ILS: An R package for statistical analysis in Interlaboratory Studies

Article *in* Chemometrics and Intelligent Laboratory Systems · August 2018

DOI: 10.1016/j.chemolab.2018.07.013

ILS: An R package for statistical analysis in Interlaboratory Studies

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

[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*
[d] *Universidad de las Fuerzas Armadas ESPE, Quito, Ecuador*

Abstract

In this paper we present an R package with routines to perform Interlaboratory Studies (ILS). The aim of the ILS package is to detect laboratories that provide not consistent results, working simultaneously with different test materials, from the perspective of the Univariate Data Analysis and the Functional Data Analysis (FDA).

The ILS package estimates the Mandel's h and k univariate statistics, based on the ASTM E691 and ISO 5725-2 standards, to identify laboratories that provide significantly different results. Cochran and Grubbs tests to evaluate the presence of outliers are also available. In addition, Analysis of Variance (ANOVA) techniques are provided, including the Tukey and F tests to evaluate differences between the means for the corresponding test variable.

One of the novelties of this package is the incorporation of tools to perform an ILS from a functional data analysis approach. Accordingly, the functional nature of the data obtained by experimental techniques corresponding to analytical chemistry, applied physics and engineering applications (spectra, thermograms, and sensor signals, among others) is taking into account by implementing the functional extensions of Mandel's h and k statistics. For this purpose, the

*Corresponding author. Phone: +34 981167000 (ext. 3210). e-mail:javier.tarrio@udc.es

ILS package also estimates the functional statistics $H(t)$ and $K(t)$, as well as the d_H y d_K test statistic, which are used to evaluate the repeatability and reproducibility hypotheses where the critical c_h and c_k values are estimated by using a bootstrap algorithm.

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

1. Introduction

An Interlaboratory Study (ILS) can be defined as a control procedure to evaluate the performance of a group of laboratories through a collaborative trial [1, 2]. In an Interlaboratory Study, an adequate number of laboratories are chosen to participate in the experiment with the aim of analysing the samples and obtain results.

Participating laboratories receive samples (previously homogenized or to be homogenized by the laboratories) for analysis, then, the measurements results of the laboratories are evaluated according to the degree of data variability. Some of the most common factors that may be a cause of variability are: the equipment of laboratories, operators, materials, temperature and humidity, among others.

Several univariate statistical techniques are frequently applied to study the consistency of test results from the different laboratories that participate in an ILS. Standard ASTM E-691 (Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method) recommends applying only one graphical technique from Mandel's k and h statistics [2], while ISO 5725-2 (Accuracy —trueness and precision— of measurement methods and results) recommends, in addition to the graphic technique, to use the Cochran and Grubbs tests [1].

Additionally, through Analysis of Variance (ANOVA), the F test can be applied taking the "laboratory" qualitative variable as the only factor to test the hypothesis of equal means between laboratory results. Moreover, the variance of repeatability and reproducibility are also estimated when ANOVA analysis is performed. On the other hand, the Tukey test is used to evaluate which laboratory provides different results taking into account the differences between the means of a test variable.

To perform consistency tests for the repeatability and reproducibility hypotheses, as well as for the detection of outliers, the values of the statistics should be compared with their corresponding critical values. If these are greater, inconsistency is detected in the results of laboratories. ISO 5725-2 provides some critical values depending on the number of laboratories L, number of measurements n and level of significance α.

At present, both ISO 5725 and ASTM-E691 do not provide a methodology for performing an ILS when data are functional, this is, in the case where the test results are curves (functional data). Functional Data Analysis (FDA) is a relatively new field of the statistic that takes curves as unit of analysis, also surfaces, and volumes defined in a continuum (such as time, or frequency's domain). Considering the recent advances in computing science, and the increasing amount of data generated by experimental techniques and sensors, the FDA has had a great development in recent years. In fact, we have many statistical methodologies that have been developed and extended to the functional case, such as exploratory analysis, regression, classification, analysis of variance, and time series [3–5]. In the specific case of ILS, FDA extensions for Mandels's h and k has been proposed and described by Flores et al. [6], in addition to other works where the FDA descriptive analysis had been introduced for ILS studies [7].

The aim of the `ILS` package [8] is, on the one hand, to facilitate the use of new tools in the FDA context and, on the other hand, to provide a comprehensive set of the more used univariate outlier test for ILS with scalar response. It is important to note that FDA techniques for Interlaboratory Studies are based on the proposals of Naya et al. [7], Flores et al. [9] and, above all, Flores et al. [6] whereby new functional extensions of h and k statistics are introduced for identifying non consistent laboratories. The functions that have been implemented in the `ILS` package can determine the Mandel's h and k statistics both in a graphical and analytical way, using a functional approach. These test statistics have also been implemented to facilitate the implementation of Repeatability and Reproducibility studies (r&R) when the data are functional. In addition, the ILS package apply the methods suggested by the norms ASTM E-691 and ISO 5725-2 for the scalar case. The ILS package is available on the Comprehensive R Archive Network at http://CRAN.R-project.org/package=ILS.

The present ILS library implement and call some of their routines in order to perform outlier detection in the framework of the Interlaboratory Studies. Thus, regarding to ILS with scalar response, there are some interesting and useful computational tools in R software. Namely, the metRology package estimates the uncertainty of the measurement, and the required statistical calculations for Interlaboratory Studies [10], whereas multcomp performs analysis of variance (ANOVA) through F and Tuckey tests [11]. On the other hand, due to the exponential increasing of FDA available techniques, there are also a continuously increasing number of R libraries devoted to this branch of statistics. Among all of them, the most important and used packages (and on which the present proposal is based) are fda.usc [12], that implements outlier detection techniques and functional ANOVA, among other tools for FDA, and the fda [13]. The present ILS package uses the applications of the multcomp and fda.usc packages before mentioned.

This work is organized as follows. In section(2), we describe examples of Interlaboratory Studies in which four sets of experimental data are obtained. Then, the ILS package is used to summarize two of these sets. In section(3), the functionality of the ILS package is illustrated through a standard ILS procedure using the Glucose dataset. Further, in section(4), the TG and DSC datasets (composed by thermogravimetric and colorimetric curves of calcium oxalate, respectively) are used to show the ILS package utilities when experimental data are curves (functional). Finally, the principal remarks of this study are summarized in the conclusion section.

2. Examples of Interlaboratory Studies

An Interlaboratory Study evaluates the analytical methods performed by laboratories, either for the evaluation of the efficiency of the laboratories involved, or for the performance of an experimental procedure, or for the validation of a standard guideline. For example, to show the application of consistency test, the `ILS`, package contains the `Glucose` dataset, avalaible on ASTM E-691 [2] that corresponds to the results of a clinical test. Likewise, from a study of the properties of the calcium oxolate material, three datasets (`IDT`, `TG`, `DSC`) were obtained, and they were incorporated into the package. These latter datasets have been extensively described in Naya et al. [7] and Flores et al. [6].

