as is the case of the 7th laboratory, $d_7^H > c_H$ (Fig. 4d). As in the previous scenario, the same result can be achieved by observing Fig. 4 b, where the central section of the $K_7(t)$ curve, shown as a dotted line, is out of the region defined by the functional quantiles (shown in dashed lines) corresponding to c_K. This interval corresponds to the region surrounding the TG inflexion point, where the variance is higher.

4.3. Effects of laboratory sample size

For each simulated scenario, the effects of laboratory sample size ($n = 10, 20$), using a signification level of $\alpha = 0.01$, on the new tests performance are studied. The aim is to evaluate the consistency of the two proposed FDA approximations for Mandel's h and k tests. Fig. 5 shows the rejection proportions under the null hypothesis for both statistics, d^H in the (a) panel, and d^K in the (b) panel. The lines of Fig. 5a and 5 b shows the power of the d^H and d^K, respectively, corresponding to the two sample sizes: segmented lines for $n = 20$ and continuous lines for $n = 10$. As can be observed in both panels, the increasing of the sample implies higher rejection proportions out of the null hypothesis, a higher test power. This is in accordance with recommendations of ASTM [1] for the univariate case. In any case, even using small samples ($n = 20$) the present FDA approximations show a good performance.

4.4. Statistic p-values study under the null hypothesis

It is important to stress that p-value of the test statistics is a random variable uniformly distributed over $[0, 1]$ interval, $U(0, 1)$, under the null hypothesis [52]. Thus, this hypothesis should be checked in order to verify the validity of the proposed FDA approximations for Mandel's h and k. Accordingly, a sample with the results for the 7 laboratories has been obtained in each simulation and, consequently, sample values for d^H and d^K and their corresponding p-values have been obtained. By the application of Kolmogorov-Smirnov test to the resulting p-values and fixing a signification level of $\alpha = 0.01$, we find that the null hypothesis that p-values are uniformly distributed in $[0, 1]$ cannot be rejected. This is supported by the results of Fig. 6, where the cumulative rejection proportion of each test is plotted as a function of different values of α. In Fig. 6 a, for the d^H statistic, the rejection proportion is practically equal to α. A similar result can be observed in Fig. 6 b. The trend is almost equal to the bisector, only in the case of small values for α is there a very slightly higher rejection proportion than expected. These results help to support the validity of the proposed tests.

4.5. Comparison between FDA and scalar approaches for ILS

The ILS studies have been currently performed using statistical scalar approaches. A scalar variable of interest is measured and analyzed. As mentioned in section 1, the univariate tests for detecting outliers are one of the most used and necessary tools in ILS to detect those laboratories that provides inconsistent results [1–5,19]. In this regard, there are several very popular scalar tests prominent among which is Mandel's h and k [19], described from a computational point of view for practitioners. However, as far as we know, there are no FDA extensions from the univariate test in order to perform laboratory outlier detection when experimental data are functional, except for the present proposal. Thus, to compare this new proposal with respect to the corresponding multivariate approaches is mandatory in order to evaluate its performance and applicability. For this purpose, synthetic functional data are used, in this case the simulated TG curves. The aim of this subsection is to check if the use of the proposed FDA extensions provides advantages with respect to the scalar approach when the data are curves (functional). Therefore, to be able to calculate univariate Mandel's h and k, extracting a representative feature of curves is necessary. In the case of TG curves, different variables have been studied: the temperature/time to lose the 5wt% of initial mass (initial decomposition temperature/time, IDT) or, alternatively, the temperature/time to lose the 10wt%. These are indices currently used to estimate the degree of thermal degradation in polymeric materials [1]. The consistency of univariate and functional approaches is compared using two different sample sizes, $n = 10$ and $n = 20$. As can be observed in Fig. 7, the power of a univariate test depends on the feature that has been extracted. In fact, for both sample sizes, the power of the univariate h statistic is higher when the extracted variable is the time/temperature for losing the 10wt%. In Fig. 7 a, the highest power corresponds to the univariate approach applied to IDT, as mentioned, but the power corresponding to the FDA extension d^H is also higher than the power of univariate h when the extracted feature is the time/temperature to lose the 5wt%. Therefore, this supports the application of the FDA approach, taking into account the performance of the univariate case depends to a high extent on the chosen feature. Moreover, Fig. 7 b shows that the tests of the power of the univariate h statistic for 5wt% and 10wt% are higher than the power of the FDA approach when $n = 20$. Summarizing, for the case of Mandel's h statistic (applied to the thermal analysis case), the use of the FDA extension could be recommended when the number of replicates is relatively small (common case in ILS) and there is more than one possible representative feature in each curve.

However, when the Mandel's k statistic is studied, we find that the

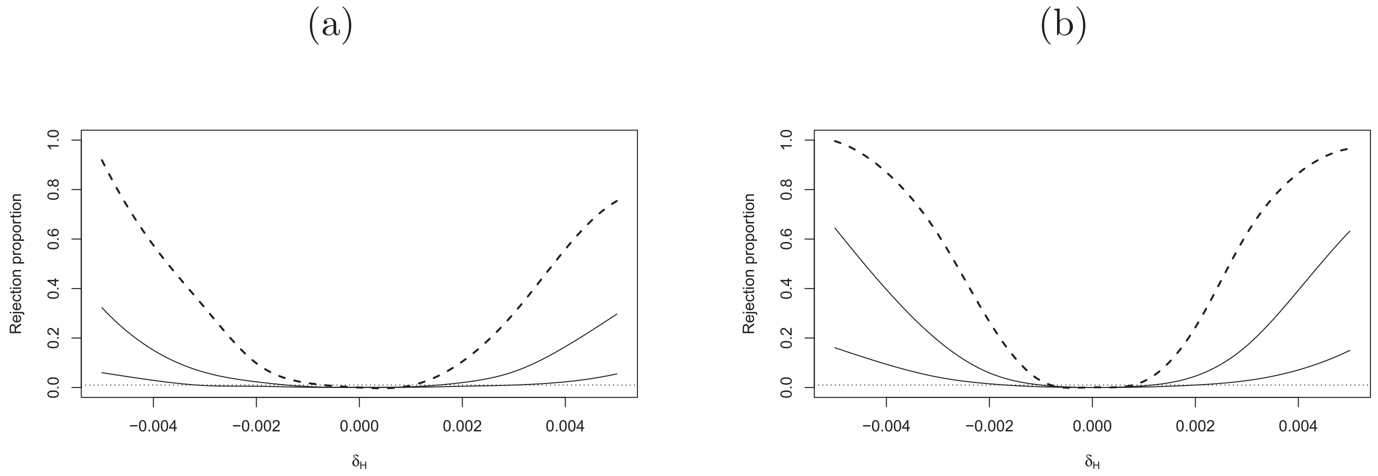

Fig. 7. (a) Test powers corresponding to univariate (5*wt*% and 10*wt*%) an FDA approaches for *h* statistic, with $n = 10$, (b) Test powers corresponding to univariate (5*wt*% and 10*wt*%) an FDA approaches for *h* statistic, with $n = 20$.

