Stockholms universitet

Conservative Estimation of the reproducibility in calibration based on bilateral inter-laboratory comparisons

Gustaf Sporrong

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2009:11
Matematisk statistik
Oktober 2009

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2009:11**
http://www.math.su.se

Conservative Estimation of the reproducibility in calibration based on bilateral inter-laboratory comparisons

Gustaf Sporrong*

Oktober 2009

Abstract

The optical properties of a paper, such as refectance, whiteness, brightness and opacity are of great importance to a paper maker. The ISO brightness is a measure of the radiance factor in the blue region. In a hierarchy of laboratories producing reference standards for the ISO brightness, it is of interest to find a measure of reproducibility between laboratories. This in order of giving a quality warranty statement to their customers. The objective in this thesis is to, through modeling, produce a statement in the form of an interval that incorporates the concepts of conservative and robust estimation. A conservative interval of measurement will be an interval that based on any one laboratorys measurement is of that length that it will contain a measurement from any other laboratory with a certain probability. The methods of estimating the key components of that interval are robust. This for the purpose of keeping the conservative concept in the statement made. A model is chosen to represent the monthly bilateral comparisons, between Authorized laboratories during the year 2005. The results produced are the estimates of reproducibility, systematic error and a conservative interval of measurement.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: gustaf.sporrong@gmail.com . Supervisor: Jan-Olov Persson.

Abstract

De optiska egenskaperna hos ett papper, t.ex. reflektans, vithet, ljushet och opacitet är av stor betydelse för en papperstillverkare. ISO- ljusstyrka är ett mått på strålningsfaktorn i den blå regionen. I en hierarki av laboratorier som producerar referensnormer för ISO- ljusstyrka, är det av intresse att hitta ett mått för reproducerbarhet mellan laboratorier. Detta för att ge ett kvalitetsuttryck som garanti till sina kunder. Syftet med denna uppsats är att, genom modellering, producera en utsaga i form av ett intervall som innefattar begreppen konservativa och robusta skattningar. Ett konservativt intervall för en mätning, kommer att vara ett intervall som, baserat på något laboratoriums mätning, är så långt att det kommer att innehålla en mätning från ett annat laboratorium med en viss sannolikhet. Metoderna för att skatta de viktigaste beståndsdelarna i detta intervall är robusta. Detta för att upprätthålla ett konservativt begrepp i uttalandet.

Contents

1 Introduction

This thesis is written at Innventia AB.

Innventia AB is one of the world's leading research and development companies in the fields of pulp, paper, graphic media, packaging and logistics. Innventia AB covers the whole value chains from the raw material to print and media, to packaging and to bio-based energy and chemicals. The activities of Innventia AB range from basic research to direct commissions, where their expert skills and know-how are utilized to find solutions for customers to apply in operations [12].

1.1 Optical characteristics of paper

The fundamental optical property of paper is its spectral radiance factor, which for papers without optical brighteners is the same as its spectral reflectance factor. From the spectral radiance factor other optical characteristics, such as whiteness, brightness, colour, opacity, light scattering coefficient and light absorption coefficient are calculated [13], [Figure 1]. Since, these characteristics are important sales points for a paper maker, it is crucial to report these values in an accurate and standardized way. An essential part of performing an accurate measurement is ensuring that the calibration of the instrument is correct [12],[24],[26].

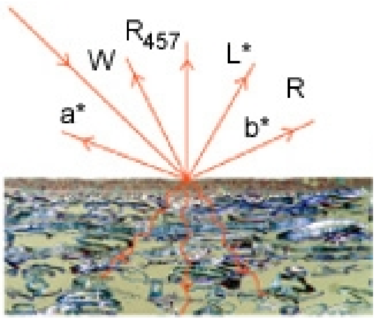


Figure 1: Optical characteristics are important properties of a paper. They are determined by detection of the light reflected and, in the case of fluorescent papers, emitted from the paper when it is illuminated in a well defined way

1.2 Calibration system

Calibration is the validation of specific measurement techniques and equipment. At the simplest level, calibration is a comparison between measurements, one of known magnitude or correctness made or set with one device, and another measurement made in as similar a way as possible with a second device. The device with the known or assigned correctness is called the reference standard

A laboratory usually calibrates to adjust the reading of the instrument to the reference standard. Optical radiance factor measurements are typically based on determination of the ratio of the instrumental reading of the test sample to the instrumental reading of a calibrated standard [14].

ISO is the International Organization for Standardization and is the recognized body that provides the guidelines for the standardization of all measurements and calibration procedures referred to in this thesis. All measurements of pulp, paper, and board within ISO, are under the jurisdiction of technical committee number 6, TC 6. ISO has established a hierarchy of calibration as follows [26]:

- Level 1- Standardizing Laboratory (abbreviated SL)

  The Standardizing Laboratories calibrate their instruments in relation to the abstract concept of the perfect reflecting diffuser, also called IR1. The standards calibrated at the Standardizing Laboratories are called IR2.

- Level 2- Authorized Laboratory (abbreviated AL)

  The Authorized Laboratories use IR2 standards to calibrate their instruments. The standards issued by the Authorized Laboratories are called IR3 standards [Figure 2].

- Level 3 - Industrial Laboratory (abbreviated IL)

  The Industrial Laboratories are also referred to as Testing Laboratories. The calibration hierarchy is illustrated in Figure 3. For the purpose of optical calibration there is in practice only one SL in the world and only five ALs. The IR2 standards are sent from the SL . The ALs use the IR2 standard to calibrate their instrument and produce IR3 standards which are sent to the ILs. The ILs can buy IR3 standards from any of the ALs. The ILs use the received IR3 standards to calibrate their instruments, which are used for the measurement of the optical properties of paper, color print, coating pigments etc.

1.3 Measurement Uncertainty

Every physical quantity with an a priori range of numerical values constituting a continuum is subject to error in its measurement. It is important to report the highest amount by which any measured quantity might be in error. There are random and systematic errors in both the instrumental readings and the calibration values. It is important to explore the origin and the nature of the errors in

Figure 2: IR 3 Reference Standards

measurement to minimize the error propagation and hopefully giving the customer more reliable measurement standards. The international guideline Guide to the expression of uncertainty in measurement [18], generally referred to as GUM, is often used in order to make uncertainty estimates in a structured way. For accredited calibration laboratories within the EU, such as the Optical Calibration Laboratory at Innventia AB, it is complemented by a document referred to as EA-4/02 [15].

These aspects will be explored in more detail in sections, 2.4 and 3.3.

1.4 The need for reliable measurements

To understand why the ILs need to know the uncertainties of the calibration data assigned to the reference standards one needs to look at the purpose of the measurements, which can be very different. The following categories cover the most common.

1. *Use of the measured data in process control.*

   In this case the absolute calibration of the instrument is of little importance since the results will not be compared to the results from another instrument. The focus here is the closeness of the agreement between the results of successive measurements of the same sample carried out under the same conditions of measurements. These are called **repeatability conditions** and include

   - the **same** measurement procedure
   - the **same** operator
   - the **same** measuring instrument

- a short interval of time

The second requirement may be relaxed if the operator dependence is negligible, e.g. when the measurement is fully automated. The last requirement may be relaxed if some means to ensure the stability of the instrument is employed.

2. *Use of the measured data for inter-IL communication*

   In this case the important thing is the agreement between the results of measurements of the same sample carried out in different ILs, i.e. under somewhat different conditions. The changed difference in conditions, called **reproducibility conditions**, may include

   - different instrument designs within the limits set by the international standard ISO 2469 [24]
   - difference between instrument age, wear etc.
   - different operators
   - different environmental conditions
   - different location
   - different time

   Therefore a calibration procedure has to be used. The traceability of the calibration affects the agreement between the ILs. In a general case when evaluating the significance of deviations found when making an inter-comparison between two ILs, one must consider all sources of deviations listed above

3. *Use of the measured data for communication between harmonized ILs.*

   If there is an especially demanding requirement on the agreement between two ILs, it is possible to take different actions to harmonize procedures in the involved ILs. Examples of such actions are the use of the same AL, the use of instruments of the same brand and model etc. This situation is special case of item 2, where one or several causes of deviation is more or less eliminated thereby improving the agreement.

4. *Use of the measured data as estimates of an absolute physical quantity*

   In most industrial applications, there is little need to know the measurement uncertainty in relation to the perfectly reflecting diffuser, which is the fundamental reference for measurements based on the reflectance factor or the radiance factor. It corresponds to the metre definition in length measurements. However, in e.g. research this fundamental physical relation may

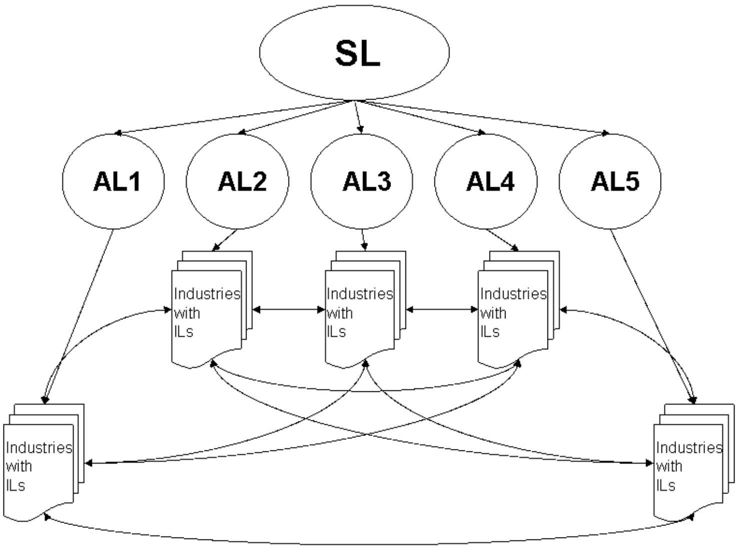


Figure 3: Directions of trade of reference standards from ALs to ILs and trade between ILs

be important. In such a case, all types of errors have to be included in the measurement uncertainty. This includes the systematic errors in the IR2 calibrations. The SL reports an expanded uncertainty of 0.4 percentage points in reflectance. At the same time the reproducibility is approximately 0.1 percentage points, which means that the IR2s sent to different ALs are impaired by common errors which will cancel out when calculating differences between the ALs.

1.5 Problem statement

As a quality warranty the ALs would like to give an estimate of the inter-AL reproducibility, to be used by the IL in their calculation of the inter-IL reproducibility. The reproducibility between the industrial labs depends on the deviations between the ALs. For the purpose of making it possible to estimate the inter-IL agreement in point 2 and 3 in section 1.4 above, the ALs are making inter-laboratory comparisons of the optical characteristics of paper every month by sending IR3 Standards to each other. Due to the fact that the number of ALs is small, the assumptions made in standard methods, [6], [4] and [5] , do not apply. Other more appropriate methods need to be found and tested.