2.1. Clinical study of blood glucose measurement

The `Glucose` dataset corresponds to the serum glucose test (measurements of the concentration of glucose in the blood used to control the diabetes). In the study, eight laboratories where involved, and five different tests were performed on blood samples labelled with different references, ranging from a low sugar content to a very high one. Three replicates were obtained for each sample.

Each of these laboratories measured five different concentration levels (A, B, C, D, E) of a given material, and at each of these levels, three measurements were taken (3 replicates). Each laboratory provided a total of 15 measurements (3 for each level), therefore, with 8 laboratories involved, 120 measurements were obtained.

In order to access this dataset, the `ILS` package installing and loading is required. Once loading is performed, the `Glucose data.frame` object is called using the following instructions.

```
R> library("ILS")
R> data("Glucose", package = "ILS")
```

The first step to perform an analysis with the ILS package consist on using the function `lab.qcdata()` (quality control data) that receives a `data.frame` as an argument. The first column of the data frame must contain the response variable, the second column accounts for the index of repetition for each laboratory, the third column includes the index of the material at which the test was performed, while the fourth column includes the index of the laboratory where the procedure was performed.

Afterward, the `qcdata` object, corresponding to the `lab.qcdata()` class, is developed. The descriptive statistics information of the dataset can be summarized by using the `summary()` function. Figure(1) shows the results of all laboratories.

```
R> qcdata <-  lab.qcdata(Glucose)
R> summary(qcdata)
```

```
       x           replicate material   laboratory
 Min.   : 39.02   1:40      A:24     Lab1   :15
 1st Qu.: 78.45   2:40      B:24     Lab2   :15
 Median :135.03   3:40      C:24     Lab3   :15
 Mean   :149.09             D:24     Lab4   :15
 3rd Qu.:196.66             E:24     Lab5   :15
 Max.   :309.40                      Lab6   :15
                                     (Other):30
```

```
R> plot(qcdata, ylab = "Laboratory", xlab = "Glucose concentration in blood")
```

In figure(1), it can be noted that the blood glucose level increases from material A to D and there is more variability between the results for each laboratory from the material C to material E.

In order to calculate the graphical and analytical statistics for the scalar (univariate) case, first, the function `lab.qcs()` (quality control statistics)has to be used. This function returns the estimation of the statistical required measures (mean, variance, etc.) for estimating the Mandel's h and k statistics, as well as the required measures to perform the Cochran and Grubbs tests.

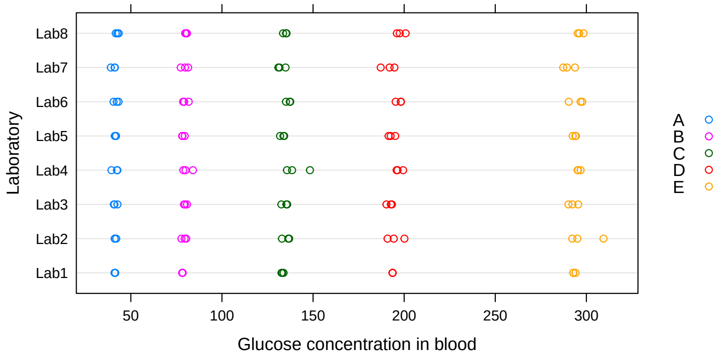

Figure 1: Measurements of glucose concentration in blood used to control diabetes.

The `qcdata` object uses the `lab.qcs()` function to create the `qcstat` object that estimates both the mean and the global deviation from the results of all laboratories and all materials. In addition, the repeatability deviation (S_r) , the deviation between the means of laboratories (S_B), and the reproducibility deviation (S_R) for each material are estimated. More information about definitions and calculation methods can be obtained in [1, 2].

```
R> qcstat <-  lab.qcs(qcdata)
R> summary(qcstat)

Number of laboratories:  8
Number of materials:  5
Number of replicate:  3
Summary for Laboratory (means):
       Lab1      Lab2      Lab3      Lab4      Lab5      Lab6      Lab7      Lab8
A  41.28333  41.44000  41.45000  41.45667  41.46333  42.02000  40.45667  42.57667
B  78.31667  79.23333  79.90333  80.96333  78.69000  79.89333  79.51667  80.34667
C 133.19667 135.40667 134.59000 140.83000 133.26667 136.61667 132.49333 134.71000
D 193.65000 195.10667 192.09000 197.21333 193.05000 197.24333 191.26000 198.12333
E 293.25333 298.91667 292.67000 295.82000 293.56333 294.95667 290.13667 296.62000


Summary for Laboratory (Deviations):
       Lab1      Lab2      Lab3      Lab4      Lab5     Lab6     Lab7      Lab8
A 0.2230097 0.4850773 1.0608016 1.8117763 0.3666515 1.408119 1.247811 0.8224557
B 0.1582193 1.3268509 0.8303212 2.7660863 0.7754354 1.636592 2.059935 0.5064912
C 0.5909597 2.1679791 1.7287857 6.6200227 1.1987215 1.287025 2.124296 1.0343597
D 0.0600000 4.6824068 1.5932043 1.9365519 1.8826311 1.649616 3.817709 2.4637844
E 0.7266590 9.1869055 2.7101107 0.8835723 0.9543759 4.034282 3.304184 1.6479078
```

```
Summary for Material:
       mean         S      S_r       S_B      S_R
A  41.51833 0.5543251 1.063224 0.6061274 1.058783
B  79.60792 0.8664835 1.496071 0.8627346 1.495481
C 135.13875 1.9071053 2.750879 2.6566872 3.478919
D 194.71708 1.4262962 2.625065 2.5950046 3.365713
E 294.49208 2.8067799 3.934974 2.6931364 4.192334
```

In figure 2, the values of S (the global deviation of all laboratories), S_r (the repeatability's deviation), S_R (reproducibility's deviation) and S_B (the deviation between the means of the laboratories) are shown for each material. A greater variability can be noted from material C to material E. Materials C and D have a greater variability between the results of the laboratories (S_R) and within them (S_r).

2.2. Characterization of materials by thermogravimetric analysis

In [7], 105 samples of calcium oxalate were analysed by Thermogravimetric (TG) techniques, obtaining 105 TG curves showing the loss of oxalate mass as a function of temperature when the oxalate samples were heated at 20°C/min. In addition, 90 samples of Calcium Oxalate were analysed by Differential Scanning Calorimetry (DSC) thermal technique, obtaining 90 DSC curves that determinate, from an SDT instrument, the difference of energy between a reference and the oxalate sample. We can observe the exchange of energy between the sample and the reference as a function of temperature when the latter vary as a linear function of the time defined by a slope of 20°C/min. Two sets of data were generated from the results, a `TG` dataset, obtained from 7 different laboratories, and a `DSC` dataset, obtained from 6 different laboratories. In each laboratory 15 curves were analysed in 1000 observations.