Table 2

d^K statistic rejection proportion (with $\alpha = 0.01$) compared with the univariate *k* statistic rejection proportion when calculated from the scalar variables time/temperature to lose the 5*wt*% and 10*wt*% of sample mass.

n	n = 10			n = 20		
δ_K	5 wt%	10 wt%	FDA	5 wt%	10 wt%	FDA
0	0.001	0.000	0.002	0.000	0.000	0.000
0.2	0.001	0.001	0.021	0.000	0.000	0.021
0.4	0.002	0.001	0.063	0.000	0.000	0.121
0.6	0.003	0.002	0.150	0.003	0.001	0.336
0.8	0.003	0.002	0.251	0.003	0.002	0.561
1	0.004	0.005	0.379	0.004	0.006	0.767
1.2	0.004	0.005	0.496	0.005	0.007	0.884
1.4	0.004	0.006	0.614	0.008	0.007	0.953
1.6	0.004	0.009	0.722	0.009	0.011	0.981
1.8	0.006	0.009	0.802	0.009	0.012	0.998
2	0.008	0.012	0.919	0.009	0.014	1

FDA approach provides the best performance in terms of power for all the studied sample sizes. In fact, Table 2 shows the rejection proportion for each shift from the null hypothesis, δ_K, corresponding to the univariate *k* statistic applied to the time/temperature to lose the 5*wt*% and 10*wt*% and to the FDA extension applied to the whole curves. As can be observed, the power of the FDA approach is rather higher than the corresponding univariate *k* statistic, independently of sample size and feature extracted (in the univariate case). Summarizing, the use of the FDA extension is recommended when intralaboratory variability is analyzed.

5. Real data application

In addition to simulation studies, the application of real data obtained by experimental techniques is needed to understand the performance and utility of the proposed methodology.

5.1. Interlaboratory study using thermogravimetric analysis

In this regard, the present functional procedure for R& R studies, described in section 3, is applied to the thermogravimetric real of calcium oxalate, presented in section2. In this real example, there are 4 laboratories that provide similar results, whereas the laboratories 1, 6 and 7 provides different results for different reasons (section2). These data have been deeply studied in [1], where the univariate *k* and *h* statistics had identified different laboratories as outliers depending on the feature extracted for TG curves. As shown below, the laboratories that provide inconsistent data are successfully identified by the application of the FDA extensions *H* and *K* (and their counterparts d^H and d^K test statistics). This approach has the goal of using the whole curve, without the feature extraction step.

As in the previous section, the first step consists of estimating the functional $H(t)$ (Fig. 9a) and $K(t)$ (Fig. 10a). From $H(t)$ (reproducibility) and $K(t)$ (repeatability) functional sample statistics, the d^H and d^K test statistics are calculated and compared with respect the c^H and c_K critical values, defined as the quantiles corresponding to $\alpha = 0.01$ (as usual in ILS studies), obtained by a bootstrap procedure (section 3). The *l* laboratory for which $d_l^H > c_H$ and/or $d_l^K > c_K$ is identified as an outlier. In addition, it is important to note that the identification of outliers in ILS is a iterative process, i.e., the univariate procedure is consecutively applied until no laboratory is identified as an outlier [1,19]. Thus, the present FDA extension is also iteratively applied, following the scheme of the scalar approach. It is also important to stress that the present FDA extensions for *h* and *k* Mandel's statistics identify the outlier laboratories, not only the outlier data within each laboratory, as in other ILS studies to deal with functional data [1].

Once *H* and *K* are calculated, the d^H and d^K distributions are estimated by bootstrap procedures. In advance, the outlier TG curves within each laboratory are removed. Thus, the 1% of the TG curves (obtained by all the laboratories) with the lowest functional mode depth [44,45] are removed. In the present study case, a TG curve corresponding to laboratory 7 has been removed. Note that laboratory 7 is one of the real outliers that we intend to detect.

Fig. 8 a shows the d^H distribution obtained by the bootstrap procedure (using $B = 500$ resamples), in addition to the c_H critical value corresponding to $\alpha = 0.01$. Fig. 8 b provides the d^H test graphical output corresponding to the first application of the functional approach (first iteration), where laboratory 7 is correctly detected as an outlier. In fact, laboratory 7 is defined by using thermogravimetric balance with a biased calibration of temperature. In the second and third iterations of FDA methodology, laboratories 1 (STA instrument with old calibration) and 6 (SDT instrument with old calibration) are also successfully detected as outliers. Fig. 8 c shows the d^H distribution in the fourth iteration or fourth application of the FDA procedure, whereas the corresponding d^H test graphical output is shown in Fig. 8 d, where no outlier laboratories are detected and the iterative process is stopped. All the outlier laboratories, from a reproducibility point of view, are successfully detected using the proposed FDA approach.

The reproducibility hypothesis can also be tested directly using the *H* functional statistic as shown in Fig. 9. Also applying an iterative process, with $\alpha = 0.01$, laboratory 7 is detected as an outlier (Fig. 9a). The *H* statistic provides additional information about the temperature interval