1.6 Research Objective

The objective of this thesis is to develop a conservative and robust model for the inter-AL reproducibility and to test the model on existing data from multi bilateral inter-laboratory comparisons.

2 Data

2.1 Variable measured

The optical properties of a paper, such as reflectance, whiteness, brightness and opacity, is described by the ALs through measurements of 9 classes of variables (see Appendix section A.1).

For the purpose of this thesis only one of those will be studied. The name of that variable is R457NF90 and represents the ISO brightness measured according to ISO 2470 on a non-fluorescent paper. The ISO brightness is a measure of the radiance factor in the blue region and is usually expressed in percentage points. From now on this variable will be referred to as Y with various indexation.

2.2 The IR3 reference standards

The IR3 standards are designed as a small booklet of optically stable and nonfluorescent paper [25]. The size is approximately 70×140 mm. The first sheet of the booklet is a protective cover. The second sheet is the sheet upon which the calibration measurements are made. The rest of the sheets in the booklet are an optical backing ensuring that the IR3 standard is opaque.

2.3 Measurement Procedure

Each AL produces IR3 standards as a batch once per month. The calibration of these IR3 standards is preceded by a careful calibration of the reference instrument and calibration checks are made regularly during the time it takes to measure the batch if IR3 standards.

The value is measured, read and recorded to the nearest 0.01% reflectance factor.

2.4 Experimental Layout

Every month 20 bilateral measurement comparisons are made. Each AL sends IR3 standard to each of the other ALs. The receiving ALs measure on these IR3s and the measurement data can be compared with the data assigned by the sending lab. This makes it possible to build up a data set comprising bilateral comparison data for all possible pairwise combinations of ALs. Normally one data set includes data from one calendar year. The data set is used to estimate the inter-AL agreement and to trace events and trends in the radiance factor scales as well as possible stability problems with the material of the IR3 standards.

	Pariwise measurements				
	AL1	AL2	AL3	AL4	AL5
AL1	–	(91.37,91.16)	(91.40,91.36)	(91.38,91.11)	(91.22,91.20)
AL2	(91.33,91.35)	–	(91.42,91.37)	(91.37,91.30)	(91.32,91.38)
AL3	(91.36,91.31)	no data	–	no data	no data
AL4	(91.15,91.32)	(91.08,91.17)	(91.13,91.31)	–	(91.15,91.34)
AL5	(91.45,91.34)	(91.85,91.70)	(91.42,91.20)	(91.15,90.89)	–

Table 1: Pairwise collected data during one month. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column.

	Number of Comparisons				
	AL1	AL2	AL3	AL4	AL5
AL1	–	12	12	11	12
AL2	12	–	12	12	12
AL3	8	8	–	8	7
AL4	12	12	12	–	12
AL5	5	10	10	12	–

Table 2: Number of comparisons made between each pair, during a whole year. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column

3 Theoretical discussion

3.1 Theoretical objective

The research objective, mentioned in section 1.6, will be divided in parts, for the purpose of finding theoretical foundations for each part. The three parts chosen are:

1. Designing a Model.

   A model that represents the result of measurement of each laboratory, has to be found.

2. Representing reproducibility.

   A way of representing the reproducibility in every model has to be found. In case an estimate of reproducibility is needed, robust methods of estimation will be used.

3. Finding a conservative Interval of measurement

   Using points 1 and 2 above, along with the concept in section 3.3.7 above, to give an interval based on a certain laboratory's measurement, that will include the measurement of any other laboratory, with a certain probability. The estimates used, e.g. reproducibility or standard deviation, will be robust. One of the purposes of that is to reduce the loss of the conservative status of the interval.

3.2 Designing a measurement model

In model fitting the aim is to replace our observed data with a set of fitted values from a model. Thinking about which model to use some delimitations have to be made. The delimitations made must have some ground in the experimental design, and knowledge about the nature of the specific problem is necessary.

The advantage of linear models and their restrictions include computational simplicity, an interpretable model form, and the ability to compute certain diagnostic information about the quality of the fit. Linear models make a set of restrictive assumptions, most importantly, that the target, dependent variable Y, is normally distributed conditioned on the value of predictors with a constant variance regardless of the predicted response value. Generalized linear models, GLM, relax these restrictions, which are often violated in practice [16][21]. The GLM is one of the most important tools in the statistical analysis of data.

There is though one more extension, of the GLM, that will be made and that is the inclusion of random effects. This extension is made by a particular type of mixed model, namely the generalized linear mixed model, GLMM. These random effects are still assumed to have a normal distribution. In classical statistics a typical assumption made is that observations are drawn from the same general population, are independent and identically distributed. Mixed model data have a more complex, multilevel hierarchical structure. Observations between levels or clusters are independent, but observation within each cluster are dependent because they belong to the same subpopulation [11][21].

A possible point of confusion has to do with the distinction between generalized linear models and the general linear model, two broad statistical models. The general linear model may be viewed as a case of the generalized linear model with identity link. As most exact results of interest are obtained only for the general linear model, the general linear model has undergone a somewhat longer historical development. Results for the generalized linear model with non-identity link are asymptotic, tending to work well with large samples.

In this thesis very few and small steps will be made towards finding a model but the ideas are derived from GLM and GLMM. The steps aim to help finding the theoretical objective mentioned above. The order will be, starting with a basic model and slowly making the necessary adjustments with the purpose of making a correct representation of the experimental design producing the monthly data in this thesis.

3.3 Definitions

When creating a model it is practical to introduce a notation that will help expressing the model to be analyzed and evaluated. The tool preceding such notation is a specific set of definitions that contribute to the formulation of the objective of this thesis.

3.3.1 True Value

The *true value* of a measurement is the value that would be obtained by a perfect measurement. It is important to realize that the true value is an idealized concept and that the exact value of the measurand is unknown.

3.3.2 Accuracy

The closeness of the agreement between the result of a measurement and a true value of the measurement will be referred to as *accuracy* of measurement.

3.3.3 Error

As mentioned in the introduction it is common for a measurement to be associated with a measurement error. *Error* is an idealized concept and errors cannot be known exactly. These errors can be seen as additions to the true value. They are referred to as, *systematic* or *random* depending on their origin and behavior. The random error arises from unpredictable variations of influence quantities. The variations of such effects give rise to variation in repeated observations of the measurand. By increasing the number of observations the magnitude of the error can be reduced. A systematic error is that part of the inaccuracy of a measurement, or statistical estimate of a parameter, that is due to a single cause or small number of causes having the same sign, and hence, in principle, is correctable, a bias or constant offset from the true value if the cause can be isolated [10]. From a practical standpoint, systematic errors are usually much more serious nuisance factors than random errors, because their magnitude cannot be reduced by simple repetition of the measurement procedure [29].

3.3.4 Factor and level

A *factor* of an experiment is a controlled independent variable, a variable whose levels are set by the experimentalist.

A *fixed level* of a factor or variable means that the effects or levels in the experiment are the only ones we are interested in. When the levels of a factor are *random*, such as operators, days, locations, where the levels in the experiment might have been chosen at random from a large number of possible levels, the model is called a random- factor model, and inferences are to be extended to all levels of the population [30][19].

3.3.5 Uncertainty

Standard deviation is a measure of *uncertainty* and is characterized by the dispersion of the values that could reasonably be attributed to the measurand. The conclusion now will be that the result of measurement is only an approximation or estimate of the value of the measurand. It gets complete only when a statement of uncertainty, of that estimate, is supplied.

3.3.6 Reproducibility

Reproducibility is a measure of agreement between independent test results under reproducibility conditions. As mentioned in section 1.4, point 2, those are conditions where test results are obtained with the same method, on identical test items, in different laboratories, with different operators, using different equipment [3].

There is no clear definition of reproducibility. One measure of reproducibility of a measurement produced by a laboratory, is the standarddeviation of the assumed error in measurement. Another common measure of reproducibility of measurement is, the standard deviation of the error of the difference between them.

If the laboratories to be compared can be thougt as drawn from a big population of laboratories then the error of measurement difference of a pair is asumed to be random. In the case where the number of laboratories constitute the whole population, then a part of the error can be considered as systematic. In this case it is not uncommom to let the absolute value of the difference, represent the reproducibility.

3.3.7 Conservative estimate and interval

A conservative estimate is a cautious, avoiding excess, approximate calculation of quantity or degree or worth of something. Avoiding excess will mean, when making statements the statement that will have high probability of being true will be chosen. The estimates included in the interval must chosen so that they contribute in making the interval conservative. This type of interval will be broader than a usual confidence interval.

3.3.8 Robustness of estimation

The idea of robustness in statistics is to make statements that will give satisfactory results even if the assumptions are violated. This means that if the assumptions made are only approximately met, the robust estimators will still have a reasonable efficiency, and reasonably small bias, as well as being asymptotically unbiased, meaning having a bias tending towards 0 as the sample size tends towards infinity [28].

An assumption could be that the data are generated from a certain model having a certain distribution. When deviation of those assumption are made and the errors are still small then distributional robust and outlier-resistant are effectively synonymous.

Trimmed estimators and Winzorised estimators are general methods to make statistics more robust [22]. One example is the median being a robust measure of central tendency, while the mean is not [1].

As seen in section 3.3.5 the standard deviation, or variability, is a measure of uncertainty. As with other robust statistics, a robust estimate of variability is minimally affected by a small fraction of outliers, at the cost of lower statistical efficiency when outliers are not present. There are several robust estimates of variability with the most familiar being, the interquartile range, IQR, and the median absolute deviation, MAD [9][22][20].

The IQR is the difference between the 75th percentile and the 25th percentile of a sample, and the MAD is the median of the absolute values of the differences between the data values and the overall median of the data set.

3.4 Model 1

3.4.1 Design

The following notation will bring structure in representing the measurement:
$i =$ the measuring laboratory, $i = 1, .., 5$.
$j =$ the measurement number of the measuring laboratory
$\mu =$ the true value of the sample
$\varepsilon_{ij} =$ the deviation of the measurement of the sample within a laboratory
$\alpha_i =$ the deviation of the i :th laboratory in a between- laboratory comparison of measurement
$Y_{ij} =$ the stochastic representation of the j :th measurement by laboratory i
The assumptions made in this model are:

1. each laboratory is seen as a sample taken from a common probability distribution of laboratories
2. every laboratory is measuring on the same sample
3. α_i is IND, independently normally distributed, $(0, \sigma_\alpha^2)\ \forall\ i$
4. ε_{ij} is IND, independently normally distributed, $(0, \sigma_\varepsilon^2)\ \forall\ i, j$
5. α_i and ε_{ij} are independent between themselves $\forall\ i, j$

The measurement will then have the form

$$Y_{ij} = \mu + \underbrace{\alpha_i + \varepsilon_{ij}}_{\text{Total random error}} \tag{3.1}$$

Expression 3.1 represents a measurement made once. If measurements are made more than once with specific time intervals between the measurements then one must add indexation representing that.