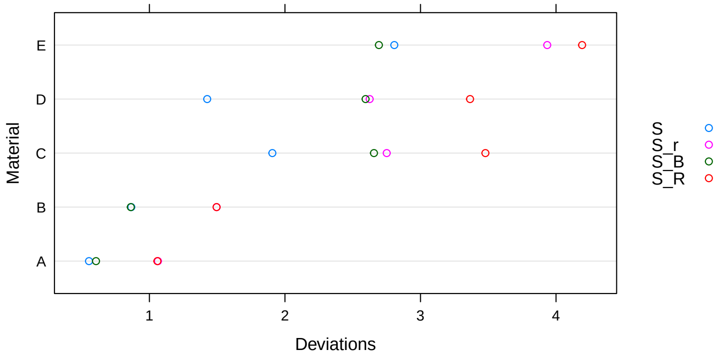

Figure 2: Measures of variability for each material obtained from the results of all the laboratories.

In addition, from the TG curves, a third set of data called IDT (Initial Decomposition Temperature) was obtained, this is a parameter defined by the temperature at which the studied material losses the 5% of its weight when it is heated at a constant rate. This dataset is composed of the IDT values of calcium oxalate obtained from 7 different laboratories that analyses (each one) 15 oxalate samples. This dataset is an example of ILS study with scalar response, obtained by extracting just only one representative feature from the TG curve. It is important to stress that when a feature extraction process is performed, there is the risk of loosing relevant information and thus making erroneous findings.

Laboratories 1, 6, and 7 presented non-consistent results. In laboratory 1 a Simultaneous Thermal Analyzer (STA) was used with an out of phase calibration program. In laboratory 6 we used a simultaneous SDT analyser with an old calibration, and finally, in laboratory 7 we used a simultaneous SDT analyser with a bias into the temperature with respect to the real values (2°C displaced with respect to the melting point of the zinc).

For the estimation of the functional statistics (for the performance of the graphical and analytical methods), the procedure was the same as for the scalar case. The ILS package provides the ils.fqcdata() (functional quality control data) function that developes an object that has the structure of a data.frame, in which each row represents a test result. The size of the data.frame is $n \times l$, where n is the number of replicates performed by each laboratory, and l accounts for the number of laboratories that participate in the study. Specific functions were implemented to make plots and summaries of these type of objects.

Then, the function ils.fqcs() (functional quality control statistical) is needed for the estimation of some important statistical functional measures: functional mean, functional variance, etc. These are necessary for the estimation of the $H(t)$ and $K(t)$ and the d_H and d_K statistics.

To built an object of class ils.fqcdata, first we defined the grid in which the observations will be obtained. In this case, the 1000 points that compose the grid accounts for temperatures ranging from 40°C to 850°C. In Figure 3, the TG curves are presented. From the fqcdata object, the fqcstat object was performed.

```
R> data(TG, package = "ILS")
R> delta <- seq(from = 40 ,to = 850 ,length.out = 1000 )
R> fqcdata <-  ils.fqcdata(TG, p = 7, argvals = delta)
R> main <- "TG curves obtained from calcium oxalate"
R> xlab <- "Temperature (C)"
R> ylab <- "Mass (\%)"
R> plot(x = fqcdata, main = main, xlab = xlab , ylab = ylab,
+       legend = TRUE,x.co = 20, y.co = 90)
```

```
R> fqcstat <- ils.fqcs(fqcdata)
R> summary(fqcstat)
```

```
Number of laboratories:  7
Number of replicates:  15
```

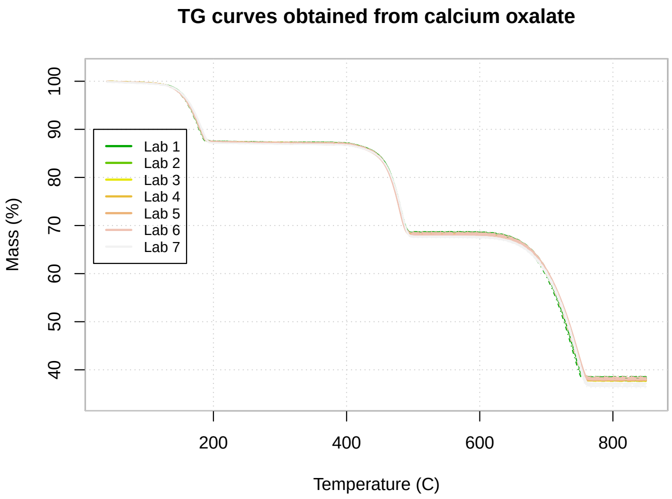

Figure 3: TG curves obtained from calcium oxalate.

```
R> xlab <- "Temperature (C)"
R> ylab <- "Mass (\%)"
R> plot(fqcstat, xlab = xlab, ylab = ylab)
```

The `plot()` function creates a panel with four graphs, in the first row we have the means and functional variances for each laboratory, while in the second row the mean and global functional variance are plotted. Figure(4) shows the different functional means and functional variances for each laboratory, as well as the overall mean and the overall variance corresponding to the complete `TG` dataset.

3. Interlaboratory Studies: Standard Approach

The `ILS` package provides two groups of functions made to detect outlying individual results (outlying replicates) and outlying laboratories: both for the scalar and the functional cases (Table 1). The `ILS` package offers graphical and analytical procedures (statistical hypothesis test) for this purpose.

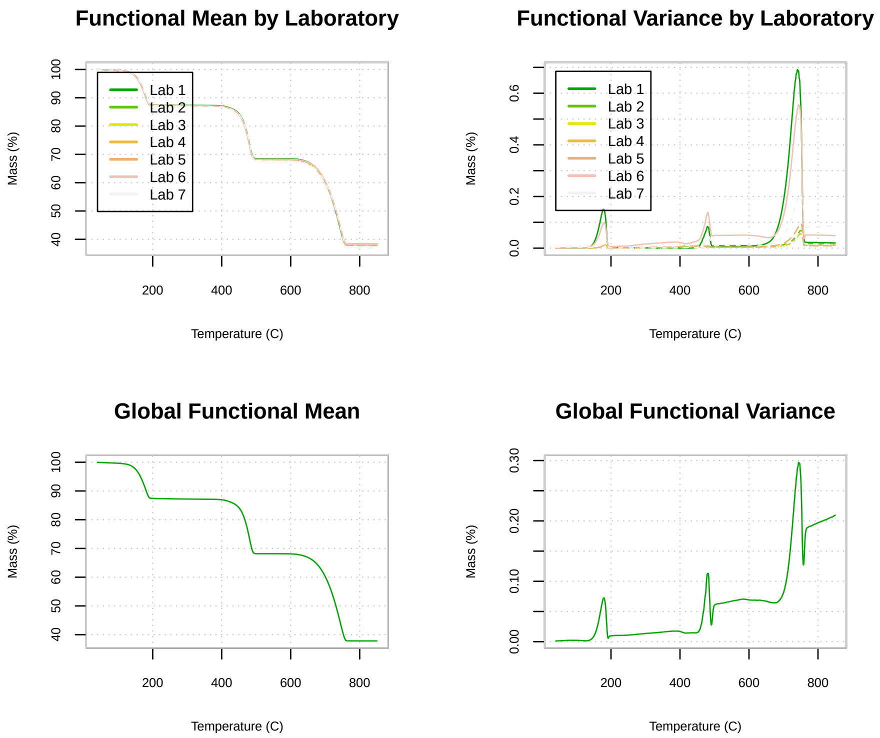

Figure 4: Functional descriptive statistics: functional means and functional variances for each laboratory as well as the overall functional mean and the overall functional variance corresponding to the TG curves obtained from calcium oxalate.

