Quality Control of High Throughput Screening

Hervé Christian Zilliox

Master of Science (by Research)
in
Pattern Analysis and Neural Networks

Supervisor: Dr Ian T. Nabney

Aston University

September 1998

Aston University

Quality Control of High Throughput Screening

Hervé Christian Zilliox

Master of Science (by Research)
in
Pattern Analysis and Neural Networks, 1998

Thesis Summary

High Throughput Screening (HTS) is an efficient way of assessing the biological activity of a large number of compounds in order to determine the few compounds that could lead to the development of a pharmaceutical product of commercial value. The process consists of screening a large number of mixtures using the standard 96-well plate featuring amongst others six specific control wells whose expected value is known. Because the measurement technique is subject to variation and because of the large number of plates involved, the quality assessment of the data is difficult and therefore automation appears to be a necessity. We propose a three-step procedure for the quality control of the data. It first consists of a study based on control wells, where a Gaussian mixture, trained with the EM algorithm, models the distribution of the control values to determine general variations on a whole screen together with any errors that would affect the control wells. The second step relies on normal wells and is based on a plate to plate comparison and an intra-plate variation detection that aims at spotting general effects such as handling mistakes or blocked jets. The Kolmogorov-Smirnov procedure was chosen to perform inter-plate comparisons whereas Siegel-Tukey and Wilcoxon tests investigate differences in spread and location in the data within a plate. The Analysis of Variance techniques complete the quality control of the screening process by focusing on the detection of systematic edge and corner effects.

Keywords: Gaussian Mixture Model, EM algorithm, Statistical tests, Analysis of Variance, High Throughput Screening.

Acknowledgements

I am grateful to Pfizer Research for funding the work described in this thesis

I would also like to thank Dr Ian Nabney for his help, the advice and useful comments he provided me with.

I must express my thanks to Wilma Keighley for the time spent in introducing me to the main issues of the project.

And eventually, I am very indebted to Bruce Williams for his help with the practical aspects of the project. I would like to thank him not only for collecting the data and spending time to show me how the whole HTS process was conducted, but most of all for answering all my questions with so much kindness.

Contents

1 **Introduction** **9**

1.1 High Throughput Screening 9

1.1.1 The process step by step 10

1.1.2 Comments on the process 13

1.2 Quality control of High Throughput Screening 13

1.2.1 The analysis of the data on a simple example 14

1.2.2 The aims of the project 18

1.3 Thesis overview 19

2 **Novelty detection** **20**

2.1 Preliminary work 20

2.1.1 Why relying on control values? 21

2.1.2 Outliers : the limitations of the traditional approach 21

2.2 A new approach : the mixture model 22

2.2.1 A brief presentation of mixture models 23

2.2.2 The *Expectation-Maximisation* (EM) algorithm 25

2.3 Mixture models in the quality control context 27

2.3.1 Training and validating the model 27

2.3.2 The novelty threshold 29

2.4 Reliability of the novelty detection algorithm 29

2.4.1 The random seed factor 29

2.4.2 Seed factor in the training set *Versus* EM seed 31

2.4.3 Influence of the size of the training set 33

2.4.4 The same ranking according to the likelihood ? 35

2.4.5 Conclusions 35

3 **Inter- and intra-plate variation** **37**

3.1 A plate to plate comparison 38

3.1.1 The Kolmogorov-Smirnov test 38

3.1.2 Outlier detection 39

3.2 Intra-plate variation 44
3.2.1 Method 46
3.2.2 The testing procedures 47

4 Edge and Corner effects **50**
4.1 Hidden assumptions 51
4.1.1 Normality of the data 52
4.1.2 Equality of variances 59
4.2 The Analysis of Variance (ANOVA) procedure 60
4.2.1 Single-Factor analysis 60
4.2.2 Multiple comparison 63

5 Application to Screens **64**
5.1 Novelty detection 64
5.1.1 Specific novelty detection: screen 2 65
5.2 Inter-plate variation 69
5.2.1 A lack of accuracy? 69
5.2.2 Limitations 71
5.3 Intra-plate variation 73
5.3.1 Two complementary procedures 74
5.3.2 Limitations 77
5.4 Edge and corner effects 77
5.4.1 Results of the detection of any systematic effect 78
5.4.2 Unbalanced samples 78

6 Conclusions **82**
6.1 A Three-step procedure 82
6.1.1 A novelty detection method based on control values 82
6.1.2 Inter- and intra-plate variations 83
6.1.3 Edge and corner effects 85
6.2 Comments 85
6.2.1 Achievements 85
6.2.2 Discussions 86
6.3 Quality control applied on real life processes 88
6.3.1 Analysing the control wells under the existing system 88
6.3.2 A fully automated procedure 90

A Reference screens **93**
A.1 Screen 1 93
A.1.1 HTA plates 93

A.1.2 Control plates 93
A.2 Screen 1b 94
A.2.1 HTA plates 94
A.2.2 Control plates 94
A.3 Screen 2 95
A.3.1 HTA plates 95
A.3.2 Control plates 95
A.4 Screen 9 96
A.4.1 HTA plates 96
A.4.2 Control plates 96
A.5 Screen 12 97
A.5.1 HTA plates 97
A.5.2 Control plates 97

B Tests of Normality **98**
B.1 Preliminary measurements 98
B.1.1 Skewness 98
B.1.2 Kurtosis 99
B.2 Graphical representations 99
B.3 Goodness-of-fit procedures 99
B.4 Conclusion 101

C Harmonic interpolation **102**
C.1 A note on interpolation 102
C.2 Harmonic interpolation 103

List of Figures

1.1 Process flow for a typical HTS operation 10

1.2 96-well plate 14

1.3 Quality control: visualisation of the control wells 16

1.4 Enzyme activity *versus* compound concentration 17

1.5 *Totals* & *NSB's* and IC_{50} plates 17

2.1 Training and validation procedure 28

2.2 Proportion of rejected plates per day (screen 2) 30

2.3 Distribution of the minimum control values 32

2.4 Effects of the seed factor on the number of novel plates for screen 2 32

2.5 Effects of the seed factor on the training and validation error (screen 2) . . . 33

2.6 Size of the Training set versus novel points and log-likelihood 34

3.1 Distribution of the number of outliers per plate in an 'ideal' case 40

3.2 Whole population versus populations purged of outliers (screen 1) 41

3.3 Distribution of the number of outliers per plate for screen 1b 45

3.4 Distribution of the variances per row (screen 9) 46

4.1 Corner, edge and middle populations 51

4.2 Distribution of corner, edge and middle populations for screen 1b 54

4.3 Corner, edge and middle populations without outliers (screen 1b) 56

5.1 Novelty detection: plates 1 to 52 (screen 2) 66

5.2 Novelty detection: plates 53 to 104 (screen 2) 67

5.3 Novelty detection: plates 105 to 156 (screen 2) 67

5.4 Novelty detection: plates 157 to 206 (screen 2) 68

5.5 Visualisation of suspicious plates (screen 9 and 12) 71

5.6 Multiple comparison procedure (result of a 100 computations) 81

6.1 Determination of the complexity of the Gaussian Mixture model 89

6.2 Inter-plate variation detection 91

B.1 Normal quantile plots: corner, edge and middle populations of screen 12 . . . 100

List of Tables

2.1 Proportion of rejected plates per assay for five different seeds (screen 2) . . . 31

2.2 Screen1, Plate 74 : Ranking for four different random seed 35

3.1 Outlier detection 45

4.1 Skewness and Kurtosis to test for Normality (screen 1b) 53

4.2 Lilliefors goodness-of-fit test for Normality 58

4.3 F test for equality of variances 61

5.1 Novelty detection applied to the tested screens 64

5.2 Proportion of novel plates per assay (screen 2) 65

5.3 Kolmogorov-Smirnov procedure to detect suspicious plates 70

5.4 Testing the sensitivity of the Kolmogorov-Smirnov procedure 72

5.5 Kolmogorov-Smirnov procedure: time *versus* accuracy 73

5.6 Preliminary Siegel-Tukey procedure 74

5.7 Testing the accuracy of the methods investigating intra plate variations . . . 75

5.8 Single-factor ANOVA 79

5.9 Multiple comparison based on Tukey's method 80

B.1 Testing Normality: measure of Skewness applied to screen 12 99

B.2 Testing Normality: measure of Kurtosis applied to screen 12 99

B.3 Lilliefors goodness-of-fit test for Normality applied to screen 12 101

Chapter 1

Introduction

Recent advances in molecular biology have resulted in a large number of new biological targets. This offers exciting opportunities for drug discovery in so far as the success of such a process mainly relies on its capacity to detect some new entities or simply to spot new properties of the existing chemical compounds. *High Throughput Screening (HTS)* technologies can be used to identify those entities that chemically react with a given therapeutic target.

This chapter provides the minimum background to understand what High Throughput Screening actually is. For further details, see (Bojanic et al., 1997). It also gives an overall presentation of the quality control of HTS, before introducing the aims of the project: the detection of any general source of errors that could alter the quality assessment of the data. The final section then consists of an overview of the thesis where some indication is given about the procedure we propose to follow to tackle the problem.

1.1 High Throughput Screening

The three elements: biological target, compound library and assay method constitute the cornerstone of screening for therapeutic drugs. Recent progress in Genomics research together with Combinatorial chemistry have provided a variety of new compounds for screening. Advances have also been made in assay technologies, robotics and computation which now enables experiments to be carried out involving a large number of mixtures featuring millions of molecules.

High Throughput Screening has absorbed these new techniques and technologies to a large extent so that a vast range of compounds can now be screened. This aims at determining the few compounds, called 'lead compounds', whose biological activity is relevant and that could be developed into a product of commercial value. The method is all the more efficient in that a huge volume of data is examined and more importantly, the relevant information is found, as quickly as possible.

1.1.1 The process step by step

The HTS process can be described as the succession of the five following steps: compounds supply, assay, data capture, data analysis and sample follow-up. Figure 1.1 depicts the basic process flow of a generic HTS operation. It should be noted that a typical screen only features one of the four following measures: luminometric, fluorimetric, radiometric and colorimetric measure.

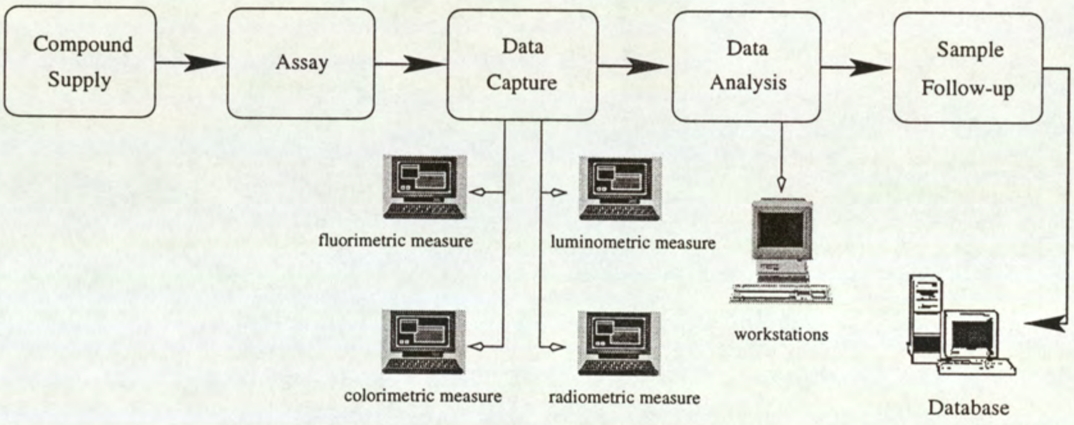


Figure 1.1: Process flow of a **HTS** operation

step 1 : supplying the compounds

The starting point of the HTS process is the supply of the compounds and the storage of the samples. First, an important consideration is to define what kind of device will be used to convey the liquid samples through the whole process. The most common format is the *96-well plate* also called 'microtitre plate' because of the small volume of liquid each well contains: it features test-tubes having a volume capacity of up to 2 ml. But microplates of 384 wells or even higher formats are also available.

Compounds are supplied from dry samples which require specific care since they are dispensed in individual tubes, weighed, formatted to fit to the microtitre plate and designed to occupy a specific position. Usually the new compounds come from large libraries held by pharmaceutical companies.

The next stage is then to dissolve the dry samples, using the adequate solvent, before putting them for storage into a liquid sample. Once again, special care is required for the storage conditions (such as temperature, atmosphere) so as to avoid any deterioration.

step 2 : constructing the assay

As HTS assays are applied to thousands of mixtures, they have specific requirements — a limited number of handling steps or an automated process so as to increase throughput — which are not necessary for those experiments involving only a few samples. There is also a potential problem with HTS assays: because a typical screen features a large number of plates, it is impossible to prepare them simultaneously. Thus, a screen is divided into several assays each constituting a certain number of plates. Due to the fact that each assay (within a screen) is constructed at different periods of time, they are not subject to the same experimental conditions (as explained in Section 5.1.1), which could result in a difference in the incubation time, that is the period of time during which the compound remains active. A means of estimating this phenomenon is required, such as control wells on each plates.

Moreover, constructing the assay implies the use of various hardwares. First, the liquid handling and assay assembly consist of 96-well pipetting devices. Robotic sample processors with disposable tips fully computer programmable are also used for their accuracy, being provided in single or multiple probe formats each of which can be independently controlled with respect to volume dispensing. However, the price to pay for flexibility is that these devices are much slower than the 96-well dispensers.

Finally, the washing steps (or 'separation' steps) involve some filtration equipment as well as plate washers. The trouble is that they are mainly manual, therefore the procedure faces some possible handling errors.

step 3 : collecting the data

Typically, each well on a plate features 20 compounds, an enzyme and a substrate. Let us consider the following simple theoretical example. An assay determines for example whether the enzyme is being inhibited by the test compounds. The enzyme catalyses the breakdown of the substrate to a coloured product. The assay measures the amount of coloured substrate. If the compounds stops the enzyme functioning then little of the coloured product is formed: this corresponds to a minimum activity well (minimum control). On the other hand, if the compounds do not affect the enzyme's function, a lot of coloured product is formed, which corresponds to a maximum activity well (maximum control). This is an example of colorimetric measures.

Thus, in most of the cases, HTS makes use of signal detection instruments or 'counters' subject to potential errors (mechanical errors due to a programming mistake for instance) to measure the colorimetric but also radiometric, fluometric or luminometric activity of the wells, so as to have an estimation of the biological activity of the mixtures. All these measures constitute the data this study is based on.

step 4 : analysing the data

The data analysis is the key step of the HTS process and for this reason will be given a more detailed description in Section 1.2.1. Software handling of the whole process flow described in Figure 1.1 is crucial for data to be effectively handled: all the elements that are part of the analysis (details of the samples, parameters describing a screen and data from the detection instruments) must be integrated into the system and combined so as to display information as clear as possible. Basically, the data analysis follows two goals:

1. assay validation to ensure the validity of the experiment
2. decision making, to determine 'hits' *i.e.* mixtures whose biological activity is considered relevant.

The assay validation relies on control wells on each plate, dedicated to the assessment of the experimental conditions, whereas some graphical representations help the operator locate the wells whose activity is greater than a fixed threshold.

step 5 : follow-up of active samples

One of the purposes of HTS data analysis is to define the active samples on a given target. Once identified, these samples are the raw material of further screening investigations so as to define as accurately as possible a mixture of biological efficiency (by testing the effects of different mixture concentrations for instance).

On the final stage, the samples are submitted to lead optimisation to improve the potency of the compounds and selectivity of the mixtures. These final experiments, contrary to HTS, rely on an existing knowledge and are usually performed on a small number of samples. If their activity is relevant enough, the lead compounds are then released to the company library and database, labelled active towards the specified target.

1.1.2 Comments on the process

Despite being an accurate means of detecting the compounds that could be developed into a product of commercial value in the general context of drug discovery, the HTS process faces some difficulties. The first and most obvious one is that the procedures involved in HTS are often time-consuming. It is therefore crucial to concentrate on improving their efficiency, for instance by focusing on the detection of false hits induced by the inherent variation of the measurement techniques.

Moreover, however successful this method is to detect the lead compounds, it is limited by the chemical diversity of the laboratory's own library, as the success of HTS procedures is entirely subject to the selected mixtures to be screened.

Finally, due to recent advances in molecular biology, HTS has to deal with more and more screens, which results in the necessity of automation. Besides, the increasing pressure to find some more new therapeutics makes effectiveness become all the more crucial. From this viewpoint, the development of a computerised system may help HTS gain in efficiency. Thus computer controls are present throughout the HTS process, ranging from integrated softwares for robots to external computers for liquid handling. But the contribution of computerisation in data management, probably the keystone of HTS for obvious reasons, is undoubtedly the more significant: not only powerful databases are needed to cope with the massive quantity of information generated by the the screening, but it is also vital that the system allows access control to data, flexible programming tools and the possibility to query the data.

On the other hand, HTS offers some advantages. In addition to effectiveness, the HTS process possesses this improvement compared to traditional chemical schemes: very little information on the biological structure of the compounds is needed to run a screen.

Especially, one of the reasons why HTS is more and more popular is that a negative result is also a result in the sense that, if successes in HTS can lead to the development of a commercial product, the records of failures in data bases give also some information about how to design further experiments.

1.2 Quality control of High Throughput Screening

This section concentrates on the quality control of HTS by focusing on the data analysis. To start with, an example of a typical data analysis gives us a better understanding of the kind of problems operators have to deal with in term of quality control. This is followed by a presentation of the different plate formats involved in the HTS process, part of which constitutes the raw material for this study. The last part presents the aims of this project.

1.2.1 The analysis of the data on a simple example

Although simplified, the example of data analysis that follows gives an idea of what it consists of together with the kind of problems that could occur in the real life HTS procedures.

the *standard 96-well plate*

The standard 96-well plate typically constitutes the basis of the HTS process. Each plate features 90 different mixtures, a mixture being composed of 20 compounds (or 'dry samples'), a substrate and an enzyme, and 6 control wells. Figure 1.2 gives a representation of the standard 96-well plate.

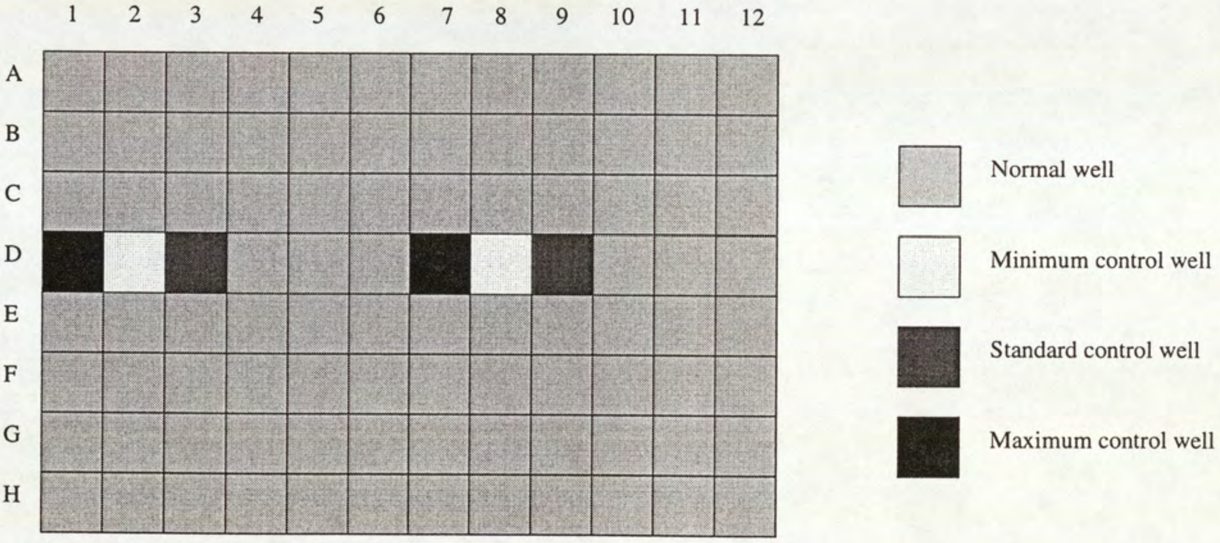


Figure 1.2: *96-well* plate

Data analysis usually features a first stage that consists of checking the control wells to ensure the quality of the data. This first step is crucial since it helps the operator assess the validity of the assay. To help him in his task, a graphical visualisation of the control wells enables him to detect major potential mistakes (such as handling mistakes) that affect a plate. These control values are then de-selected giving an end to the assay assessment.

To make things clearer, suppose one wishes to detect a novel enzyme inhibitor (*i.e.* a compound that stops an enzyme functioning). As mentioned in the previous section, in the context of colorimetric measures for instance, the enzyme catalyses the breakdown of a substrate to a coloured product. The six control wells on a plate have the following composition and are the same on each plate of a given screen.

- Position D1 & D7:

  the enzyme's functions are not affected in any manner. A lot of coloured product is formed; this corresponds to a *maximum activity* well.

- Position D2 & D8:

  the enzyme is fully inhibited. Little coloured product is formed: it corresponds to a *minimum activity* well.

- Position D3 & D9:

  the enzyme is partially inhibited by a compound known to have an 'average' effect. This corresponds to an *average activity* well or *standard control* well.

For an ideal screen, all the maximum (respectively minimum and standard) controls on all the plates of the screen should be the same (since they contain the same mixtures). In practice however, the activity boundaries (0% and 100%) that evaluate the actual inhibition of the enzyme are taken as an average of the maximum and minimum controls of the assay, as an assay[1] is subject to great variation (the role of standard controls is restricted to the quality control of the normal wells).

The control wells involved in this computation are selected or de-selected by the operator who relies on a graphical visualisation similar to Figure 1.3, the de-selection being conducted on the wells that differ by more than a standard deviation from the mean. The mean of the data is then re-computed (without the removed controls). All the controls are checked again and de-selected if necessary, with regards to the new data and so on, until all the values are considered as satisfying, which gives an end to the assay assessment.

The activity boundaries (0% and 100%) are then computed as an average of the maximum and minimum control values. The operator then sets a threshold above which a well is considered as a 'hit'. Each plate is then manually checked and all the wells whose activity are upon the threshold are kept for further studies, in order to determine the lead compounds.

Note: the de-selection of a maximum or minimum control well only affects the computation of the activity boundaries. In practice, no action is taken over other values of the plate even if this may indicate that some errors have taken place. It is only if the 6 control wells are suspicious that the corresponding plate is de-selected.

[1] an assay features a set of plates screened on the same date as stated in Section 1.1.1.

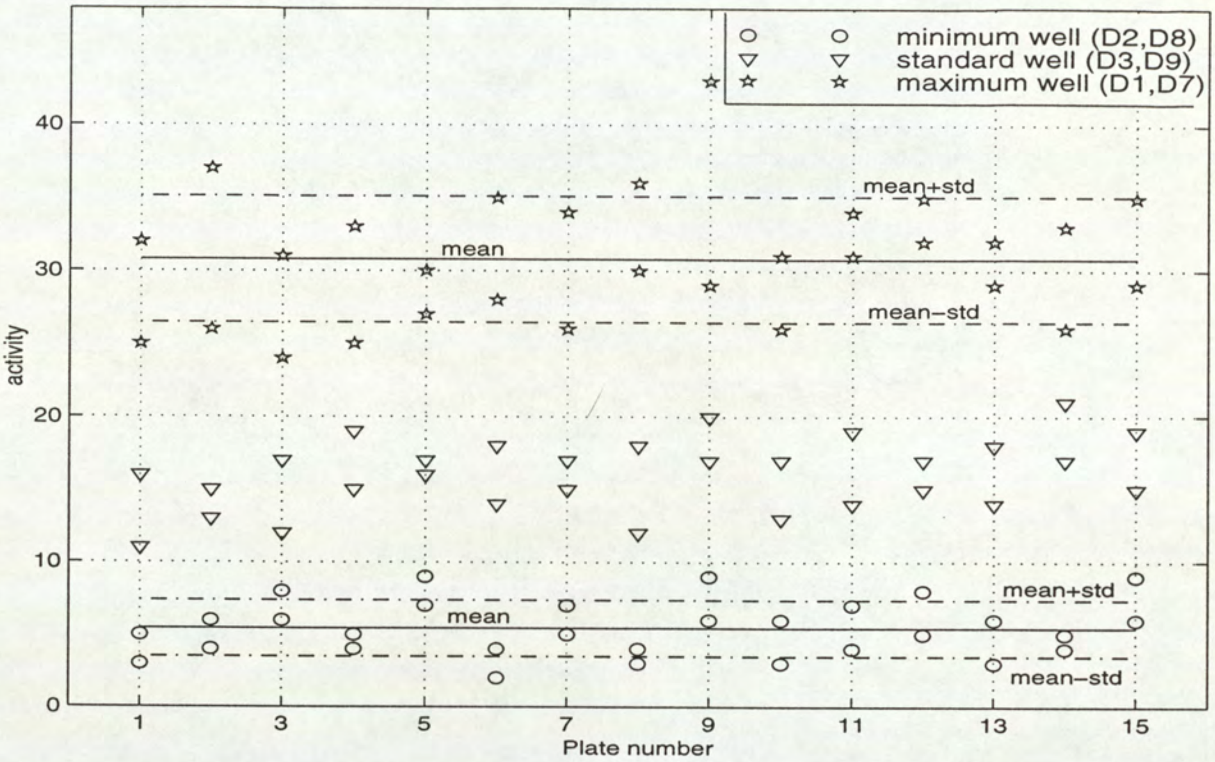


Figure 1.3: Quality control

the other plate formats used for this project

Other kinds of format apart from the 96-well plate are involved in HTS. Here is a brief presentation of the IC_{50} plates and the *Totals* & *NSB's* (where *Totals* stands for *totally inhibited* referring to the maximum controls and *NSB* for *Non Specific Binding i.e.* the minimum controls.