3.4.2 Reproducibility

As a consequence of assumptions 1-5, in section 3.4.1 above and the definition made in 3.3.6, the reproducibility is the standarddeviation of the differences between laboratories. This is commonly represented by the standard deviation of the same difference and is $\sqrt{2}\sqrt{\sigma_\alpha^2 + \sigma_\varepsilon^2}$.

Without any analysis of data one can see that this model does not represent the conditions of the experimental layout the data are based on, and for that reason no more work will be done based on this model. The importance of this model lies in seeing the notation of a simple model and the reproducibility representation. This for the purpose of extending to models more close to the specific layout of the experiment.

3.5 Model 2

3.5.1 Design

Let
i = the sending laboratory which =1, .., 5
j = the receiving laboratory which =1, .., 5
k = the measuring laboratory, i or j from above
μ_{ij} = the true value of the sample measured by i and j
ε_{ijk} = the random error of the measurement on the sample sent by i, received by j and measured by k, $k = i, j$ and $i \neq j$
α_k = the systematic error of measurement, produced by sending laboratory k
Y_{ijk} = the stochastic representation of the sample measurement from i to j measured by k, $k = i, j$ and $i \neq j$

The assumptions made are:

1. the laboratories are measuring pairwise on the same sample
2. there is a systematic error among the measuring laboratories with size α_k for laboratory k, $k = 1, .., 5$.
3. ε_{ijk} is IND, independently normally distributed, $(0, \sigma^2)$ $\forall$ i, j, k

The measurement will then have the form

$$Y_{ijk} = \mu_{ij} + \alpha_k + \varepsilon_{ijk}, \quad \forall i, j, k, \; k = i, j \text{ and } i \neq j \tag{3.2}$$

The representation of measured data, pairwise compared, will be important to visualize for the purpose of finding an estimate of the reproducibility (see Table 3).

Measurement of samples					
	AL1	AL2	AL3	AL4	AL5
AL1	–	(Y_{121}, Y_{122})	...	...	...
AL2	(Y_{212}, Y_{211})	–	...	...	...
AL3	...	...	–	...	...
AL4	...	...	...	–	(Y_{454}, Y_{455})
AL5	...	...	...	(Y_{545}, Y_{544})	–

Table 3: Representation of the pairwise collected data during one month. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column.

Expression 3.2 represents a measurement made once. If measurements are made more than once with specific time intervals between the measurements then one must add indexation representing that.

It should also be noted that the assumptions, for the representation of measurement, in this model are far less than those in model 1. Assumptions 1-3 for model 1, in section 3.4.1, are here replaced by the true layout of the experiment. Here an assumption about a systematic error is made. Further on assumption 4, in 3.4.1, still holds. Namely that the random errors, for every sample measured by every laboratory in every pairwise combination, have the same distribution with expected value zero and variance σ^2.

3.5.2 Reproducibility

Since a systematic error is assumed to be present, based on section 3.3.6, the pairwise reproducibility will be represented by the absolute value of the difference between two laboratories measuring on identical material.

The difference will have the form

$$D_{ij} = Y_{ijj} - Y_{iji} = \alpha_j - \alpha_i + \varepsilon_{ijj} - \varepsilon_{iji} \text{ for } i \neq j, \tag{3.3}$$

and the pairwise reproducibility for the pair (i, j) will be denoted as

$$|E(D_{ij})| = |\alpha_j - \alpha_i|\,, \tag{3.4}$$

Table 4 is symmetrical in the sense of

$$E(D_{ij}) = -E(D_{ji}) \tag{3.5}$$

which is a condition that must be satisfied by the differences for the model 2 to be used.

Pairwise differences					
	AL1	AL2	AL3	AL4	AL5
AL1	–	$\alpha_2 - \alpha_1$	...	...	...
AL2	$-(\alpha_2 - \alpha_1)$	–	...	...	...
AL3	...	...	–	...	...
AL4	...	...	...	–	$\alpha_5 - \alpha_4$
AL5	...	...	...	$-(\alpha_5 - \alpha_4)$	–

Table 4: Representation of the expected values of the pairwise differences, $E(D_{ij})$.The table organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column.

3.5.3 Estimation of reproducibility

The difference in expression (3.3) has the form of the general linear model, for which methods of estimates are derived and presented in [19]. The method of Least Squares is chosen for this specific purpose.

Let
$\mathbf{W}^T = (W_1, .., W_N)$ =the vector representing the stochastic quantity of interest
$\mathbf{w}^T = (w_1, .., w_N)$ =the vector of observed values of $\mathbf{W}$
$\boldsymbol{\theta}^T = (\theta_1, .., \theta_k)$ =the vector of the k variables to be estimated
$\mathbf{e}^T = (e_1, .., e_N)$ =the vector of random errors
A =$N \times k$- sized matrix with coefficients representing the linear combinations of the estimates of interest.

$\mathbf{I}_N = N \times N$ identity matrix
The general linear model has the form $\mathbf{W} = \mathbf{A} \cdot \boldsymbol{\theta} + \mathbf{e}$, where $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ where
The Least Square, LS, estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{w} \tag{3.6}$$

Converting the differences in expression (3.3) to the notation used in the general linear model, leads to the following notations:
$\mathbf{W}^T = (W_1, .., W_{20}) = (D_{12}, ..D_{15}, .., D_{51}, .., D_{54})$ = the stochastic representation of differences
$\mathbf{w} = (w_1, .., w_{20}) = (d_{12}, .., d_{15}, .., d_{51}, .., d_{54})$ =observed differences
$\boldsymbol{\theta}^T = (\theta_1, .., \theta_5) = (\alpha_1, .., \alpha_5)$
$\mathbf{e}^T = (e_1, .., e_{20}) = (\varepsilon_{122} - \varepsilon_{121}, .., \varepsilon_{544} - \varepsilon_{545}) = (\sqrt{2}\sigma, .., \sqrt{2}\sigma)$
A is the 20×5 matrix that for which the, 20×1 sized, matrix $A \cdot \theta$ satisfies

$$(A \cdot \theta)^T =$$
$$= (\alpha_2 - \alpha_1, .., \alpha_5 - \alpha_1, .., \alpha_5 - \alpha_4, -(\alpha_2 - \alpha_1), ..., -(\alpha_5 - \alpha_4)) \tag{3.7}$$

The estimates, $(\widehat{\alpha}_1, .., \widehat{\alpha}_5) = \widehat{\boldsymbol{\theta}}^T$, are derived by using expression (3.6) under the constraint $\Sigma_i \alpha_i = 0$. These estimates are used to estimate the reproducibility, $E(D_{ij})$, with

$$\widehat{E}(D_{ij}) = \widehat{\alpha}_j - \widehat{\alpha}_i \tag{3.8}$$

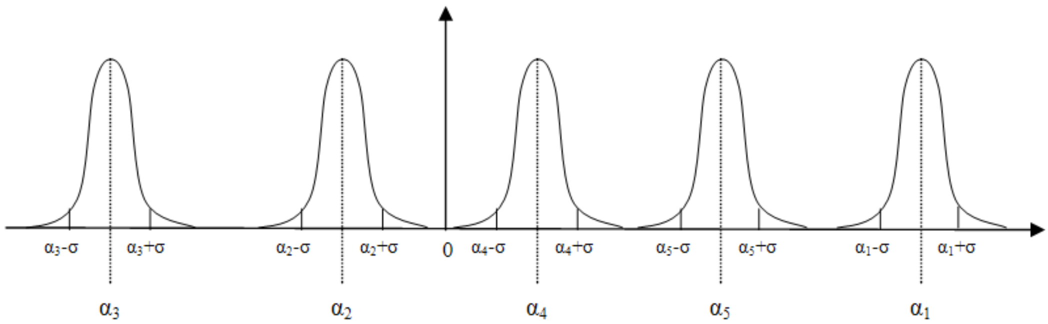


Figure 4: The figure shows one possible combination of order between the α_i, $i = 1, .., 5$, that are to be estimated

3.5.4 Estimation of dispersion

In a conservative interval of measurement, the next key component will be the standard deviation $\mathrm{SD}(D_{ij})$ which has the following form:

$$\mathrm{SD}(D_{ij}) = \sqrt{\mathrm{Var}(D_{ij})} = \sqrt{\mathrm{Var}(\varepsilon_{ijj} - \varepsilon_{iji})} =$$
$$= [\text{assumption 3 in section 3.5.1}] = \sqrt{\sigma_j^2 + \sigma_i^2} =$$
$$= \sqrt{\sigma^2 + \sigma^2} = \sqrt{2}\sigma \tag{3.9}$$

The estimation of σ is is done by using the residual in the estimation of the general linear model. The estimate is given by

$$\widehat{\sigma} = \sqrt{\frac{\widehat{\mathbf{e}}^T \cdot \widehat{\mathbf{e}}}{2 \cdot 16}} = \sqrt{\frac{(\mathbf{w} - \mathbf{A}\widehat{\boldsymbol{\theta}})^T(\mathbf{w} - \mathbf{A}\widehat{\boldsymbol{\theta}})}{2 \cdot 16}} \tag{3.10}$$

3.5.5 Estimation of conservative interval of measurement

Following the outlines drawn above in section 3.3.7 and 3.1, a conservative interval of measurement will be an interval that based on any one laboratory's measurement is of that length that it will contain a measurement from any other laboratory with a certain probability, say $1 - \beta$.

Now let

M = a measurement made by any laboratory at any time

A conservative interval, with level 1-β, of measurement can have the following form:

$$I_C = (M - (\max_i \alpha_i - \min_i \alpha_i) - z_{\frac{\beta}{2}}\sqrt{2}\sigma,\ M + (\max_i \alpha_i - \min_i \alpha_i) + z_{\frac{\beta}{2}}\sqrt{2}\sigma)$$

Based on the estimations of reproducibility, found in section 3.5.3, a conservative interval will be estimated by

$$\widehat{I}_C = (M - (\max_i \widehat{\alpha}_i - \min_i \widehat{\alpha}_i) - z_{\frac{\beta}{2}}\sqrt{2}\widehat{\sigma},\ M + (\max_i \widehat{\alpha}_i - \min_i \widehat{\alpha}_i) + z_{\frac{\beta}{2}}\sqrt{2}\widehat{\sigma})$$

3.6 Model 3

3.6.1 Design

Let

i = the sending laboratory which =$1, .., 5$

j = the receiving laboratory which =$1, .., 5$

k = the measuring laboratory, i or j from above

μ_{ij} = the true value of the sample measured by i and j

ε_{ijk} = the random error of the measurement of the sample between i and j measured by k, $k = i, j$ and $i \neq j$

α_{ik} = the systematic error of measurement, produced by sending laboratory i and measured by k, where $k = 1, .., 5$

Y_{ijk} = the stochastic representation of the sample measurement from i to j measured by k, $k = i, j$ and $i \neq j$

Visualizing then the representation of pairwise collected measurement, see table 3, the measurement will have the form

$$Y_{ijk} = \mu_{ij} + \alpha_{ik} + \varepsilon_{ijk}, \quad \forall i, j, k, \; k = i, j \tag{3.11}$$

The assumptions made in this model are:

1. the laboratories are measuring pairwise on the same sample
2. $\forall\ i, j, k$ the ε_{ijk}:s are independently normally distributed with $E(\varepsilon_{ijk}) = 0$, and $\text{Var}(\varepsilon_{ijk}) = \sigma_{ik}^2$

Expression 3.11 represents a measurement made once. If measurements are made more than once with specific time intervals between the measurements then one must add indexation representing that.