Table 1: Functions incorporated to the ILS package to perform Interlaboratory Studies with a scalar and functional approach.

Scalar	Technique	Function	Description
approach	Plot	`h.qcs`, `k.qcs`	Mander's h and k statistics
	Test	`test.cochran`,	Cochran test
		`test.grubbs`	Grubbs test
		`lab.aov`	ANOVA
Functional	Plot	`h.fqcs`, `k.fqcs`	Mandel's $H(x)$ y $K(x)$ functional statistics
approach	Test	`mandel.fqcs`	Mandel's functional test

As above mentioned, among the methodologies used to evaluate the consistency of laboratory results, we must highlight the r&R studies, which quantify the variability between laboratories (reproducibility) and variability between results (repeatability). The repeatability is the variability between the results of the independent tests obtained for each individual laboratory, i.e. the evaluation of the variability produced by the measurement system since the results are obtained by a single operator in each laboratory and in a short interval of time. On the other hand, the reproducibility refers to the variability between the results of individual tests obtained in different laboratories, allowing to determine the bias.

Accordingly with the repeatability and reproducibility concepts, Mandel's h and k statistics are used in ILS to detect laboratories that provide inconsistent results. The h statistic explains the variability between the laboratories, that is, estimates the bias, which is the difference of the means of each laboratory with respect to the global mean, while the k statistic estimates the variability within the laboratories, comparing the repeatability corresponding to each laboratory.

The decision rule for detecting whether a laboratory is inconsistent is based on the comparison between the value of the h or k statistic and the critical value calculated with a significance level of 0.5%, which is the one recommended by ASTM E-691 [2].

On the other hand, the `ILS` package performs the Cochran test to examine the consistency within a laboratory, whereas the Grubbs test is commonly used to examine consistency between laboratories. The Grubbs test can also be used as a consistency test for the results obtained in a laboratory using identical materials. These tests are recommended by ISO 5725-2 [1].

3.1. Consistency tests

The basic statistical model proposed on ISO 5725-2 that estimate the accuracy and precision of an analytical method is:

$$y = m + B + \epsilon \tag{1}$$

Table 2: ANOVA approach for the estimation of S_r^2 and S_B^2.

Source	Mean squares	Estimate of
Laboratory	$MS_B = \frac{\sum_{i=1}^{L} n_i(\bar{y}_i - \bar{(\bar{y})}^2)}{(L-1)}$, $S_B^2 = \frac{MS_B - MS_r}{\bar{n}}$	$\sigma_r^2 + \bar{n}\sigma_B^2$
Residual = repeatibility	$MS_r = \frac{\sum_{i=1}^{L} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{(N-L)}$, $d.f. = (N-L)$	σ_r^2

Where m is the general mean for the material under analysis, B is the laboratory bias component under repeatability conditions, and ϵ is the random error occurring in each measure under repeatability conditions. The repeatability variance σ_r^2 is estimated by S_r^2, which is the within-laboratory variance. On the other hand, the between-laboratory variance σ_B^2 is estimated by S_B^2, this variance is related to laboratory bias. The reproducibility variance σ_r^2 is given by:

$$\sigma_R^2 = \sigma_r^2 + \sigma_B^2 \tag{2}$$

Whereby m is the overall mean for the material under analysis, B is the laboratory bias component under repeatability conditions, and ϵ accounts for the random error occurring in each measure under repeatability conditions. The repeatability variance σ_r^2 is estimated by S_r^2, which is the within-laboratory variance. On the other hand, the between-laboratory variance σ_B^2 is estimated by S_B^2; this variance is related to laboratory bias. The reproducibility variance σ_r^2 is given by:

Using the `ILS` package, one-way ANOVA analysis and mean comparison test can be performed. However, laboratories that present non-consistent results should be excluded from the ILS in advance [1]. Accordingly, consistency tests and identification of atypical results must be performed in advance of ANOVA analysis.

In table(2), the one-way ANOVA approach results are shown, with $\bar{n} = \frac{1}{L-1}(\sum_{i=1}^{L} n_i - \frac{\sum_{i=1}^{L} n_i^2}{\sum_{i=1}^{L} n_i})$, and $N = \sum_{i=1}^{L} n_i$. If $S_B^2 < 0$, set $S_B^2 = 0$.

There are two possible scenarios in which the presence of outliers can be evaluated: the first is that the results of one laboratory deviates from the others in terms of precision, that is, when the measurements made by a laboratory differ significantly with respect to the measurements obtained by other laboratories. The second scenario is related with the identification of outliers in a laboratory for a certain level. The statistics and tests recommended by ISO 5725-2 and ASTM E-691 are described below.

3.1.1. Mandel's h statistic and Grubbs test

Let $(x_1, x_2, \dots, x_L)$ a sample of L observations. The $x_l; l = 1, \dots, L$ are modelled as realizations of random variables $X_l; l = 1, \dots, L$ being identically and independently distributed according to the normal distribution $N(\mu, \sigma^2)$. We denote:

$$\bar{X} = \frac{\sum_{l=1}^{L} X_l}{L} \tag{3}$$

as the mean of the X_l,

$$S^2 = \frac{1}{L-1} \sum_{l=1}^{L} (X_l - \bar{X})^2 \tag{4}$$

as the sample variance of the X_l.

Mandel's h statistic is defined by [14]:

$$H_l = \frac{X_l - \bar{X}}{S}; l = 1, \dots, L \tag{5}$$

Which has the same distribution for all $l = 1, \dots, L$. The critical value [15] is:

$$h_{l;1-\frac{\alpha}{2}} = \frac{(L-1)t_{L-2;1-\frac{\alpha}{2}}}{\sqrt{L(t^2_{L-2;1-\frac{\alpha}{2}} + L - 2)}} \tag{6}$$

Whereby $t_{L-2;1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the t distribution with $v = L-2$ degrees of freedom.