IC_{50} is defined as the concentration of compound causing a 50% reduction in the effect of the enzyme under study (see Figure 1.4). The more potent the compound, the less is required to produce the 50% inhibition. Thus, the IC_{50} plates are generally used as the last step of the screening procedure. One specific compound is under investigation at different concentrations (close to the $IC50$ concentration) so as to determine its optimum concentration before it is registered as active towards the given therapeutic target (the enzyme). Indeed an IC_{50} plate features 12 columns: the same compound is disposed on 2 successive columns, each paired column being at a different concentration, and the first two are respectively dedicated to maximum and minimum controls.

The data analysis is conducted in a similar way as described previously, a hit being this time characterised by a specific compound together with the most efficient concentration.

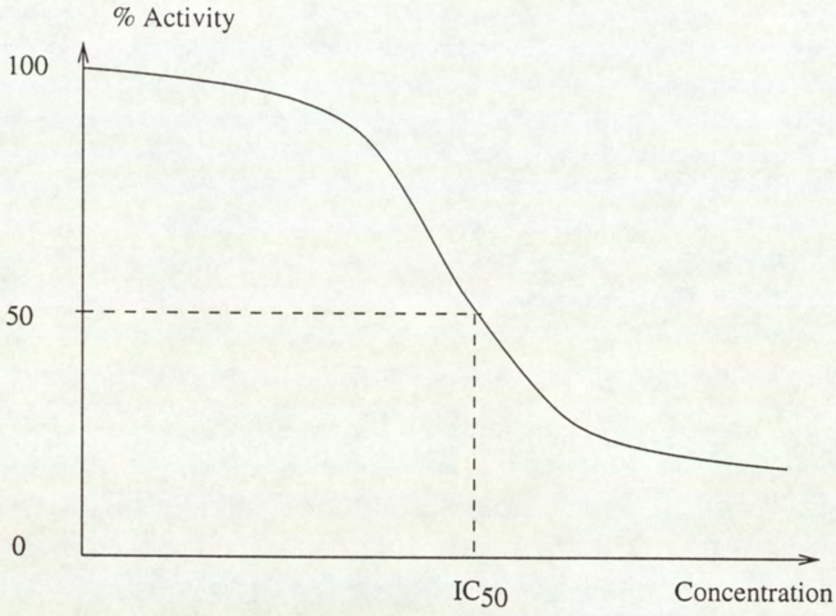


Figure 1.4: Enzyme activity *versus* compound concentration

The *Totals* & *NSB's* plates have been generated especially for this study, since they are a specific requirement of the model that was built up last year and which constitutes the first step of our procedures. It basically consists of a 96-well plate, half of which is dedicated to minimum control wells, the other half featuring maximum controls. Therefore, the designation *control plates* will be used in the rest of the thesis to mention *Totals* & *NSB's* plates, whereas a typical 96-well will be called *normal plate*. Figure 1.5 gives a representation of all the different plate formats mentioned above.

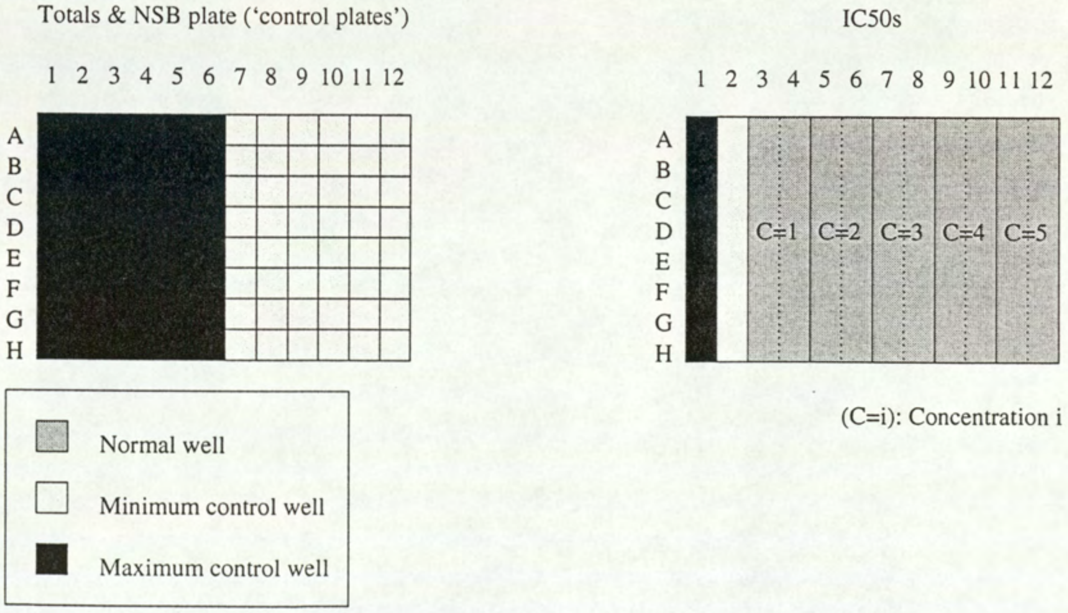


Figure 1.5: ***Totals*** & ***NSB's*** and IC_{50} plates

1.2.2 The aims of the project

Because the variations in the measurement techniques are not well understood, it is difficult to assess the quality of the data. As emphasised previously, currently the assessment of the plates is manual and greatly subjective. Besides, a typical screen features several hundreds of plates. These are the reasons why the quality assessment of HTS appears as a necessity.
The study carried out aims at assessing the quality of the HTS process as objectively as possible. Its aim is not to provide a fully automated quality assessment method, but should help the operator in his task by pointing out a few numbers of plates on which we suspect some errors have taken place. By doing this, the quality assessment of the data, conducted manually up to now, would not require to check every single plate on a screen (that can feature several hundreds plates in practice), which is a rather tedious task.
The study especially concentrates on the detection of any kind of mistakes or effects that could alter the data. This implies a twofold aspect: not only is this work based on control wells (that help the operator validate an assay as stated in Section 1.2.1) but it also considers the normal wells so as to detect some potential handling mistakes or hits, by thoroughly examining plate to plate variations. In addition, mechanical errors like blocked jets are also subject to investigation, for instance through an intra-plate variation detection.
Finally, the detection of edge and corner effects, with the goal of taking any action to improve the design of the experiments involved in the HTS process if necessary, completes the aims of this study and the quality control procedures.

It is to be noted that the method to assess the quality of the data should respect a double constraint:

1. The built-up procedures should not be too time-consuming, so as to enable some 'on-line' quality assessment of the data.

2. The procedures should be flexible enough to adapt quite easily to the different kind of situations they will have to cope with. They should take into account the various devices used to construct the assay (and that represent potential sources of errors), but also offer some possibility of automation without preventing any manual intervention.

1.3 Thesis overview

This thesis consists of six parts that follow the different steps we propose for the quality control of HTS. The second chapter presents the first step: the novelty detection, a study exclusively based on control values. The first three sections give a brief summary of the work that was done last year. A preliminary study justifies why we can rely on control values to assess the quality of the data. Then the reasons why a Gaussian Mixture Model was chosen to achieve the quality control are presented. It is followed by an explanation of what Gaussian mixtures are together with some details about the *Expectation-Maximisation* algorithm used to train the model. The final section investigates some sources of variations in last year's model.

The third chapter presents the statistical approach we propose to tackle the detection of effects altering the data values of normal wells. A first section details the plate to plate comparison that is used to detect any *general* effects (like handling mistakes) or hits. It introduces the general statistical procedure we followed to achieve that (the Kolmogorov-Smirnov test), together with the method computed. Some other methods dealing with outlier detection are also presented: we explain in particular why we did not follow the standard approach and chose a method based on an empirical definition of outliers.

The second section focuses on blocked jets detection. The testing procedures investigate both a difference in spread and a difference in location, through statistical methods such as the Siegel-Tukey and Wilcoxon tests.

Chapter 4 concentrates on the detection of edge and corner effects. We first present a study investigating whether the assumptions hidden behind the procedure applied, the Analysis of Variance, held. This study is based on some visual representations, numerical evaluations (with some measurements such as the Skewness and the Kurtosis), and some statistical tests. Then the two-step method itself is described: a single factor analysis to determine whether there is any evidence for edge or corner effects and a multiple comparison to point out which effect is predominant.

The results of these different steps are presented in Chapter 5, where the difficulties encountered are discussed. The final chapter gives a summary of the study, some comments about the method used. The results achieved and the main limitations of the built-up procedures are summed up. The last sections discuss some possible ways of implementing the procedures so as to apply them on real life processes.

Chapter 2

Novelty detection

In any industrial task, novelty detection is used to determine an unusual output; in the context of quality control of HTS, novel observations can be defined as data points whose value significantly differ from other data taken under the same experimental conditions.

This chapter presents the first step of the quality control procedure. It is based on the model that was built up last year whose main characteristic is that it only relies on control wells to assess the HTS process. The first section explains why we can concentrate only on the control values to assess the data quality. It also puts forward the limitations of the standard approach dealing with outliers.

The second part gives a general presentation of the model. In particular, it explains what motivated the choice for a Gaussian mixture model but also describes the main algorithm used to train the Gaussian mixture: the *Expectation-Maximisation (EM)* algorithm.

Mixture models are then presented in the quality control context. This section gives a *technical* presentation of the model implemented for HTS: how to use it in practice, in particular how to train it on the data set to achieve the novelty detection and what its parameters are.

The final section presents the work carried out at the beginning of this year: an investigation on the reliability of the novelty algorithm by studying some sources of variation in the mixture model.

2.1 Preliminary work

The point is not to explain in detail the testing procedure carried out last year but just to explain what motivated the choice of a method based on density inference to tackle the detection of abnormal plates.

2.1.1 Why relying on control values?

To start with, some statistical procedures such as the χ^2 and Kolmogorov-Smirnov tests were carried out to investigate the goodness-of-fit on a Gaussian distribution. These experiments focused on a comparison between the observed[1] and expected distribution, comparison based both on a visual and numerical analysis.

The results obtained were rather mixed : for a given screen, whereas the χ^2 statistical test rejected the normality of the data, the Kolmogorov-Smirnov procedure gave some evidence for Normality at the same time. But even if the Kolmogorov-Smirnov test is said to be the right test[2] to apply in such a situation as it is mentioned in (Neave and Worthington, 1988), Normality was subject to uncertainty and the hypothesis was rejected.

A second statistical procedure tested the correlation between the various control wells. The tests were carried out using the Pearson's sample correlation coefficient and gave positive results : all the controls of the 96-well plate were mutually correlated with a strong correlation between maximum and standard controls. Given that there's no difference between the standard controls and the normal wells on a normal plate[3], this strong correlation tends to validate, from a statistical point of view, a study based on control values only to assess the quality of the data. And such a correlation also allows a detection of unusual variations on a specific plate in so far as an unusual control well gives evidence that something went wrong for the whole studied plate. To tackle the detection of unusual values, a first and natural idea was to use the techniques proposed by the traditional approach.

2.1.2 Outliers : the limitations of the traditional approach

As mentioned in (Barnett and Lewis, 1978), the detection of outliers is far from being an easy task. A standard approach proposes the following two steps to detect outliers:

1. Use visual techniques to spot extreme observations.

[1] the samples tested consist of different screens, each of which features three control plates : 144 minimum and 144 maximum wells

[2] applied to unbinned distributions, it is more reliable than the χ^2 test since it does not require any arbitrary categories.

[3] the only difference between a standard control and a normal well is that the activity of the former is known.

2. Apply some 'discordancy tests' *e.g.* tests for outliers to determine whether the points selected in the first step significantly differ from the rest of the sample (with regard to the chosen significance level).

The literature gives a plethora of tests to detect outliers ((Barnett and Lewis, 1978) gives a quick overview of more than 20 just for the Gaussian case): it is however to be noted that most of the techniques these tests refer to assume an underlying Gaussian density.
Besides, even if this two-step procedure seems at first sight quite efficient and relevant, it is weakened by some drawbacks. First, the procedure is greatly subjective since a set of outliers to be tested has to be chosen at first hand, according to some graphical visualisation that does not necessarily show a good accuracy. Besides, this graphical inspection hardly enables an automated process that seems to be a necessity in the context of HTS, given the increasing number of data involved in the process.
In addition, (Barnett and Lewis, 1978) underlines that many of the discordancy statistics have to deal with the problem of *'masking'*, which alters the discordancy of some extreme observations under investigation as outliers, because of the presence of some other *less extreme* observations that were not considered as outliers.
Finally and this is surely the most frustrating aspect of the traditional approach, the method is not based on a density estimation. This means that no description of the data is provided in terms of probability (contrary to the novelty detection) and therefore it is impossible to determine which extreme value is more likely to be an outlier, *e.g.* it is impossible to obtain a ranking of the extreme observations.

All these reasons explain why we chose an alternative and simple method rather than following the standard approach that is far too complicated, time-consuming and too inaccurate to satisfy our needs.

2.2 A new approach : the mixture model

This section introduces the theory behind last year's model. First we shall recall why mixture models have been chosen to detect novel plates.

2.2.1 A brief presentation of mixture models

Why mixture models?

Many arguments motivate the choice of mixture models to infer the density of the HTS control values. First, the mixture models have a kind of universal approximation property, since they can fit any probability density. Besides, from a practical point of view, an advantage of mixture models lies in their speed of evaluating the density at a new data point, which is undoubtedly an asset in a quality control procedure.

What are mixture models?

Mixture models represent the density function $p(\mathbf{x})$ of the data as a linear combination of M basis functions in the form :

$$p(\mathbf{x}) = \sum_{j=1}^{M} p(\mathbf{x}|j)P(j) \quad , \tag{2.1}$$

where $p(\mathbf{x}|j)$ is the probability density or likelihood that $\mathbf{x}$ is from component j and $P(j)$ the *mixing parameters* or *prior* probability of the data point being generated from the component j of the mixture.

- The priors are chosen to satisfy the constraints

$$\begin{cases} 0 \leq P(j) \leq 1 \\ \\ \sum_{j=1}^{M} P(j) = 1 \quad . \end{cases} \tag{2.2}$$

- Similarly, the component densities $p(\mathbf{x}|j)$ of the mixture are :
  - normalised :

$$\int p(\mathbf{x}|j)\, d\mathbf{x} = 1 \ , \ \forall j \ , \ j = 1 \ , \ \ldots \ , \ M \quad . \tag{2.3}$$

  - chosen to be Gaussian density functions :

$$p(\mathbf{x}|j) = \frac{1}{(2\pi)^{(d/2)}|\mathbf{\Sigma}_j|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)^T \Big\} \ , \tag{2.4}$$

  where $\boldsymbol{\mu}_j$ is the mean and $\mathbf{\Sigma}_j$ the covariance matrix of the component j.

The covariance matrix $\mathbf{\Sigma} = [\sigma_{i,k}]_{i,k=1\ldots d}$ of the Gaussian mixture models the covariance[4] of the underlying random variables X_i, X_k. Usually the covariance matrix used in mixture models can be of three different types :

1. Full covariance matrix
2. Diagonal matrix : $Diag(\sigma_1^2, \ldots, \sigma_d^2)$ where the σ_i are not necessary equal
3. Covariance matrix $\sigma^2 I$ where I is the Identity matrix

with the following properties :

- Full covariance matrix
  - no constraint is made about the model
  - the inversion of $\mathbf{\Sigma}$ in Equation 2.4 is computationally expensive
  - the curves of equal density are ellipses without constraint on their axes
- Diagonal matrix
  - the model ignores potential correlations between the variables[5]
  - the inversion of $\mathbf{\Sigma}$ is easy since there are only d parameters
  - the curves of equal density are ellipses whose axes are directed by the vectors defining the axes of the graph
- The $\sigma^2 I$ covariance matrix
  - the model imposes that the elements on the diagonal are all equal the other ones being equal to zero
  - the inversion of $\mathbf{\Sigma}$ is trivial since there is no more than 1 parameter
  - the curves of equal density are circles

This type of model is generally known as a *Gaussian Mixture model.*
The density estimation therefore consists of determining the parameters of the model : $\{P(j), \boldsymbol{\mu}_j, \mathbf{\Sigma}_j, \ j = 1, \ldots, M\}$, especially to choose M, the number of basis functions in the model and the structure of the covariance matrix $\mathbf{\Sigma}$.

[4] $cov(X_i, X_k) = \mathcal{E}[(X_i - \mathcal{E}[X_i])(X_k - \mathcal{E}[X_k])]$ where $\mathcal{E}[X]$ represents the expectation of X.

[5] $cov(X_i, X_k) = \sigma_{i,k} = 0$ for $i \neq k$

2.2.2 The *Expectation-Maximisation* (EM) algorithm

The *Expectation Maximisation* algorithm (or EM algorithm) is the answer to the question of how to define the parameters of our model as it provides a simple method for estimating the mixture parameters.

- The estimation of the parameters

Many procedures have been developed for determining the parameters of a Gaussian Mixture Model, given a set of data. Most of them are maximum likelihood techniques and consist of maximising the *likelihood* $\mathcal{L} = \prod_{n=1}^{N} p(\mathbf{x}^n)$ of the parameters, which is equivalent to minimising the negative log-likelihood given by

$$E = -\ln \mathcal{L} = -\sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{M} p(\mathbf{x}^n|j)P(j) \right\} \quad . \tag{2.5}$$

that can be considered as an error function.

The EM algorithm then alternates two steps starting from a preliminary initialisation :

E-step* or *Expectation step : Determine $Q(\theta, \theta^{old})$, where $Q(\theta, \theta^{old}) = \mathcal{E}[\ln \mathcal{L}|\theta^{old}]$ The expectation is computed using the current, fixed values of the parameters.

M-step* or *Maximisation step: Define θ^{new} so as to maximise $Q(\theta, \theta^{old})$.

$\theta = \{P(j), \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j,\ j = 1, \ldots, M\}$ represents the set of parameters to be determined, and $Q(\theta, \theta') = \mathcal{E}[\ln \mathcal{L}|\theta']$, a function of the observed data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$. It consists of choosing θ^{new} for minimising the expectation $\mathcal{E}[E|\theta^{old}]$ and this leads to a new error E^{new}.
For Gaussian mixtures, the new error E^{new} admits[6] an upper bound such that

$$E^{new} - E^{old} \leq -\sum_{n=1}^{N}\sum_{j=1}^{M} P^{old}(j|\mathbf{x}^n) \ln \left\{ \frac{P^{new}(j)p^{new}(\mathbf{x}^n|j)}{p^{old}(\mathbf{x}^n)P^{old}(j|\mathbf{x}^n)} \right\} \quad . \tag{2.6}$$

If we let Q be the the right-hand side in (2.6) then we have $E^{new} \leq E^{old} + Q$ and so we can seek to minimise the upper bound $E^{old} + Q$ with respect to the *new* values of the parameters. Minimising Q (E^{old} is fixed) will necessarily lead to a decrease in the value of

[6] Jensen's inequality : $\ln\left(\sum_j \lambda_j x_j\right) \geq \sum_j \lambda_j \ln(x_j)$ if $\lambda_j \geq 0$ and $\sum_j \lambda_j = 1$ combined to the Equation (2.5) lead to the relation (Details are given in(Bishop, 1995)).

E^{new} unless it is already a local minimum.

For $\mathbf{\Sigma} = Diag(\sigma_1^2, \dots, \sigma_d^2)$ (the choice of such a covariance matrix is explained at the end of this section), the minimum of Q is obtained after differentiating E with respect to the parameters $P(j)$, μ_j and σ_j, which leads to the relations:

$$\mu_j^{new} = \frac{\Sigma_n P^{old}(j|\mathbf{x}^n)\mathbf{x}^n}{\Sigma_n P^{old}(j|\mathbf{x}^n)} , \tag{2.7}$$

$$(\sigma_k^{(j)\ new})^2 = \frac{\Sigma_n P^{old}(j|\mathbf{x}^n)(x_k^{(n)} - \mu_k^{(j)\ new})^2}{\Sigma_n P^{old}(j|\mathbf{x}^n)} , \tag{2.8}$$

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|\mathbf{x}^n) , \tag{2.9}$$

for $k = 1, \dots, d$ and $j = 1, \dots M$. Details and proof are available in (Bishop, 1995).

- The initialisations

  The initialisation procedure that sets the starting values of the updating relations (2.7), (2.8) and (2.9) is not to be neglected in so far as it can alter the performance of the EM algorithm when minimising the negative log-likelihood error (2.5). For the training procedure, the parameters take the following initialisations :

  Priors $P(j)$: $\frac{1}{M}$ for all j

  Centres μ_j : they are randomly chosen in the interval $[\min(\mathcal{T}) \max(\mathcal{T})]$ where $\mathcal{T} = \{\mathbf{x}_n\}_{n=1,\dots,N}$ with $\mathbf{x}_n = (x_1^{(n)}, \dots, x_d^{(n)})$ a d-dimensional vector representing the data set.

  The variance is set to

$$\min_{i \neq j} || \ \mu_i - \mu_j \ || \ .$$

Besides, as far as the implementation is concerned, the components σ_j^2 (for $j = 1, \dots, d$) of the covariance matrix $\mathbf{\Sigma}$ are checked at each step of the EM algorithm and re-initialised if needed so as to avoid ill-conditioned matrices (if variance collapses to zero).

Now that the parameters of the Gaussian mixture are defined, it is to be added that some testing procedures carried out last year showed that the diagonal matrix, as a structure for the covariance matrix, is the best compromise when applied to novelty detection on HTS. And finally, some empirical studies based on the performance of the trained mixture model set the number of basis functions to two. So our mixture model is now completely defined.

2.3 Mixture models in the quality control context

Since there is no difference between the standard controls and the normal wells, the correlation tests mentioned in Section 2.1.1 validated an approach based on control values only to assess the quality of the data, which is obviously convenient since the studied data points are far less numerous (6 controls on a normal plates) than normal wells (90 on a normal plate). Thus the new approach focuses on the control values rather than the normal wells.

In addition, if we know what the distribution of 'normal' values looks like, from a probabilistic point of view, the points considered as novel are the ones that are unlikely for this distribution. So the novelty criterion is only a matter of defining properly a certain threshold.

This section briefly explains how the distribution of the control wells is modelled (it especially describes the technique used) and also how the novelty threshold is defined.

2.3.1 Training and validating the model

The approach basically consists of 'learning' the distribution of control values. To do that three control plates (featuring only maximum and minimum control values as stated in Section 1.2.1) are added at the beginning of each screen. The Gaussian Mixture model is then trained and validated on these data as explained in the following sections.

Cross-validation

The Gaussian mixture model is trained by minimisation of the error function (2.5) with respect to a set of data. But we can not be sure that the minimisation of this error function for a single data set gives a good performance of the model when applied on new screens. It is the reason why we proceed by *cross-validation* to determine the model with the best performance.

What does cross-validation consist of? Basically the three additional control plates, that appears to be references with regard to the distribution of control values, are used to generate two independent sets of data, as explained in Figure 2.1: a *training* set and a *validation* set. From the initial 2-dimension vector featuring the 144 control values added at the beginning of each screen, a random selection is first made to split it into two independent 2-dimension vectors V_{training} and $V_{\text{validation}}$. Each of these vectors V_{training} and $V_{\text{validation}}$ are then used to generate a $4-tuple$, where each maximum column (respectively minimum column) represents a random selection of maximum values in V_{training} (respectively minimum value in $V_{\text{validation}}$), that constitute the training set $\mathcal{T}$ and the validation set $\mathcal{V}$.