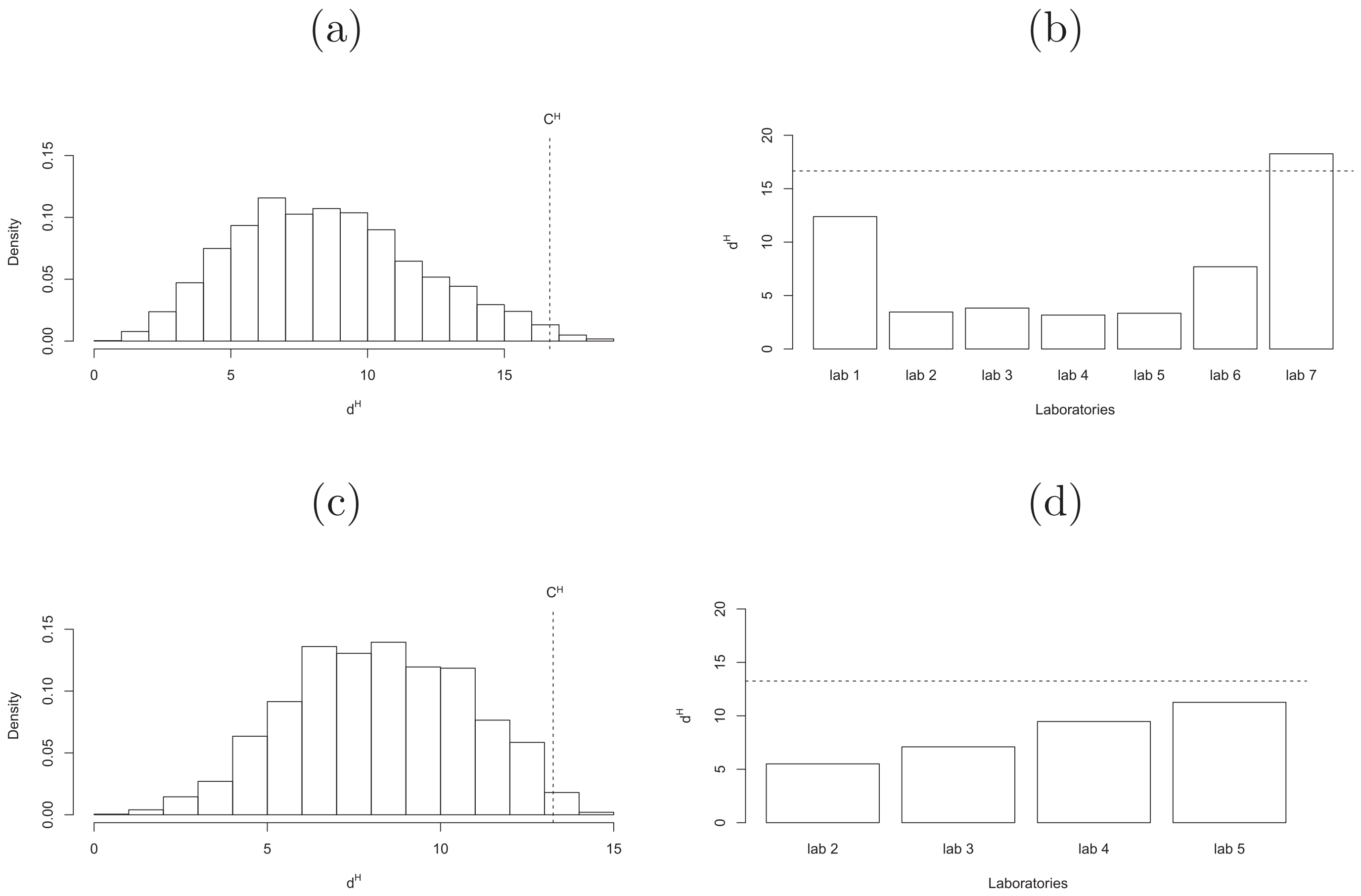

Fig. 8. (a) Histogram for the d^H sample statistic (first iteration), (b) d^H statistic for each laboratory and C_H for $\alpha = 0.01$ (first iteration), (c) Histogram for the d^H sample statistic (4th iteration), and (d) d^H statistic for each laboratory and C_H for $\alpha = 0.01$ (4th iteration).

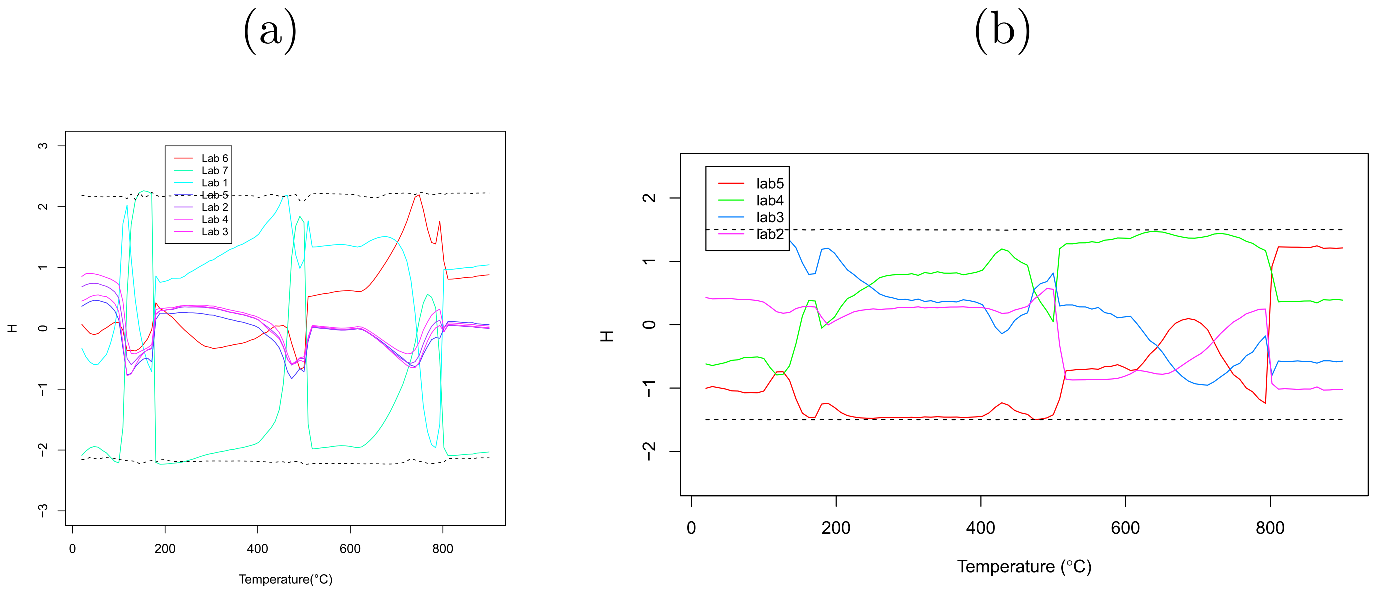

Fig. 9. (a) $H(t)$ functional statistic with 99% confidence bands (first iteration), (b) $H(t)$ functional statistic with 99% confidence bands (4th iteration).

where laboratory 7 becomes outlier: the region corresponding to the first degradation step of calcium oxalate is outside of the 99% confidence bands. Laboratories 1 and 6 are detected as outliers in the 2nd and 3rd methodology iterations. Fig. 9 b shows the H statistic corresponding to each laboratory in the 4th iteration. These laboratories correspond to the four consistent laboratories from the real data, as expected.