For the case defined by L laboratories that obtain n replicates each one, the h statistic is defined by:

$$h_l = \frac{d_l}{S_{\bar{x}}} = \frac{\bar{x}_l - m}{\sqrt{\sum_{l=1}^{L} \frac{(\bar{x}_l - m)^2}{L-1}}}; l = 1, \ldots, L \tag{7}$$

Whereby $\bar{x}_l$ is the mean of the n results of each laboratory, and m is the global mean of the results of the L laboratories.

A laboratory is detected as inconsistent when the value of the statistic h_l is greater than the critical value, i.e. when $h_l \geq h_{l;1-\frac{\alpha}{2}}$.

On the other hand, if we want to determine if the observation $X_{\max} = \max(X_1, \ldots, X_n)$ is an outlier, the Grubbs test [16] is used. The statistic corresponding to this test is defined by the following expression:

$$G = \max_{i=1,\ldots,n} H_i = \max_{i=1,\ldots,n} \frac{X_i - \bar{X}}{S} = \frac{X_{\max} - \bar{X}}{S} \tag{8}$$

If we want to determine if the smallest observation $X_{\min} = \min(X_1, \ldots, X_n)$ is an outlier, the test statistic is:

$$G = \max_{i=1,\ldots,n} - H_i = \max_{i=1,\ldots,n} \frac{X_i - \bar{X}}{S} = \frac{X_{\min} - \bar{X}}{S} \tag{9}$$

The critical value [15] for this test is defined by:

$$g_{n;1-\alpha} \approx \frac{(n-1)t_{n-2;1-\frac{\alpha}{n}}}{\sqrt{n(n-2+t^2_{n-2;1-\frac{\alpha}{n}})}} \tag{10}$$

For the special problem where there are L laboratories and n replicates obtained for each one, the statistic $g_{L;1-\alpha}$ is defined. In this case, the observations must be replaced by the means of the results corresponding to each laboratory, whereas the mean of the observations is also replaced by the global mean obtained as the mean of laboratories mean.

If a laboratory is identified as an outlier, after applying the h statistic and the Grubbs test to different levels within a laboratory, this is an evidence of the presence of a laboratory high bias (due to a high systematic error in calibration, or errors in the equations when the results were computed).

3.1.2. Mandel's k statistic and Cochran test

Let $(S_1^2, S_2^2, \ldots, S_L^2)$ be a series of L sample variances with each one based on n observed values. Under the assumption that the observed values $x_{ji}; j = 1, 2 \ldots, L, i = 1, 2, \ldots, n$ are realizations of random variables X_{ji} identically and independently distributed according to a normal distribution $N(\mu_i, \sigma^2)$ for each j, the sample variances $S_j^2; j = 1, \ldots, L$ divided by their expectation σ^2 follow a χ^2/v with $v = n - 1$ degrees of freedom. Mandel's k statistic [14] is defined by:

$$k_l = \frac{S_l}{\sqrt{\bar{S^2}}}; j = 1, 2, \ldots, L \tag{11}$$

with

$$\bar{S^2} = \frac{\sum_{l=1}^{L} S_l^2}{L} \tag{12}$$

with the same distribution for all $l = 1, \ldots, L$. The critical value [15] is:

$$k_{l,n;1-\alpha} = \sqrt{\frac{L}{(1 + \frac{L-1}{F_{v_1,v_2;\alpha}})}} \tag{13}$$

Where $F_{v1,v2;\alpha}$ is the α-quantile of the distribution F with $v_1 = (L-1)(n-1)$ and $v_2 = n - 1$ degrees of freedom.

When L laboratories with n replicates are studied, the k statistic is defined by:

$$k_l = \frac{S_l}{S_r} = \frac{\sqrt{\sum_{l=1}^{L} \frac{(x_l - \bar{x})^2}{n-1}}}{\sqrt{\frac{\sum_{l=1}^{L} S_l}{L}}} \tag{14}$$

Where S_l is the standard deviation of the replicates of each laboratory for a given material. A laboratory is detected as inconsistent when the value of the statistic k is greater than the critical value, this is, $k_l \geq k_{l,n;1-\alpha}$.

On the other hand, to determine if the highest variance $S_{max}^2 = max(S_1^2, S_2^2, \ldots, S_L^2)$ is an outlier, we used the Cochran test:

$$C = \frac{S_{max}^2}{\sum_{j=1}^{L} S_j^2} = \frac{1}{L} \max_{i=1,\cdots,L} \frac{S_j^2}{\bar{S^2}} \tag{15}$$

For this test, the critical value, given in [15], follows the expression:

$$c_{l,n;1-\alpha} \approx \frac{1}{(1+(L-1)F_{v_1,v_2;\frac{\alpha}{L}})} \tag{16}$$

Where $F_{v1,v2;\frac{\alpha}{L}}$ is the $\frac{\alpha}{L}$-quantil of the F distribution with $v_1 = (L-1)(n-1)$ and $v_2 = n-1$ degrees of freedom.

The Cochran test is a one tail test for outliers, because it only evaluates the highest value in a series of variances. If a laboratory is detected as an outlier, using the k statistic or with the Cochran test, this indicates that the variance within the laboratory is high (due to lack of familiarity with the analytical method, differences of appreciation among operators, inadequate equipment, equipment in poor state, or careless execution), in which case, the total of results collected by this laboratory, should be rejected and taken out of the study.

The detection of inconsistent laboratories must be repeated until laboratories stop reporting outliers. However, the consistency tests should be used with caution, because if this process is carried out in excess, could lead to false outlier identification.

3.2. ILS: `Glucose` *study*

In this section, we will use the `qcdata` and `qcstat` objects `lab.qcdata()` and `lab.qcs()` created in subsection 2.1 from the `Glucose` dataset.

First, an analysis of the variability for each laboratory will be performed. For this purpose, the k statistic (`k.qcs()`) and the Cocharn test (`cocharn.test()`) will be used to identify if there is any laboratory with non-consistent results. Subsequently, the h statistic (`h.qcs()`) and the Grubbs test (`grubbs.test()`) will be used to perform an analysis to evaluate inter-laboratory variability.