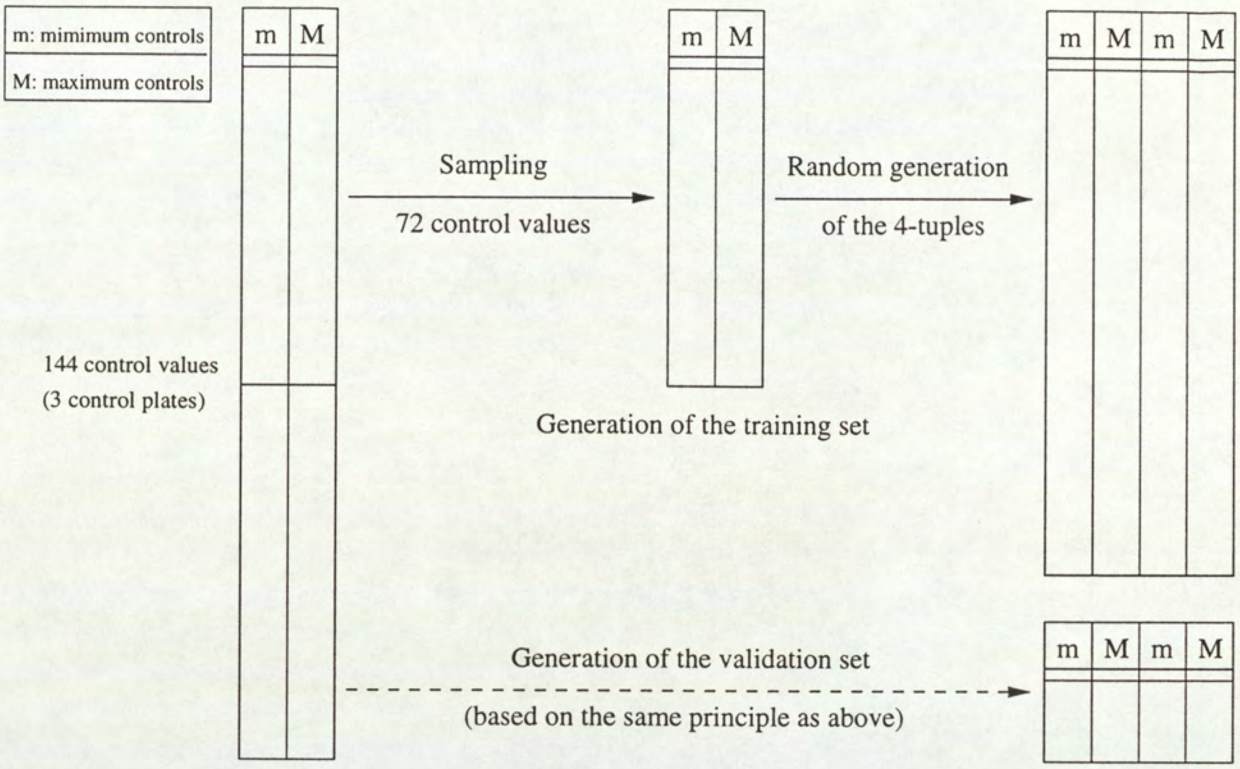


Figure 2.1: Generation of the training and validation set.

The procedure is then repeated several times. More complex cross-validations are mentioned in (Bishop, 1995) but give similar results, referring to the work that was carried out last year.

Note: a 4-tuple is a 4-dimension vector. It features 4 columns of control values (2 columns of minima, 2 of maxima) that are used to train and validate the model, with regards to the control wells (D_1, D_2, D_3, D_4) on a normal plate, whose validity is tested.

The selection criterion

To choose which one is the *best* model, we first have to define a selection criterion. Thus, it is usually the best fit with regard to the error function that is kept. Indeed, the performance of the different models is compared by evaluating the error (2.5) on the validation set $\mathcal{V}$. The model that has the smallest error with respect to the validation set $\mathcal{V}$ is then considered as the best model. This approach is known as the *hold out method*.

Note: (Bishop, 1995) emphasises that sometimes, another set of data is chosen as a test set to confirm the performance of the selected model, so as to avoid over-fitting the data of the validation set.

Last year's approach also considered a second means of selecting a model: the model is chosen with regards to the number of novel points detected in the validation set. The smaller this number, the better the model. In practice, these two strategies gave similar results on the tested screens, hence our decision to keep the first criterion for our procedure.

2.3.2 The novelty threshold

The novelty threshold is chosen to be the minimum value of the density function or *likelihood*[7] of the validation set. This means in practice that all the controls (on a normal plate) that have a smaller probability than the smallest probability of the control plates values are considered as abnormal.

Another possibility to define this threshold, rather than using the validation set, is to subsample from the Mixture model density function. The likelihood for this sample is then computed and a threshold is set to a certain percentile, which corresponds in fact to the definition of a new novelty threshold. This definition takes advantage of the probabilistic definition of the data provided by the density function.

2.4 Reliability of the novelty detection algorithm

Last year's results showed variations after several runs of the main program: the number of plates flagged as abnormal according to the chosen criterion (the novelty threshold is defined as the minimum value of the density function of the validation set) did not have a constant value, this phenomenon is investigated by focusing on the random seed factor.

The first part of this section explains what the random seed factor is, where it appears and to what extent it influences the results. It then presents how the study led us to consider three different aspects of the problem : the influence of the seed in the generation of the training set versus the seed factor in the *Expectation-Maximisation algorithm (*EM), the influence of the size of the training set, and finally an investigation to see whether plates had the same ranking according to their likelihood, despite the sources of variation.

2.4.1 The random seed factor

What is a *seed factor* ?

Last year's study is based on Gaussian mixture models to infer the density of the HTS control values. The training procedure was described, focusing especially on how the training set is generated as a random selection of the control values featured by the added control plates.

[7] the value $p(x)$ taken by the density function p at a point x is generally called the 'likelihood' of this point

What is called *random seed factor* is the starting point of the number generator that specifies a random sequence. This means that for a given seed factor, the randomly generated sequence of numbers will always be the same. So all the procedures using a random sequence depend on this seed factor. In particular, both the generation of the training set and the initialisation for the EM algorithm feature a random process.

The influence of the seed factor on the novelty detection

Last year's study focused on the control values to determine the number of plates declared 'novel'. The tests carried out mainly concerned Screen 2 : the results are summarised in Figure 2.2. The same screen has been used this year to compare the results by using five different random seeds (see Table 2.1).

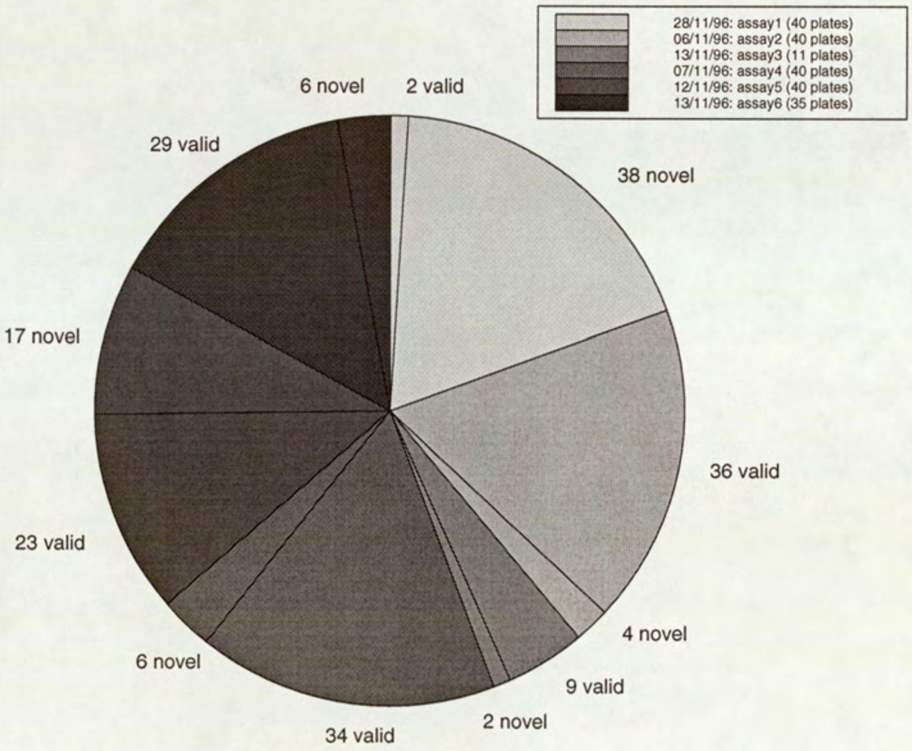


Figure 2.2: *Totals & NSBs*

Last year's results : Proportion of rejected plates per day or assay (screen 2).

Even if the number of rejected plates is more or less the same as far as the average is compared to last year's, one can notice that for a given screen and several runs of the program, results show big variations: the number of rejected plates for 13/11/96 for instance ranges from 3 to 13 out of 35 that is a difference of 10 (30%). This phenomenon is due to the seed factor. However, since a random process appears both in the generation of the training set

Date	Number of plates	Number of rejected plates	Average	Proportion
28/11/96	40	38-38-38-38-38	38	95%
06/11/96	40	3-2-8-6-3	4	10%
13/11/96	11	0-2-3-5-0	2	18%
07/11/96	40	4-3-9-8-3	5	13%
12/11/96	40	19-13-20-21-16	18	45%
13/11/96	35	11-3-13-12-7	9	26%

Table 2.1: *Totals & NSBs*

Proportion of rejected plates per day (assay) for five different seeds (screen 2)

and the EM algorithm, it is important to understand which one has the greatest effect.

2.4.2 Seed factor in the training set *Versus* EM seed

The first aspect considered in this investigation deals with the effects of a fixed random seed in the generation of the training set versus a fixed random seed in the EM algorithm.
A first problem arose that had to deal with the zeros in the minimum control values in the control plates. Since plotting the distribution of the minimum values hardly allows us to determine whether the zeros should be considered as outliers or not (as shown in Figure 2.3), the experiments are carried out both with the zeros and without them.

The starting point of our study concerns the influence of the random seed factor in the generation of the test set. First the novelty detection procedure is computed with a fixed seed in the initial conditions of the EM algorithm, the training set being each time generated with different seeds. A second experiment then consists of fixing a seed for the generation of the training set and initialising the EM algorithm with different seeds for each run of the novelty detection.
All the tests carried out on the different screens lead to the same results : when the seed varies in the generation of the training set, the number of plates declared 'novel' for a given screen show big fluctuations. On the contrary, when it is fixed, the number of abnormal plates only takes two different values (close to each other) as shown in Figure 2.4.

The tests lead to the same conclusion as far as both the training and validation error per point, used as indicators to see the performance of the mixture model, are considered (as Figure 2.5 proves it).

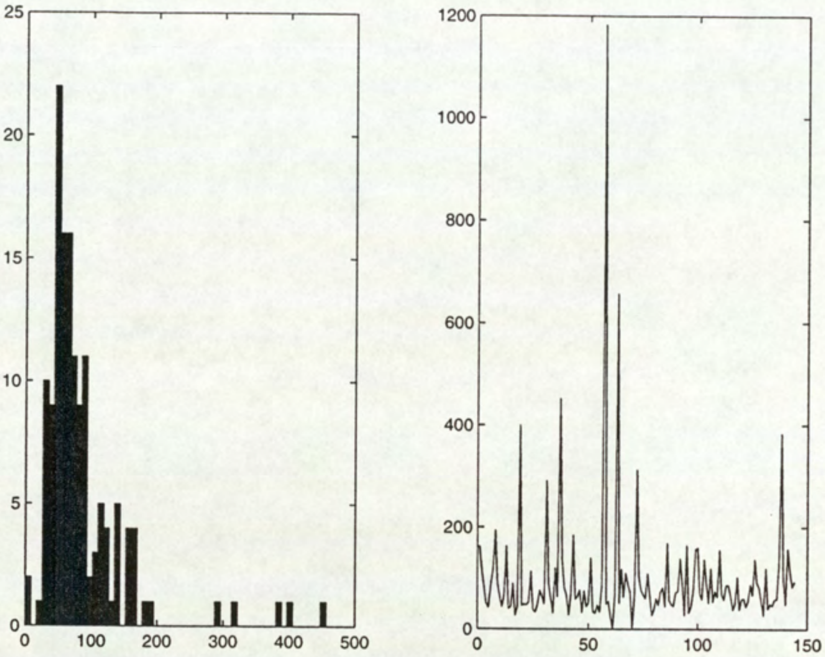
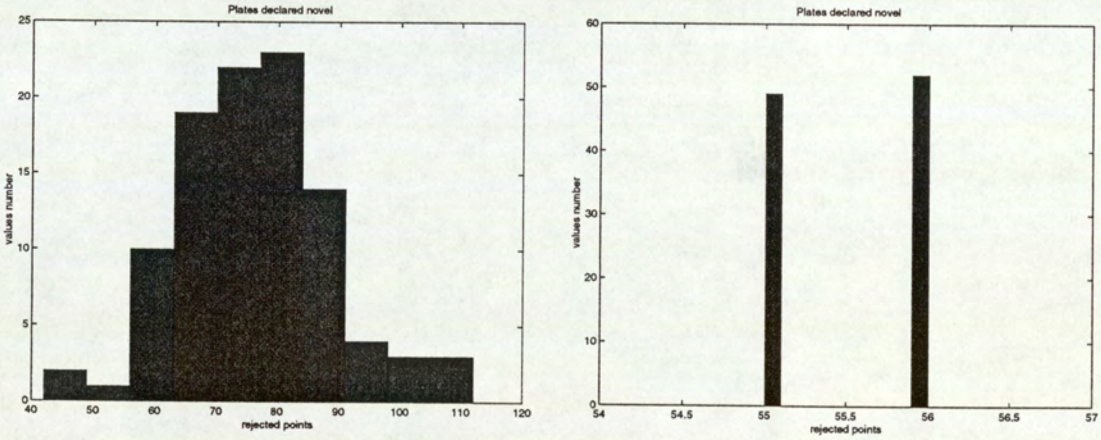


Figure 2.3: Distribution of the minimum control values

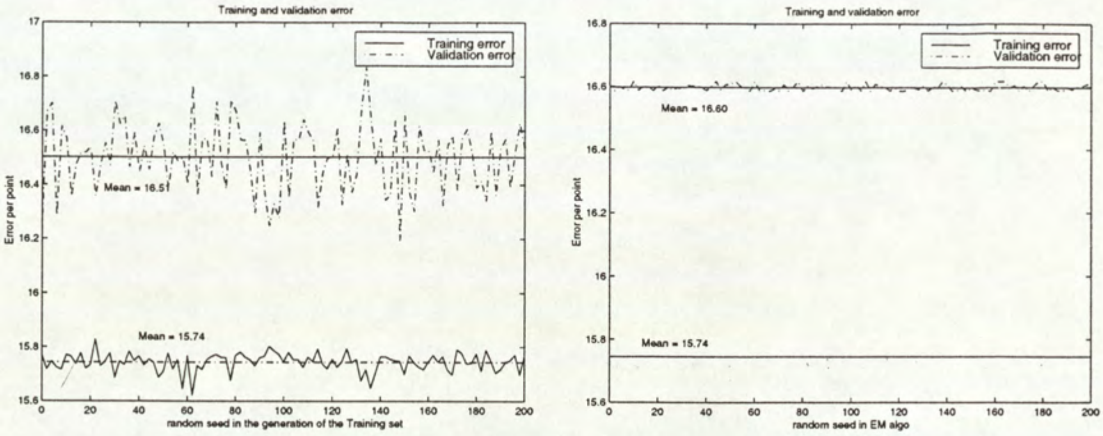


(a) Fixed EM seed

(b) Fixed seed in the Training set

Figure 2.4: Effects of the seed factor on the number of novel plates for screen 2

Thus the results obtained for the first part of this investigation clearly prove that the random seed factor has an influence on the number of plates declared novel that should not be neglected, when it varies in the generation of the training set, whereas it does not influence anything when intervening in the initialisations of the centres in the EM algorithm. Such variations are obviously undesirable as the results should not depend on the chosen seed value. Further investigation therefore appears to be necessary.



(a) Fixed EM seed

(b) Fixed seed in the Training set

Figure 2.5: Effects of the seed factor the training and validation error per point for screen 2.

2.4.3 Influence of the size of the training set

The variation in the results can be due to a training set that was not big enough. We therefore investigate this issue. Indeed it is expected that the number of rejected plates will be less affected by the random seed factor, or at least not in such a large way, if the mixture model is trained with a larger data set, bearing in mind that the training procedure should not be too computationally expensive.

As seen before, it is the random generation of the training set that influences the results. Therefore, to concentrate only on its effects, the random seed factor is fixed in the EM algorithm. We then study both the number of plates declared 'novel' and the negative log-likelihood (validation error per point), for varying random seed factors in the generation of the training set, as a function of the size of the training set. The results expected are smaller and smaller error-bars together with a decreasing curve (while the size of the training set is increasing), which would clearly prove that the influence of the random seed in the training set can be neglected for a training set that is large enough.

The first step of this investigation considers a reasonable size for the training set: the mixture model is trained with data sets whose size ranges from 144 to 1000 as shown in the top row of Figure 2.6. Since no general trend arises from these curves, the investigation is carried out on larger training sets (see the bottom row of Figure 2.6). The results do not show any improvement. Even if the use of a larger training set would have led to better results (as far as the validation error per point is considered), it would have been rather unrealistic to keep on training the mixture model with such large sets of data since the learning procedure took

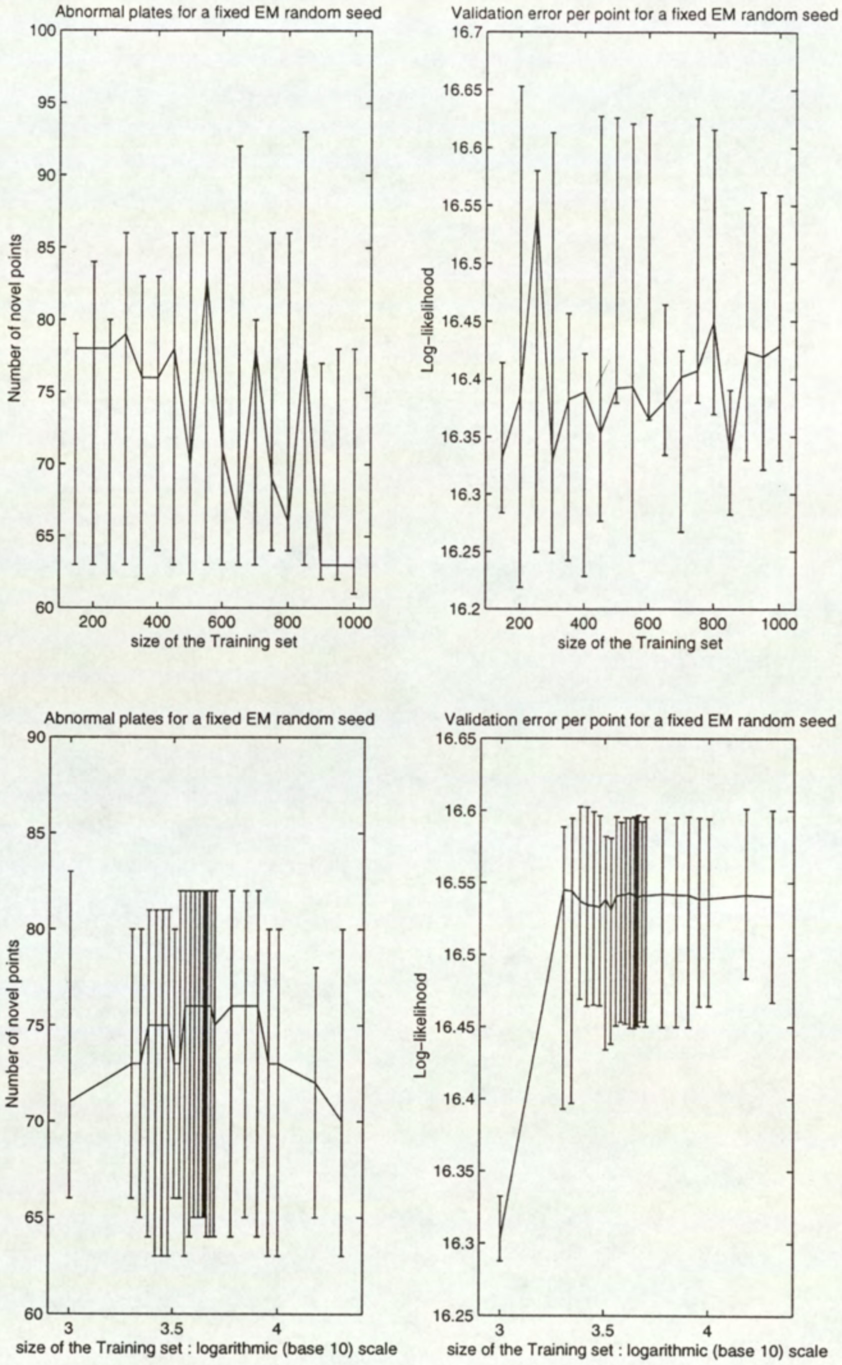


Figure 2.6: **Size of the Training set versus novel points and log-likelihood :** for each size of the training set, both the number of abnormal plates and the validation error are computed 60 times. The solid line joins the median of the values which range over the error bars.

no less than several hours.

The results obtained are far from our expectation in the sense that there is no ideal size for the training set that improves the results, since the curves do not follow any general trend that is easy to interpret. Therefore, a size is chosen so as to avoid a time consuming algorithm, as time is a crucial issue in term of quality control in the HTS process.

2.4.4 The same ranking according to the likelihood ?

It is now clear that the number of plates flagged as 'novel' depends on the random seed factor in the generation of the training set, whatever its size. We investigate the novelty order of the standard plates according to their likelihood, for different random seeds in the generation of the training set : this is the measurement used to flag a plate as 'novel' since the novelty threshold has been defined as the minimum value of the validation set density function. Experiments are carried out on each screen, for different random seeds while generating the training set.

Random seed	With the zeros in the distribution of the minima	Without the zeros
20	5	4
100	3	2
350	2	5
500	2	5

Table 2.2: Screen1, Plate 74 : Ranking for four different random seed

The results are rather satisfying, since the ranking of the plates according to their likelihood do not show huge differences for various random seeds (as shown in Table 2.2). But also, plates known to be invalid have a high novelty value *i.e.* a low likelihood in all the tested cases. Indeed, let's consider the example of plate **74 in screen 1**. It contains a zero as a maximum control value (maximum activity) whereas the corresponding minimum activity shows a positive value. This plate should be highly considered as highly abnormal, since one of its minimum values is higher than the corresponding maximum value. For each of four different random seeds, plate 74 has a ranking lower than 5 (out of 115) (see Table 2.2), which clearly shows a really low likelihood and hence a high probability of being abnormal.

2.4.5 Conclusions

While investigating some sources of variation in the novelty detection algorithm, we made clear that the random seed factor, that is the starting point of the number generator that specifies a random sequence, in the generation of the training set is the cause for these unexpected variations.

Moreover, since our expectations that a bigger training set would undermine the variations were not fulfilled, the size for the training set was chosen so as to avoid a time consuming training procedure for the Gaussian mixture model.

Finally, we came to the conclusion that, although the random seed factor in the generation of the training set influences the number of plates declared 'novel' with the chosen criterion, this phenomenon is not worrying since the ranking of the plates according to their likelihood does not show significant differences (in particular, plates that are most likely to be flagged as abnormal have the same low likelihood whatever the random seed factor and hence a high probability of being considered as novel).

Chapter 3

Inter- and intra-plate variation

This chapter presents the second step of the quality control procedure. Whilst the novelty detection (introduced in the previous chapter) only relies on control values to assess the quality of the data, this second stage of the procedure focuses on normal wells.

The aim here is to detect any general effect that could affect a plate. Not only does this concern some potential problems occurring during the physical preparation of the experiment, such as handling mistakes where, for instance, an inaccurate volume was dispensed or a contamination between plates or wells, due to a failure in the washing step occurred, but also some potential problems in robots: it can be a blocked tip affecting all the wells in a given row or column on a plate (depending on the kind of devices used), or an electronic mistake in the measurement devices. More generally, this step aims to determine any potential errors inducing wrong values in the data.

It first consists of a plate to plate comparison where a general statistical method such as the Kolmogorov-Smirnov test is used to detect any general difference, together with some other methods dealing with the presence of extreme values (or outliers). This procedure aims to spot some suspicious plates by comparing the distribution of values of the plates and pointing out the ones that significantly differ from the other ones. The quality control procedure is then completed by an intra-plate variation to detect any differences within a plate. This part of the procedure is more specially designed to detect blocked jets by investigating differences in spread and location between the data coming from different rows, within the same plate.

3.1 A plate to plate comparison

Considering the first set of valid plates (determined as a result to the novelty criterion chosen in the first step of the quality control method and based on control wells), we wish to detect invalid plates where the control values are unaffected. Because we do not know what any single value on a plate should be, we are forced to compare the distribution of values on the whole plates to detect inter-plate variations. The idea is to find some general procedure that is able to deal with any type of particularity that could possibly distinguish one plate from another. To do this, we are using a well known statistical procedure: the Kolmogorov-Smirnov test.

3.1.1 The Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test used as a general procedure to investigate differences between two population distributions is even considered by (Neave and Worthington, 1988) as the best known among several distribution-free procedure to test differences *of any kind*, in location, or spread, or more general differences in shape, between two sample populations.

Presentation

Several computations of the Kolmogorov-Smirnov test exist, based on different definitions of the cumulative distribution function. The one computed here is given by

$$F_n(x) = \frac{\displaystyle\sum_{i,\, x_i \leq x} x_i}{\displaystyle\sum_{i} x_i} \quad .$$

The procedure consists of testing two hypotheses:

H_0: the two samples come from the same distribution

H_a: the two samples come from different distributions.