The same procedure is applied to test the hypothesis of repeatability, but instead using the K functional statistic and d^K test statistic. Laboratory 6 is detected as an outlier (Fig. 10a and Fig. 10b) with $\alpha = 0.01$. The region corresponding to the first, second, and third degradation steps of

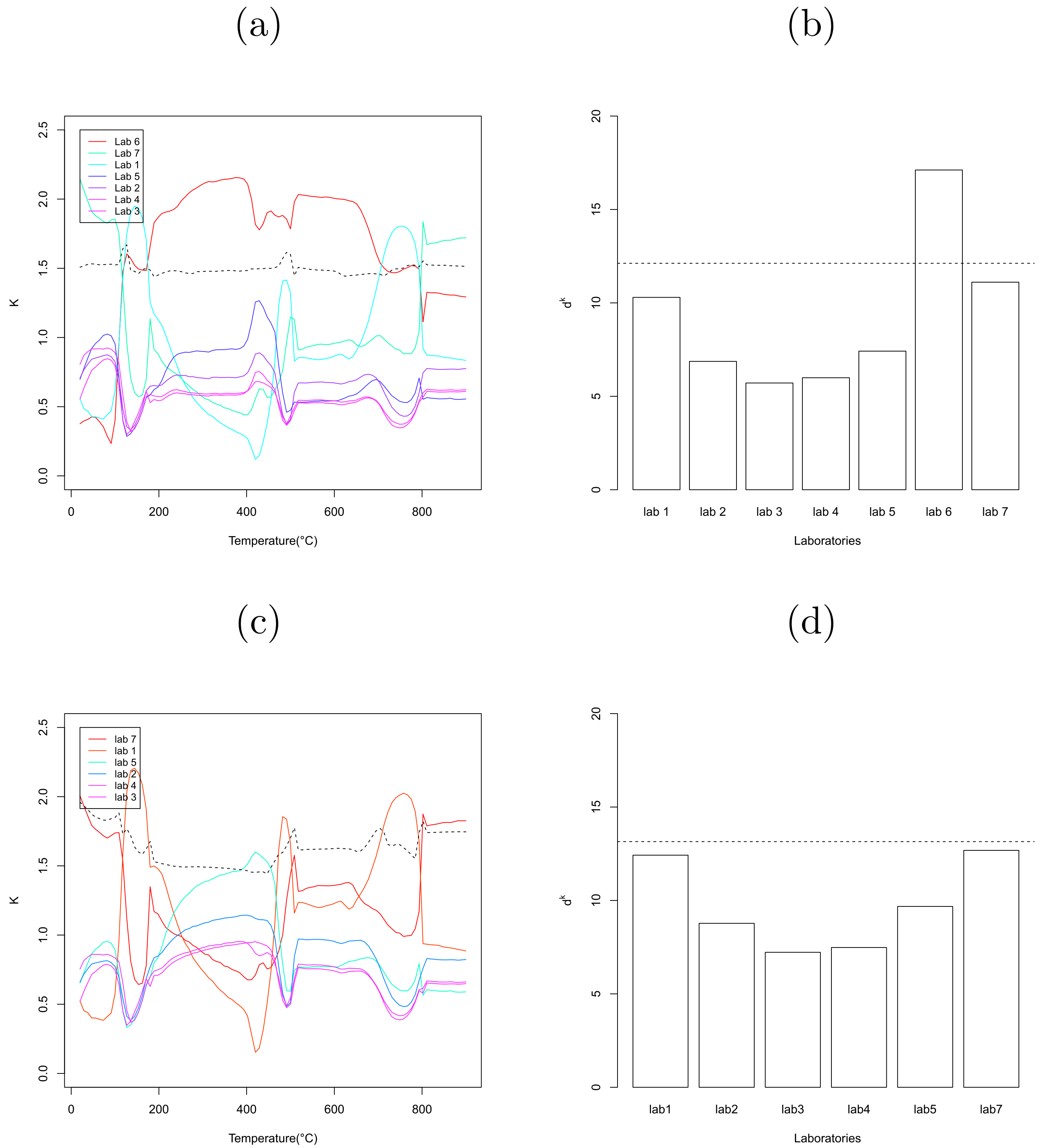

Fig. 10. (a) Functional K statistic for each laboratory and confidence band for $\alpha = 0.01$ (1st iteration), (b) d^K test statistic for each laboratory and critical level corresponding to $\alpha = 0.01$ (1st iteration), (c) Functional K statistic for each laboratory and confidence band for $\alpha = 0.01$ (2nd iteration), and (d) d^K test statistic for each laboratory and critical level corresponding to $\alpha = 0.01$ (2nd iteration).

calcium oxalate is outside of the 99% confidence bands. The iterative process stop at this iteration because no outlier laboratories are detected in the second methodology application (Fig. 10c and Fig. 10d). Although laboratories 1 and 7 are candidates to be outliers (Fig. 10c), there is not enough evidence at $\alpha = 0.01$ (Fig. 10d). Thus, at a confidence level of $\alpha = 0.01$ (commonly used in ILS) only laboratory 6, where a too old calibration had been used, is detected as an outlier taking into account the intralaboratory variability (repeatability).

In conclusion, laboratories 1, 6, and 7 have been properly identified as outliers. They provide inconsistent results if compared with the remaining four laboratories. The FDA H functional statistic and d^H test statistic have detected laboratories 1, 6 and 7 taking into account the