The difference between assumption 2 above and assumption 3 in 3.5.1 is the abandonment of limitations. It should be noticed that the assumption of equal distribution is gone along with the assumption on equality of variance. This is a big step towards making more robust statements about the reproducibility of measurements.

3.6.2 Estimation of reproducibility

As for model 2 in section 3.5.3, the choice of representing reproducibility will be by the use of the pairwise differences. The difference in this model has the form

$$D_{ij} = Y_{ijj} - Y_{iji} = \alpha_{ij} - \alpha_{ii} + \varepsilon_{ijj} - \varepsilon_{iji} \tag{3.12}$$

and the pairwise reproducibility is

$$|E(D_{ij})| = |\alpha_{ij} - \alpha_{ij}| \tag{3.13}$$

The estimation of $E(D_{ij})$ is done by estimating the difference, $\alpha_{ij} - \alpha_{ii}$. When the values of the differences are available a linear system of equations will give those estimates, under the condition $\Sigma_j \alpha_{ij} = 0$.

3.6.3 Estimation of dispersion

For the purpose of creating a conservative interval a way of estimating σ_{ij} must be chosen.

If there is no possibility of estimating the σ_{ij}:s separately, the variance of the difference must be examined more closely as to find approaches towards conservative estimates. The variance of the difference can be written as follows

$$\begin{aligned} \text{Var}(D_{ij}) = \text{Var}(Y_{ijj} - Y_{iji}) = \text{Var}(\varepsilon_{ijj} - \varepsilon_{iji}) = \\ = [\text{assumption 2 in section 3.6.1}] = \sigma_{ij}^2 + \sigma_{ii}^2 \end{aligned} \tag{3.14}$$

Since $\sigma_{ij} \leq \sqrt{\sigma_{ij}^2 + \sigma_{ii}^2} = \sqrt{\widehat{\text{Var}}(D_{ij})}$and $\sigma_{ii} \leq \sqrt{\sigma_{ijj}^2 + \sigma_{iji}^2} = \sqrt{\widehat{\text{Var}}(D_{ij})}$,one way of making a conservative estimation $\widehat{\sigma}_{ij}$, of σ_{ij}, is by setting

$$\widehat{\sigma}_{ij} = \sqrt{\widehat{\text{Var}}(D_{ij})} \tag{3.15}$$

3.6.4 Estimation of conservative interval of measurement

As mentioned in section 3.1 and seen in section 3.5.5, a conservative interval, of level 1-β, of a measurement will be one that based on any one measurement, taken from any laboratory, will contain a measurement from any laboratory with a probability of 1-β

Let

M = a measurement made by any laboratory at any time

A level 1- β, conservative interval of a measurement, could have the form of

$$I_C = \\ = (M - (\max_{ij} \alpha_{ij} - \min_{ij} \alpha_{ij}) - z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} \sigma_{ij}, M + (\max_{ij} \alpha_{ij} - \min_{ij} \alpha_{ij}) + z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} \sigma_{ij}) \tag{3.16}$$

and the estimate, based on robust estimates of the dispersion, of that conservative interval will be

$$\widehat{I}_C = \\ = (M - (\max_{ij} \widehat{\alpha}_{ij} - \min_{ij} \widehat{\alpha}_{ij}) - z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} \widehat{\sigma}_{ij}, M + (\max_{ij} \widehat{\alpha}_{ij} - \min_{ij} \widehat{\alpha}_{ij}) + z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} \widehat{\sigma}_{ij}) \tag{3.17}$$

4 Analysis and Results

Now an analysis of the collected data is required as to determine which model will be applied and why.

4.1 Model 1

The experminetal layout behind this model is to use a sample, measure it and send it to the next laboratory who measures and sends to the next. This method of collecting data is called Round Robin. This method was actually used at some point by Innventia but is now abandoned in favor for the the method being used now namely, the bilateral design of the experiment where laboratories measure pairwise on the same sample. Model one is thereby rejected as a candidate for representing the data collected.

4.2 Model 2

4.2.1 Assumptions

Assumption 1 in section 3.5.1 about the laboratories measuring pairwise on the same sample, is valid for this model.

Since the five laboratories participating in the experiment constitute the whole population of laboratories, it is reasonable to assume a systematic error between the laboratories. Innventia recognizes this fact studying the data and sees e.g. that they always get higher measurement values than certain other laboratories.

Assumption 3 in section 3.5.1, about the distribution of errors being normal, is accepted as valid without any specific tests.

Both assumptions in model 2 are valid leading to the next of estimating reproducibility.

4.2.2 Estimation of reproducibility

Under the assumptions of model 2, the relation $E(D_{ij}) = -E(D_{ji})$, must be seen in the data. Examining Table 5, pairwise hypotheses tests can be set up to check if $E(D_{ij}) = -E(D_{ji})$, is valid for all pairs. The result of those tests, seen in Appendix section A.3, suggest that it can not be said that, $E(D_{ij}) = -E(D_{ji})$ for every pair since there are three cases of rejection.

Model 2 is therefore discarded as a candidate and no more validation tests, e.g. for equality of standard deviations in assumption 3, are necessary.

	Mean of difference					Standard deviation of difference				
	AL1	AL2	AL3	AL4	AL5	AL1	AL2	AL3	AL4	AL5
AL1	–	-0.22	-0.040	-0.32	0.040	–	0.03	0.11	0.04	0.05
AL2	0.050	–	0.050	-0.080	0.070	0.02	–	0.09	0.02	0.08
AL3	-0.040	-0.10	–	-0.22	-0.12	0.04	0.04	–	0.05	0.12
AL4	0.15	0.080	0.14	–	0.20	0.03	0.03	0.09	–	0.12
AL5	-0.080	-0.16	-0.15	-0.24	–	0.06	0.05	0.07	0.04	–

Table 5: The tables show the mean of estimates of the systematic erros and the standard deviation of the same, for a whole year. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column

4.3 Model 3

4.3.1 Assumptions

Model 3 makes the extension of representing a systematic error separately for each pair of laboratories. This is an additional step in trying to capture different patterns in different pairs of difference. Since the difference is chosen as a measure of reproducibility this is a step of giving more accurate description of reproducibility.

Further on assumptions 1 and 2 in section 3.6.1 for model 3, are valid for the same reason as for model 2 in section 4.2.1. These assumptions concern the experimental layout and the distribution of errors.

4.3.2 Estimation of reproducibility

Using data collected for the whole year the need to use the mean value over the year arises, and with that also the need to estimate that mean. Experience shows that there are usually a few errors in the reporting of the bilateral data due to mistakes by the persons inserting the data into the Excel sheets. Most of these outliers are detected by studying time plots of the data, i.e. by a manual validation of the data. However, robust methods for the estimation of the means and standard deviations reduce the need for this time-consuming work. Another reason was given in 3.1. For the estimation of the mean difference, reproducibility, the choice of estimate will be the median, as mentioned in section 3.3.8.

Let
$D_{ij} = \alpha_{ij} - \alpha_{ii}$= the monthly difference between i sending
and j receiving, $i \neq j$
d_{ij} = the observed difference, during a month, between
sending laboratory i and j receiving, $i \neq j$.

Clearly the reproducibility between pair (i, j) can be estimated by

$$\widehat{\alpha}_{ij} - \widehat{\alpha}_{ii} = \underset{\text{whole year}}{\text{median}}\, d_{ij} \tag{4.1}$$

Median of Differences					
	AL1	AL2	AL3	AL4	AL5
AL1	-	-0.22	0.050	-0.32	0.050
AL2	0.040	-	0.040	-0.080	0.080
AL3	-0.050	-0.10	-	-0.22	-0.13
AL4	0.14	0.090	0.14	-	0.20
AL5	-0.080	-0.15	-0.15	-0.25	-

Table 6: The table shows $\widehat{\mathbf{E}}(D_{ij})$. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column.

In Table 6 the estimates of reproducibility are presented.

4.3.3 Estimation of dispersion

It was seen in section 3.6.3 and 3.6.4, that there is need to estimate the dispersion of differences between measurements. There is no possibility of estimating the σ_{ij}:s separately due to insufficient amount of data. Therefore the choice of a conservative estimate of σ_{ij}, as seen in section 3.6.3, is made to be

$$\widehat{\sigma}_{ij} = \sqrt{\widehat{\text{Var}}(D_{ij})} \tag{4.2}$$

As in section 4.3.2, and for the same reasons, a robust estimate will be used. Using the notation of section 4.3.2 a small addition in notation will be made, namely

s_{ij} =estimation of standard deviation of the difference, using yearly data, between sending laborataory i and measuring j.

Innventia required an estimate of dispersion being more resilient to outliers in a data set than the standard deviation. In the standard deviation, the distances from the mean are squared, so on average, large deviations are weighted more heavily, and thus outliers can heavily influence it. In the MAD, the magnitude of the distances of a small number of outliers is irrelevant. Therefore MAD is choosen to estimate the standard deviation of the difference. For a symmetric distribution, as in this thesis, the MAD is the distance between the 1st and 2nd (equivalently, 2nd and 3rd) quartiles, so for a symmetric distribution about the mean, the MAD is the 3rd quartile. Thus MAD for the normal distribution can be seen in,

$$\frac{MAD}{\sigma} = z_{3/4} \Rightarrow$$
$$\Rightarrow \sigma = \frac{MAD}{z_{3/4}}$$

In order to use the MAD as a consistent estimator for the estimation of the standard deviation σ_{ij}, a correction must be made so that

$$s_{ij} = \frac{1}{z_{3/4}} \cdot MAD_{ij} = 1.48 \cdot \underset{\text{all year}}{\text{median}} \left| d_{ij} - \widehat{E}(D_{ij}) \right| \tag{4.3}$$

The results of estimation are shown in Table 7

MAD estimation of difference					
	AL1	AL2	AL3	AL4	AL5
AL1	−	$s_{12} = 0.030$	$s_{13} = 0.030$	$s_{14} = 0.030$	$s_{12} = 0.059$
AL2	$s_{21} = 0.030$	−	$s_{23} = 0.059$	$s_{24} = 0.030$	$s_{12} = 0.059$
AL3	$s_{31} = 0.059$	$s_{32} = 0.044$	−	$s_{34} = 0.030$	$s_{12} = 0.13$
AL4	$s_{41} = 0.030$	$s_{42} = 0.015$	$s_{43} = 0.044$	−	$s_{12} = 0.059$
AL5	$s_{51} = 0.044$	$s_{52} = 0.015$	$s_{53} = 0.0889$	$s_{54} = 0.044$	−