Given samples of size n_1 and n_2, the Kolmogorov-Smirnov statistic then consists of comparing the maximum value of the absolute differences between the two cumulative distribution functions of the two populations:

$$D = \max_{-\infty < x < \infty} |F_{n_1}(x) - F_{n_2}(x)| \quad .$$

with some critical values D_α (where α is the significance level of the test) given by tables[1]. The critical region is of the form $D \geq D_\alpha$, which means that in such a region the *Null* hypothesis H_0 that the 2 samples come from the same distribution is rejected in favour of the *alternative* hypothesis H_a.

Note: as mentioned in (Neave and Worthington, 1988), for large sample sizes (*e.g.* more or equal than 35 as it is the case in the context of HTS), critical values are only the result of approximations.

Method

The idea behind such a procedure is based on the simple fact that an invalid plate shows significant differences when compared to a normal plate. Thus, we expect that an invalid plate should differ from quite a large number of plates, which would clearly prove that the testing procedure is actually relevant and enables us to spot suspicious plates *i.e.* plates whose distribution of values differs in a significant way from the other plates. Hence this could be a good means of detecting general effects such as handling mistakes or mechanical problems such as blocked jets, assuming that they affect the distribution of values significantly enough. Practically, the implemented procedure randomly chooses a certain number of reference plates (among the plates considered as valid by the novelty algorithm) whose distribution of values are compared to all the other plates of the subset of 'valid' plates, using the Kolmogorov-Smirnov test. Referring to the statement mentioned below, a suspicious plate then is a plate that differs from an unusual high number of other plates.

Note: the advantages and limitations of this method will be discussed in Chapter 5. Some remarks will especially be made concerning the choice of the number of reference plates.

3.1.2 Outlier detection

A second method is investigated whose aim is to spot some suspicious plates. The point is not here to focus on the distribution of values any more but to make a study of extreme values on each 'valid'[2] plate of a screen. To achieve that, rather than making use of the traditional approach to deal with outliers (presented in Section 2.1.2), which is far too inaccurate and time-consuming to satisfy our needs, we choose an alternative and simple method.

[1] Details of these critical values can be found in (Neave and Worthington, 1988).

[2] according to the novelty algorithm

Procedure applied

The aim of this procedure is simply to detect suspicious plates by focusing on the number of extreme values. Indeed, relying on the simple idea that a suspicious plate is a plate showing a significantly high number of extreme values, what is expected here is to find a small number of plates that would show an unusual high number of outliers. Figure 3.1 gives a visual representation of our expectations.

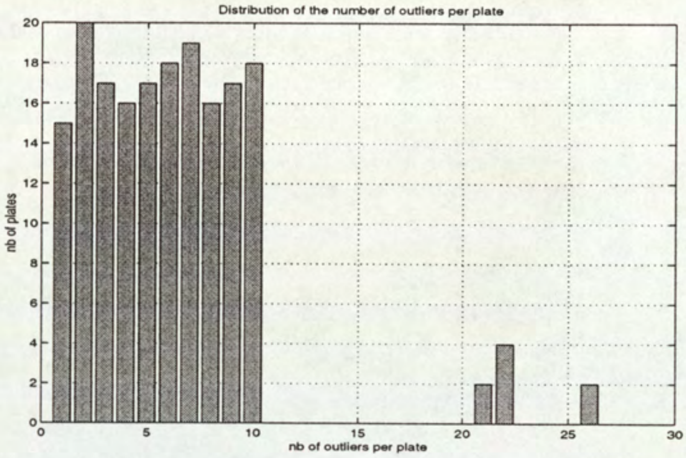


Figure 3.1: Distribution of the number of outliers per plate in an 'ideal' case

To do that a rather experimental but well-accepted and intuitive process of detecting outliers is computed, and this is the method used in the remainder of the thesis to determine whether an extreme value can be considered as an outlier:

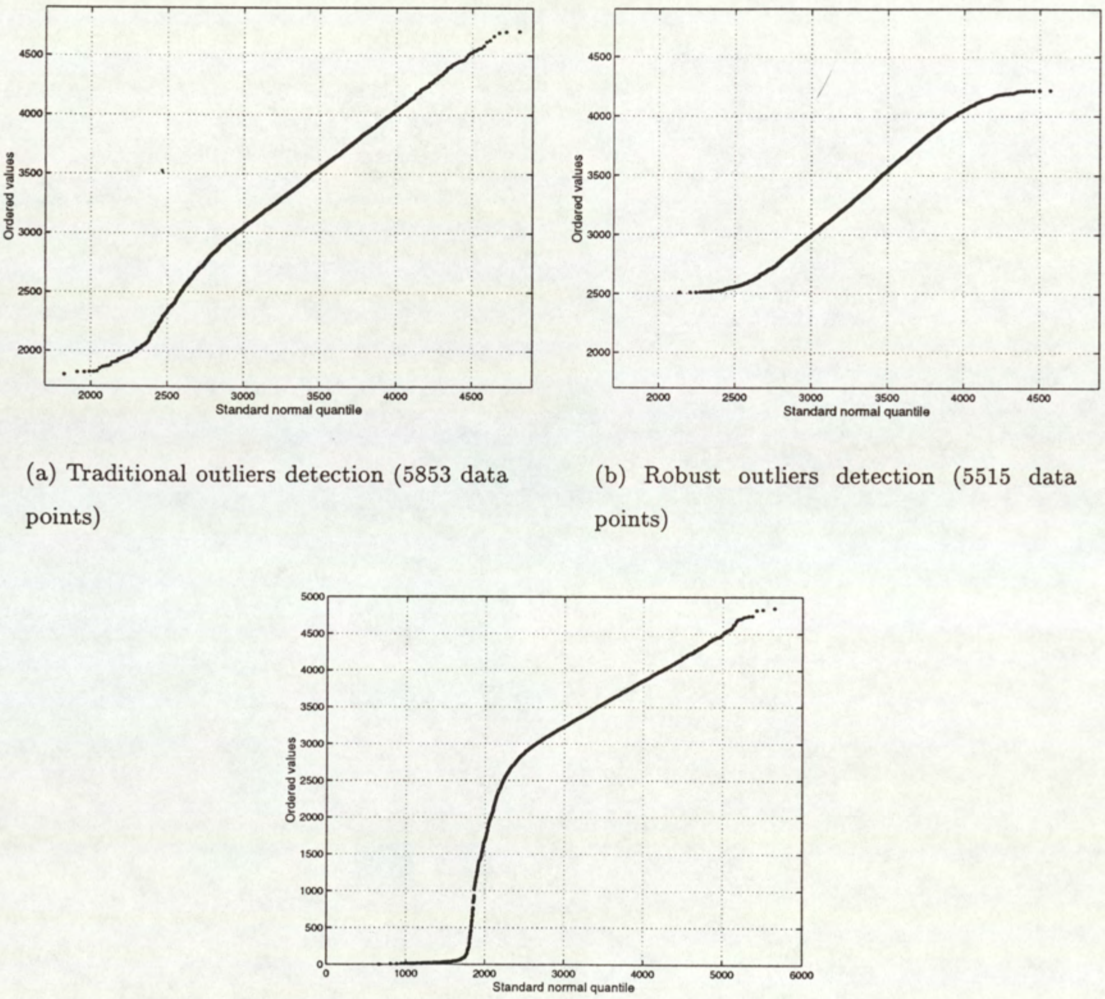
> *An outlier is a value that differs by more than* 2 standard deviations *from* the mean *of the data.*

It is however to be noted that this definition can be slightly modified : the standard deviation and the mean respectively can be replaced by any other measure of spread such as the *Median Absolute Deviation* that will be introduced in the following section (respectively any other measure of the mean).

The built-up procedure exploits the two possible means of detection:

1. A traditional detection using the mean and the standard deviation of the data
2. A robust detection using a robust measure of the mean and a robust measure of the spread

Figure 3.2 gives an indication of the differences between the two methods used to detect outliers (the graphs are presented here using the *quantile-quantile* technique which is described



(a) Traditional outliers detection (5853 data points)

(b) Robust outliers detection (5515 data points)

(c) Whole population (6144 data points)

Figure 3.2: **1st set of valid plates from Screen** 1*b*: the three graphs respectively represent the data without the outliers and the whole data population

in Section 4.1.1). An examination of these graphs shows that, though similar, the two plots representing the two methods of detecting outliers give slightly different results. The reason for this is that, when applied, the robust detection is less affected by outliers than the traditional detection, as far as the computation of spread is concerned. Therefore the former method detects a larger number of outliers as indicated on the previous figures.

- The traditional detection

The first method computed removes the extreme values using the traditional outliers detection. The definition presented before is applied on the tested screens with the sample mean defined by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \, ,$$

where $x_1, \ldots, x_n$ denotes n data values, and the standard deviation given by

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \; .$$

- A robust detection

A second possibility to deal with outliers is to make use of the tools developed by robust detection methods. As emphasised in (Mason et al., 1989), unlike the traditional summary statistics measuring the centre (the sample mean) or the spread (the standard deviation) of a set of data, robust statistics do not suffer that much from extreme values. That is why they are often preferable to the traditional statistics when analysing real world data.

The *m-estimator* is the robust statistics used as an alternative to the sample mean.
M-estimators are nothing but weighted averages of data values. They are defined by:

$$m = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

where

$$w_i = \begin{cases} \frac{-tv}{x_i - m} & \text{if } x_i < m - tv \ , \\ 1 & \text{if } m - tv \leq x_i \leq m + tv \ , \\ \frac{tv}{x_i - m} & \text{if } m + tv < x_i \ . \end{cases}$$

t being the *tuning constant* (usually chosen to be 1.345 or 1.5 depending on how severely one wishes to limit the influence of extreme values) and v a robust measure of spread, usually the *Median Absolute Deviation* presented in the following.

M-estimators serve two crucial purposes. First, as said before, they are a robust alternative to averages whose greatest merit is to use, unlike the *Sample Median*, all the data values. Second, the weights w_i help to identify when the traditional summary statistics may be influenced by outliers. Indeed, if all the observations in a data set are sufficiently well-behaved, the weights equal 1 and the m-estimator is equal to the sample mean, whereas extreme data values are given weights less than 1.

In practice, m-estimators are computed iteratively, the *Sample Median* being usually used as an initial estimate of m.

The *Median Absolute Deviation (MAD)* is a robust measure of variations in the data values. It is defined by

$$MAD = \frac{median(|x_i - M|)}{0.6745}$$

where M is the sample median, a number that divides ordered data values into two groups of equal size, and determined as follows:

1. Order the data from the smallest to the largest values

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

2. Determine the median such as

$$M = \begin{cases} x_{(q)} & \text{if } n \text{ is odd, where } q = \frac{(n+1)}{2} \ , \\ \frac{x_{(q)} + x_{(q+1)}}{2} & \text{if } n \text{ is even, where } q = \frac{n}{2} \ . \end{cases}$$

The constant 0.6745, used in the definition of the MAD estimator, enables a good estimate of the population standard deviation with large samples from so-called 'well-behaved' populations *e.g.* populations associated with a Normal probability distribution.

Results and comments

Both traditional and robust methods were applied to the tested screens. The first results obtained with a traditional detection method were however far from what was expected: not only was it impossible to isolate a few plates significantly differing from the other ones because of a larger number of outliers, but it was even hard to find a common pattern or to classify them according to their number of extreme values.

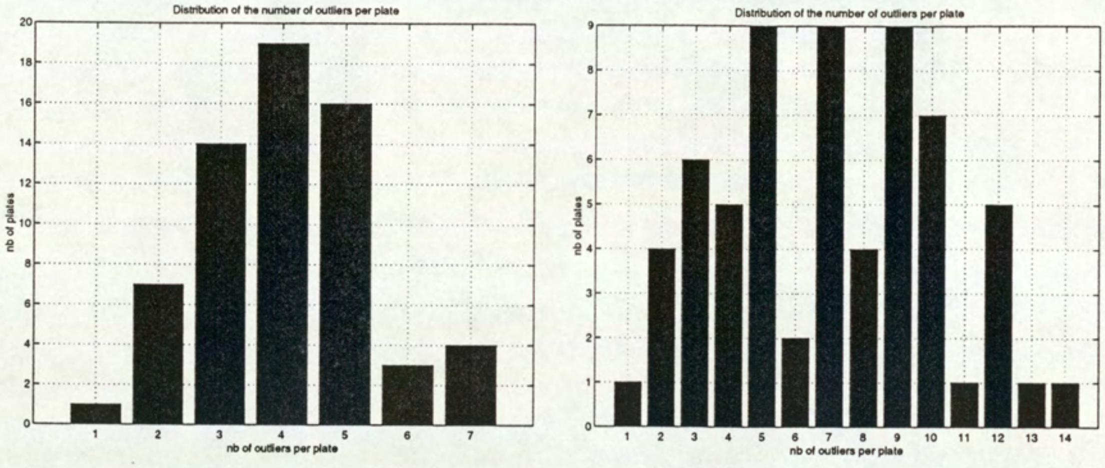
One can then think that these results are simply due to the fact that the mean and the standard deviation are altered themselves by the possible presence of outliers. It is the reason why we then opted for a detection based on robust statistics. It was at least hoped that this robust detection of outliers would enable us to determine some similarities between the plates and spot some plates among the studied set that would show an unusual high numbers of extreme values, with this new robust definition of an outlier.

Unfortunately, this new attempt failed: the results were even worse than the ones obtained with a traditional detection. Not only did the different plates show as many fluctuations as previously, but they also showed a larger range of outliers (see Figure 3.3), which made it even harder to identify some potentially suspicious plates. Table 3.1 presents an analysis of the results for the traditional and the robust detection of outliers : for each screen on which the number of outliers per 'valid' plate was estimated, the table presents the mean and spread of the values obtained (for both methods) together with the minimum and maximum number of outliers observed.

Note: despite the negative results obtained for this attempt, such an investigation is not useless, since Section 4.1.1 computes the same method to dispose of the outliers.

3.2 Intra-plate variation

The quality control procedure presented in the previous section is completed by an intra-plate variation detection. Contrary to the Kolmogorov-Smirnov method, the statistical experiments carried out aims more specifically at determining mechanical problems such as blocked jets. However, it can also be used as a general procedure to detect any differences in location or spread within a plate.



(a) Traditional detection

(b) Robust detection

Figure 3.3: Distribution of the number of outliers per plate for screen 1b

Screens	Analysis of the number of outliers per 'valid' plate							
	Mean		Spread		[min (m), max (M)]			
	Td	Rb	Td	Rb	m_{Td}	M_{Td}	m_{Rb}	M_{Rb}
Screen 1	4.2	4.6	1.1	2.5	2	5	0	8
Screen 1b	4.0	7	1.4	3.2	1	7	1	14
Screen 2	3.9	5.4	1.5	2.9	1	10	0	14
Screen 9	3.1	5.0	1.3	3.0	1	6	0	13
Screen 12[a]	4.7	8.4	1.7	4.4	0	11	0	21

[a] The analysis has been carried out for all the plates. Since no control plates are available for this screen, it is impossible to run the novelty detection and therefore to determine a first set of valid plate.

Table 3.1: Robust (Rb) and traditional (Td) outlier detection for the 'valid' plates of the tested screens

3.2.1 Method

A rather simple method is computed to achieve the intra-plate variation detection. To start with, because most of the 96-well plates are in practice filled column per column with 8 head dispensers, we decided to develop a procedure based on the distribution of values of rows. Indeed, if one of the tip is blocked, since plates are filled column per column, it affects all the wells located on the same row. As a result, the affected row is more likely to show unusual low or high values[3] (at least lower or higher than the values of the other rows on a plate). Thus, the idea is to investigate differences in spread between the distributions of rows within a plate. Figure 3.4 gives an idea of what the distribution of variances per row looks like, for a normal HTS screen.

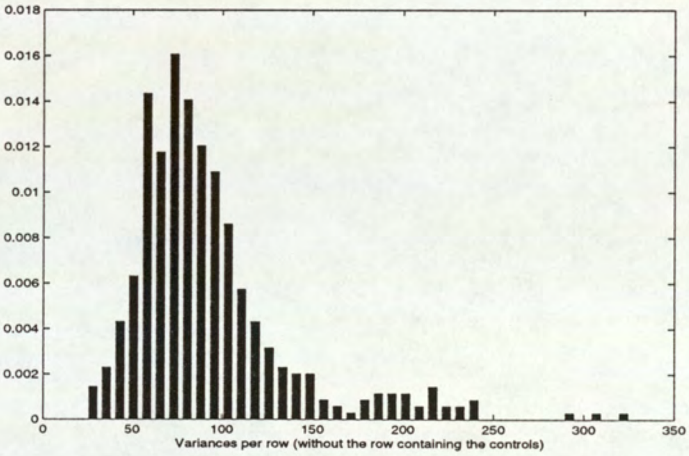


Figure 3.4: Distribution of the variances per row for all the plates of screen 9 (all the rows excepted the row containing the control values)

As expected, the large majority of the rows have a similar variance, if no specific mechanical problems have occurred. Therefore, the idea is that a row altered by a blocked tip should show a smaller variance than the other rows of the plate since the corresponding wells would show either a minimum or a maximum activity.

Besides, to increase the accuracy of the method one might think that it may be worth also investigating a difference in location since a row featuring minimum activity wells would surely show a difference in mean compared to the other rows.

For all these reasons, to complete the inter-plate variation procedure, we propose to detect any intra-plate effects by investigating differences in spread together with differences in location between the distribution of values of the rows within a plate. As suggested in (Neave and

[3] In the case of a blocked jet, no concentration of reagent for instance is added. From a chemical point of view, this corresponds to either a minimum or a maximum activity well.

Worthington, 1988), the Siegel-Tukey and Wilcoxon procedures are used to achieve this. It is to be noted that the procedure described in this section remains quite general and requires some modifications to be adapted to the different situations. Thus the intra-plate variation detection is developed here to be used with plates filled per columns, since such a process is the most widespread. But the procedures could as well be adapted to detect potential errors occurring with devices filling the plates per rows (12 head dispensers). Besides, two different procedures can be designed whether one wishes to investigate more specifically some potential errors due to blocked jets or simply some intra-plate variations that are briefly presented in Section 6.3.2.

3.2.2 The testing procedures

The two procedures we propose to use to investigate differences in location and spread are both distribution-free and therefore no specific assumptions have to be verified before applying these two methods. Besides, both of them are based on the same principle. Finally, (Neave and Worthington, 1988) describes them as "extremely good and widely used".

The Wilcoxon test

The version presented in this section is often referred to as a *rank-sum* test, because the statistic of this two-sample test is computed by adding together certain ranks. But it is in fact better known under the name of *Mann-Whitney* test.

The procedure consists of testing two hypotheses:

> H_0: *on average, the two populations A and B are the same.*
>
> H_a: *on average, the two populations are different.*

Given two samples $\mathcal{T}_A$ and $\mathcal{T}_B$ of size n_A and n_B, the Wilcoxon procedure consists of computing the statistic

$$U = min(U_A, U_B)$$

where

$$U_i = R_i - \frac{1}{2}n_i(n_i+1), \quad i \in \{A,\ B\}\ ,$$
$$R_i = \ sum\ of\ the\ ranks\ of\ the\ A\ (or\ B)\ \ .$$

The ranks are obtained using the following method: the data are first arranged into the ascending orders. It is then easy to write down a list of 'A's and 'B's corresponding to the origins of the numbers in the ordered sequence, thus obtaining a letter sequence. The ranks of the observations in the two samples are finally obtained by simply numbering the letters in the letter sequence, from 1 to N with $N = n_A + n_B$ as follows:

A	A	A	B	A	A	B	...	B
1	2	3	4	5	6	7	...	N

If tied (*i.e.* equal) observations occur between observations from both samples, the letter sequence is no longer uniquely defined and some generalisation of the test statistic needs to be defined. Tied observations are simply given the average rank of the positions they cover in the letter sequence.

Thus let us consider the following example where 3 data points (2 'A's and 1 'B') have the same value (1.40):

actual data values	1.24	1.35	**1.40**	**1.40**	**1.40**	2.00	5.20
letter sequence	A	A	**A**	**B**	**A**	A	B
Position	1	2	**3**	**4**	**5**	6	7

$R_A = 1 + 2 + \{(3+4+5)/3\} + \{(3+4+5)/3\} + 6 = 17.$

The computed statistic U is then compared to some critical value U_{critic} given by tables (see (Neave, 1978)). If $U_{computed} \leq U_{critic}$, the Null hypothesis H_0 that the 2 populations are equal on average is rejected in favour of the alternative hypothesis H_a.

The Siegel Tukey test

The Siegel-Tukey procedure is based on the same principle. It also consists of testing two hypotheses:

H_0: *there is no difference in spread between the two populations A and B.*

H_a: *the two populations show differences in spread.*

The test statistic is the same : $U = min(U_A, U_B)$ (with the same notations as before), but the test is converted to a test for *differences in spread* rather than location by reordering the ranks of the data so as to reflect the above argument. Indeed, the idea is to emphasise both ends of the letter sequence. Thus, the reordering scheme ranks the smallest value as 1, then the two largest as 2 and 3, the next two on the left-hand side as 4 and 5, and so on until the middle of the sequence.

$$\begin{array}{cccccccc} \mathrm{A} & \mathrm{A} & \mathrm{A} & \dots & \mathrm{B} & \mathrm{A} & \mathrm{A} & \mathrm{B} \\ 1 & 4 & 5 & \dots & 7 & 6 & 3 & 2 \\ & \longrightarrow & & & & & \longleftarrow & \end{array}$$

Tied observations are treated in a similar manner as before.

Even if The Siegel-Tukey procedure has some weaknesses compared to other tests for differences in spread such as Mood's or Ansari-Bradley's, for instance it is not exactly symmetric with respect to the letter sequence and its power is a little less than Mood's test, this is outweighed by the convenience of being able to use in practice the Mann-Whitney tables to compute the critical values.

Chapter 4

Edge and Corner effects

The third and final step of the quality control procedure is a kind of mix of the first two procedures described in this thesis. It is similar to the Novelty detection in the sense that it assesses the quality of the experimental conditions by investigating what is happening on the edges of the plates that constitutes the raw material of the HTS process. On the other hand, like the quality control step that investigates inter- and intra-plate variations, it exclusively concentrates on normal wells, without taking into account the controls.

The procedure focuses on the normal wells located at the edges of the plates to detect potential variations with comparison to the inner wells. Since each well features biological mixtures, the point is to investigate whether the way the experiments are designed has an influence or not on the data values. For instance, it is interesting to investigate if the fact that some wells are directly in contact with the air (the ones located on the edges) present different patterns from the ones in the middle of a plate or if the number of neighbours induces some specific features. Therefore, we consider three different populations:

1. A population of *corner values*, the most exposed to the air, and with only 3 neighbours.
2. A population of *edge values* (slightly less exposed to the air than the corner wells) but with 5 neighbours.
3. A population of *middle values* featuring the inner wells that have 8 neighbours.

Figure 4.1 gives a representation of the different populations considered and their location on the 96-well plate.

To tackle the problem of detecting some significant edge and corner effects, we propose to use the *Analysis of variance (ANOVA)* method. The point is in fact to detect any *systematic* variation in looking at the population means. We therefore eliminate the extreme values due

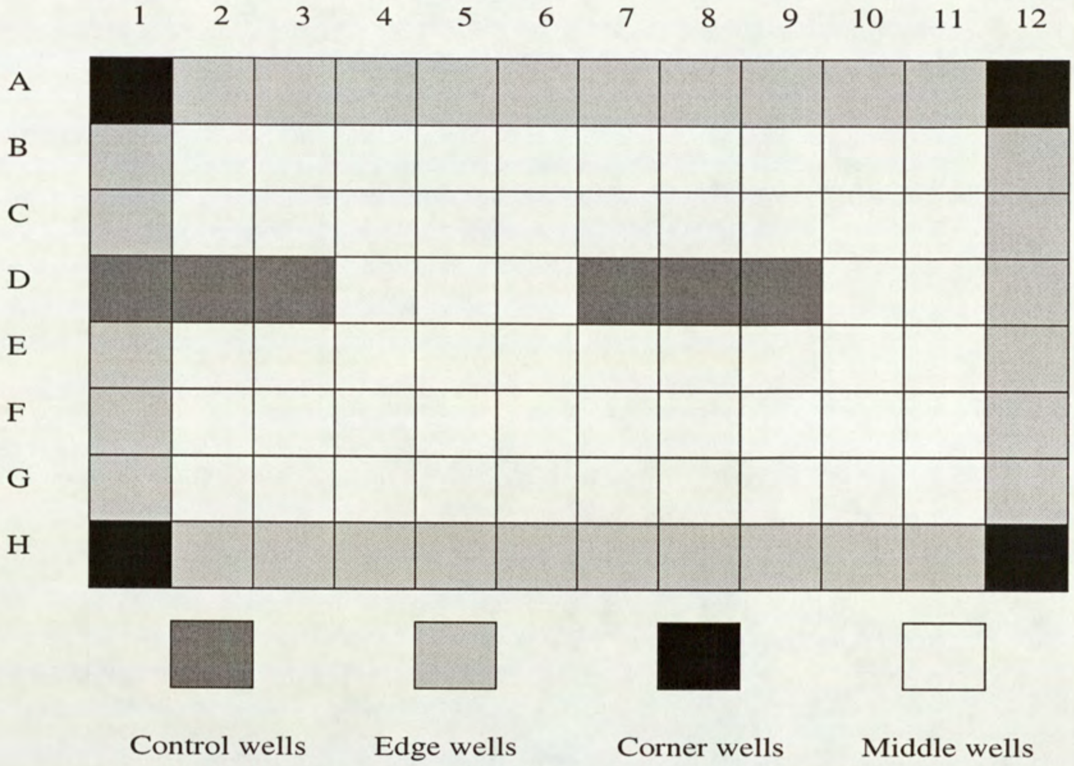


Figure 4.1: A 96-well plate featuring the different populations considered to detect edge and corner effects

to mistakes or hits and apply the ANOVA procedures from standard statistics on the three different populations mentioned above.