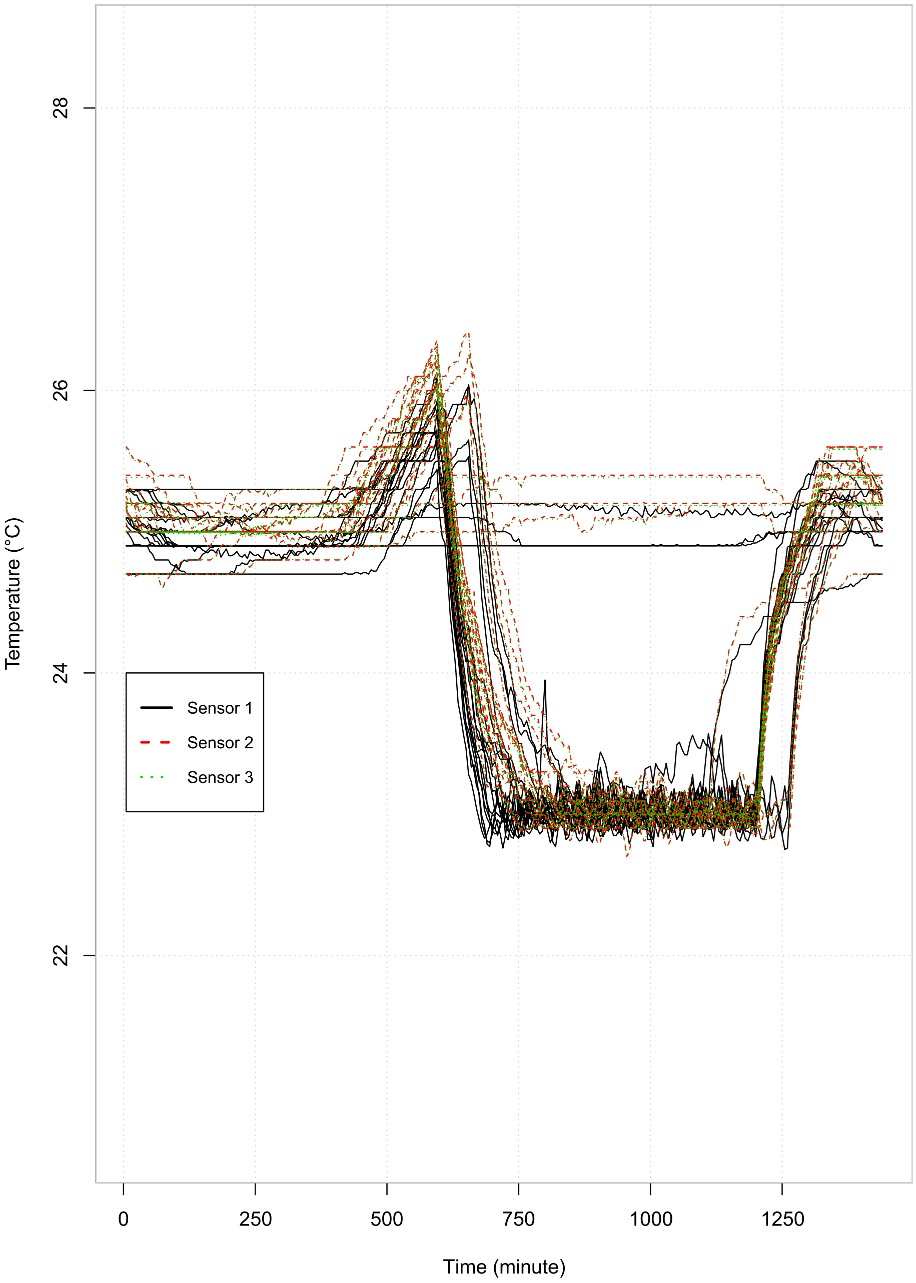

Fig. 11. Plan of the store, situation of the three temperature sensors associated with the three AHUs.

reproducibility hypothesis, whereas the FDA K and d^K statistics detect laboratory 6 in the context of the repeatability hypothesis. Thus, the proposed FDA extensions for Mandel's h and k statistics provide a new, useful way to perform outlier detection in ILS studies when dealing with functional data, without the additional step of representative feature extraction. Not only in the context of thermal analysis but in different domains of analytic chemistry, applied physics and engineering.

5.2. Identification of anomalous sensors

There is a record of the temperature of a room in a store (from August 8, 2017 to October 6, 2017) obtained from the measurements of three sensors. Taking into account that the three sensors are measuring the temperature of the same room, the objective is to detect if the results obtained by the three sensors are comparable. In particular, sensor 1 is used as a reference for correct operation (by the maintenance managers). Therefore, it is intended to evaluate whether sensors 2 and 3 provide consistent measurements. Fig. 11 shows the original data measured at a frequency of 5 min, the curves of temperature. The timetable of the store is comprised between 10 and 21, when the ventilation and air conditioning installation are running. The effect is the temperature lowering during working hours (in Panama the installations are of ventilation and air conditioning, never heating).

A short FDA descriptive analysis of temperature curves is developed by means of the ILS R package, developed by the authors so that practitioners could apply the proposed h and k FDA statistics. Functional means and variances are provided (for each laboratory or sensor, and also taking into account all the curves) as shown in Fig. 12.

In Fig. 13, the typical output for h and k scalar statistics are emulated for the functional case. Taking into account the intralaboratory variability (d_h statistic), sensor 1 is identified as an outlier if compared with sensors 2 and 3, as shown in Fig. 11. Otherwise, there are no differences when intralaboratory variability is studied (d_k statistic). Thus, the results support the statement of store managers: sensors 2 and 3 are