Table 7: Table of estimates of the dispersion using yearly data. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column

4.3.4 Conservative interval of measurement

In 3.6.4,

$$\widehat{I}_C = $$
$$= (M - (\max_{ij} \widehat{\alpha}_{ij} - \min_{ij} \widehat{\alpha}_{ij}) - z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} s_{ij}, M + (\max_{ij} \widehat{\alpha}_{ij} - \min_{ij} \widehat{\alpha}_{ij}) + z_{\frac{\beta}{2}} \sqrt{2} \max_{ij} s_{ij}) \tag{4.4}$$

was derived as the 1-$\beta-$level conservative interval to be used. The components remaining to be estimated are the $\alpha_{ij} : s$. The α_{ij}:s are solved using Table 6, see appendix section 4.3.2, and the results are shown in Table 8

Using Table 8 and Table 7 the following important results are seen

Estimates of systematic errors					
	AL1	AL2	AL3	AL4	AL5
AL1	$\widehat{\alpha}_{11} = 0.108$	$\widehat{\alpha}_{12} = -0.112$	$\widehat{\alpha}_{13} = 0.058$	$\widehat{\alpha}_{14} = -0.212$	$\widehat{\alpha}_{15} = 0.158$
AL2	$\widehat{\alpha}_{21} = 0.024$	$\widehat{\alpha}_{22} = -0.016$	$\widehat{\alpha}_{23} = 0.024$	$\widehat{\alpha}_{24} = -0.096$	$\widehat{\alpha}_{25} = 0.064$
AL3	$\widehat{\alpha}_{31} = 0.05$	$\widehat{\alpha}_{32} = 0$	$\widehat{\alpha}_{33} = 0.1$	$\widehat{\alpha}_{34} = -0.12$	$\widehat{\alpha}_{35} = -0.03$
AL4	$\widehat{\alpha}_{41} = 0.026$	$\widehat{\alpha}_{42} = -0.024$	$\widehat{\alpha}_{43} = 0.026$	$\widehat{\alpha}_{44} = -0.114$	$\widehat{\alpha}_{45} = 0.086$
AL5	$\widehat{\alpha}_{51} = -0.206$	$\widehat{\alpha}_{52} = -0.276$	$\widehat{\alpha}_{53} = -0.376$	$\widehat{\alpha}_{54} = 0.276$	$\widehat{\alpha}_{55} = 0.126$

Table 8: Table of estimates of systematic errors. The table is organized such that each sending laboratory corresponds to a row, and each receiving laboratory corresponds to a column

Expression 4.4 in 4.3.4 needs to be evaluated. Using Table 8, Table 7 and choosing $\beta = 0.05$ so that $z_{\frac{\beta}{2}} = 1.96$, the following can be derived

$$\min_{ij}\widehat{\alpha}_{ij} = -0.38\,, \max_{ij}\widehat{\alpha}_{ij} = 0.276 \text{ and } \max_{ij} s_{ij} = 0.13 \;\Rightarrow$$

$$(\max_{ij}\widehat{\alpha}_{ij} - \min_{ij}\widehat{\alpha}_{ij}) + 1.96 \cdot \sqrt{2}\max_{ij} s_{ij} = 0.656 + 1.96 \cdot \sqrt{2} \cdot 0.13 = 1.0163 \quad (4.5)$$

Expression 4.5, leads to the resulting interval being,

$$(M - 1.02, M + 1.02) \quad (4.6)$$

4.3.5 Final Statement

The objective in this thesis was, through modeling, to produce a statement in the form of an interval that incorporates the concepts of conservative and robust estimation.

Expression 4.6 is a conservative interval that, based on any one laboratory's measurement M, will contain a measurement from any other laboratory with a 95% probability. This would be the statement given to Innventia, based on the data set from 2005.

The objective of this thesis is thereby achieved.

5 Summary and discussion

Using 20 monthly bilateral comparisons, between ALs during the year 2005, it was of interest to find a measure of reproducibility between the authorized laboratories. This for the purpose of giving a quality warranty statement to their customers. In this thesis the goal was, through modeling, to derive a statement in the form of a conservative interval that, based on any one laboratory's measurement, is of that length that it will contain a measurement from any other laboratory with a certain probability. Robust methods of estimation of the key components of that interval were used. This for the purpose of keeping the conservative concept in the statement made and for the reason mentioned in section 4.3.2, namely to reduce the amount of time-consuming work.

The choice of three models was made to suggest a working procedure in finding robust estimators of reproducibility which would lead to conservative intervals. Of those three models models the first two were rejected in favor for the third one. The third model was true to experimental design with more realistic assumptions about the systematic error and the variance of the random error not being equal for all laboratories. One can speculate that there may be practices in the handling of the paper before and after measurement, that could be the cause of the systematic error. That different laboratories have different regulatory climate conditions, routines in measuring, or transportation arrangements.

A more conservative interval, than the one derived in expression 4.6, can be created by examining the standard deviations of the estimates of the systematic errors. Incorporating those in the estimation of the total standard deviation will give rise to a broader interval.

One can argue that the interval produced is too broad. Maybe Innventia is interested in finding a more precise and effective interval for prediction instead of a broad one. A more narrow interval could be produced, by choosing the maximum and the minimum systematic error, along with the maximum standard deviation estimated for each row of sending laboratories. This would also produce o conservative interval. More different adjustment can be made to produce different conservative intervals for measurement made by laboratories.

Another important aspect is time. Could time be of significance for the values produced? Is there a drift and does this drift spread? Just by checking the figures in Appendix section A.2, a suggestion could be made to check for e.g. the existense of a drift leading to higher or lower values for certain pair of laboratories. A model including the aspect of time could be included to bring light to these questions. Time series analysis would be a tool for that.

One could also examine robust estimation in more detail as to use the most effective robust estimators.

Doing all of the above would of course demand more time than was set aside for this thesis.

References

[1] *A Festschrift for Erich L. Lehmann in Honor of His Sixty Birthday,* David L Donoho, Peter J DL Donoho, PJ Huber, 1982

[2] *Accuracy and precision,* http://en.wikipedia.org/wiki/
/Accuracy_and_precision

[3] *Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions,* ISO 5725-1:1994, http://www.iso.org/iso/iso_catalogue

[4] *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,* ISO 5725-2:1994, http://www.iso.org/iso/iso_catalogue

[5] *Accuracy (trueness and precision) of measurement methods and results - Part 3: intermediate measures of the precision of a standard measurement method,* ISO 5725-3:1994, http://www.iso.org/iso/iso_catalogue

[6] *Accuracy (trueness and precision) of measurement methods and results - Part 4: Basic methods for the determination of the trueness of a standard measurement method,* ISO 5725-4:1994, http://www.iso.org/iso/iso_catalogue

[7] *Accuracy (trueness and precision) of measurement methods and results - Part 5: Alternative methods for the determination of the trueness of a standard measurement method,* ISO 5725-5:1994, http://www.iso.org/iso
/iso_catalogue

[8] *Accuracy (trueness and precision) of measurement methods and results - Part 6: Use in practice for accuracy values*, ISO 5725-6:1994, http://www.iso.org/iso/iso_catalogue

[9] *Alternatives to the Median Absolute Deviation*, Rousseeuw, Peter J., Croux, Christophe, Journal of the American Statistical Association, 1993 (December)

[10] *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements,* John R. Taylor, 1999

[11] *Approximate Inference in Generalized Linear Mixed Models,* N. E. Breslow and D. G. Clayton, Journal of the American Statistical Association, Vol. 88, No. 421 (Mar., 1993), pp. 9-25.

[12] *About Innventia*, Innventia AB, http//www.innventia.com

[13] *Basic equations for optical properties,* SCAN-G 5:03 (Scandinavian pulp, paper and board testing committee, 2003).

[14] *Calibration,* http://en.wikipedia.org/wiki/Calibration

[15] *Expression of the uncertainty of measurement in calibration, EA-4/02*, European cooperation for Accreditation, 1999.

[16] *Generalized Linear Models*, http://download.oracle.com/docs/cd/ /B28359_01/datamine.111//b28129/algo_glm.htm

[17] *Generalized Linear Models, Second Edition.* McCullagh, Peter; Nelder, John, 1989

[18] *Guide to the expression of uncertainty in measurement – Corrected and reprinted 1995.*, BIPM/IEC/IFCC/ISO/IUPAC/IUPAP/OIML Guide (Bureau International des Poids et Mesures, International Electronic Commission, International Federation of Clinical Chemistry, International Organization for Standardization, International Union of Pure and Applied Chemistry, International Union of Pure and Applied Physics, International Organization of Legal Metrology, Genève, 1993).

[19] *Kompendium i tillämpad matematisk statistik*, Rolf Sundberg, 1997

[20] *Measures of Scale*, http://www.itl.nist.gov/div898/handbook/eda/section3/ /eda356.htm

[21] *Mixed Models: Theory and application,* Eugene Demidenko, 2004

[22] *Outliers in statistical data,* Vic Barnet & Toby Lewis, 1978

[23] *Paper and board, Determination of CIE whiteness, D65/10°(outdoor daylight), Second edition 2004-11-15*, ISO 11475:2004, http://www.iso.org/iso /iso_catalogue

[24] *Paper, board and Pulps - Measurement of diffuse reflectance factor, Third edition 1994-12-01*, ISO 2469:2007, http://www.iso.org/iso/iso_catalogue

[25] *Paper, board and pulps – Measurement of diffuse blue reflectance factor (ISO brightness),* ISO 2470:1999, http://www.iso.org/iso/iso_catalogue

[26] *Paper Board and pulps - International calibration of testing apparatus - Nomination and acceptance of standardizing and authorized laboratories, Second Edition 2005-04-15*, ISO 4094:2005, http://www.iso.org/iso/iso_catalogue

[27] *Pulps – Preparation of laboratory sheets for the measurement of diffuse blue reflectance factor (ISO brightness), Third Edition,* 1999-10-15.

[28] *Robust Statistics*, http://en.wikipedia.org/wiki/Robust_statistics

[29] *Systematic Error,* http://www.statistics.com/resources/glossary/s/
/syserror.php

[30] *Variansanalys och försöksplanering*,Ulla Engstrand, Ulf Olsson, 2003

A Appendix

A.1 Classes of Variables

A.1.1 CIE Whiteness

The CIE whiteness according to ISO 11475 measured on a fluorescent paper. Illuminant: CIE standard illuminant D65. Observer: 10° (CIE 1964 standard observer). The value 100 corresponds to the perfect reflecting diffuser. A normal office paper has a whiteness value in the range 140–165.