This chapter is divided into two parts. We first test the assumptions behind the standard method of analysis on the studied data. Indeed all the populations the ANOVA procedure is applied to have to be normal with equal variances. The second part describes the procedure itself: single factor ANOVA first to determine if the tested populations are different, and Tukey's multiple comparison to investigate which populations significantly differ from the other ones (if the first step showed any variation between the populations.)

4.1 Hidden assumptions

As mentioned previously, before computing the Analysis of Variance procedure, some strong assumptions have to be verified : the I populations or treatment distributions (here $I = 3$: corners, edges and middles) must all be normal with the same variance σ^2. Let X_{ij} be the random variable that denotes the jth measurement taken from the ith population. It means that X_{ij} has to be normally distributed with

$$\begin{aligned}\mathcal{E}(X_{ij}) &= \mu_i \ , \\ Var(X_{ij}) &= \sigma^2 \quad .\end{aligned}$$

Hence, as suggested in (Devore, 1991a), it is highly recommended to first test the normality of the data, and then the equality of the variances, to be sure that the pre-requisite conditions for ANOVA hold.

4.1.1 Normality of the data

Skewness and Kurtosis

As a preliminary step, before computing any sophisticated procedure, a basic idea to test whether a sample comes from a normal distribution or not is to look at the data, especially to check the symmetry of the data set, since it is a strong property of the normal distribution. To characterise this, rather than simply estimating the percentage of data points higher and lower than the mean, as suggested by (Miller and Ruppert, 1986), a more useful measure is given by the Skewness defined by :

$$\gamma_1(\mathbf{x}) = \frac{\mathcal{E}[(x-\mu)^3]}{\sigma^3} \ ,$$

and the Kurtosis

$$\gamma_2(\mathbf{x}) = \frac{\mathcal{E}[(x-\mu)^4]}{\sigma^4} - 3 \ ,$$

for zero mean data ($\mathcal{E}[\mathrm{x}]$ is the expectation of $\mathbf{x}$), that give a better idea of what the data look like. Indeed, a distribution with a right tail heavier than the left one has a Skewness γ_1 positive. Similarly, when the tails of the distribution contain more mass than the Normal distribution, the Kurtosis γ_2 is positive, whereas $\gamma_1 = \gamma_2 = 0$ for the Normal distribution. Table 4.1 presents the values obtained on one of the tested screens (screen 1b, see Appendix A).

The values of the Skewness and Kurtosis for all the populations give quite poor results, that seem to be far from the ideal zero expected in the case of Normal distributions. However, since the point of this investigation is to detect any *systematic* edge or corner effects, extreme values (due to mistakes or hits) that are more likely to alter the data are not to be taken into account. Therefore, the same measures are applied to the populations without outliers. The method used to dispose of the outliers is the one described in Section 3.1.2. This seems to work quite well and the improvements are remarkable: the symmetry of all the populations is almost perfect. It is confirmed by the following graphical visualisations and goodness-of-fit tests.

Populations	Number of data points		Skewness		Kurtosis	
	with outliers	without[a]	with	without	with	without
Corners	256	248	-1.68	-0.15	8.78	-0.01
Edges	1984	1912	-17.29	-1.23	7.78	-0.25
Middles	3520	3400	-0.48	-0.04	7.85	-0.16

[a] The outliers represent here all the data that differ by more than 2 standard deviation from the mean.

Table 4.1: Measure of Skewness and Kurtosis for the different populations involved in the ANOVA procedure for **Screen 1b**. The 2 measures have been applied on the whole populations and the populations purged of outliers.

Graphical visualisation

To illustrate the results of the investigation carried out previously, some graphical representations are used. For obvious reasons, we can not reproduce here all the results obtained for the different tested screens. That is why we choose to detail the whole testing procedure on one particular screen (screen 1*b*) for which the hypotheses for ANOVA (Normality and equality of variances) seem to hold.

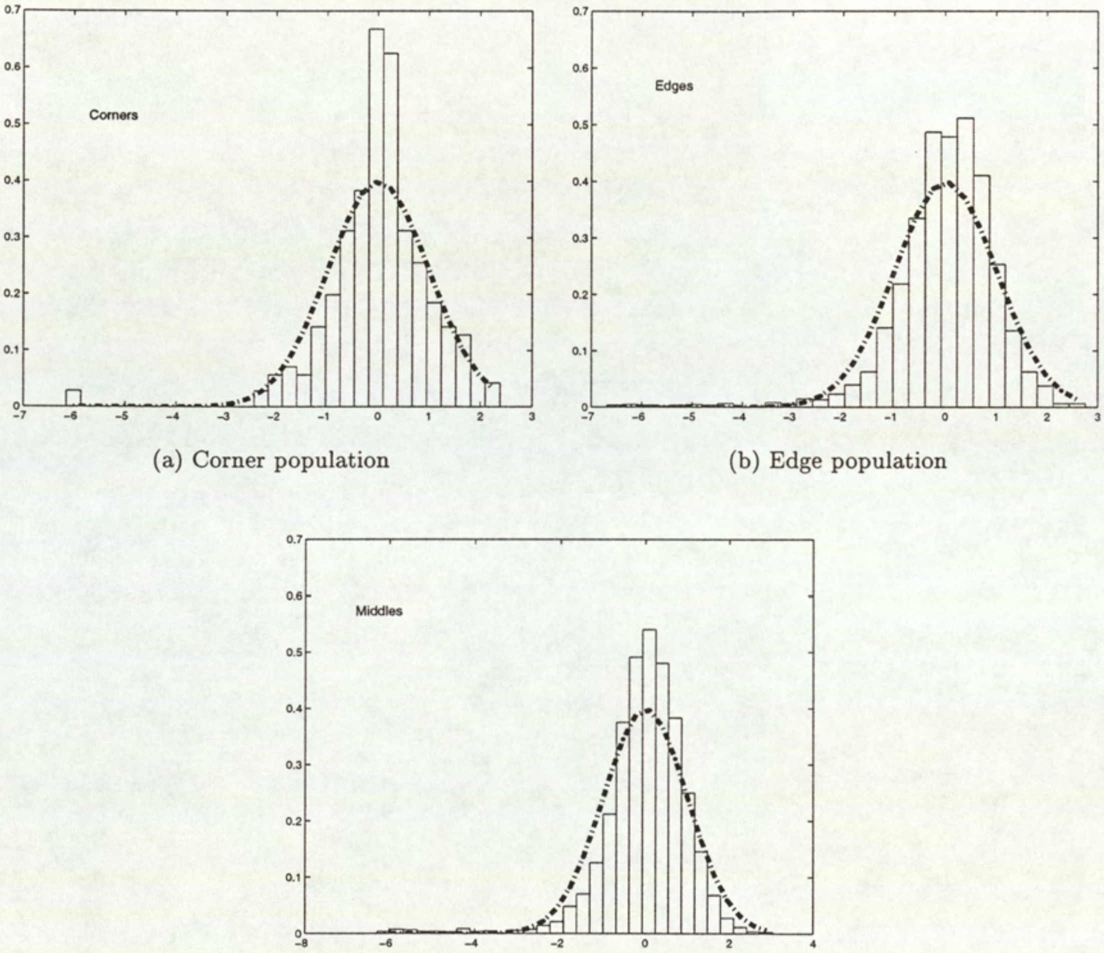
First, Figure 4.2 presents the distribution of values of all three populations (a population of corner, edge, and middle values). The graphs represent the normalised data, *i.e.* the data are set to zero mean and unit variance, together with a Gaussian distribution. As mentioned before and suggested by the results obtained with the Skewness and Kurtosis measurements, the plots show left skewed tails, which corresponds to a negative Skewness, and also heavy tails (as stated by the positive values of the Kurtosis).

Normal quantile-quantile plots are then used to show how close to a Normal distribution the populations without outliers are, justifying at the same time why we disposed of the extreme values. First, we present the philosophy of quantile plots and define a specific category: the normal quantile-quantile plots. Details can be found in (Mason et al., 1989).

- Presentation of quantile-quantile plots

Quantile plots display many distributional features of a set of data. They can in particular be used to assess the fidelity of some data to a hypothesised generating probability distribution.

First, a quantile (denoted $Q\{f\}$) is a number that divides a population into two groups. Thus, a specified fraction f of the data are less than or equal to the value of the quantile. To compare two sample distributions using quantile plots, the following procedure can be



(a) Corner population

(b) Edge population

(c) Middle population

Figure 4.2: Distribution of corner, edge and middle populations for screen 1b

followed. Given two ordered samples $\{y_i\}, i = 1, \dots, n$ and $\{x_i\}, i = 1, \dots, m$, $m \geq n$, for each data fraction $f_i = i/n$ in the smaller sample, the aim is to find an *interpolated quantile* $x'(f_i)$ for the larger sample, defined by:

$$x'(f_i) = \begin{cases} x_i & \text{if } n = m \ , \\ (1-g)x_k + g x_{k+1} & \text{if } n < m \ . \end{cases}$$

k being the integer portion of $h = (m+1)f_i$, with $g = h - k$ (if $k \geq m$, $x'(f_i) = x_m$). The quantile-quantile plot techniques consist of plotting the quantile $Q_y\{f_i\} = y_i$ versus $Q_x\{f_i\} = x'(f_i)$, $i = 1, \dots, n$. All the plotted points lie on or near to the same line if the two distributions are identical.

Quantile-quantile plots can also be used to compare a sample distribution with a theoretical reference distribution such as the Normal probability distribution. It therefore consists of plotting $Q_y\{f_i\} = y_i$ versus the standard normal quantile $Q_{SN}\{f_i\}$ whose approximation[1] is given by $Q_{SN}(f) = 4.91(f^{0.14} - (1-f)^{0.14})$, f_i being defined by $f_i = (i - \frac{3}{8})/(n + \frac{1}{4})$.

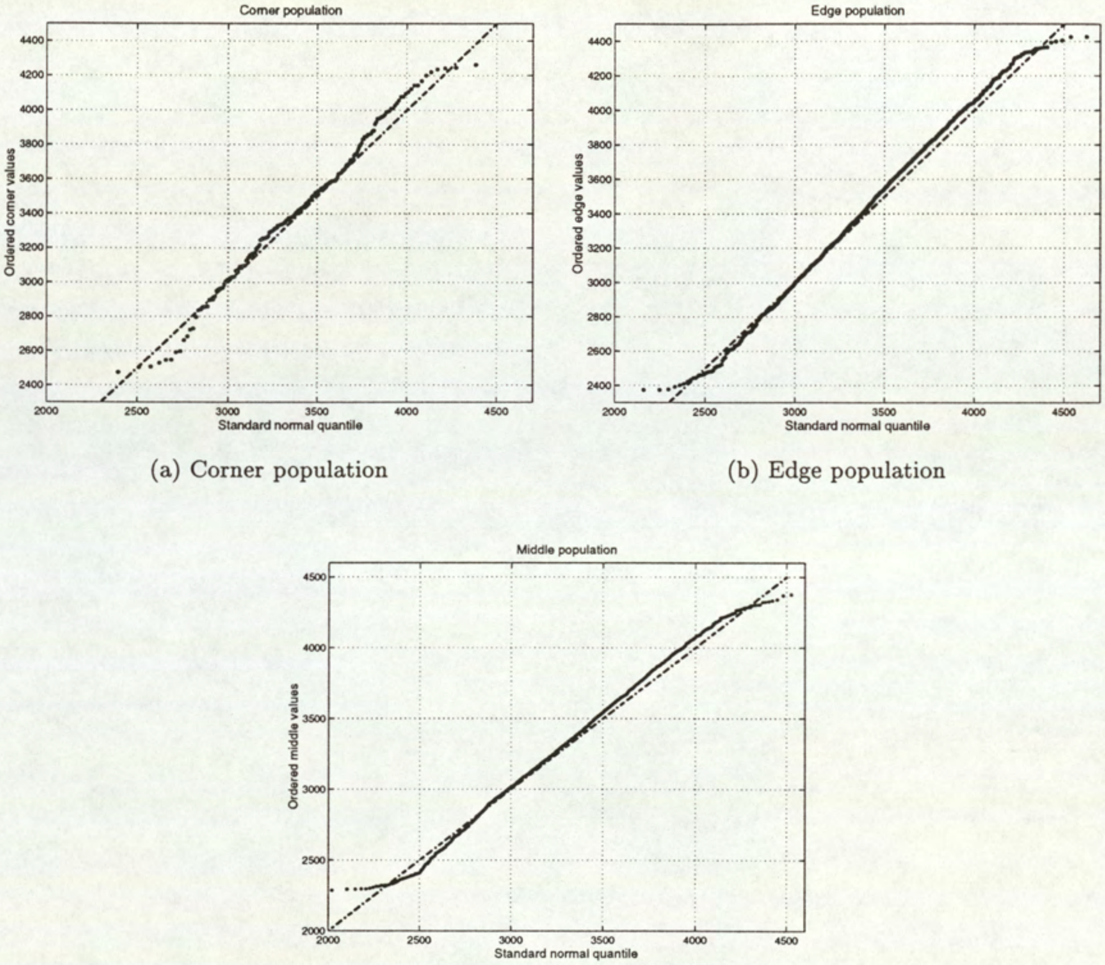
- Quantile-quantile plots applied to screen 1*b*

The aim is to give a graphical representation of what was underlined previously: when we eliminate the extreme values, the distribution of data is close to a Normal distribution. The quantile-quantile plot technique is applied to screen 1*b*. Figure 4.3 represents the corner, edge and middle populations of screen 1*b* without the extreme values. The line $y = x$ is superimposed on the plot because all the plotted points should lie on or near this line if the three populations are identical to the Normal distribution. The results are rather satisfying: most of the data points are close to this line, which tends to prove that the Normality of the three distributions is accepted.

Goodness-of-fit test

In order to give a numerical estimation of how close the different populations are to a Normal distribution, a statistical procedure has been applied: the Lilliefors goodness-of-fit test.

[1] This approximation, often used in statistical software, corresponds to the 'standard' normal distribution ($\mu = 0$, $\sigma = 1$). The quantile $Q_N(f)$ for a Normal distribution with any mean μ and variance σ^2 is derived from $Q_{SN}(f)$ such that $Q_N(f) = \sigma Q_{SN}(f) + \mu$.



(a) Corner population

(b) Edge population

(c) Middle population

Figure 4.3: Distribution of Corner, edge and middle populations without outliers for screen 1b

- Presentation of the Lilliefors test

The Kolmogorov-Smirnov procedure to test whether a sample is drawn from an hypothesised distribution can not be used unless the hypothesised distribution is *completely specified* (*i.e.* the exact value of σ and μ are required in the case of a Gaussian). The method proposed by W. H. Lilliefors to test whether a population has some unspecified Normal distribution is therefore to estimate μ by the sample mean $\bar{X}$ and σ by the classical estimator $\widehat{\sigma}$ defined in Section 3.1.2.

Given a sample of size n, the Lilliefors goodness-of-fit test consists of testing two hypotheses:

H_0: ***the sample is from a Normal distribution***

H_a: ***the sample is not from a Normal distribution.***

The statistic compares the maximum value D of the absolute differences between the cumulative distribution function $F_n(x)$ of the population and the hypothesised cumulative function $F_0(x)$ defined by

$$F_n(x) = \frac{number\ \ of\ \ observations \leq x}{n} \ ,$$
$$F_0(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} \exp\{-\frac{(y-\widehat{\mu})^2}{2\widehat{\sigma}^2}\}dy \ \ .$$

with some critical values D_α. As with the Kolmogorov-Smirnov test (see Section 3.1.1, the critical region is of the form $D \geq D_\alpha$. Details can be found in (Neave and Worthington, 1988).

- Results of the Lilliefors test

Table 4.2 presents the results of the Lilliefors goodness-of-fit test on the screens the ANOVA procedure is applied to. The corner, edge and middle populations mentioned in this table are purged of extreme values, with the method presented in Section 3.1.2.

If for some screens the populations are undoubtedly Normal, for some others, it seems that this hypothesis is not accepted. These results should however be taken with caution. First, since the tested screens feature large data populations (4 corner, 31 edge and 55 middle values per plate), the critical values are only the results of approximations given by (Neave and Worthington, 1988). (Neave and Worthington, 1988) also underlines that Lilliefors' table of critical values was rather inaccurate and therefore suggests the use of other reference

Lilliefors test	Corners		Edges		Middles	
Critical values[a]	5%	1%	5%	1%	5%	1%
	0.0571	0.0667	0.0206	0.0240	0.0154	0.0180
$D_{Screen\ 1b}$	0.0542		0.0213		0.0108	
Conclusion	H_0 *accepted*		*Some evidence*		H_0 *accepted*	
Critical values	5%	1%	5%	1%	5%	1%
	0.1440	0.1681	0.0526	0.0614	0.0392	0.0458
$D_{Screen\ 1}$	0.0747		0.0385		0.0410	
Conclusion	H_0 *accepted*		H_0 *accepted*		*Some evidence*	
Critical values	5%	1%	5%	1%	5%	1%
	0.0379	0.0443	0.0136	0.0158	0.0102	0.0119
$D_{Screen\ 2}$	0.0698		0.0419		0.0545	
Conclusion	H_0 *rejected*		H_0 *rejected*		H_0 *rejected*	
Critical values	5%	1%	5%	1%	5%	1%
	0.0585	0.0683	0.0210	0.0246	0.0158	0.0185
$D_{Screen\ 9}$	0.0645		0.0403		0.0441	
Conclusion	*Some evidence*		H_0 *rejected*		H_0 *rejected*	

[a] For large samples (n>50), the critical values D_α are approximations :
for a significance level α of 5%, $D_\alpha = \frac{0.899}{\sqrt{n}}$,
for $\alpha = 1\%$, $D_\alpha = \frac{1.050}{\sqrt{n}}$.
Details can be found in (Neave and Worthington, 1988), page 103.

Table 4.2: Lilliefors test on corner, edge and middle populations. It is to be noted that the Lilliefors test was applied to the 'valid' set of plates obtained after computation of the Novelty algorithm.

values.

In addition, due to this large number of data points, the possible presence of some outliers (even after we disposed of the extreme values by eleminating every data point that differs by more than 2 standard deviation from the mean), may alter the results.

Finally, despite the fact that some screens did not show the expected results, we however applied the ANOVA procedure to them for various reasons (excepted for screen 12 that features populations showing a too large departure from Normality as we can see in Appendix B). First, graphical representations such as Figure 4.3 gave quite satisfying results and did not show any striking departure from Normality. But above all, ANOVA is a rather popular, efficient and simple method to investigate differences between a given number of populations. With more time, non parametric methods could have been investigated.

4.1.2 Equality of variances

The second hypothesis that has to be tested before computing the ANOVA procedure is the equality of variance. Since the Normality of the data is verified, the procedure chosen to test the equality of variances between the different populations on which the ANOVA method is applied is the well known F test.

Fischer test

Details of this procedure can be found in any statistical book, for example (Devore, 1991a). Let X_1, X_2, …, X_m and Y_1, Y_2, …, Y_n be two independent samples from a Normal distribution with variance σ_X^2, respectively σ_Y^2. S_X^2 and S_Y^2 denote the corresponding sample variances. The method consists of validating one of the hypotheses:

> H_0: *The 2 samples have equal variances ($\sigma_X^2 = \sigma_Y^2$)*
>
> H_a: *The variances of the 2 samples are not equal ($\sigma_X^2 \neq \sigma_Y^2$).*

To do that the F test proposes to compare the statistic value $f = S_X^2/S_Y^2$ to some critical values. The alternative hypothesis H_a that $\sigma_X^2 \neq \sigma_Y^2$ is then accepted in both cases (given the choice of the significance level α):

$$\begin{cases} f \geq F_{\alpha/2,m-1,n-1} \ , \\ f \leq F_{1-\alpha/2,m-1,n-1} \ . \end{cases}$$

It is however not necessary to tabulate both critical values since

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}$$

Results

The results are presented in Table 4.3 for the screens on which the ANOVA procedure is applied. In general they are rather satisfying: for most of the populations, the equality of variances is verified. However, since the F test requires that the populations are normal, any departure from Normality affects the results of this statistical test. This could explain why H_0 is rejected for the edge-middle populations of screen 2 and 9, since the results obtained in the previous section shows that the Normality is far from obvious.

Besides, the same problem as before is encountered: since the populations feature a large number of data values, the critical values are only the results of approximations. As suggested by (Lindley and Scott, 1984), for ν_1 and ν_2 not too high, linear interpolation is accurate enough to determine the critical values (if the latter are not given by the tables), but for large values, harmonic interpolation should be used (it is this method that was used to determine the critical values mentioned in Table 4.3). The procedure is detailed in Appendix C.

4.2 The Analysis of Variance (ANOVA) procedure

The *Analysis of Variance* is the procedure we propose to apply to detect any systematic edge and corner effects and to complete the quality control of HTS. This method was chosen for many reasons. First, the Analysis of variance is a popular method that proved its relevance over the years, and it is probably the reason why every single statistical book mentions it. In addition, the method is consistent and powerful: more than being a common statistical test, it features a twofold analysis, taking into account both mean and variance of the data. Above all, the strength of this method is its simplicity.

As mentioned above, any statistical literature describes this well-known method: (Devore, 1991b) was used as a reference to compute this procedure.

4.2.1 Single-Factor analysis

The simplest problem is referred to *single-factor* or *single-classification*. For this study, it involves the analysis of samples from three populations: corners, edges and middles as explained in the beginning of this chapter.

F test	Corners-Edges		Corners-Middles		Edges-Middles	
Critical values[a]	10%	2%	10%	2%	10%	2%
	1.1289	1.1937	1.1265	1.1819	1.0195	1.0294
	0.9388	0.9096	0.9446	0.9178	0.9848	0.9772
$D_{Screen\ 1b}$	1.0711		1.0943		1.0216	
Conclusion	*H_0 accepted*		*H_0 accepted*		*Some evidence*	
Critical values	10%	2%	10%	2%	10%	2%
	1.5473	1.8754	1.5369	1.8577	1.1253	1.1891
	0.7200	0.6254	0.7347	0.6432	0.9110	0.8709
$D_{Screen\ 1}$	0.4869		0.5589		1.1477	
Conclusion	*H_0 rejected*		*H_0 rejected*		*Some evidence*	
Critical values	10%	2%	10%	2%	10%	2%
	1.0583	1.0876	1.0572	1.0858	1.0088	1.0133
	0.9712	0.9568	0.9741	0.9610	0.9931	0.9895
$D_{Screen\ 2}$	1.0032		1.0302		1.0269	
Conclusion	*H_0 accepted*		*H_0 accepted*		*H_0 rejected*	
Critical values	10%	2%	10%	2%	10%	2%
	1.1354	1.2035	1.1329	1.1996	1.0205	1.0308
	0.9360	0.9055	0.9420	0.9140	0.9841	0.9761
$D_{Screen\ 9}$	1.1320		1.0911		0.9639	
Conclusion	*H_0 accepted*		*H_0 accepted*		*H_0 rejected*	

[a] the first line gives $F_{\alpha/2,m-1,n-1}$, the second one $F_{1-\alpha/2,m-1,n-1}$

Table 4.3: F test, equality of variances tested on corner-edge, corner-middle and edge-middle populations for the different screens

Single-Factor ANOVA consists of comparing the two hypotheses:

H_0: $\mu_{corner} = \mu_{edge} = \mu_{middle}$

H_a: *at least two of the* μ_i*'s are different.*

where the μ_i's, $i \in$ {corner, edge, middle}, refers to the corresponding population means. The test statistic computed to perform this comparison is given by:

$$f = \frac{MSTr}{MSE} \ .$$

Let's denote I the number of treatments (or populations) under investigation ($I = 3$ in this study) and J_i, $i \in$ {corner, edge, middle}, the number of observations in each sample. f is the ratio of two quantities called *mean squares* that are simply the sum of squares divided by their number of degrees of freedom:

$$MSTR = \frac{SSTr}{I-1} \qquad MSE = \frac{SSE}{n-I}, \quad \text{with } n = \sum_{i=1}^{I} J_i$$

The *total sum of squares (SST)*, *treatment sum of squares (SSTr)*, and *error sum of squares (SSE)* are defined for samples of unequal sizes[2] by:

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij}^2 - \frac{1}{n}X_{..}^2$$

$$SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{I}\frac{1}{J_i}X_{i.}^2 - \frac{1}{n}X_{..}^2$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \bar{X}_{i.})^2 = SST - SSTr$$

where

$$X_{i.} = \sum_{j=1}^{J_i} X_{ij} \qquad X_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij}$$

The Null hypothesis H_0 is then rejected if $f \geq F_{\alpha,I-1,n-I}$, where $F_{\alpha,I-1,n-I}$ is some critical values[3] tabulated from an F distribution.