The following statements creates a `k.qcs()` object and the corresponding graph for the k statistics for each laboratory and material (see Figure 5).

```
R> k <- k.qcs(qcdata, alpha = 0.005)
R> plot(k)
R> summary(k)
```

```
Number of laboratories:  8
Number of materials:  5
Number of replicate:  3
Critical value:  2.06084
Beyond limits of control:
        A    B     C    D     E
Lab1 TRUE TRUE  TRUE TRUE  TRUE
Lab2 TRUE TRUE  TRUE TRUE FALSE
Lab3 TRUE TRUE  TRUE TRUE  TRUE
Lab4 TRUE TRUE FALSE TRUE  TRUE
Lab5 TRUE TRUE  TRUE TRUE  TRUE
Lab6 TRUE TRUE  TRUE TRUE  TRUE
Lab7 TRUE TRUE  TRUE TRUE  TRUE
Lab8 TRUE TRUE  TRUE TRUE  TRUE
```

```
R> cochran.test(qcdata)
```

```
Test Cochran


 Critical value: 0.5156875


 Alpha test: 0.00625
  Smax Material           C p.value
1 Lab4        A 0.20033869  0.0231
2 Lab4        B 0.15447962  0.0102
3 Lab4        C 0.10935197  0.0029
4 Lab2        D 0.08493741  0.0010
5 Lab2        E 0.07416440  0.0005
```

In Figure(5), the discontinuous line represents the critical value obtained at a significance level of 0.005. Hence, outliers were detected for the material 5 of laboratory 2, and for material 3 of laboratory 4, since the corresponding

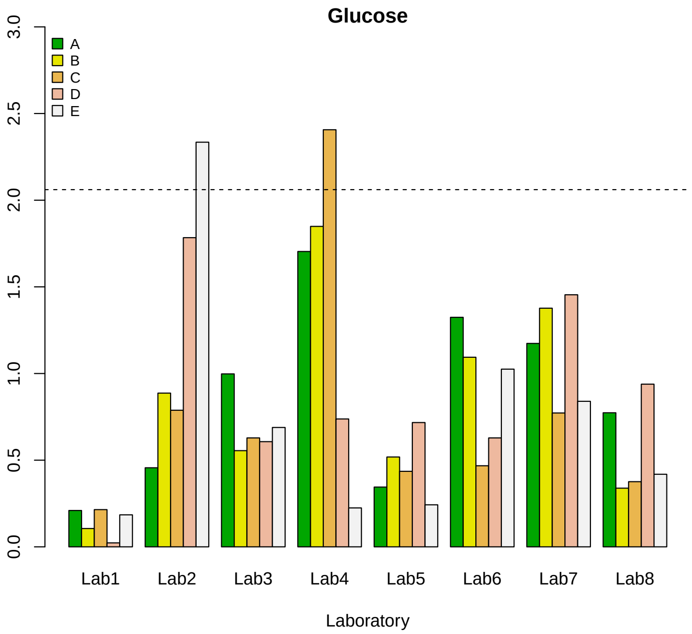

Figure 5: values of Mandel's k, classified by laboratory and material, corresponding to glucose measurements in blood available in the `Glucose` database.

values of the k statistics were greater than the critical value obtained for $L = 8$, $n = 15$ and $\alpha = 0.005$ (following the ASTM standard).

The `k.qcs()` function computes the following objects:

- k: The k statistic for each laboratory and material.
- k.critical: The critical value for the α defined in the function `k.qcs()`.
- violations: Matrix of $L \times R$ dimension (number of laboratories by number of materials).

The matrix of `violations` contains logical values resulting from comparisons between the critical value and the k value. If this comparison is `FALSE`, the laboratory reports outlying results at a certain level, this is, the critical value is less than the statistic value. In this example, the critical value is 2.06.

We performed the Cochran test using the `cocharn.test()` () function. In this case study, with the maximum variance for each material, no laboratory was considered inconsistent, since the critical value was 0.52 and the p-values in each material did not exceed the 5% significance level.

We proceeded to use the functions `h.qcs()` and `plot(h)()` to estimate and plot the h statistics for each laboratory and material. Subsequently, the Grubbs test was applied. The critical value was 2.15, therefore, from this result it can be seen in figure 5 that laboratories 4, 7 and 8 presented non-consistent results at a significance level of $\alpha = 0.005$. Moreover, laboratories with very extreme results were detected by using the Grubbs test, i.e. laboratories defined by very large and very small results (glucose content).

```
R> h <- h.qcs(qcdata, alpha = 0.005)
R> plot(h)
R> summary(h)

Number of laboratories:  8
Number of materials:  5
Number of replicate:  3
Critical value:  2.152492
Beyond limits of control:
        A    B     C     D    E
Lab1 TRUE TRUE  TRUE  TRUE TRUE
Lab2 TRUE TRUE  TRUE  TRUE TRUE
Lab3 TRUE TRUE  TRUE  TRUE TRUE
Lab4 TRUE TRUE FALSE  TRUE TRUE
Lab5 TRUE TRUE  TRUE  TRUE TRUE
Lab6 TRUE TRUE  TRUE  TRUE TRUE
Lab7 TRUE TRUE  TRUE FALSE TRUE
Lab8 TRUE TRUE  TRUE FALSE TRUE

R> grubbs.test(qcdata)

Test Grubbs
```

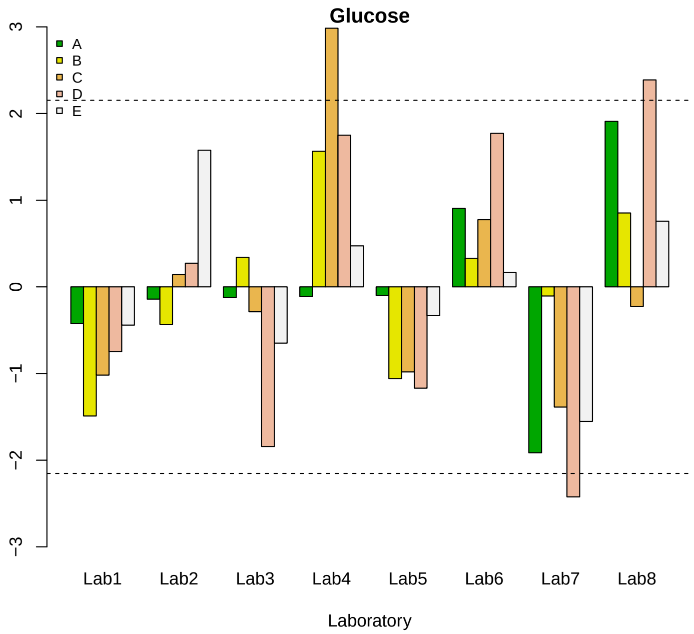

Figure 6: Mandel's h statistics,classified for material and laboratory.

```
 Critical value: 1.154478


 Alpha test: 0.00625
  Material Gmax    G.max p.value.max Gmin    G.min p.value.min
1        A Lab8 1.909229      0.0489 Lab7 1.915242      0.0485
2        B Lab4 1.564273      0.0809 Lab1 1.490219      0.0899
3        C Lab4 2.984235      0.0102 Lab7 1.387137      0.1040
4        D Lab8 2.388179      0.0241 Lab7 2.423819      0.0229
5        E Lab2 1.576391      0.0795 Lab7 1.551749      0.0823
```

4. Interlaboratory Studies: New FDA Approach

A random variable χ is a functional variable if it takes values in a functional space F (full normed or semi-normed). A particular case occurs when the functional variable $\chi = \{\chi(t) : t \in T\}$ where T is an interval $T \subset \mathbb{R}$ that belongs to a Hilbert space, as is the case of continuous functions in an interval [3].