A.1.2 LabC2

The CIE L*, a* and b* value. L* corresponds to the lightness on a scale from 0 to ~100. Positive values of a* is a measure of the redness. Negative a* corresponds to greenness. Positive values of b* is a measure of the yellowness. Negative b* corresponds to blueness. The colorimetric evaluation is made with the C/2° condition (CIE illuminant C and CIE 1931 standard observer). Dimension 4 is: L*, a*, b*. Total of three variables

A.1.3 LabD6510

The colorimetric evaluation is made with the D65/10° condition (CIE standard illuminant D65 and CIE 1964 standard observer). Dimension 4 is: L*, a*, b* Total of three variables

A.1.4 R10N

The spectral reflectance factor measured every 10 nm with a spectral bandwidth of 10 nm. Dimension 4 is: R(400), R(410),. . . ,R(700). Total of 31 variables

A.1.5 R20N

The spectral reflectance factor measured every 20 nm with a spectral bandwidth of 20 nm. Dimension 4 is: R(400), R(420),. . . ,R(700). Total of 16 variables

A.1.6 R457FL90

The ISO brightness measured according to ISO 2470 on a fluorescent paper. The ISO brightness is a measure of the radiance factor in the blue region and is usually expressed in percentage points. When the measurement is made on fluorescent samples, the illumination in the instrument needs to be adjusted with regard to its UV content.

A.1.7 R457NF90

This is the variable studied in this thesis

It is the ISO brightness measured according to ISO 2470 on a non-fluorescent paper. The ISO brightness is a measure of the radiance factor in the blue region and is usually expressed in percentage points. When the measurement is made on non-fluorescent samples, the illumination in the instrument does not need to be adjusted with regard to its UV content.

A.1.8 TristimC2

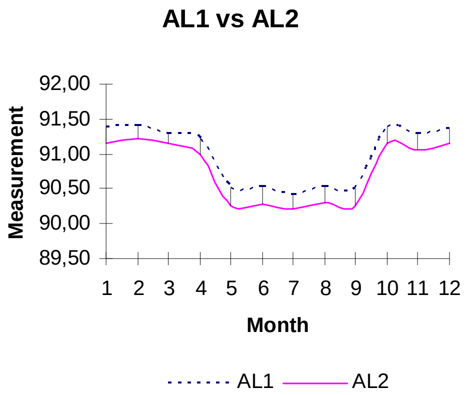
The Zeiss Elrepho tristimulus values Rx, Ry, Rz and CIE tristimulus values X, Y, Z. The tristimulus values are weighted averages of the spectral reflectance factors, e.g. . The colorimetric evaluation is made with the C/2° condition (CIE illuminant C and CIE 1931 standard observer). Dimension 4 is: Rx, Ry, Rz, X, Y, Z. Total of 6 variables

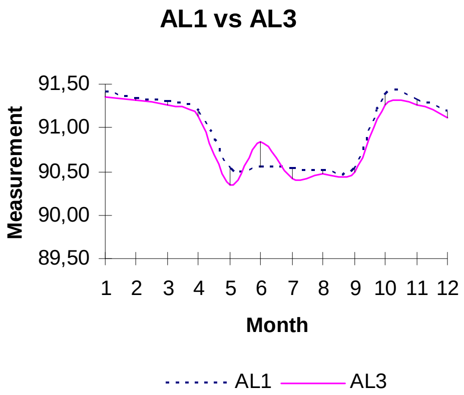
A.1.9 TristimD6510

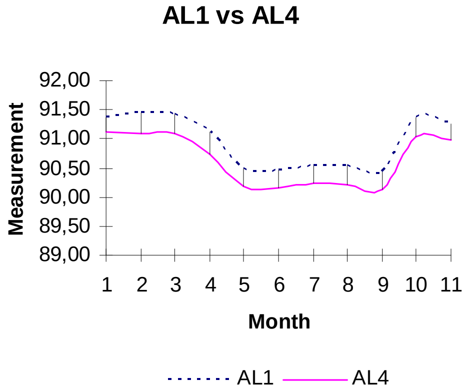
The Zeiss Elrepho tristimulus values Rx, Ry, Rz and CIE tristimulus values X, Y, Z. The tristimulus values are weighted averages of the spectral reflectance factors, e.g. . The colorimetric evaluation is made with the D65/10° condition (CIE standard illuminant D65 and CIE 1964 standard observer). Dimension 4 is: Rx, Ry, Rz, X, Y, Z.. Total of 6 variables