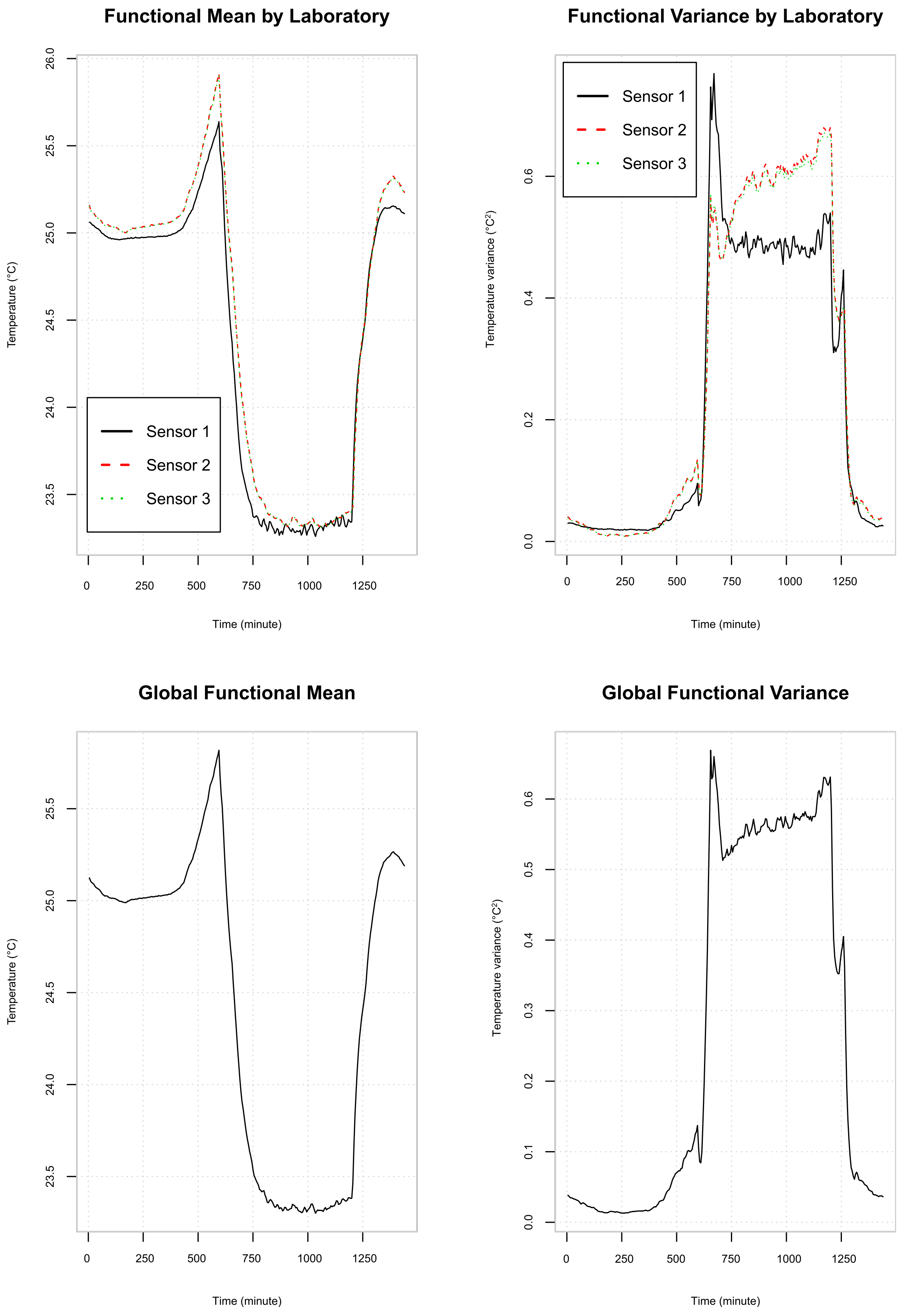

Fig. 12. Plan of the store, situation of the three temperature sensors associated with the three AHUs.

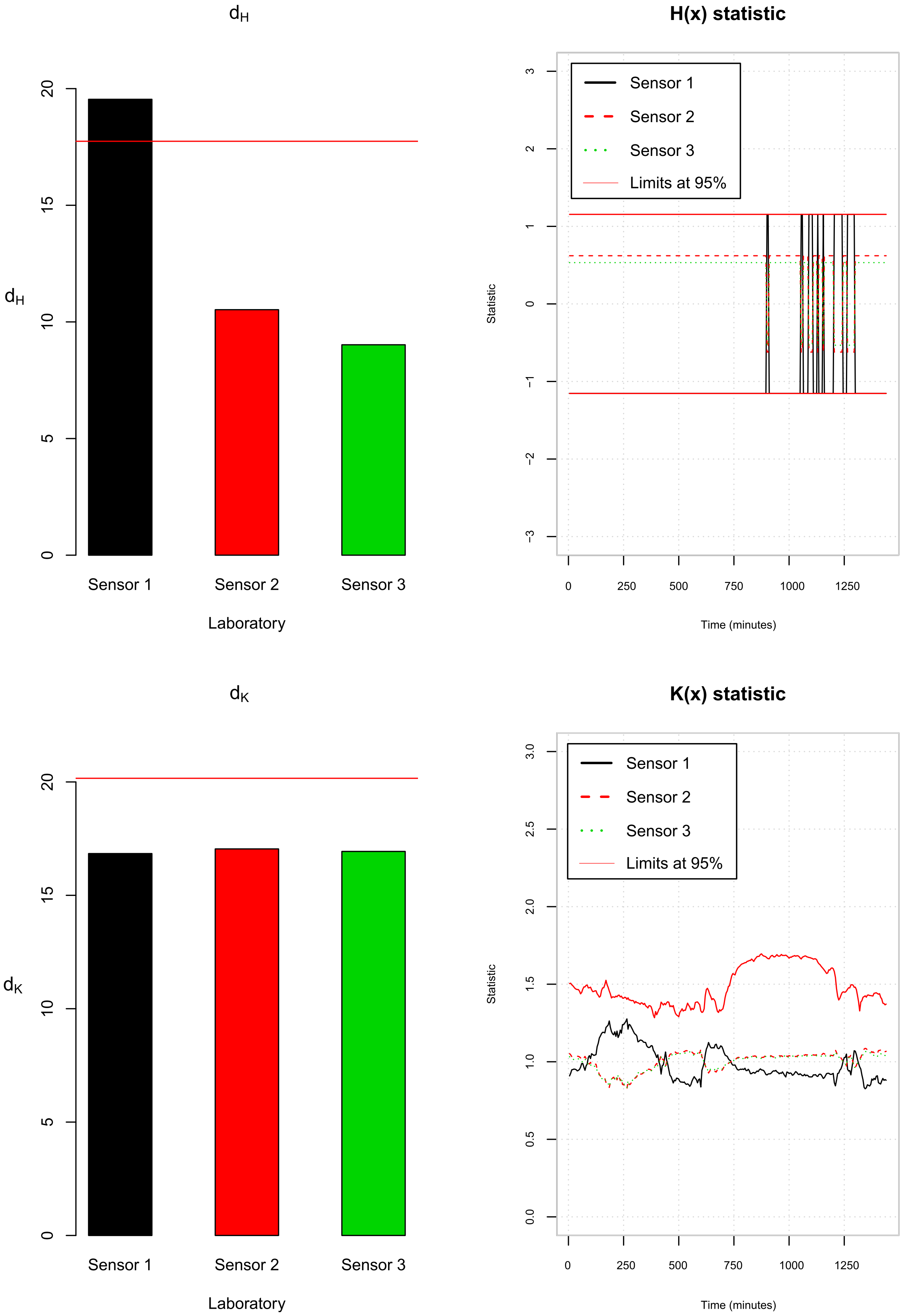

Fig. 13. Plan of the store, situation of the three temperature sensors associated with the three AHUs.

malfunctioning and corrective actions are needed.