A set of functional data $\chi_1, ..., \chi_n$ is the observation of n functional variables $\chi_1, ..., \chi_n$ with the same distribution as χ. Where χ is usually assumed to be an element of:

$$L_2(T) = \{f : T \to R, \int_T f(t)^2 dt < \infty\}$$

With the inner product $(f, g) = \int_T f\,(t)\,g\,(t)\,dt$.
The norm of $\chi(t)$ is defined by:

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

In this context, the `ILS` package is used to apply consistency tests (outlying detection) in an Interlaboratory Study. For this purpose, the `TG` dataset composed of the Thermogravimetric (TG) curves described in subsection 2.2 is used.

In this section, we use the `ILS` package to perform the estimations and graphical representation of the statistics $H(t)$, $K(t)$, d_H and d_K, with the aim to perform a r&R study for the datasets composed of functional data `TG` and `DSC` that are also included in the `ILS` package.

4.1. Hypothesis of reproducibility and repeatability

In the ILS studies, each laboratory performs n samples experimentally, obtaining n different curves of observations $\{X_1^l(t), \ldots, X_n^l(t)\}$, which are obtained for each, $l = 1, \ldots, L$. Functional statistics $H_l(t)$ and $K_l(t)$ are calculated for each laboratory assuming the corresponding null hypothesis that there are no statistically different measurements between the laboratories.

The null hypothesis of reproducibility states that:

$$H_0 : \mu_1(t) = \mu_2(t) = \cdots = \mu_p(t), \tag{17}$$

Where $\mu_l(t)$, $l = 1 \ldots L$ is the functional mean of the population for each laboratory l. To evaluate the reproducibility of the laboratory results, the $H(t)$ statistic is calculated as follows:

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

Where $\bar{X}_l(t)$ and $S_l(t)$ are the mean and the functional point-to-point variance calculated for the l laboratory.

The null hypothesis of repeatability can be defined by:

$$H_0 : \sigma_1^2(t) = \sigma_2^2(t) = \cdots = \sigma_L^2(t), \tag{18}$$

Where $\sigma_l(t)$, $l = 1 \ldots L$ are the theoretical functional variances corresponding to each laboratory l. The repeatability test is based on the statistic $(K(t))$, expressed as:

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

Where, $\bar{S^2}(t) = \frac{1}{L} \sum_{l=1}^{L} S_l^2(t)$.

On the other hand, to test the reproducibility hypothesis, the test statistic $d - H$ is defined as:

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

Considering that the larger values of d_K correspond to non-consistent laboratories, for the repeatability hypothesis, we define $d_l^K = \big\|K_(t)\big\|$, and likewise, the large values of d_K correspond to non-consistent laboratories.

4.2. ILS: Thermogravimetric Study

The techniques developed in [6] to check if inconsistent laboratories are detected either by outliers in the within-laboratory or in between-laboratory variability, have been implemented in the `ILS` package. As above mentioned, laboratories 1, 5 and 6 have provided different results from the remaining laboratories and should be detected as outliers. We use the datasets described in 2.2, the TG dataset that contains Thermogravimetric test results from 7 laboratories, while the DSC dataset contains results from 6 laboratories (excluding laboratory 1).

First you estimate the functional statistics $H(t)$ and $K(t)$ by the function mandel.fqcs(), then you make the corresponding graphs in the defined functional space. Figure 7 shows both the $K(t)$ and $H(t)$ statistics for each laboratory, as well as the d_K and d_H contrast statistics. The control limit between short lines is constructed at a significance level $\alpha = 0.01$ corresponding to the critical values c_K and c_H. The following code refers to the use of the ILS package into the TG dataset.

```
R> mandel.tg <- mandel.fqcs(fqcdata,nb = 10)
R> plot(mandel.tg,legend = T,col=c(rep(3,5),1,1))
```

The reproducibility hypothesis is tested by using the d_H test statistic and the d_K test statistic. In addition, a graphical interpretation is proposed by plotting the $K(x)$ and $H(x)$ functional statistics in the results (curves) domain. Figure 7 shows the statistics $K(x)$ and $H(x)$ corresponding to each laboratory, using a signification level of $\alpha = 0.01$. Laboratories 1, 6 and 7 were detected as outlying laboratories in the first iteration of the methodology. The region corresponding to the first, second and third stages of calcium oxalate degradation are outside the 99% confidence bands.

Additionally, in Figure 7 the test statistics d_H and d_K are plotted and compared with their corresponding critical values c_H and c_K (control limits), defined as the quantile concerning a signification level of $\alpha = 0.01$. In the case of d_H statistic, it is concluded that the laboratory 7 does not meet the reproducibility hypothesis. According with [6], the laboratories 1, 6, and 7 were detected as outliers through an iterative process, whereas when the repeatability hypothesis was tested by the d_K test statistic, the laboratory 6 was identified as an outlier.

Finally, the ILS package was used to perform outlier detection tasks in the

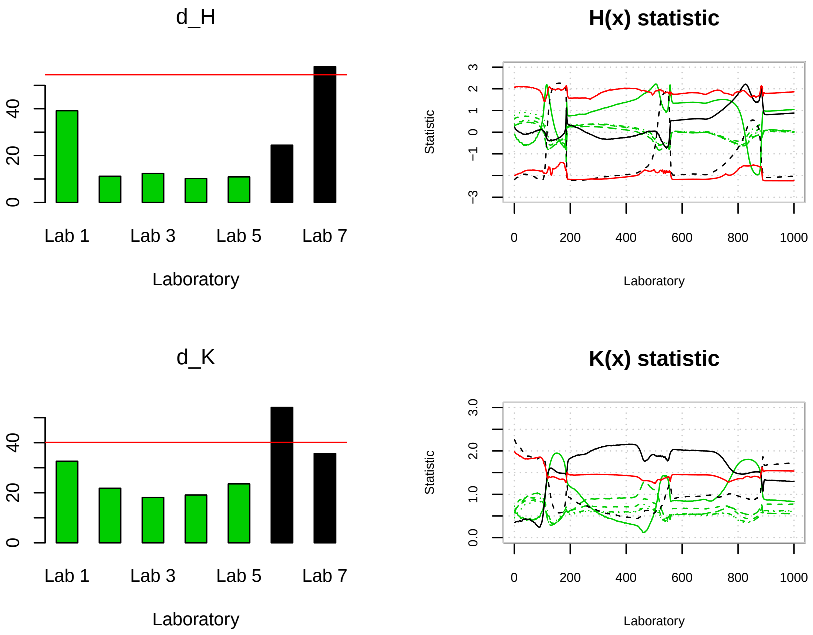

Figure 7: TG dataset: The right panels show the functional statistics $H(x)$ (up) and $K(x)$ (below) for each laboratory, whereas the left panels show the d_H (up) and d_K (below) test statistics for each laboratory.