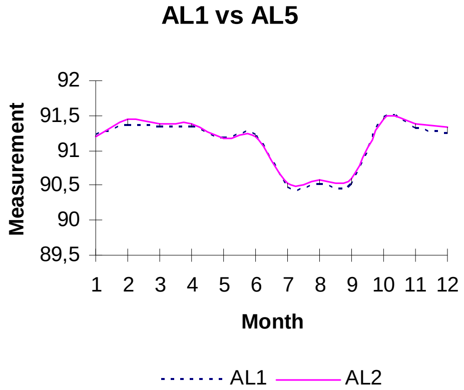
A.2 Graphs

A.2.1 AL1 sender

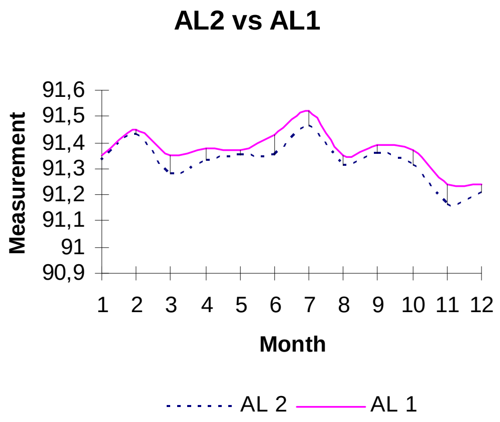


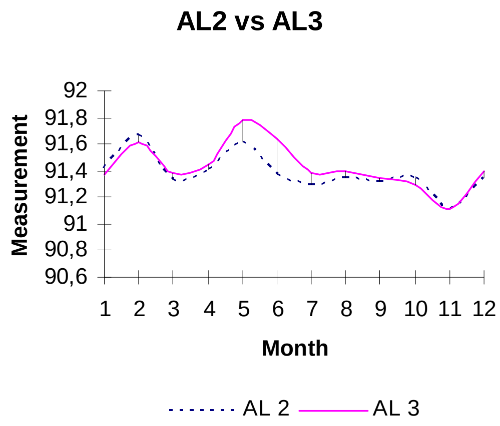


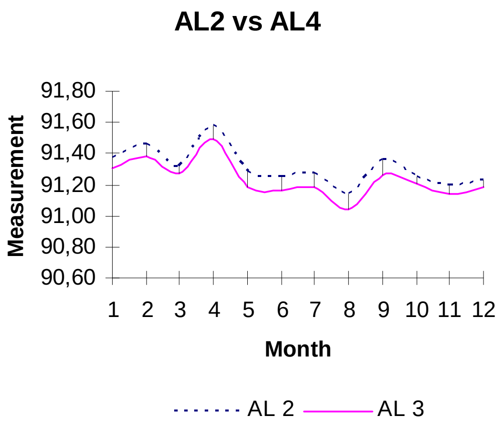


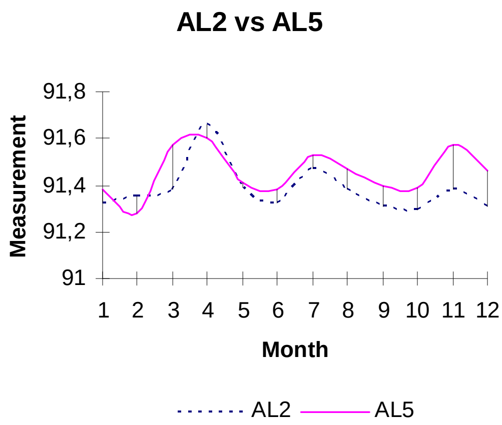


A.2.2 AL2 sender

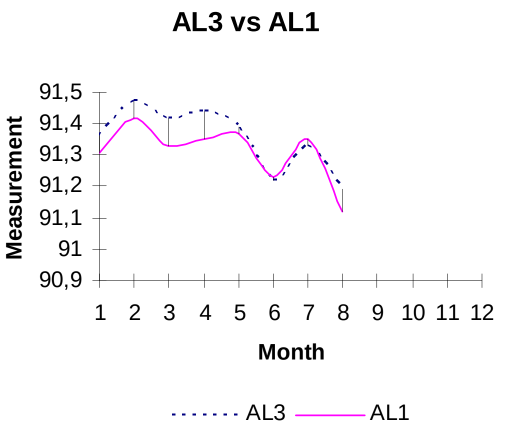


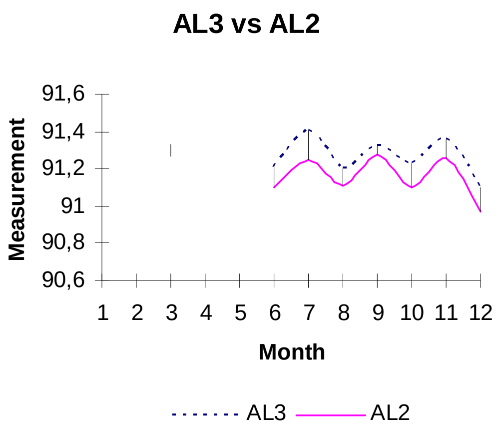


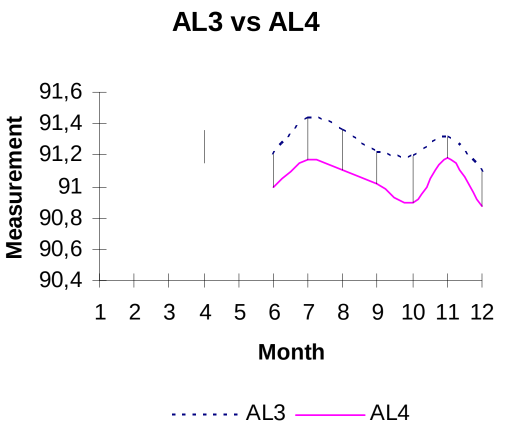


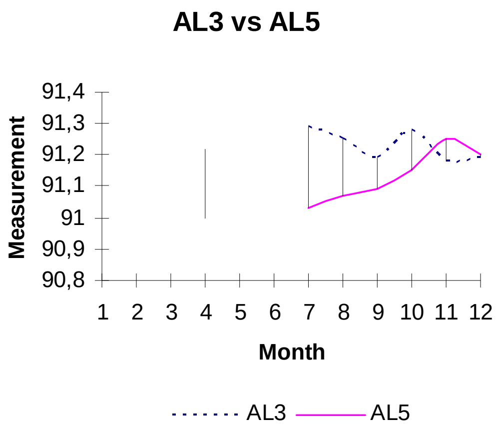


A.2.3 AL3 Sender

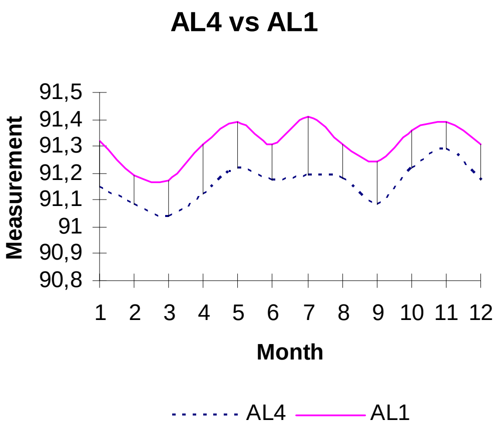


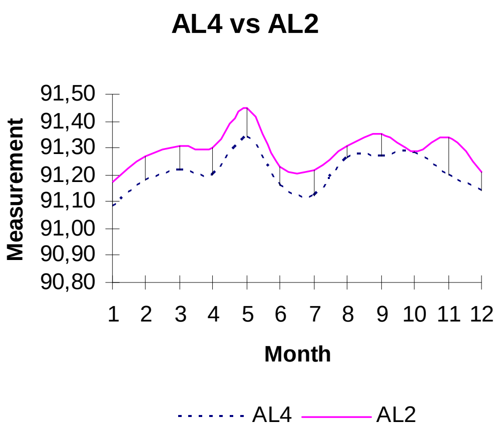


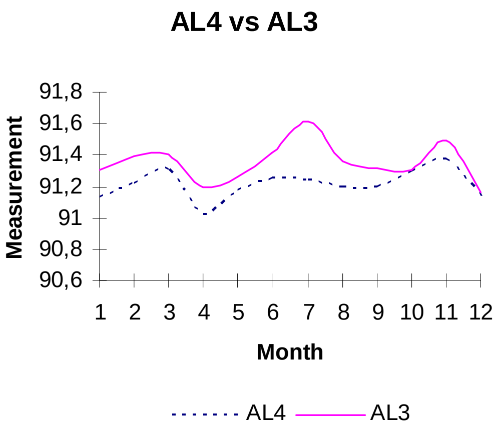


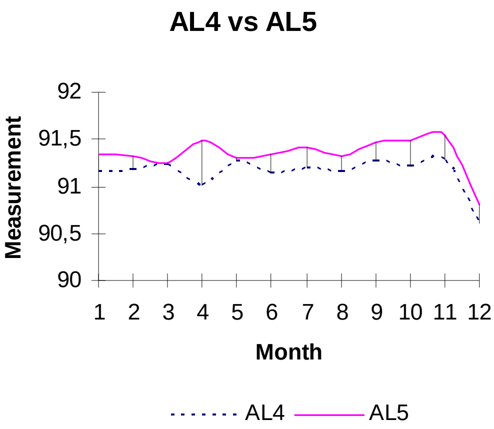


A.2.4 AL4 Sender

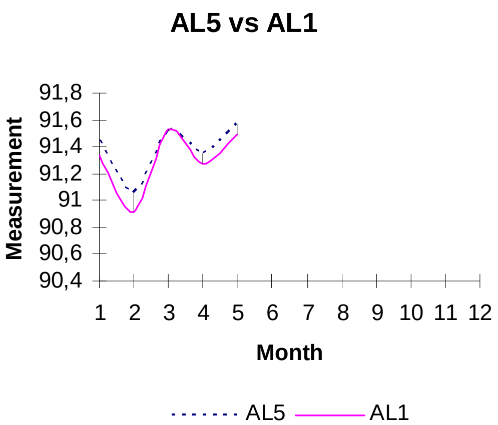


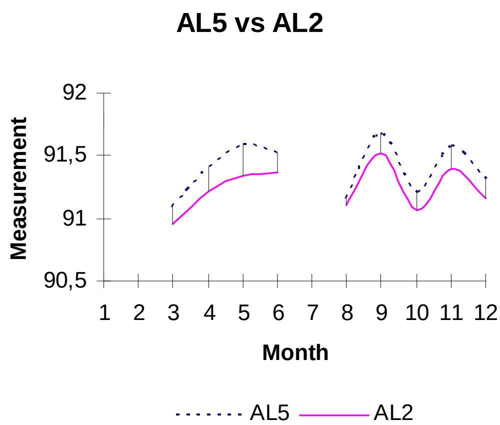


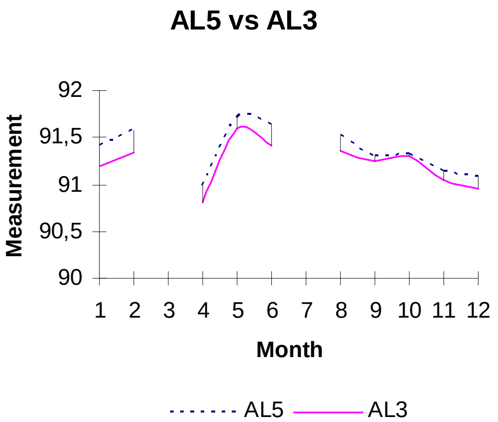


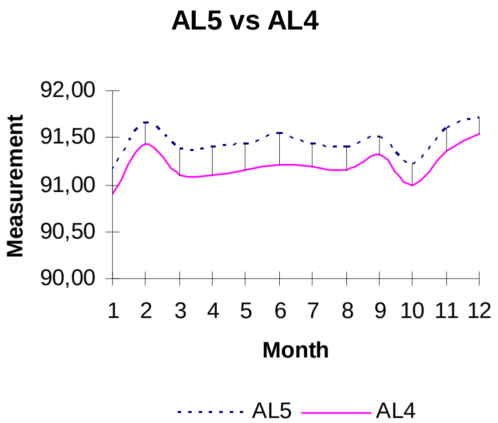


A.2.5 AL5 Sender


AL5 vs AL1
Measurement
91,8
91,6
91,4
91,2
91
90,8
90,6
90,4
1 2 3 4 5 6 7 8 9 10 11 12
Month
AL5 AL1


AL5 vs AL2
Measurement
92
91,5
91
90,5
1 2 3 4 5 6 7 8 9 10 11 12
Month
AL5 AL2


AL5 vs AL3
Measurement
92
91,5
91
90,5
90
1 2 3 4 5 6 7 8 9 10 11 12
Month
AL5 AL3


AL5 vs AL4
Measurement
92,00
91,50
91,00
90,50
90,00
1 2 3 4 5 6 7 8 9 10 11 12
Month
AL5 AL4

A.3 Results for section 4.2.2

Assumptions here are inherented from section 3.5.1

Here follows pariwise tests of the notion of symmetri mentioned in expression 3.4 in section 3.5.2. Double sided t-tests are made with a 5% level of significance.

The data needed for obtaining the test statistics are taken from Table 2 and 5.

1. $H_0 : \alpha_2 - \alpha_1 = 0$ vs $H_1 : \alpha_2 - \alpha_1 \neq 0$
   reject if $\frac{|\widehat{\alpha}_2 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12 + 12 - 2)$=2.074
   $\sqrt{2}\widehat{\sigma} = \sqrt{2} \cdot \sqrt{\frac{11 \cdot (0.03^2 + 0.02^2)}{(12-1)+(12-1)}} = \sqrt{2} \cdot 2.5495 \times 10^{-2} = 3.6055 \times 10^{-2}$
   $\frac{|\widehat{\alpha}_2 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} = \frac{0.22-0.05}{3.6055 \times 10^{-2}} = 4.715 > t_{0.025}(12 + 12 - 2) = 2.074$
   Decision : reject H_0

2. $H_0 : \alpha_3 - \alpha_1 = 0$ vs $H_1 : \alpha_3 - \alpha_1 \neq 0$
   reject if $\frac{|\widehat{\alpha}_3 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(11 + 7 - 2)$=2.120
   $\sqrt{2}\widehat{\sigma} = \sqrt{2} \cdot \sqrt{\frac{11 \cdot 0.11^2 + 7 \cdot 0.04^2}{(12-1)+(8-1)}} = \sqrt{2} \cdot 8.9536 \times 10^{-2} = 0.12662$
   $\frac{|\widehat{\alpha}_3 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} = \frac{2 \cdot 0.04}{0.12662} = 0.63181 < t_{0.025}(11 + 7 - 2) = 2.120$
   Decision: Do not reject H_0

3. $H_0 : \alpha_4 - \alpha_1 = 0$ vs $H_1 : \alpha_4 - \alpha_1 \neq 0$
   reject if $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(11 + 12 - 2) = 2.080$
   $\sqrt{2}\widehat{\sigma} = \sqrt{2} \cdot \sqrt{\frac{10 \cdot 0.04^2 + 11 \cdot 0.03^2}{(11-1)+(12-1)}} = \sqrt{2} \cdot 3.5119 \times 10^{-2} = 4.9666 \times 10^{-2}$
   $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} = \frac{0.32-0.15}{4.9666 \times 10^{-2}} = 3.4229 > t_{0.025}(11 + 12 - 2) = 2.080$
   Decision : reject H_0

4. $H_0 : \alpha_5 - \alpha_1 = 0$ vs $H_1 : \alpha_5 - \alpha_1 \neq 0$
   reject if $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12 + 5 - 2)$=2.131
   $\sqrt{2}\widehat{\sigma} = \sqrt{\frac{11 \cdot 0.05^2 + 4 \cdot 0.06^2}{(12-1)+(5-1)}} = \sqrt{2} \cdot 5.2852 \times 10^{-2} = 7.4744 \times 10^{-2}$
   $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_1|}{\sqrt{2}\widehat{\sigma}} = \frac{0.08-0.04}{7.4744 \times 10^{-2}} = 0.53516 < t_{0.025}(12 + 5 - 2) = 2.131$
   Decision: Do not reject H_0

5. $H_0 : \alpha_3 - \alpha_2 = 0$ vs $H_1 : \alpha_3 - \alpha_2 \neq 0$
   reject if $\frac{|\widehat{\alpha}_3 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12+8-2)$=2.101
   $\sqrt{2}\widehat{\sigma} = \sqrt{\frac{11 \cdot 0.09^2 + 7 \cdot 0.04^2}{(12-1)+(8-1)}} = \sqrt{2} \cdot 7.4647 \times 10^{-2} = 0.10557$
   $\frac{|\widehat{\alpha}_3 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} = \frac{0.05}{0.10557} = \mathit{0.47362} < t_{0.025}(\mathit{12} + \mathit{8} - \mathit{2}) = \mathit{2.101}$
   Decision: Do not reject H_0

6. $H_0 : \alpha_4 - \alpha_2 = 0$ vs $H_1 : \alpha_4 - \alpha_2 \neq 0$
   reject if $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12+12-2) = \ 2.074$
   $\sqrt{2}\widehat{\sigma} = \sqrt{2} \cdot \sqrt{\frac{11 \cdot (0.02^2 + 0.03^2)}{(12-1)+(12-1)}} = \sqrt{2} \cdot 2.5495 \times 10^{-2} = 3.6055 \times 10^{-2}$
   $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} = \mathit{0} < t_{0.025}(\mathit{12} + \mathit{12} - \mathit{2}) = \ \mathit{2.074}$
   Decision: Do not reject H_0