[2] Since each plate features 4 corner, 31 edge and 55 middle values, the three populations ANOVA is applied to have unequal sizes.

[3] The critical values for this method are computed with the same method as in Section 4.1.2. For large samples, harmonic interpolation is required.

4.2.2 Multiple comparison

When single-factor analysis accepts the Null hypothesis H_0, the analysis is terminated, because no significant differences were found between the populations under investigation. However, in the case where H_0 is rejected, one usually wishes to know which sample population differs from the other ones. *Multiple comparison* procedures aim to carry out this further investigation.

As stated in (Devore, 1991b), the statistics literature present a large variety of such procedures. When sample sizes are equal, most statisticians recommend the method based on Tukey's procedure to determine whether $\mu_i = \mu_j$. However, there is more controversy among them when deciding which method should be applied when sample sizes are different. The procedure presented here for unequal sample sizes is a variation of Tukey's method, recommended in (Miller and Ruppert, 1986).

The multiple comparison procedure we chose has the advantage of being quite straight forward and rather easy to compute. It is based on the *Studentized Range Distribution* to obtain confidence intervals for the considered pairwise differences under investigation. Thus, the procedure can be implemented as follows:

- Given a significance level α, a critical value $Q_{\alpha,I,n-I}$ is determined[4], using some tables featuring Studentized range distribution values.
- The second step consists of calculating the w_{ij}'s whose aim is to determine whether two sample means are different or not. The method to compute the w_{ij}'s is recommended in (Miller and Ruppert, 1986). Two cases are distinguished:

$$w_{ij} = \begin{cases} Q_{\alpha,I,I(J-I)} * \sqrt{MSE/J} & \text{for samples of equal sizes ,} \\ Q_{\alpha,I,n-I} * \sqrt{\frac{MSE}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)} & \text{for unequal sample sizes .} \end{cases}$$

- Pairs whose sample means differ by more than w_{ij} are then considered as significantly different.

Note: a more detailed description of this procedure can be found in (Devore, 1991b).

[4] Because of the large sizes of the data involved, harmonic interpolation is often used to obtain the critical values

Chapter 5

Application to Screens

This chapter presents the main results obtained for the different steps we propose to follow to achieve the quality control of HTS which were presented in the previous chapters. For each procedure applied, some comments are made emphasising the interesting aspects of the chosen methods together with the problems encountered and the limitations of the approach.

5.1 Novelty detection

This first section presents the results obtained for the novelty detection that constitutes the first step of our procedure. Table 5.1 gives an overview of the general results after testing the algorithm on the available screens (for further details see (Fouquart, 1997)). It is to be noted that this procedure was not applied to screen 12 for the simple reason that no additional control plates had been generated for this screen. It was therefore impossible to train and validate the Gaussian mixture model.

Screens	Number of plates	Number of novel plates	Proportion
Screen 1b	263	199	76%
Screen 1	115	105	91%
Screen 2	206	63	31%
Screen 9	206	140	68%

Table 5.1: Proportion of plates declared novel for all the tested screens for a given random seed.

As mentioned in Section 2.3.2, two thresholds are possible for the novelty algorithm that give similar results according to (Fouquart, 1997). As a reminder, we chose to define the novelty threshold as the minimum value of the density function of the validation set. The results in Table 5.1 are obtained for a fixed random seed (randomly chosen). Section 2.4 discussed in detail the variations in the results due to the initial parameters (and especially the random seed factor). Nothing *a priori* justifies the choice of any particular random seed factor since it does not have a specific meaning or interpretation. The only reason why a choice is made is to provide some fixed results that can be used for further studies. But as explained in Section 2.4, the variation in the data likelihood is not worrying since the revised version of the novelty detection algorithm generates a stable ranking of the plate according to their abnormality. Indeed, fixing a random seed gives an indication of the number of invalid plates (for different random seeds, the results vary within a limited range of values) so that the operator can consider all the 'invalid plates' to which he can add the next $n\%$ most invalid ones to be sure that his selection features the plates that are more likely to show some anomalies.

5.1.1 Specific novelty detection: screen 2

Results

A more specific analysis, assay by assay, of the revised novelty detection is proposed in this section. The results presented in Table 5.2 concern screen 2, chosen because it features the smallest daily variation (see Appendix A.3), and therefore is quite close to the type of data generated by an automated screening device.

Date	Number of plates	Number of novel plates	Proportion
28/11/96	40	34	85%
06/11/96	40	3	8%
13/11/96	11	0	0%
07/11/96	40	6	15%
12/11/96	40	16	40%
13/11/96	35	4	11%

Table 5.2: Proportion of rejected plates per day (or assay)

These results are also illustrated by the following four groups of graphs presenting the variations for the different assays (separated by the vertical dashed lines): the top two graphs of each series are dedicated to the maximum and minimum values, whereas the third plot is the negative log-likelihood of the corresponding 4-tuples (the 2 minima $D1 - D7$, the 2 maxima $D2 - D8$).

For the graphs representing maximum and minimum control values, the symbols:

- $\square$ and $\circ$ denote the accepted values
- '$*$' and '$+$' denote the suspicious controls

These results are given by the *Decision graph* that features the novelty threshold corresponding to the minimum value of the likelihood function of the validation set. Therefore, all the points that are above this threshold are declared novel.[1]. In practice, it is the *Decision graph* that is proposed to the operator and on which his decision is based. The advantage of such a graph is obviously its simplicity but also the fact that it presents the information in a clear manner.

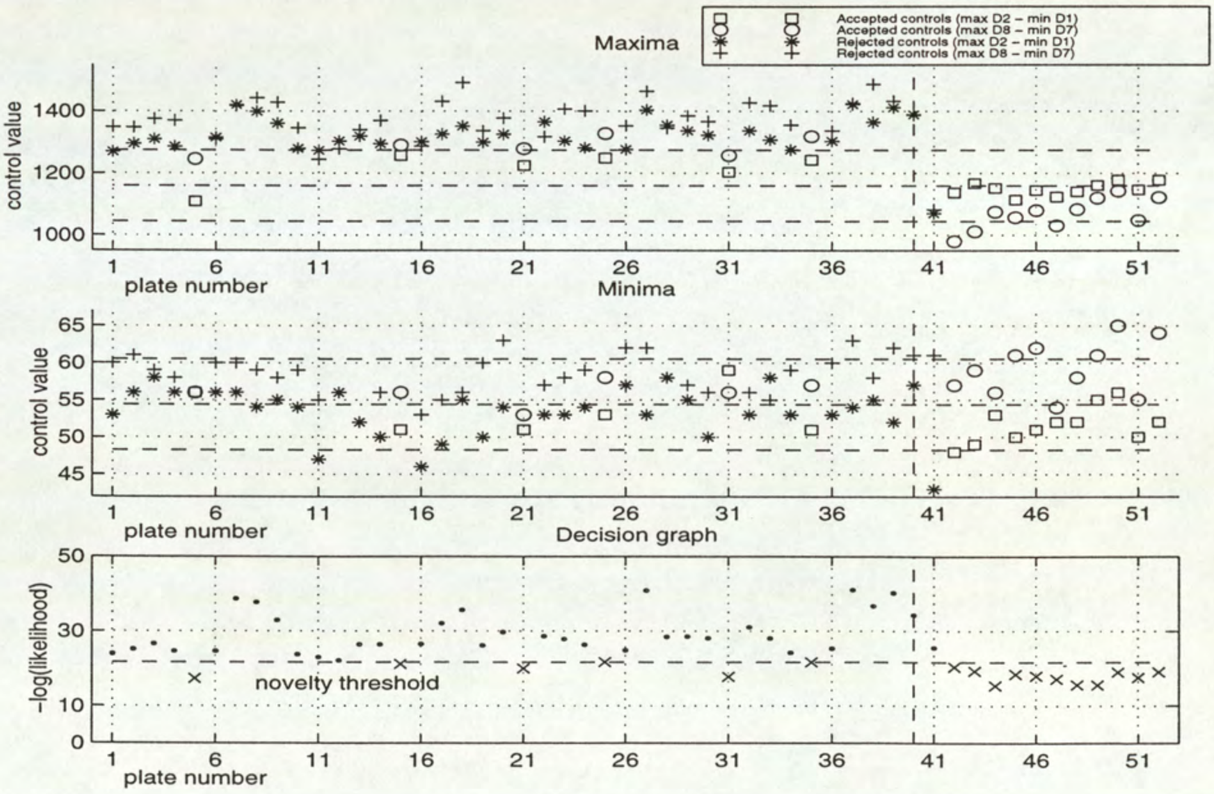


Figure 5.1: **Novelty detection on HTS screen:** plates 1 to 52

[1] Novel points are the ones whose probability is lower than the minimum probability of the validation set with the chosen definition of the threshold. Since the *Decision graph* represents the negative log-likelihood $-log(p(x))$ that can be interpreted as an error function, the lower the probability, or likelihood as stated in 2.3.2, at the point x, the greater the error and hence the more likely the point is to be abnormal

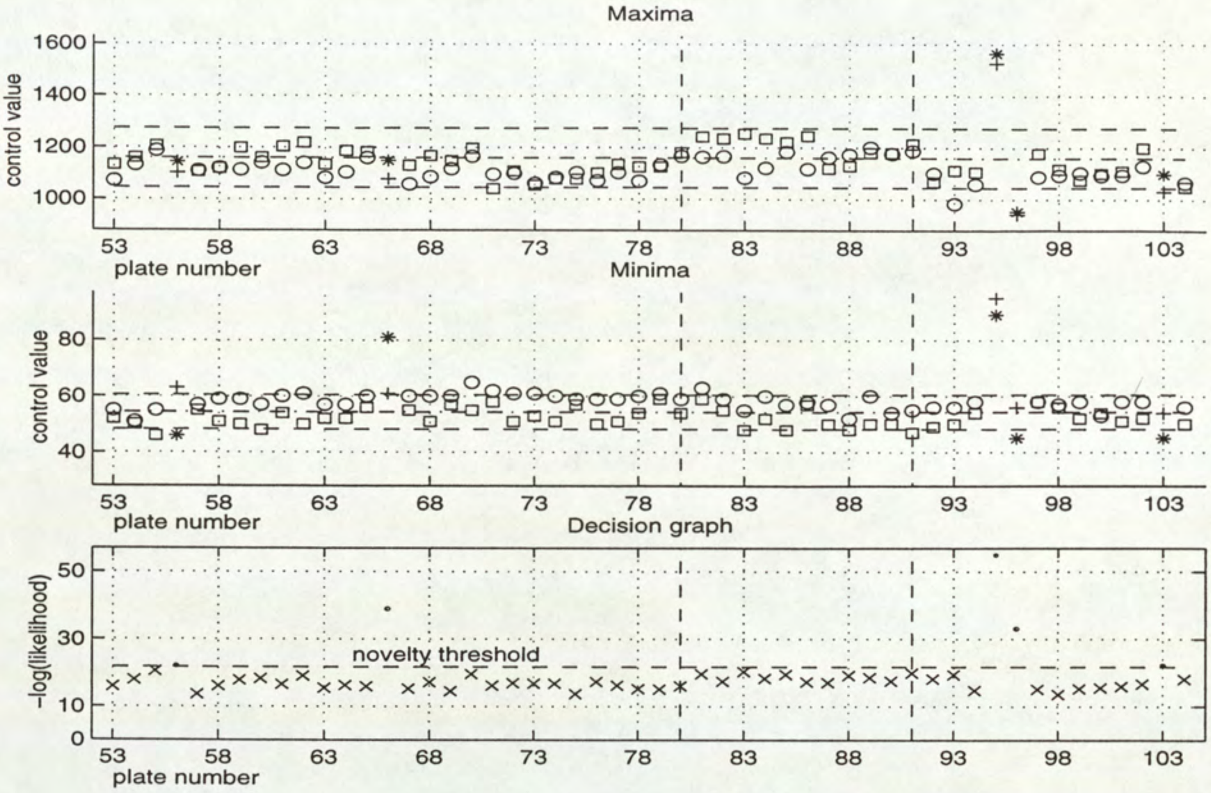


Figure 5.2: **Novelty detection on HTS screen:** plates 53 to 104

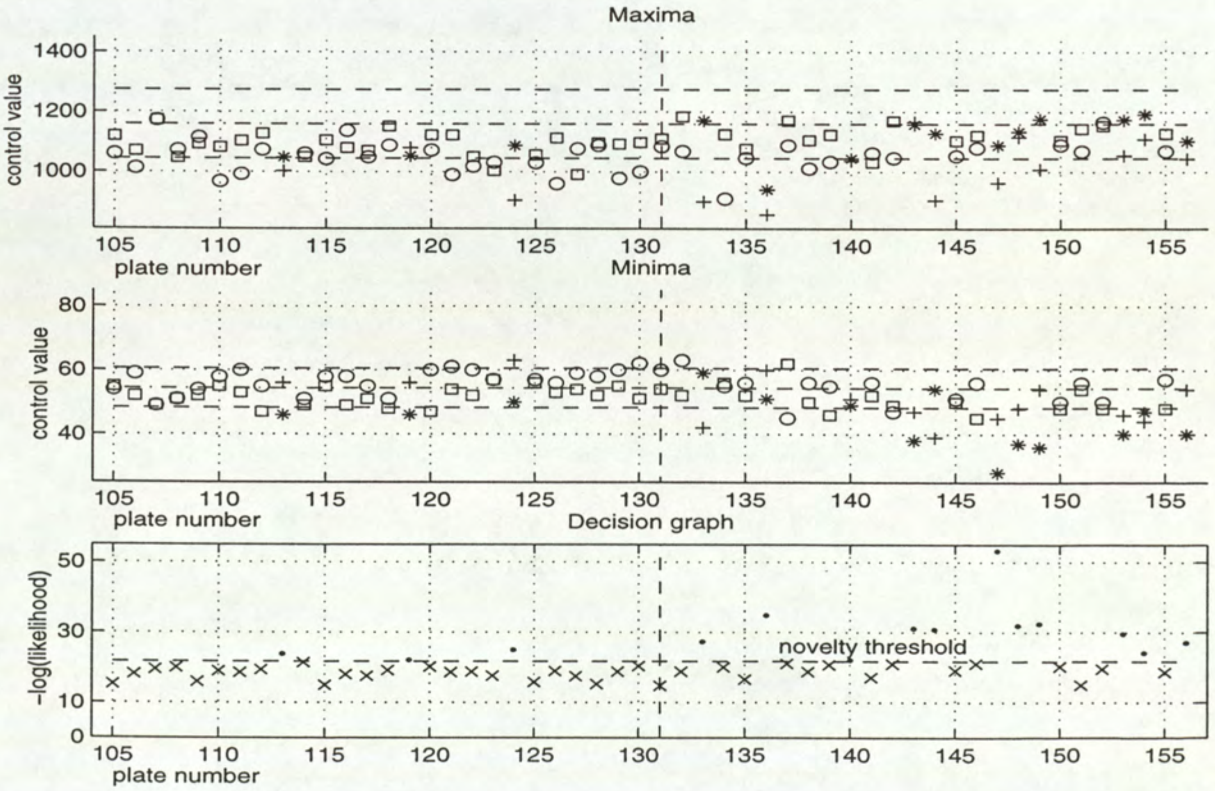


Figure 5.3: **Novelty detection on HTS screen:** plates 105 to 156

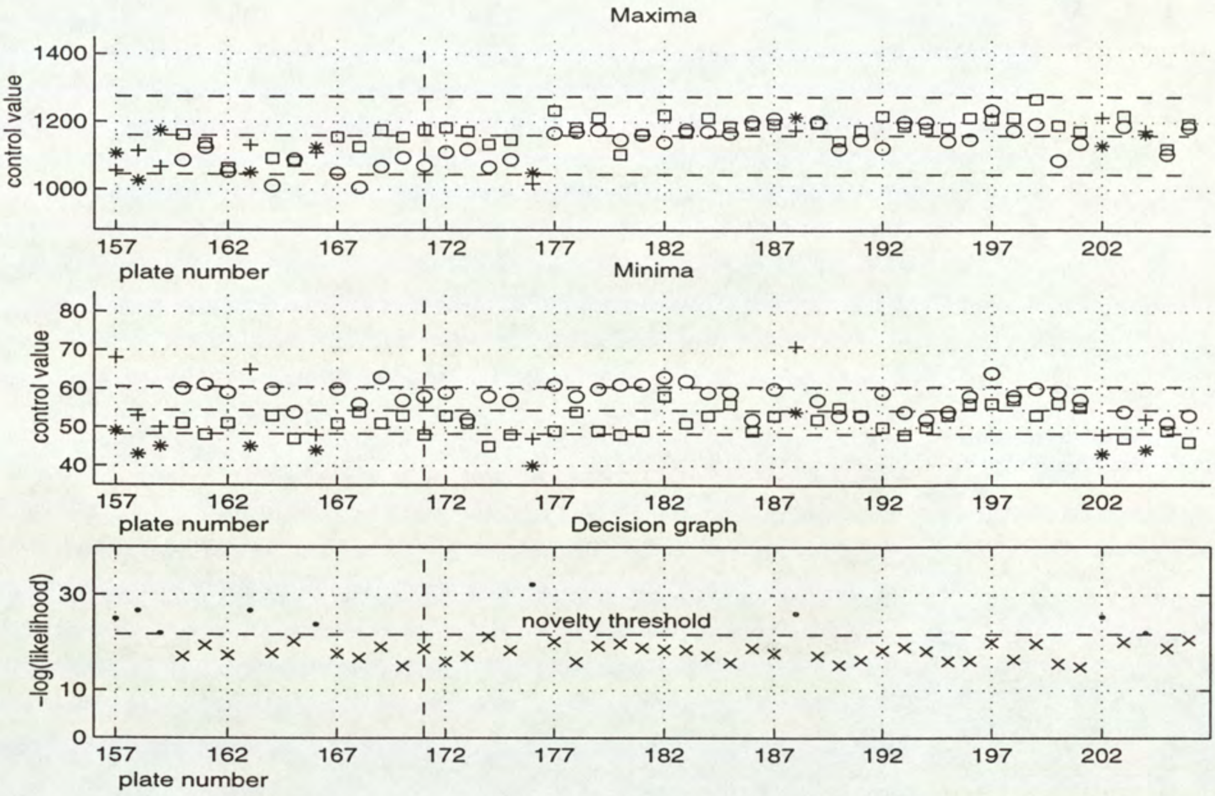


Figure 5.4: **Novelty detection on HTS screen:** plates 157 to 206

Comments

The novelty detection points out 63 abnormal plates out of 206 on screen number 2, as reflected by Table 5.2.

First, as mentioned in Section 1.2.1, it is to be noted that the rejection of a plate *i.e.* a plate considered as abnormal does not imply that some anomalies have been detected on all four control wells of this plate (and would therefore have to be de-selected for the computation of the activity boundaries of the mixture, as explained in Section 1.2.1), but simply that either at least one of the controls is suspicious or that the combination of the four values is unusual. Further investigations are then necessary to determine the kind of problems that occurred.

Concerning the results themselves, Table 5.2 clearly shows that an unusually high proportion of plates are rejected in the first assay (85%). Figure 5.1 indicates that it is the values of the maximum controls that are much higher for this assay than any other. In fact, this is typically the kind of problem these graphs aim to detect. Indeed, such a variation reflects some differences in the experimental conditions: the plates prepared for the first assay have been left to incubate[2] longer than for the others. The mixtures feature therefore a more advanced reaction and the maximum control values are higher than any other maxima prepared in the usual experimental conditions.

Finally, some values such as the ones featured by plates 204 and 206 are rather close to each other and one of them is considered as valid whereas the other one is flagged as suspicious. Such a case, when the corresponding values are close to the novelty threshold, typically reflects the kind of situation when the operator's judgement could be required to choose whether the decision taken by the software is appropriate or whether both plates should be kept or rejected. His judgement is based on the *Decision graph* presented at the beginning of this section.

5.2 Inter-plate variation

The second step of our procedure, based on the normal wells on a plate, first features an inter-plate variation detection relying on a well known general procedure: the Kolmogorov-Smirnov test. Results obtained for this statistical procedure are presented in the following section together with its limitations.

5.2.1 A lack of accuracy?

We first briefly recall the method that was applied in practice to realise the inter-plate variation detection. As stated in Section 3.1.1, a random number of reference plates are chosen among the 'valid' plates detected by the novelty algorithm. Their distribution of values are then compared to the distribution of values of all the other 'valid' plates, using the Kolmogorov-Smirnov test.

Two means of detecting invalid plates can then be applied. Since a suspicious plate should differ from quite a large number of plates, the idea is to both examine the number of plates a reference plate differs from (so as to determine which ones can be considered as suspicious *e.g.* the ones that differ with an unusually large number of plates) and also how many reference plates each plate differs from, given that a suspicious plate shows significant differences from a large number of reference plates.

Table 5.3 presents the results obtained. The number of suspicious plates are obtained with the following method: the plates differing (at a $\alpha\%$ significance level) from more than $\alpha\%$ of

[2] The incubation period is the time during which the interaction of enzyme, substrate and test compound takes place.

the chosen reference plates are considered as suspicious. Such a threshold can obviously be modified depending on how severely one wishes to assess the quality of a plate.

Screens	Nb of valid[a] plates	Nb of reference plates	Number of suspicious plates		
			$\alpha = 20\%$	$\alpha = 10\%$	$\alpha = 5\%$
Screen 1	10	6	0	0	0
Screen 1*b*	64	50	0	0	0
Screen 2	143	100	0	0	0
Screen 9	66	50	2	0	0
Screen 12	796	790	65	58	109[b]

[a] 'Valid' refers here to a study exclusively based on the control values. Hence the valid plates are not necessarily 'valid' in the sense that all the normal wells do not show any suspicious value

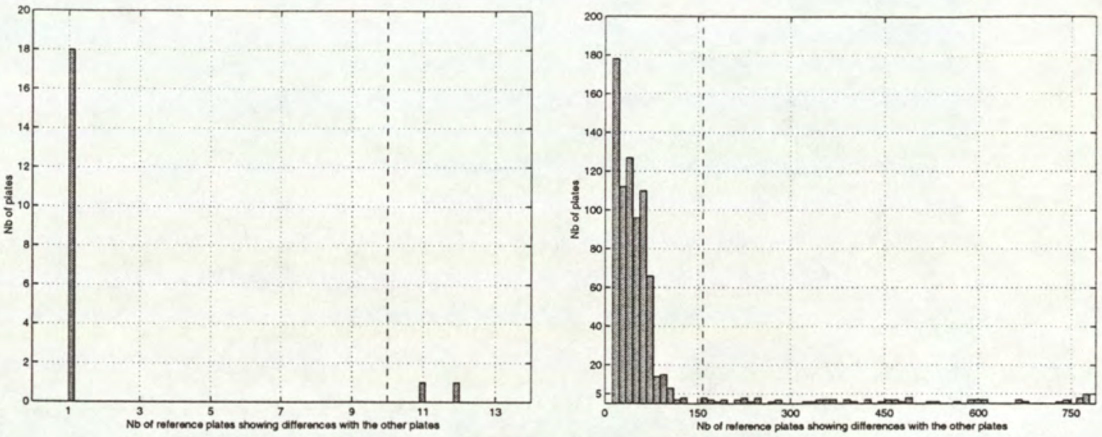
[b] 109 plates are declared suspicious in the sense that they differ from more than $\alpha\%$ *i.e.* 40 reference plates (referring to the chosen definition for the threshold). This result should be undermined by the fact that only 40 plates out of 796 are suspicious, with the same threshold as for a 10% significance level, *i.e.* only 5% of the plates differ from more than 79 reference plates.

Table 5.3: Kolmogorov-Smirnov test applied on the different tested screens so as to detect any suspicious screens

Note: since no control plates are available for screen 12, the novelty detection was not applied to the screen and therefore all the plates are taken into account in the inter-plate variation detection.

The following graphs (see Figure 5.5) give a visual representation of the results obtained for Screen 9 and Screen 12 with a significance level $\alpha = 20$ %. The vertical dashed line represents the threshold defined in the previous paragraph: all plates above this threshold are considered as suspicious.

The results presented in Table 5.3 are quite surprising. If for some screens the procedure applied seems to point out some abnormal plates, for most of the tested screens the distribution of values of the different plates do not seem to show any significant differences. However, this is not worrying. First the plates on which the procedure was tested were already selected by the novelty detection algorithm (obviously the selection criterion only concerns the control wells, but plates showing differences in the experimental conditions are already rejected as explained in the previous section); as a result the remaining plates should not show many anomalies, unless some mechanical problems such as blocked jets or the presence of a large number of 'hits' consequently alter the distribution of data, which does not seem to be the case.



