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 box-plots [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 intra-laboratory 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.

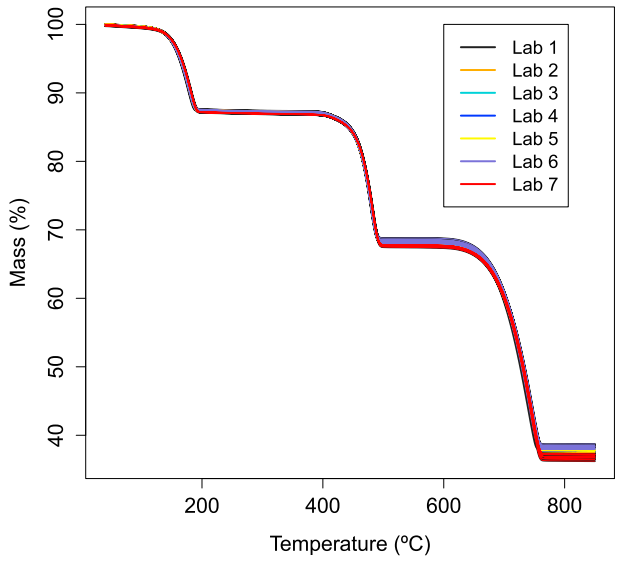


**Fig. 1.** TG data obtained by the different emulated laboratories.

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).