6. Conclusions

Functional extensions for Mandel's h and k statistics have been proposed for dealing with curves obtained by experimental techniques. They identify the non-consistent laboratories in an ILS study using the information provided by the whole curve. Accordingly, an alternative approach to performing the outlier laboratory detection task in ILS studies when the experimental data are functional (curves) is provided, as opposed to the current scalar approaches. In fact, using this new proposed methodology, the feature extraction step from experimentally obtained curves is prevented. Moreover, another of the goals of the present FDA ILS procedure for outlier detection is to directly identifies inconsistent laboratories as outliers, while previous FDA attempts for ILS only identified atypical curves within each laboratory. Thus, the present work provides an improvement that works with functional data in the same manner as the popular scalar h and k approach. The present FDA approach consists of the calculation of functional (pointwise obtained in the interval where curves are defined) H and K statistics, from which d^H and d^K test statistics are obtained using the L_2 distance. The d^H test statistic is defined to test the reproducibility hypothesis, evaluating the interlaboratory variability, whereas the repeatability hypothesis (intralaboratory variability) is tested by d^K. The d^H and d^K probability distributions are estimated by a bootstrap procedure, and thus the test of the critical values C_H and C_K can be obtained as the quantiles corresponding to $\alpha = 0.01$. Note that, prior to the application of bootstrap resampling, the computation of functional data depth is required to remove outlier curves.

The simulation study performed using TG curves has provided information about the validity and power of the new FDA extensions of the h and k tests, in addition to the effects of sample size. Namely, the d^H and d^K p-values under the null hypotheses of reproducibility or repeatability are uniformly distributed, supporting the validity of the proposed test statistics. Further, all the simulated outlier laboratories have been successfully identified by the application of, on the one hand, H and d^H to detect interlaboratory changes and, on the other hand, K and d^K to identify intralaboratory differences. Concerning the test power, when the number of laboratory replicates is not very high ($n = 10$) the d^H test power is higher than its h scalar counterpart when features such as time/temperature to lose the $5wt$% are extracted. Depending on the feature extracted, the power of functional approach can be higher or not than the corresponding scalar approach. However, the d^K power is always higher than the corresponding scalar k statistics, whatever the extracted feature is. Finally, as expected, increasing the number of laboratory replicates significantly increases the functional extension tests powers.

The FDA approach application to thermogravimetric real data has shown that the proposed methodology is able to detect all the real outlier laboratories. For this task, the use of both d^H and d^K test statistics or H and K functional statistics has been necessary. At this point, note that the use of H and K functional statistics also shows the time/temperature interval where the laboratories become outliers. The identification of these intervals can provide relevant information about the physical or chemical processes that induce more differences between the measurements of the different laboratories.

The application of d^H and d^K test statistics or H and K functional statistics to environmental variables that are continuously monitored has provided for the successful identification of non-consistent indoor temperature sensors. These FDA approximations can be applied in the context of thermal comfort, energy efficiency, and also variable monitoring in environmental science and technology.

The above-mentioned results derived from simulation and real data studies support the use of this methodology, not only in the context of thermal analysis, but also in different domains of analytic chemistry, applied physics, and engineering.

Acknowledgments

The research of Salvador Naya, Javier Tarrío-Saavedra and Rubén Fernández-Casal has been supported by MINECO grants MTM2014-52876- R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de InvestigaciÃ^{3}n de Galicia ED431G/01), all of them through the ERDF. The research of Miguel Flores has been partially supported by Grant PII-DM-002-2016 of Escuela Politécnica Nacional of Ecuador. The authors also thank Sonia Zaragoza, Fridama, Nerxus and $\sum$ qus companies their help in providing real data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chemolab.2018.03.016.

References

[1] S. Naya, J. Tarrío-Saavedra, J. López-Beceiro, M. Francisco-Fernández, M. Flores, R. Artiaga, Statistical functional approach for interlaboratory studies with thermal data, J. Therm. Anal. Calorim. 118 (2014) 1229–1243.
[2] E. Hund, D.L. Massart, J. Smeyers-Verbeke, Inter-laboratory studies in analytical chemistry, Anal. Chimica Acta 423 (2000) 145–165.
[3] P.-T. Wilrich, Critical values of mandelsh and k, the grubbs and the cochran test statistic, AStA Adv. Stat. Anal 97 (2013) 1–10.
[4] E. Maier, P. Quevauviller, B. Griepink, Interlaboratory studies as a tool for many purposes: proficiency testing, learning exercises, quality control and certification of matrix materials, Anal. Chimica Acta 283 (1993) 590–599.
[5] Y. Vander Heyden, J. Smeyers-Verbeke, Set-up and evaluation of interlaboratory studies, J. Chromatogr A 1158 (2007) 158–167.
[6] I. O, For Standardization, Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method, 1994. International Organization For Standardization.
[7] P. Kelly, Outlier detection in collaborative studies, J. Assoc. Off. Anal. Chem. 73 (1990) 58–64.
[8] S. Uhlig, P. Lischer, Statistically-based performance characteristics in laboratory performance studies, Analyst 123 (1998) 167–172.
[9] V. Dvorkin, Data processing in the interlaboratory test by analysis of covariance, Chemom. Intell. Lab. Syst 22 (1994) 127–146.
[10] W. Cochran, The distribution of the largest of a set of estimated variances as a fraction of their total, Ann. Eugen 11 (1941) 47–52.
[11] F. Grubbs, Sample criteria for testing outlying observations, Ann. Math. Stat. 21 (1950) 27–58.
[12] F. Grubbs, G. Beck, Extension of sample sizes and percentage points for significance tests of outlying observations, Technometrics 14 (1972) 847–854.
[13] U. Graf, P.T. Wilrich, H.J. Henning, K. Stange, Formeln und Tabellen der angewandten mathematischen Statistik, Springer, 1997.
[14] P.L. Davies, Statistical evaluation of interlaboratory tests, Fresenius' Zeitschrift anal, Chemie 331 (1988) 513–519.
[15] A.M. Committee, Robust statistics–how not to reject outliers. part 1. basic concepts, Analyst 114 (1989) 1693–1697.
[16] J.C. Miller, J.N. Miller, Statistics for Analytical Chemistry, John Wiley and Sons, New York, NY, 1988.
[17] W. Youden, Industrial Quality Control, 1959.
[18] P. Minkkinen, Estimation of variance components from the results of interlaboratory comparisons, Chemom. intell. lab. syst 29 (1995) 263–270.
[19] E. ASTM, 691-99. standard practice for conducting an interlaboratory study to determine the precision of a test method, Annu. Book ASTM Stand. 14 (2003) 203–224. ASTM E691-16, Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method, ASTM International, West Conshohocken, PA, 2016, www.astm.org.
[20] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer, 2006.
[21] W. Saeys, B. De Ketelaere, P. Darius, Potential applications of functional data analysis in chemometrics, J. Chemom. 22 (2008) 335–344.
[22] J.S. Morris, P.J. Brown, R.C. Herrick, K.A. Baggerly, K.R. Coombes, Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models, Biometrics 64 (2008) 479–489.
[23] G. Montana, M. Berk, T. Ebbels, Modelling Short Time Series in Metabolomics: a Functional Data Analysis Approach, in: Software Tools and Algorithms for Biological Systems, Springer, 2011, pp. 307–315.
[24] M. Plichta, S. Heinzel, A.-C. Ehlis, P. Pauli, A. Fallgatter, Model-based analysis of rapid event-related functional near-infrared spectroscopy (nirs) data: a parametric validation study, Neuroimage 35 (2007) 625–634.
[25] J. Tarrio-Saavedra, S. Naya, M. Francisco-Fernandez, R. Artiaga, J. Lopez-Beceiro, Application of functional anova to the study of thermal stability of micro-nano silica epoxy composites, Chemom. intell. lab. syst 105 (2011a) 114–124.