(a) Screen 9, α = 20 %

(b) Screen 12, α = 20 %

Figure 5.5: Each bar represents the number of times y plates (y given by the y-axis differ from a given number x of reference plates. Thus, for screen 9 for instance, 18 plates significantly differ from 1 reference plate

In addition, the testing procedure seems to be quite relevant. Not only does the statistical literature agree on the fact that the best results are obtained when the samples are sufficiently large (see (Kanji, 1993)), which is the case in our study as all of our samples feature 96 data points, but (Neave and Worthington, 1988) also underlines that it is one of the best known distribution-free procedure in order to test for *general* differences between two sets of data. Finally, on a screen on which no data pre-processing was performed (Screen 12 for instance), the Kolmogorov-Smirnov tests detects many differences between the plates and it points out some suspicious plates. However, the small number of invalid plates detected by the procedure (except for screen 12, $\alpha = 1\%$ where the threshold defined before needs to be reconsidered) seems to suggest that, applied to this specific context, the Kolmogorov-Smirnov method could suffer from a lack of sensitivity that needs to be investigated.

5.2.2 Limitations

Procedure requires a threshold

A simple procedure was carried out to test the viability of our procedure and assess it. In the randomly selected reference plates, some wrong values[3] were progressively artificially generated. The same method as before is then applied: for each wrong data point generated, the distribution of values of the reference plate is compared to the distribution of values

[3] In the case of a blocked jet or any other major robot failures, *wrong values* are more likely to be either maximum or minimum values.

of all the other plates. Table 5.4 presents the result obtained when the wrong values are respectively maxima and minima. It is to be noted that concerning screen 12, to avoid a time-consuming procedure the method has been applied on a subset of 50 reference plates.

Screens	Number[a] of minimum values added before detection	Number of maximum values added before detection
Screen 1	10	96
Screen 1b	10	93
Screen 2	9	96
Screen 9	11	78
Screen 12	9	96

[a] on average, out of 96 values

Table 5.4: Testing the sensitivity of the Kolmogorov-Smirnov procedure

Note: these results have been obtained for a 20% significance level. For a lower value, they are slightly higher, but remain in the same range of value.

The Kolmogorov-Smirnov test is supposedly a powerful catch-all test according to (Neave and Worthington, 1988) and the right test to apply when it is unclear what kind of differences to expect between the populations, these results clearly illustrate some of the limitations of the procedure. The first obvious conclusion that can be drawn from this table is that the Kolmogorov-Smirnov procedure does not seem to work efficiently with maximum values: hardly any outliers being maxima (whether they are hits or due to a mistake) would be detected by this general method. The second limitations concern the number of extreme values that can be detected by the test: the Kolmogorov-Smirnov procedure requires a minimum threshold of 10% of unusual values on a plate to detect that something suspicious has occurred.

Time-accuracy trade-off

Another limitation is inherent to the method applied to carry out the Kolmogorov-Smirnov procedure. Because of the large number of plates involved in the HTS process, a method that consists of randomly choosing some reference plates whose distribution of values are then compared to the distribution of values of the other plates can present some limitations. First, the larger the number of reference plates chosen, the more accurate the method. Indeed, even if the statistical test compares the distribution of values of a given reference plate to all the other plates, hence limiting the risks of leaving aside an invalid plate, idealistically all the plates should be taken as a reference. By doing this, the results are undoubtedly more accurate. The drawback to using this procedure, however, is that it is time consuming. Table 5.5 gives an idea of how long the procedure takes for screen number 12, chosen because it contains the greatest number of plates (796).

Number of reference plates	Time
50	20s
100	40s
200	1min 10s
500	3min 00s
790	4min 40s

Table 5.5: Timing of the Kolmogorov-Smirnov procedure for screen number 12 for different numbers of reference plates.

Note: the procedure was performed on a Sparc (Sun 5).

First, these results are quite reasonable, given that the Kolmogorov-Smirnov procedure is applied nref *796 times where nref is the number of reference plates that was chosen. In addition, it is not worrying since the procedure is not in general applied to a whole screen, but as explained in Section 6.3.2 rather to a subset of plates which passed the mixture model based novelty detection test. Finally, if such a procedure has however to be applied on a whole screen featuring more than 500 plates, the difficulty would lie in finding a trade-off between a computationaly expensive method and an accurate procedure.

5.3 Intra-plate variation

Intra-plate variation detection completes the second step of the quality control procedure. It was originally more specifically aiming at determining blocked jets, but in practice can be applied as a general procedure to detect any differences in location or spread. The method applied was carried out on the different 'valid' plates obtained after computation of the novelty algorithm.

5.3.1 Two complementary procedures

A preliminary investigation

As stated in Section 3.2, the procedures applied for the detection of intra-plate variation focus on the distribution of values *per row*, as most of the devices used to fill the plates are 8 head dispensers *i.e.* dispense the mixtures in the wells column after column.

A preliminary investigation has been carried out. *A priori*, no plates among the tested plates were affected by blocked jets, and no mechanical problems were reported whilst collecting the data. This was confirmed by a simple experiment. A row was randomly picked up whose variance was compared to the variance of all the other rows on the plate (excepted the row featuring the control values). This was carried out for each plate among the 'valid' plates and repeated several times with each time a different reference row, the idea being that, if some errors such as a blocked jet had occurred whilst collecting the data, the same row on each plate would show differences with the other rows of the plate.

Not only did this experiment enable us to conclude that no jets were blocked, since the same given row chosen as a reference on all the other plates did not show differences with the other rows of the plate, but it also helped us spot some plates on which the procedure detected some anomalies as underlined by Table 5.6.

Screens	Number of plates	Proportion of plates	Proportion[a] of rows
Screen 1	10	20%	9%
Screen 1b	64	40%	9%
Screen 2	143	30%	8%
Screen 9	66	31%	10%
Screen 12	731	50%	12%

[a] percentage over the total number of rows of the 'valid' set of plates.

Table 5.6: Proportion of rows and plates on which the Siegel-Tukey procedure detected some difference in spread with a 10% significance level.

Note: despite the fact that quite a large number of plates seem to show some anomalies according to this preliminary detection, the low percentages of rows tend to raise a certain number of questions. Since most of the plates only feature a single row with an unusual spread, one can wonder whether it is not simply due to some variation inherent to the testing procedure rather than the presence of some unusual values.

Presentation of the results

Since no specific mechanical problems were detected by this preliminary investigation, to test the accuracy of the statistical procedures applied, a row was randomly chosen on each plate and artificially filled with wrong values (either minimum or maximum values subsampled from the control plates, or directly from the control wells on the plates for screen 12.). Then both the Siegel-Tukey and the Wilcoxon procedures were applied between the artificially generated reference row and the other rows on a plate to investigate differences in location and spread. By doing this, we should hopefully have a better idea of what the actual limitations of our procedures are.

Table 5.7 summarises the results obtained.

Screens	Proportion of non detected artificial row								
	$\alpha_{Wilcoxon}$			α_{Siegel}			$\alpha_{None\ of\ the\ procedures}$[a]		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Screen 1	68%	85%	90%	30%	63%	85%	15%	14%	60%
Screen 1*b*	40%	48%	61%	64%	73%	92%	16%	25%	53%
Screen 2	46%	52%	65%	90%	96%	99%	37%	49%	64%
Screen 9	22%	28%	42%	94%	96%	99%	19%	25%	41%
Screen 12	26%	31%	43%	64%	71%	80%	14%	20%	37%

[a] These three columns indicate the proportion of artificially generated wrong rows that are not detected by any of the procedures computed, for three different significance level.

Table 5.7: Proportion of the artificially generated rows that are not detected by the statistical methods applied for different significance levels, when the wrong added values are maxima.

Analysis of the results

First, it is to be noted that this table presents the results obtained when the wrong values added on the chosen reference row are maximum values. The reason why we did not mention the experiment featuring wrong minima is that, the Wilcoxon test that investigates for differences in location happened to give perfect results as expected: 100% of the artificial wrong minimum rows were detected, the reason being that for all the tested screens, the distributions of values are in general closer to a maximum than a minimum activity well. Therefore minimum values generated by any kind of mistakes or hits are most likely to be detected by a procedure seeking differences in location.

In addition, it is quite obvious from Table 5.7 that when applied separately, both procedures do not perform very well on the different tested screens. If the results are not too bad as far as screens 9 and 12 are concerned for the Wilcoxon procedure (investigating differences in location), the results are however rather poor for the other screens. Besides, however surprising that may seem, contrary to our expectations, the Siegel-Tukey procedure does not work very well or at least it is far less efficient than an investigation for differences in location (excepted for screen 1) and the results obtained with the Siegel-Tukey procedure for screen 2 and 9 are even dramatic.

However, this is not worrying since the method that consists of combining both procedures to investigate differences in spread together with a difference in location gives rather satisfying results, at least much better results than any method taken alone. Indeed, for a significance level of 10%, on average 18% of the wrong rows are not detected by any of the two methods and 24% for $\alpha = 5\%$, with a screen presenting 'extreme' results (screen 2 features more than 30% of invalid rows that are not detected). Even if the results vary from one screen to another it is to be noticed that the results are quite similar and acceptable for a significance level of 10% and 5%, but definitely unsatisfying for $\alpha = 1\%$.

These results clearly show that, in fact both procedures are complementary. As underlined by (Neave and Worthington, 1988), both of these testing procedures are designed to be fully efficient to detect whether the two populations are different solely in the specified manner, *i.e.* either in location or in spread. If the populations differ in any other manner (or both in spread and location), the tests suffer a severe reduction in power. This can be one of the reasons why, when considering screens 9 and 2 for instance, the results are quite satisfying with the Wilcoxon method and so poor with the Siegel-Tukey procedure: the Siegel-Tukey test gives awful results simply because there is not any difference in spread in the data. In fact while in most of the comparisons between an artificial invalid row and a normal row, the main feature is a difference in location, a situation in which the Siegel-Tukey procedure is known to be very inefficient.

5.3.2 Limitations

The procedures described in Section 3.2 do not show the expected accuracy. Despite the fact that the results are quite satisfying, as shown previously the statistical procedures we propose to apply seem to be limited to a certain threshold (on average 20% of the artificially generated wrong rows featuring maxima for a 10% significance level).

First, these limitations are inherent to the procedures themselves. Indeed, both the Siegel-Tukey and Wicoxon methods are only fully efficient when applied in their own specialities, *i.e.* either to detect differences in spread only (for the Siegel-Tukey test) or differences in location only (for the Wilcoxon procedure). Any other kind of differences (especially differences in both spread and location) are less likely to be spotted.

In addition, because of the nature of the procedures (investigating differences in location or spread by comparing the data populations of the rows on a plate), the method can become rather time consuming, especially when applied to a typical HTS screen featuring hundreds of plates. However, this should not be too worrying if the statistical tests are integrated in a automated procedure such as the one described in Section 6.3.2, where an 'on-line' intra-plate detection is carried out, which means that it would only concern a set of plates in practice, not a whole screen.

Finally, one the fundamental reasons why the procedures show the threshold mentioned above concerns the data values themselves. One can argue that, in most of the cases, the artificial rows featuring maxima are not detected because the statistical methods computed suffer from a lack of accuracy whilst detecting the pattern under study (the variability of the data or a difference in the mean). But in fact, the invalid rows are not detected because the artificial values added before detection do not introduce enough spread in the data. But also they do not affect the mean of the data significantly enough. For this reason, there is no way in which such values can be detected (even by any other procedure).

5.4 Edge and corner effects

The last part of the quality control procedures investigates what is happening at the edges of the plates, in order to spot some systematic alteration, due to the surrounding air for instance, so as to improve, if necessary, the design of HTS experiments. This section presents the results of this investigation.

5.4.1 Results of the detection of any systematic effect

The Analysis of Variance procedure was applied to cells on the set of 'valid plates' obtained so far, divided into three different populations: edge, corner and middle values. As stated in the previous chapter, because we are only interested in systematic effects, all three populations have been purged of extreme values, all the values that differ by more than 2 standard deviations from the mean in fact.

Table 5.8 presents the results obtained for the preliminary investigation, single-factor analysis. The first observation that can be made on these results is that all the tested screens excepted screen 9 show differences in at least two of the populations for which the method was applied, which suggests that there are some edge or corner effects.

The second observation concerns screen 2. The value obtained for the statistic f seems to be rather high. And if we recall that the Normality of edge and middle populations was subject to uncertainty (see Section 4.1.1), this result should be interpreted with caution.

On the screens showing some differences, we apply the multiple comparison procedure described in Section 4.2.2 to investigate which populations significantly differ from the other ones. The results obtained are presented in Table 5.9.

All three pairs of populations for screens 1 and 2 seem to be significantly different, which suggests that there are both edge and corner effects. The results obtained for screen 1b lead to slightly different conclusions. Edge and middle values clearly show significant differences, whereas corner and edge populations don't show any. However, if corner and edge populations show similarities with a 5% significance level, a further investigation carried out with $\alpha = 1\%$ might suggest to consider this result with caution.

In addition, (Devore, 1991a) underlines that the multiple comparison procedure described in Section 4.2.2 should be computed on samples whose sizes are reasonably close to one another.

5.4.2 Unbalanced samples

The samples under investigation are unbalanced (4 corner, 31 edge and 55 middle values per plate). And the multiple comparison method applied is only based on approximations which are valid for samples that are not too badly unbalanced, as recommended in (Miller and Ruppert, 1986). An intuitive idea is thus to randomly subsample from the larger populations (edge and middle values) so as to apply the multiple comparisons procedure on samples of equal sizes. Such a process is then computed 100 times so as to give a representation as close

Critical values[a]	10%	5%	1%
	2.3039	2.9976	4.6089
$f_{\text{Screen 1b}}$	25.5391		
Conclusion	H_0 *rejected at every level*		
Critical values	10%	5%	1%
	2.3092	3.0067	4.6306
$f_{\text{Screen 1}}$	20.7339		
Conclusion	H_0 *rejected at every level*		
Critical values	10%	5%	1%
	2.3034	2 .9967	4.6068
$f_{\text{Screen 2}}$	169.16		
Conclusion	H_0 *rejected at every level*		
Critical values	10%	5%	1%
	2.3040	2 .9977	4.6091
$f_{\text{Screen 9}}$	1.4011		
Conclusion	H_0 *accepted at every level*		

[a] obtained after computing a harmonic interpolation of tabulated values

Table 5.8: Single factor ANOVA applied to the different screens

as possible of the whole data set.

The results obtained are presented in Figure 5.6, where the missing parts of the pie charts represent some irrelevant results (for instance, when the Null hypothesis that all the populations do not show any significant differences is rejected and the multiple comparison procedure leads to conclude that all the three populations are the same) that constitute a low percentage. Obviously, subsampling from the larger populations has its limitations, in particular a lack of accuracy inherent to the method itself, even if it is computed a hundred times, but it gives quite a reasonable idea of the phenomenon under investigation: edge and corner effects detection.

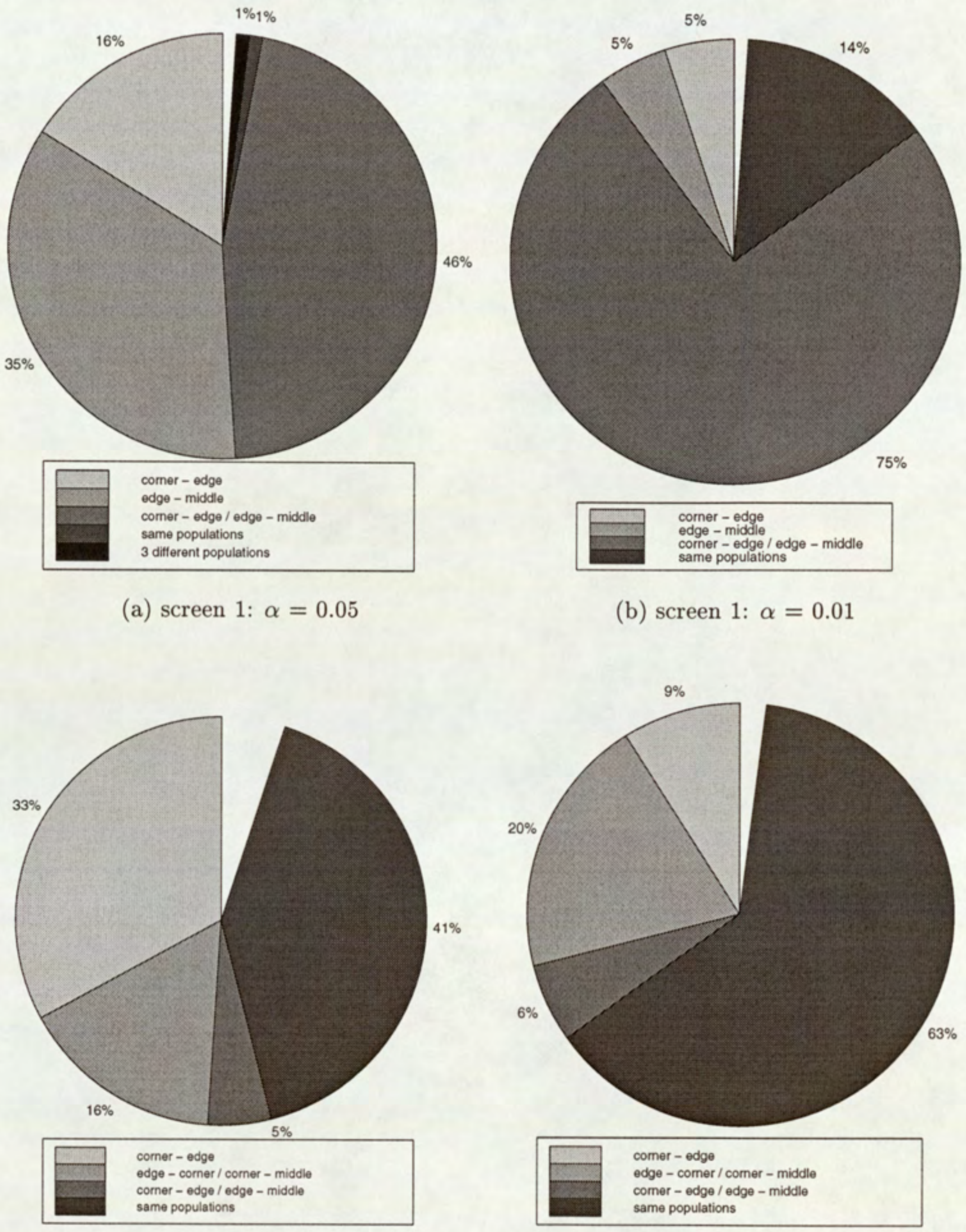
The main conclusion that we can draw from this experiment is that the results obtained for unbalanced samples are confirmed. The procedure computed on subsamples of equal sizes gives evidence of a similarity between corner and edge values for screen 1b. It is first to

	Corner-Edge		Corner-Middle		Edge-Middle	
W_{ij}	5%	1%	5%	1%	5%	1%
	61.6593	76.7553	60.0902	74.8020	26.1148	32.5085
$\lvert\bar{X}_{i.} - \bar{X}_{j.}\rvert_{\text{Screen 1b}}$	6.6036		70.8391		77.4427	
Conclusion	No differences		Some similarities		Significant differences	
W_{ij}	5%	1%	5%	1%	5%	1%
	138.9047	173.0011	135.2241	168.4171	59.4794	74.0796
$\lvert\bar{X}_{i.} - \bar{X}_{j.}\rvert_{\text{Screen 1}}$	190.3048		303.9476		113.6428	
Conclusion	Significant differences		Significant differences		Significant differences	
W_{ij}	5%	1%	5%	1%	5%	1%
	6.9544	8.6566	6.7795	8.4389	2.9468	3.6681
$\lvert\bar{X}_{i.} - \bar{X}_{j.}\rvert_{\text{Screen 2}}$	19.5577		38.1306		18.5729	
Conclusion	Significant differences		Significant differences		Significant differences	

Table 5.9: Multiple comparison ANOVA applied to the different screens

be noted that the large majority of the sampled populations have no significant differences. Nevertheless, for each result that did not feature 3 similar populations, edge and corner values show similarities (61% of the cases where 3 similar subsamples have not been detected for a 5% significance level) and there are also some similarities between corner and middle values as the procedure performed on unbalanced samples seemed to suggest.

Earlier results showed three different populations as far as screen 1 is concerned. The results obtained after subsampling tend to be slightly different as 46% of the subsamples with a 5% significance level and 75% with a 1% significance level show this particular feature: similarities between corner-edge populations on the one hand and edge-middle populations on the other hand. So, maybe the fact that there are both corner and edge effects could be subject to uncertainties, the results obtained seem however to prove that populations on the corner of a plate show differences with the middle populations.



(a) screen 1: $\alpha = 0.05$

(b) screen 1: $\alpha = 0.01$

(c) screen 1b: $\alpha = 0.05$

(d) screen 1b: $\alpha = 0.01$

Figure 5.6: result of 100 computations of the multiple comparisons procedure.

Note: the results obtained for screen 2 are not depicted on Figure 5.6 for the simple reason that 100% of the subsamples show significant differences. This confirms the result obtained for unbalanced samples and clearly suggests that screen 2 features both edge and corner effects, bearing in mind however that the Normality of the data was subject to uncertainty.

Chapter 6

Conclusions

This chapter presents the conclusions of the thesis. It first recalls the three step procedure we propose to follow to tackle the quality assessment of the data involved in High Throughput Screening.

The novelty detection based on control values to spot potential variations between the control and the normal plates is the first step. Quality control is then completed by an inter- and intra-plate variation detection to highlight potential general mechanical errors or experimental mistakes. The final step especially focuses on the detection of corner and edge effects.

Some comments are then made about the different procedures: discussions about their advantages and particularities together with their weak points and limitations.

The last section deals with some possible ways of presenting the results to users. It introduces the practical context in which the implemented procedures are more likely to exist.

6.1 A Three-step procedure

Three different steps have been proposed to carry out the quality control of HTS. The first one concentrates on control values only and consists of pointing out the unusual control wells on a plate.

6.1.1 A novelty detection method based on control values

First, a preliminary study validated the procedure based solely upon the control wells to assess the quality of the data. The traditional approach to deal with outliers suffered from some severe limitations: a lack of robustness while testing the hypothesis whether an extreme value is suspicious or not, no possibility of automation, and an assumption of normally distributed data.

The novelty detection method, the first step of our procedure, is based on density inference. Three plates are added at the beginning of each screen, featuring only control values (minimum and maximum controls)[1]. Since these additional plates feature wells that are filled with the same mixture as the control wells on a normal plate, they can be used to assess the quality of the controls on normal plates.

A Gaussian Mixture model was chosen (because of some practical constraints such as computational efficiency) to model the unconditional probability density of the control values. The data of control plates are separated into 2 sets: a training set to train the Gaussian mixtures and a validation set that determines the different parameters of the model, its complexity and the structure of the covariance matrix, using cross-validation techniques. A threshold is then set to decide whether a value is novel or not in terms of density estimation. It is chosen as the minimum value of the density function of the validation set.

The final stage of the method is the novelty detection: the control wells on a normal plate that show a lower probability than the lowest probability of the validation set, the novelty threshold, are declared novel and the plate is flagged as abnormal.

6.1.2 Inter- and intra-plate variations

The second step of the implemented method focuses on the normal wells only. It aims at determining any general effect (such as handling mistakes or problems in a robot) that could alter the HTS data. The procedures implemented are carried out on the first set of 'valid' plates that the novelty detection has determined.

Inter plate variation

We wish to detect invalid plates where the control values are unaffected. But since we do not know what any single value on a plate should be — because of the inherent nature of the data (biological mixtures) — we are forced to compare the distribution of values on the whole plate to detect any inter-plate variation.

The procedure randomly chooses among the set of plates considered valid by the procedure outlined in Section 6.1.1 a certain number of plates as a reference. We can then compare their distribution of values with the distribution of values of all the other plates from the set, each plate being taken one after another. To see how significantly the distribution of

[1] These three additional plates are called 'control plates'.

values from the plates taken as a reference differ from each plate, a general significance test — the Kolmogorov-Smirnov test, said to be the right test to use for detecting *any* difference (especially of a more general nature than a difference in location only or in spread only) between two samples — is used.

How is the determination of invalid plates achieved? The idea behind such a procedure is that an invalid plate shows significant differences with quite a large number of reference plates. Thus, two ways of detecting invalid plates are considered: in an automated procedure the invalid plates would be determined by setting a threshold, for instance the plates differing by more than α per cent from the reference plates, α being the significance level of the statistical test, are suspicious. As far as a manually conducted or half automated procedure is concerned, by inspecting the number of reference plates a plate differs from, the operator can decide whether the plate is suspicious or not and hence is worth being investigated further.

Intra-plate variation

The quality control procedure is completed by the detection of variations within a plate. This investigation was performed with the particular aim of detecting mechanical problems like blocked jets.