Interlaboratory Study defined by the `DSC` dataset. Thus, Figure 8 shows that repeatability hypothesis was not reject. Otherwise, the reproducibility's hypothesis was rejected in the case of laboratory 6 (see Figure 8), that is properly detected as an outlier.

```
R> data(DSC, package = "ILS")
R> fqcdata.dsc <-  ils.fqcdata(DSC, p = 6,
+  index.laboratory = paste("Lab",2:7), argvals = delta)
R> mandel.dsc <- mandel.fqcs(fqcdata.dsc,nb = 10)
R> plot(mandel.dsc,legend = F,col=c(rep(3,4),1,3))
```

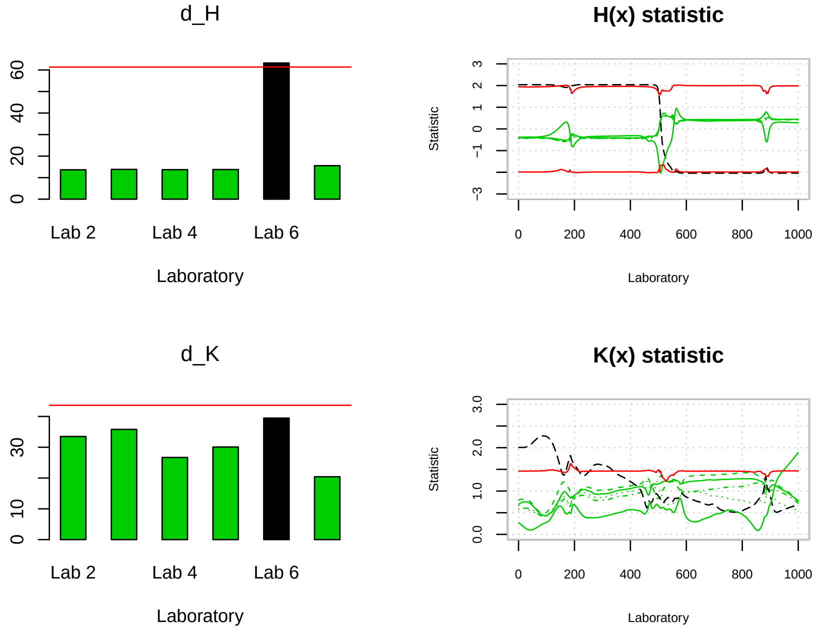

Figure 8: DSC dataset: The right panels show the functional statistics $H(x)$ (up) and $K(x)$ (below) for each laboratory, whereas the left panels show the d_H (up) and d_K (below) test statistics for each laboratory.

5. Conclusions

The present ILS library has been implemented in R software to provide practitioners of Academia and Industry an open source computational tool to perform the statistical analysis dealing with Interlaboratory Studies. The development and presentation of this library fills a gap within the software alternatives to carry out this type of analysis.

In fact, this package provides the main descriptive and outlier detection tools dealing with the Interlaboratory Studies and recommended by the ISO 5725-4-1994 and ASTM E-691 standards. Namely, Mandels's h and k test (including their graphical output), Grubbs' test, Cochran's test, in addition to ANOVA utilities.

On the other hand, apart from standard univariate statistical techniques, the ILS package provides FDA techniques to deal with functional data (curves), as data unit, in the framework of Interlaboratory Studies. Indeed, the main novelty of this computational proposal is the implementation of the functional extensions of the h and k Mandel's statistics when data results are curves, preventing to lose relevant information derived from reduction dimension and feature extraction processes. These new methods can identify successfully the outlier laboratories directly from data curves.

Thus, different study cases dealing with analytical chemistry and applied physics (thermal analysis), and also clinical studies have been presented and solved by using the univariate and FDA approaches provided by the `ILS` computational tool.

The proposal of these computational tools dealing with functional data is important and necessary since the increasing number of processes that are actually sensorized, and the complexity of data obtained by laboratory experimental techniques. They produce in many cases a huge volume of complex data that had to be properly analysed for their future exploitation. In this paper we have propose new computational tools that allows to perform the statistical analysis of ILS studies when working with this new paradigm of data.

6. Acknowledgements

This research/work of Salvador Naya, Javier Tarrío-Saavedra and Rubén Fernández-Casal have 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 Investigación 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.

References

[1] ISO-5725, International Standard ISO 5725-4-1994: 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, Geneva, Suiza, 1994.

[2] ASTM-E691, Practice for conducting and interlaboratory study to determine the precision of a test method., West Conshohocken, USA, 2004.

[3] J. O. Ramsay, B. W. Silverman, Functional Data Analysis, 2nd ed., Springer-Verlag, New York, 2002.

[4] J. O. Ramsay, B. W. Silverman, Applied Functional Data Analysis: Methods and Case studies, Springer-Verlag, New York, 2002.

[5] F. Ferraty, P. Vieu, Nonparametric functional data analysis: theory and practice, Springer Science & Business Media, 2006.

[6] M. Flores, J. Tarrío-Saavedra, R. Fernández-Casal, S. Naya, Functional extensions of mandel's h and k statistics for outlier detection in interlaboratory studies, Chemometrics and Intelligent Laboratory Systems (2018).

[7] S. Naya, J. Tarrío-Saavedra, J. López-Beceiro, M. Francisco-Fernández, M. Flores, R. Artiaga, Statistical functional approach for interlaboratory studies with thermal data, Journal of Thermal Analysis and Calorimetry 118 (2014) 1229–1243.

[8] M. Flores, R. Fernandez, S. Naya, J. Tarrio-Saavedra, R. Bossano, ILS: Interlaboratory Study, 2018. URL: `https://CRAN.R-project.org/package=ILS`, r package version 0.2.

[9] M. Flores, S. Naya, J. Tarrío-Saavedra, R. Fernández-Casal, Functional Statistics and Related Fields, Springer, 2017, pp. 123–130. doi:`10.1007/978-3-319-55846-2_16`.

[10] S. L. R. Ellison, metRology: Support for Metrological Applications, 2017. URL: https://CRAN.R-project.org/package=metRology, R package version 0.9-26-2.

[11] T. Hothorn, F. Bretz, P. Westfall, R. M. Heiberger, A. Schuetzenmeister, S. Scheibe, multcomp: Simultaneous Inference in General Parametric Models, 2017. URL: https://CRAN.R-project.org/package=multcomp, R package version 1.4-7.

[12] M. Febrero-Bande, M. Oviedo, Statistical computing in functional data analysis: The r package fda.usc, Journal of Statistical Software, Articles 51 (2012) 1–28.

[13] J. Ramsay, G. Hooker, S. Graves, Functional data analysis with r and matlab, 2010.

[14] J. Mandel, A new analysis of interlaboratory test results, In: ASQC Quality Congress Transaction-Baltimore (2014) 360–366.

[15] P. Wilrich, Critical values of mandel's h and k, the grubbs and the cochran test statistic, AStA Adv Stat Anal 97 (2013) 1–10.

[16] F. Grubbs, G. Beck, Extension of sample sizes and percentage points for significance tests of outlying observations, Technometrics 14 (1972) 847–854.