7. $H_0 : \alpha_5 - \alpha_2 = 0$ vs $H_1 : \alpha_5 - \alpha_2 \neq 0$
   reject if $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12+10-2)$=2.086
   $\sqrt{2}\widehat{\sigma} = \sqrt{2} \cdot \sqrt{\frac{11 \cdot 0.08^2 + 9 \cdot 0.05^2}{(12-1)+(10-1)}} = \sqrt{2} \cdot 6.8154 \times 10^{-2} = 9.6384 \times 10^{-2}$
   $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_2|}{\sqrt{2}\widehat{\sigma}} = \frac{0.16-0.07}{9.6384 \times 10^{-2}} = 0.93376$
   Decision: Do not reject H_0

8. $H_0 : \alpha_4 - \alpha_3 = 0$ vs $H_1 : \alpha_4 - \alpha_3 \neq 0$
   reject if $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_3|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(8+12-2)$=2.101
   $\sqrt{2}\widehat{\sigma} = \sqrt{\frac{7 \cdot 0.05^2 + 11 \cdot 0.09^2}{(8-1)+(12-1)}} = \sqrt{2} \cdot 7.6956 \times 10^{-2} = 0.10883$
   $\frac{|\widehat{\alpha}_4 - \widehat{\alpha}_3|}{\sqrt{2}\widehat{\sigma}} = \frac{0.22-0.14}{0.10883} = \mathit{0.73509} < t_{0.025}(\mathit{8} + \mathit{12} - \mathit{2}) = \mathit{2.101}$
   Decision: Do not reject H_0

9. $H_0 : \alpha_5 - \alpha_3 = 0$ vs $H_1 : \alpha_5 - \alpha_3 \neq 0$
   reject if $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_3|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(7+10-2)$=2.110
   $\sqrt{2}\widehat{\sigma} = \sqrt{\frac{6 \cdot 0.12^2 + 9 \cdot 0.07^2}{(7-1)+(10-1)}} = \sqrt{2} \cdot 9.3274 \times 10^{-2} = 0.13191$
   $\frac{|\widehat{\alpha}_5 - \widehat{\alpha}_3|}{\sqrt{2}\widehat{\sigma}} = \frac{0.15-0.12}{0.13191} = \mathit{0.22743} < t_{0.025}(\mathit{7} + \mathit{10} - \mathit{2}) = \mathit{2.110}$
   Decision: Do not reject H_0

10. $H_0 : \alpha_5 - \alpha_4 = 0$ vs $H_1 : \alpha_5 - \alpha_4 \neq 0$

reject if $\frac{|\widehat{\alpha}_5-\widehat{\alpha}_4|}{\sqrt{2}\widehat{\sigma}} > t_{0.025}(12+12-2)$=2.074
$\sqrt{2}\widehat{\sigma} = \sqrt{\frac{11\cdot(0.12^2+0.04^2)}{(12-1)+(12-1)}} = \sqrt{2}\cdot 8.9443\times 10^{-2} = 0.12649$
$\frac{|\widehat{\alpha}_5-\widehat{\alpha}_4|}{\sqrt{2}\widehat{\sigma}} = \frac{0.27}{0.12649} = \mathit{2.1346} > t_{0.025}(\mathit{12}+\mathit{12}-\mathit{2}) = \mathit{2.074}$
Decision: Reject H_0

Three hypothesis are rejected. Enough for not accepting the symmetri conditions. Joint test can be performed to investigate this more

A.4 Results for section 4.3.2

Now solving in Table 6 for:

1. row AL1 will give :

$$\begin{aligned} &\widehat{\alpha}_{12}-\widehat{\alpha}_{11}+\widehat{\alpha}_{13}-\widehat{\alpha}_{11}+\widehat{\alpha}_{14}-\widehat{\alpha}_{11}+\widehat{\alpha}_{15}-\widehat{\alpha}_{11} = \\ &= [\Sigma_j\alpha_{ij}=0] = -\widehat{\alpha}_{11}-\widehat{\alpha}_{11}-\widehat{\alpha}_{11}-\widehat{\alpha}_{11}-\widehat{\alpha}_{11} = \\ &= -5\widehat{\alpha}_{11} = -0.22-0.05-0.32+0.05 \end{aligned} \tag{A.1}$$

   $\widehat{\alpha}_{11} = 0.108$ and thereby every column in row 1 gets a solution. The solutions are

   $\widehat{\alpha}_{12} = -0.22+0.108 = -0.112$

   $\widehat{\alpha}_{13} = -0.05+0.108 = 0.058$

   $\widehat{\alpha}_{14} = -0.32+0.108 = -0.212$

   $\widehat{\alpha}_{15} = 0.05+0.108 = 0.158$

2. row AL2 will give :

$$\begin{aligned} &\widehat{\alpha}_{21}-\widehat{\alpha}_{22}+\widehat{\alpha}_{23}-\widehat{\alpha}_{22}+\widehat{\alpha}_{24}-\widehat{\alpha}_{22}+\widehat{\alpha}_{25}-\widehat{\alpha}_{22} = \\ &= [\Sigma_j\alpha_{ij}=0] = --\widehat{\alpha}_{22}-\widehat{\alpha}_{22}-\widehat{\alpha}_{22}-\widehat{\alpha}_{22}-\widehat{\alpha}_{22} = \\ &= -5\widehat{\alpha}_{22} = 0.04+0.04-0.08+0.08 \end{aligned} \tag{A.2}$$

   $\widehat{\alpha}_{22} = -0.016$

   $\widehat{\alpha}_{21} = 0.04-0.016 = 0.024$

   $\widehat{\alpha}_{23} = 0.04-0.016 = 0.024$

   $\widehat{\alpha}_{24} = -0.08-0.016 = -0.096$

   $\widehat{\alpha}_{25} = 0.08-0.016 = 0.064$

3. row AL3 will give :

$$\begin{aligned}\widehat{\alpha}_{31}-\widehat{\alpha}_{33}+\widehat{\alpha}_{32}-\widehat{\alpha}_{33}+\widehat{\alpha}_{34}-\widehat{\alpha}_{33}+\widehat{\alpha}_{35}-\widehat{\alpha}_{33}= \\ =[\Sigma_j\alpha_{ij}=0]=-\widehat{\alpha}_{33}-\widehat{\alpha}_{33}-\widehat{\alpha}_{33}-\widehat{\alpha}_{33}-\widehat{\alpha}_{33}= \\ =-5\widehat{\alpha}_{33}=-0.05-0.10-0.22-0.13\end{aligned} \tag{A.3}$$

$\widehat{\alpha}_{33}=0.1$

$\widehat{\alpha}_{31}=-0.05+0.1=0.05$

$\widehat{\alpha}_{32}=-0.10+0.1=0.0$

$\widehat{\alpha}_{34}=-0.22+0.1=-0.12$

$\widehat{\alpha}_{35}=-0.13+0.1=-0.03$

4. row AL4 will give :

$$\widehat{\alpha}_{41}-\widehat{\alpha}_{44}+\widehat{\alpha}_{42}-\widehat{\alpha}_{44}+\widehat{\alpha}_{43}-\widehat{\alpha}_{44}+\widehat{\alpha}_{45}-\widehat{\alpha}_{44}=$$

$$=[\Sigma_j\alpha_{ij}=0]=-\widehat{\alpha}_{44}-\widehat{\alpha}_{44}-\widehat{\alpha}_{44}-\widehat{\alpha}_{44}-\widehat{\alpha}_{44}= \tag{A.4}$$

$$=-5\widehat{\alpha}_{44}=0.14+0.09+0.14+0.20 \tag{A.5}$$

$\widehat{\alpha}_{44}=-0.114$

$\widehat{\alpha}_{41}=0.14-0.114=0.026$

$\widehat{\alpha}_{42}=0.09-0.114=-0.024$

$\widehat{\alpha}_{43}=0.14-0.114=0.026$

$\widehat{\alpha}_{45}=0.20-0.114=0.086$

5. row AL4 will give :

$$\begin{aligned}\widehat{\alpha}_{51}-\widehat{\alpha}_{55}+\widehat{\alpha}_{52}-\widehat{\alpha}_{55}+\widehat{\alpha}_{53}-\widehat{\alpha}_{55}+\widehat{\alpha}_{54}-\widehat{\alpha}_{55}= \\ =[\Sigma_j\alpha_{ij}=0]=-\widehat{\alpha}_{55}-\widehat{\alpha}_{55}-\widehat{\alpha}_{55}-\widehat{\alpha}_{55}-\widehat{\alpha}_{55}= \\ =-5\widehat{\alpha}_{55}=-0.08-0.15-0.15-0.25\end{aligned} \tag{A.6}$$

$\widehat{\alpha}_{55}=0.126$
$\widehat{\alpha}_{51}=-0.08-0.126=-0.206$
$\widehat{\alpha}_{52}=-0.15-0.126=-0.276$
$\widehat{\alpha}_{53}=-0.15-0.126=-0.276$
$\widehat{\alpha}_{54}=-0.25-0.126=-0.376$

A.5 Results for section 5

Groups of Systematic errors	
$\lvert\alpha_{ijj}\rvert \leq \max_{ij} s_{ij}$	$1.6 \cdot \max_{ij} s_{ij} \leq \lvert\alpha_{ijj}\rvert \leq 2.8 \cdot \max_{ij} s_{ij}$
$\widehat{\alpha}_{11} = 0.11$	$\widehat{\alpha}_{14} = -0.21$
$\widehat{\alpha}_{12} = -0.11$	$\widehat{\alpha}_{15} = 0.16$
$\widehat{\alpha}_{13} = 0.058$	$\widehat{\alpha}_{51} = -0.21$
$\widehat{\alpha}_{21} = 0.024$	$\widehat{\alpha}_{52} = -0.28$
$\widehat{\alpha}_{222} = -0.016$	$\widehat{\alpha}_{53} = -0.37$
$\widehat{\alpha}_{23} = 0.024$	$\widehat{\alpha}_{54} = 0.28$
$\widehat{\alpha}_{24} = -0.096$	
$\widehat{\alpha}_{25} = 0.064$	
$\widehat{\alpha}_{311} = 0.05$	
$\widehat{\alpha}_{32} = 0$	
$\widehat{\alpha}_{33} = 0.10$	
$\widehat{\alpha}_{344} = -0.12$	
$\widehat{\alpha}_{35} = -0.03$	
$\widehat{\alpha}_{41} = 0.026$	
$\widehat{\alpha}_{42} = -0.024$	
$\widehat{\alpha}_{43} = 0.026$	
$\widehat{\alpha}_{44} = -0.11$	
$\widehat{\alpha}_{45} = 0.086$	
$\widehat{\alpha}_{55} = 0.13$	

Table 9: The systematic errors are here divided in three groups, relating the size of their absolute values to the estimate of the standard deviation