[26] J. Tarrio-Saavedra, S. Naya, M. Francisco-Fernandez, J. Lopez-Beceiro, R. Artiaga, Functional nonparametric classification of wood species from thermal data, J. Therm. Anal. Calorim. 104 (2011b) 87–100.

[27] M. Francisco-Fernandez, J. Tarrio-Saavedra, A. Mallik, S. Naya, A comprehensive classification of wood from thermogravimetric curves, Chemom. Intell. Lab. Syst 118 (2012) 159–172.

[28] M. Francisco-Fernández, J. Tarrío-Saavedra, S. Naya, J. López-Beceiro, R. Artiaga, Statistical classification of early and late wood through the growth rings using thermogravimetric analysis, J. Therm. Anal. Calorim. 127 (2017) 499–506.

[29] M. Francisco-Fernández, J. Tarrío-Saavedra, S. Naya, J. López-Beceiro, R. Artiaga, Classification of wood using differential thermogravimetric analysis, J. Therm. Anal. Calorim. 120 (2015) 541–551.

[30] J. Janeiro-Arocas, J. Tarrío-Saavedra, J. López-Beceiro, S. Naya, A. López-Canosa, N. Heredia-García, R. Artiaga, Creep analysis of silicone for podiatry applications, J. Mech. Behav. Biomed. Mater 63 (2016) 456–469.

[31] B.J.A. Mertens, Logistic regression modeling on mass spectrometry data in proteomics case-control discriminant studies, in: S. Datta, B.J.A. Mertens (Eds.), Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry, Springer International Publishing, 2017, pp. 213–238.

[32] J.O. Ramsay, Functional Data Analysis, Wiley Online Library, 2006.

[33] R. Fraiman, G. Muniz, Trimmed means for functional data, TEST 10 (2001) 419–440.

[34] A. Cuevas, M. Febrero, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Comput. Stat. Data Anal. 51 (2006) 1063–1074.

[35] P. Raña, G. Aneiros, J. Vilar, P. Vieu, Bootstrap confidence intervals in functional nonparametric regression under dependence, Electron. J. Stat 10 (2016) 1973–1999.

[36] G. Aneiros, J.M. Vilar, R. Cao, A.M. San Roque, Functional prediction for the residual demand in electricity spot markets, IEE Trans. Power Syst 28 (2013) 4201–4208.

[37] W. Cho, S. Kim, S. Park, Human action classification using multidimensional functional data analysis method, in: Adv. Multimed. Ubiquitous Eng, Springer, 2016, pp. 279–284.

[38] J. Cuesta-Albertos, M. Febrero-Bande, A simple multiway anova for functional data, TEST 19 (2010) 537–557.

[39] G. Aneiros, J. Vilar, P. Raña, Short-term forecast of daily curves of electricity demand and price, Int. J. Electr. Power Energy Syst. 80 (2016) 96–108.

[40] R.L. Buckner, D. Head, J. Parker, A.F. Fotenos, D. Marcus, J.C. Morris, A.Z. Snyder, A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume, Neuroimage 23 (2004) 724–738.

[41] C.B. Embling, J. Illian, E. Armstrong, J. van der Kooij, J. Sharples, K.C. Camphuysen, B.E. Scott, Investigating fine-scale spatio-temporal predator–prey patterns in dynamic marine ecosystems: a functional data analysis approach, J. Appl. Ecol. 49 (2012) 481–492.

[42] R. Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.

[43] S. Graves, G. Hooker, J. Ramsay, Functional Data Analysis with R and MATLAB, Springer, New York, 2009.

[44] M. Febrero, P. Galeano, W. González-Manteiga, Outlier detection in functional data by depth measures, with application to identify abnormal nox levels, Environmetrics 19 (2008) 331–345.

[45] M. Febrero-Bande, M. de la Fuente, Statistical computing in functional data analysis: the r package fda.usc, J. Stat. Softw 51 (2012).

[46] M. Flores, S. Naya, J. Tarrío-Saavedra, R. Fernández-Casal, Functional data analysis approach of mandel's h and k statistics in interlaboratory studies, in: G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu (Eds.), Functional Statistics and Related Fields, Springer, 2017, pp. 123–130.

[47] K. Daly, E. Gai, J. Harrison, Generalized likelihood test for fdi in redundant sensor configurations, J. Guid. Control Dyn 2 (1979) 9–17.

[48] I. Barbeito, S. Zaragoza, J. Tarrío-Saavedra, S. Naya, Assessing thermal comfort and energy efficiency in buildings by statistical quality control for autocorrelated data, Appl. energy 190 (2017) 1–17.

[49] A. Cuevas, M. Febrero, R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Comput. Stat. 22 (2007) 481–496.

[50] M. Flores, ILS: Interlaboratory Study, 2016. R package version 0.1.0.

[51] J. Tarrío-Saavedra, J. López-Beceiro, S. Naya, M. Francisco-Fernández, R. Artiaga, Simulation study for generalized logistic function in thermal data modeling, J. Therm. Anal. Calorim. 118 (2014) 1253–1268.

[52] E.L. Lehmann, J.P. Romano, Testing Statistical Hypotheses, Springer Science & Business Media, 2006.