Since this study depends on the kind of robots that are used to fill the HTS plates, we developed a method to be applied to 8 head dispensers *i.e.* robots that fill a plate column per column, since they are the most commonly used in the HTS process. It can be easily adapted to other robots by changing the subset of values that are considered as a group. The underlying idea is that a blocked jet induces values that should be close to one another: indeed, a lack of reagent (*i.e.* the added compounds to be tested on the biological target) in a well corresponds to either a minimum or a maximum activity mixture, depending on the nature of the target. Since a blocked tip affects a given row (8 head dispensers fill the plates column per column so that a given row is associated to a specific tip), the method we developed is based on a comparison between the rows of a plate: a row is taken as a reference and compared to the other rows of the plate. The procedure is all the more reliable since many reference rows on the plate are chosen, but also more time -consuming.

The procedure relies on statistical methods to investigate both a difference in spread and a difference in location. Since a blocked jet generates values quite close to one another, the wells that are not affected by the faulty tip show a larger dispersion per row, hence an investigation for differences in spread. Moreover, as the wrong values are either minimum or maximum values, a detection of differences in location can also be powerful in spotting mechanical problem.

6.1.3 Edge and corner effects

The final step of the quality control procedures, like the novelty detection, assesses the experimental conditions in the sense that it investigates what is happening to the wells located on the edges of the HTS plates. In particular, a distinction is made between corner and edge values so as to detect any specific alteration of the data due to the the exposure to the surrounding air or the number of neighbours (a well located on the corner of a plate has less neighbours than the wells on an edge or the middle wells).

The procedure is based on the Analysis of Variance method. Three populations are under investigation: a population of corner, edge and middle values. To start with, Single Factor Analysis aims at detecting whether the populations are similar or not. Then if any significant difference is detected, multiple comparisons procedures are computed so as to give a better idea of which population differs from the other populations under investigation.

By seeking any systematic edge and corner effects, the procedure gives indications about how to improve the results of HTS, for instance by redesigning the experiments or recalibrating values. Indeed the detection of significant differences between edge, corner and middle values on many screens might suggest that the exposure to the air for instance has an influence that should be further investigated or that the activity of a well is more likely to be altered if there are many neighbours.

6.2 Comments

6.2.1 Achievements

First, the procedures developed in the general context of quality control encompass different techniques and procedures whose result is a rather complete assessment of the HTS process. The experimental conditions are subject to investigation, which is all the more vital since plates aren't screened the same day due to the large number of data a typical HTS experiment involves. Thus, the role of novelty detection is to spot any anomalies in the experimental conditions, such as the variations detected on screen 2 (see Section 5.1.1) most probably due to the fact that the plates of the first assay were incubating much longer than the other plates of the screen. In addition, edge and corner effects detection aims at assessing the design of the experiment itself by investigating whether the specific location of some wells (on the edges of a plate) induces some systematic effects. The results obtained tend to prove that a majority of the screens feature wells whose activity might be influenced to some degree by the position they occupy on the plate.

Moreover, both the control positions on a plate and the normal wells are examined by the different steps we propose to follow. First, novelty detection assesses the *control wells*, which is crucial for the HTS process for the reason mentioned above. In addition, a general procedure such as the Kolmogorov-Smirnov test to investigate differences of any kind between the plates, added to procedures seeking differences in spread and location between the rows (so as to detect any intra-plate variations) complete the quality control task by detecting suspect values on *normal wells*. If the Kolmogorov-Smirnov method does not seem to be accurate enough to determine suspicious values or hits when they are close to a maximum, the work carried out to detect variations within a plate gives satisfying results as stated in Section 5.3.1. Indeed when rows are artificially filled with minimum values the procedure gives a perfect result as it spots 100% of the wrong rows; when the wrong values are maximum values, an average of 20% of the artificial rows are not detected by any of the two procedures investigating differences in location and spread.

6.2.2 Discussions

The procedures however suffer from some drawbacks. First, as far as the novelty detection is concerned, the main limitation is inherent to the method applied to carry out this detection. Indeed, the cornerstones in this procedure are the three control plates added to the beginning of each screen. Since they are taken as a reference, it is *crucial* that the screening of these plates is conducted with maximum care and accuracy. Even if the outlier detection techniques can cope with some insignificant extreme values, if a major problem occurs during the screening of the control plates, the whole novelty detection would fail and spot a high number of invalid plates, because the references the method relies on are wrong.

Besides, the screening process is conducted assay by assay (*i.e.* day per day). Therefore, if the experimental conditions are not steady enough, the distributions of the data of each assay could display significant differences. This could lead to a high number of novel plates since the control plates can not capture the entire variation on the whole screen unless they are generated for each assay. But this is rather idealistic and not really achievable in practice since it would be too time- consuming and most of all it would involve extra costs. It is therefore up to the operator to decide whether the systematic variations between the control and the normal HTS plates are acceptable or not.

Several limitations prevent the inter- and intra-plate variation detection from being fully efficient. The first problem encountered concerns a lack of sensitivity. As mentioned in Section 5.2.2, despite being the right test to use to detect differences of *any* kind between the distribution of values of normal plates, the Kolmogorov-Smirnov procedure first turns out to be unable to spot invalid maximum values, and requires a minimum threshold of 10% suspicious values to detect an anomaly.
However, the problem of invalid maximum values is partly solved by the intra-plate variation detection that can be considered as an extension of the general Kolmogorov-Smirnov procedure, in the sense that it features a more detailed investigation, focusing on differences within a plate. Indeed, suspicious maximum values are detected by a combination of two statistical procedures that investigate both differences in spread and location. Nevertheless, as before, the methods applied do not seem to be fully efficient: an experiment was designed where a randomly chosen row on each plate is filled with artificial wrong values. On average over the tested screens, 20% of the invalid rows were not be spotted by any of the procedures. In most of the cases, the artificial rows featuring maxima are not detected, not because of a lack of accuracy of the computed statistical methods, but simply because the invalid data points generated neither introduce enough variability nor enough weight to alter the mean of the data. In other words the values in the artificial rows are completly indistinguishable from real data. For this reason, there is no way to detect them in principle.
Finally, the methods applied involve a comparison between the data populations of the rows on a plate or a comparison of the distribution of values of two different plates. It can therefore become rather time consuming (especially when applied to a typical HTS screen featuring hundreds of plates): obviously, the more comparisons, the more accurate the procedure is. However, this should not be too worrying when the procedure is integrated in a automated procedure (such as the one described in Section 6.3.2, where an 'on-line' intra-plate detection is carried out) as it only concerns a set of plates in practice and not a whole screen.

The final step of our method seems to give rather satisfying results. However, the main difficulty lies in validating the hypotheses hidden behind the ANOVA procedure. Indeed, for some of the tested screens, the Normality of the data was rejected. Is that due to the fact that there are too many outliers, or does it come from an inaccuracy of the computed goodness-of-fit test? Despite this, the ANOVA procedure was performed on such a 'suspicious' screen, but because of this uncertainty while validating the hypotheses, the results obtained should be considered with caution.

Another problem encountered concerns the second step of the ANOVA procedure. Indeed, some statistics literature mention that the multiple comparisons procedure that was computed to detect potential edge or corner effects should not be run on badly unbalanced samples. The three sample populations under investigation (corner, edge and middle values) are not balanced since an HTS plate features 4 corner, 31 edge and 55 middle values. Therefore the procedure was carried out a second time, subsampling from the larger populations so as to run the multiple comparisons on balanced samples. The results obtained with unbalanced samples were slightly modified but their main pattern remained: values on the edges of a plate (corners and edges) seem to show differences with the middle population.

It is to be noted that an alternative to the ANOVA procedure could have been to investigate some non-parametric methods. The main drawbacks of such methods are that they are much more complicated to manipulate than parametric methods such as the ANOVA procedure, and most of all, they do not allow for interpretability.

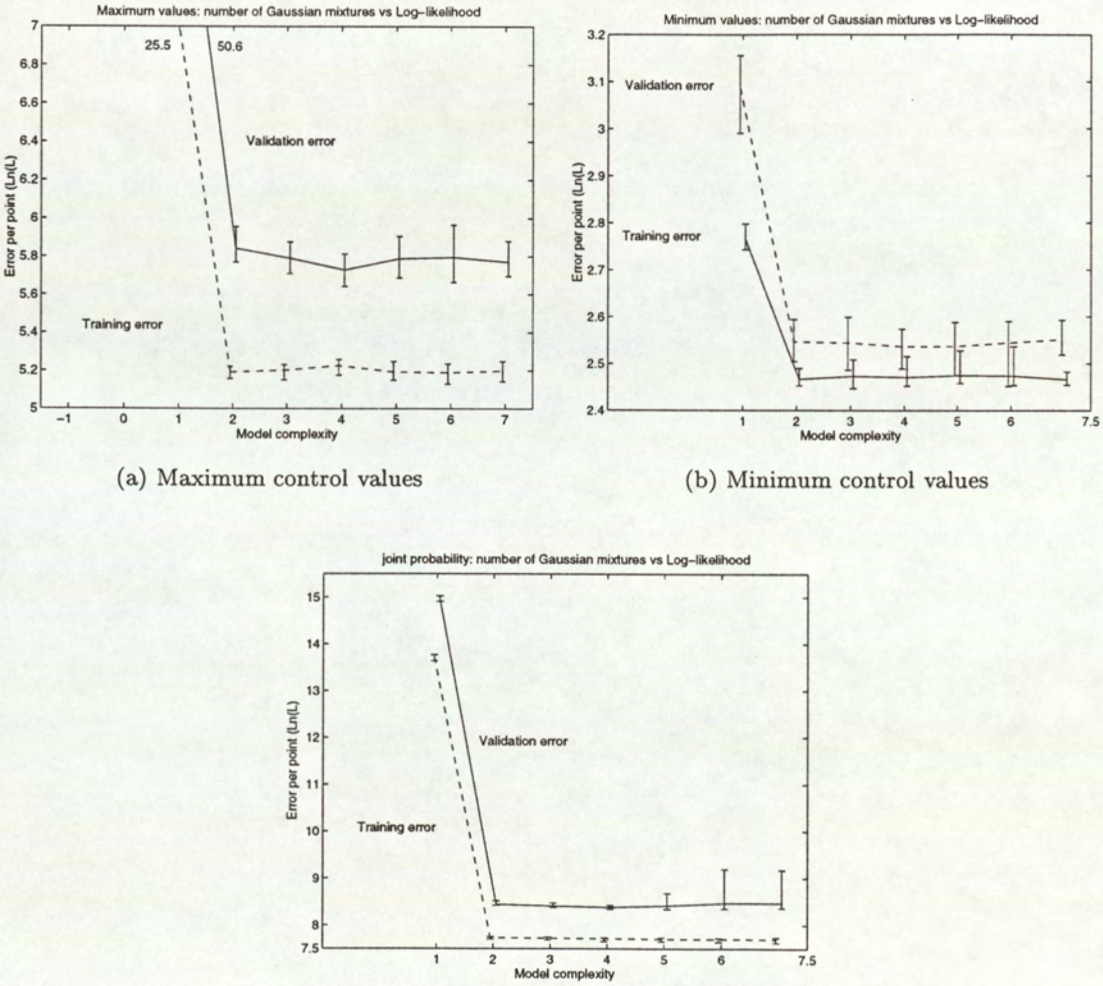
6.3 Quality control applied on real life processes

This section gives some ideas about how the procedures that were built up could be adapted to be applied on two different processes Pfizer is currently working on (or plans to use) to perform the quality control of High Throughput Screening.

6.3.1 Analysing the control wells under the existing system

The system Pfizer is currently using to validate a screen features an assay by assay analysis and concentrates on control wells. This means that variations from one assay to another, because the screening was conducted under unsteady experimental conditions for instance, are not detected. Indeed, the high number of novel plates detected for screen 2 (see Section 5.1.1) is probably due to a significant difference during the incubation period, and it would not have been detected by an assay by assay analysis, since most of the plates flagged as 'abnormal' are not suspicious when compared to the other plates of the same assay. The first step of our procedure aims at assessing the experimental conditions so as to validate the different assays. Thus, the novelty detection is advisable to carry out the analysis of the results on a whole screen.

A preliminary stage can be considered, which consists of defining the parameters of the mixture models for each screen *e.g.* mainly the number of Gaussian mixtures the model should feature to fit the data. Figure 6.1 depicts the kind of graphs that can be used to achieve this selection. Since each plot represents the error of the model as a function of the complexity, the complexity chosen should correspond to the lowest error, all graphs being taken into account.



(a) Maximum control values

(b) Minimum control values

(c) Joint probability

Figure 6.1: Determination of the complexity of the Gaussian Mixture model for screen 2. For each degree of complexity, the likelihood of the model is computed 10 times. The 2 plots representing respectively the training and validation errors joins the average of all the values computed that range over the error bars.

Once the parameters of the model are defined, the novelty detection procedure can be used to assess the control values of the whole screen. Graphs such as Figures 5.1 to 5.4 in Section 5.1.1 point out the plates featuring suspicious control values.

6.3.2 A fully automated procedure

The Robolab procedure is a fully automated screening system. When programmed appropriately, it undertakes the whole assay from the very beginning of the HTS process through to data capture. A copy of the data is then sent to the existing system for analysis.

The advantages of automation are considerable. Such a robot can operate 24 hours a day, and achieve greater throughputs than that achievable by any operator. The second advantage is that the variations between the different samples are minimised, since the procedures are controlled by a robot. It is therefore a necessity to make sure that the Robolab functions are fully efficient. In this respect, any kind of mechanical errors like blocked jets should be detected as early as possible so as to avoid conducting the whole HTS process a second time. Besides, the aim is also to screen all the plates in a row, without interrupting the robot's work, so as to avoid the necessity of reprogramming its functions. Thus, plates on which some errors are detected can be put at the end of the screen and go through a second run (featuring wells with different mixtures for instance so as to carry out further experiments), all this without stopping the robot. Therefore, an 'on-line' detection is advisable. This method could be carried out as follows:

- Plate selection

The first step consists of selecting a set of plates on which the different quality control treatments are applied. However, a trade-off must be found between a sufficient and a reasonable number of plates: *sufficient* so as it makes sense to run the procedures (especially as far as the inter-plate variation detection is concerned) and *reasonable* so as to avoid time-consuming methods.

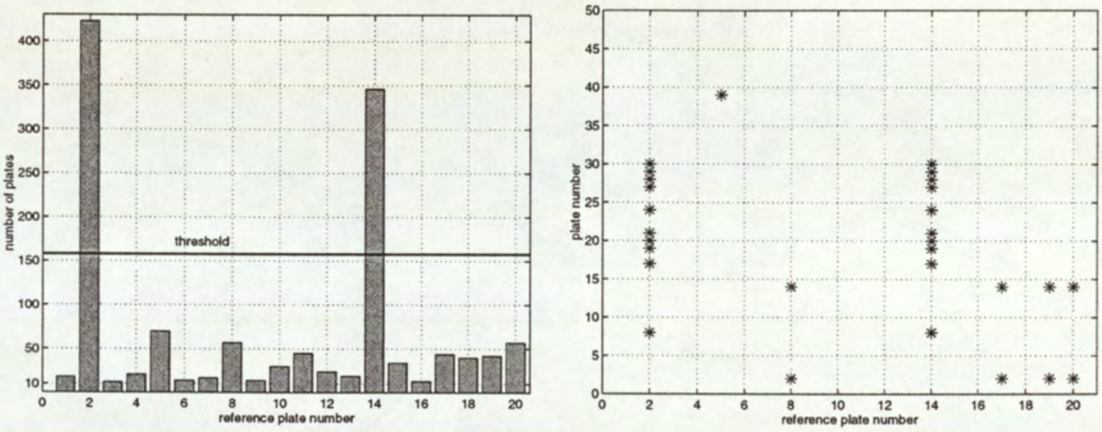
- Novelty detection

The novelty detection is conducted on the same principle as in the previous section. The number of times cross-validation is run to select the best model (as explained in Section 2.3) can be adapted so as to reduce the learning time of the novelty detection. Besides, the model only needs to be trained once (on the control plates), it then can be used on any single plate of the screen.

Novelty detection rejects some plates. From now on, we only consider the 'valid' plates.

- Inter-plate variation detection

The Kolmogorov-Smirnov procedure can then be used as a general method to detect suspicious plates. To do that, each plate is respectively taken as a reference and its distribution of values is compared to the distribution of values of the other plates. Since the method is computed on a small number of plates, it should not be too time consuming. Then graphs like Figure 6.2 help the operator to detect the suspicious plates. The suspicious plates are for instance the plates that show differences with more than α% of the total number of reference plates, α being the significance level of the Kolmogorov-Smirnov test. It's then up to the operator to determine whether or not this threshold (depicted by the horizontal line) is acceptable.



(a) How many plates are different

(b) Which plates are different

Figure 6.2: Determination of the plates presenting suspicious distribution of values, based on the Kolmogorov-Smirnov procedure. The graph on the left hand side represents the first 20 plates of screen 12. Each bar indicates how many plates the corresponding reference plate differs from, whereas the one on the right hand side depicts the plate numbers each reference plate differs from (the figure represents the first 20 plates (taken as references) and the plates among the first 50 plates, they differ from.

- Intra-plate detection

Both Wilcoxon and Siegel-Tukey procedures are then computed so as to complete the detection of unusual values, by focusing on intra-plate variations. Two slightly different methods can be considered, one focusing on the detection of intra-plate variations in general, the other one on the detection of blocked jets only. It is to be noted that the former method is more likely to be used.

The first method consists of investigating differences in spread and location between all the rows of a plate, each row being respectively taken as a reference and compared to the distribution of the other rows of the plate. The process is then repeated for the other plates of the 'valid' set of plates. The results are then available per plate. On a small set of plates, such a procedure should not take too long.

The second method concentrates on the detection of a blocked jet only. The principle is to take the same row as a reference on each plate and to compare its distribution of values with the distribution of values of the other rows of the plate. If such a row, taken as a reference, appears to differ from many other rows on the plate and if such a phenomenon appears to be a trend all over the plates of the set, it might suggest that the corresponding tip is blocked.

Even if these procedures can be time consuming when applied on a large number of plates, it should not be worrying in the context of the Robolab procedure. First they are only applied to a subset of plates, and most of all, one should bear in mind that some plates are incubating while the intra-plate variation detection procedure is applied on another subset of plates, so that time is not such a crucial issue any more in practice.

- Corner and edge effects detection

The final stage could be an edge and corner effects detection, so as to see whether or not wells on the edges of the plates are significantly affected by their location on the plate. The procedure can be applied on a specific set of plates or on the whole screen, provided the assumptions of Normality and equality of variances are verified.

A summary of all the different problems encountered can be made as a report at the end of the screen.

All the procedures described in this thesis are implemented under Matlab. Due to the lack of time and the variety of problems that can occur in the HTS process some procedures such as the ones involved in intra-plate variation detection were built up so as to test their efficiency and accuracy on artificial invalid data. Thus, they need some more testing and adaptation before being used as real life procedures.

Appendix A

Reference screens

A.1 Screen 1

A.1.1 HTA plates

Screen number	1
Number of plates	115
Counter used	Packard 9912V Microplate Topcount
Normal plate	HTA: 1-115
Invalid plates	35 & 57 : Double ligand
	77 & 78 : No assay window
Date/s of Assay/s	1: 1-16 = 01/07/96
	2: 17-46 = 10/07/96
	3: 47-86 = 17/07/96
	4: 87-106 = 18/07/96
	5: 107-114 = 17/09/96
	6: 115 = 03/10/96

A.1.2 Control plates

Screen number	1
Number of plates	3
Counter used	Packard 9912V Microplate Topcount
Control plate	Totals & NSBs: 1-3
Date/s of Assay/s	1: 1-3 = 13/02/97

A.2 Screen 1b

A.2.1 HTA plates

Screen number	10
Number of plates	263
Counter used	Packard 9912V Microplate Topcount
Normal plates	cHTA: 231
	HTA: 32
Date/s of Assay/s	1: 1-10 = 30/07/96
	2: 11-16 = 14/08/96
	3: 17-26 = 15/08/96
	4: 27-56 = 21/08/96
	5: 57-93 = 28/08/96
	6: 94-95 = 1/11/96
	7: 96-155 = 5/11/96
	8: 156-209 = 6/11/96
	9: 210-224 = 11/11/96
	10: 240-263 = 21/11/96

A.2.2 Control plates

The plates of this screen have the same controls as Screen 1.

Screen number	1
Number of plates	3
Counter used	Packard 9912V Microplate Topcount
Control plate	Totals & NSBs: 1-3
Date/s of Assay/s	1: 1-3 = 13/02/97

A.3 Screen 2

A.3.1 HTA plates

Screen number	2
Number of plates	206
Counter used	Anthos HTII
Normal plate	cHTA: 1-206
Date/s of Assay/s	1: 1-40 = 28/11/96
	2: 41-80 = 06/11/96
	3: 81-91 = 13/11/96
	4: 92-131 = 07/11/96
	5: 132-171 = 12/11/96
	6: 172-206 = 13/11/96

A.3.2 Control plates

Screen number	2
Number of plates	3
Counter used	Anthos HTII
Control plate	Totals & NSBs: 1-3
	Standards Only: 5+6
	Max/Min/Standards: 4
Date/s of Assay/s	1: 1-3 = 19/11/96
	2: 4-6 = 14/05/97

A.4 Screen 9

A.4.1 HTA plates

Screen number	9
Number of plates	206
Counter used	Wallac LKB 1205-001 Beta Plate LSC
Normal plate	HTA: 1-206
Invalid plates	173-299 : A6 ALL ACTIVE (not relevant)
Date/s of Assay/s	1: 9/1/97 = 1-20
	2: 15/1/97 = 21-50
	3: 16/1/97 = 51-74
	4: 22/1/97 = 75-108
	5: 23/1/97 = 109-134
	6: 24/1/97 = 135-142
	7: 05/2/97 = 143-172
	8: 06/2/97 = 173-206

A.4.2 Control plates

Screen number	9
Number of plates	3
Counter used	Wallac LKB 1205-001 Beta Plate LSC
Control plate	Totals & NSBs: 3
Date/s of Assay/s	1: not provided = 1-3

A.5 Screen 12

A.5.1 HTA plates

Screen number	12
Number of plates	796
Counter used	Anthos
Normal plate	cHTA: 1-1602
Date/s of Assay/s	1: 15/04/96 = 1-160
	2: 16/04/96 = 161-320
	3: 19/04/96 = 321-476
	4: 29/04/96 = 477-516
	5: 01/05/96 = 676-676
	6: 22/05/96 = 677-684
	7: 30/05/96 = 685-692
	8: 18/06/96 = 693-696
	9: 01/07/96 = 697-844
	10: 09/07/96 = 845-1244
	11: 15/07/96 = 1245-1256
	12: 16/07/96 = 1257-1276
	13: 24/07/96 = 1277-1280
	14: 08/08/96 = 1281-1532
	15: 16/08/96 = 1533-1540
	16: 19/08/96 = 1541-1602

Blank plates (time zero reading):

1-40 81-120 161-200 241-250 261-280 311-349 389-427
467-486 507-546 587-626 667-668 671-672 675-678 683-684
687-723 761-797 835-864 895-924 955-974 995-1014 1095-1124
1155-1174 1195-1214 1235-1240 1247-1248 1251-1254 1259-1262 1267-1268
1271-1284 1299-1312 1327-1329 1333-1335 1339-1347 1357-1365 1375-1384
1395-1404 1415-1423 1433-1441 1451-1482 1515-1516 1519-1520 1523-1526
1531-1560 1591-1596

Note: Blank plates are a base line reading. They are read before the assay starts (or just after it starts) to give a reference reading, which is subtracted from the final reading to give a true response: this process is called blank subtraction.

A.5.2 Control plates

No control plates were provided for this screen.

Appendix B

Tests of Normality

This chapter displays the results obtained for **screen 12**, to test the Normality of the data. Different methods are carried out to test the Normality of the 3 distributions under investigation (corner, edge and middle populations).

First, some preliminary measurements evaluate the symmetry of the populations. In addition, some graphical representations give an indication of how close the distribution of the data is from a Normal distribution. Finally the Lilliefors goodness-of-fit test displays a numerical evaluation of the Normality of the distributions under investigation. The theory behind these different ways of estimating the Normality of the data is presented in Section 4.1.1. Therefore, we are just displaying the results obtained for screen 12 that justifiy why the ANOVA procedure was not applied to this screen.

B.1 Preliminary measurements

We briefly recall in this section that the Skewness and the Kurtosis give an indication of the behaviour of a population in comparison to the Normal distribution. Indeed, the Skewness is positive for a distribution with a right tail heavier than the left one. Similarly, the Kurtosis is positive when the tails of a distribution contain more mass than the Normal distribution. Both Skewness and Kurtosis are equal to zero for the Normal distribution.

B.1.1 Skewness

One particular feature is to be noted here: despite the fact that the populations are purged of extreme values, the edge distribution seems to show a heavy left tail. Besides, the results are apparently quite satisfying for both corner and middle populations. One should however

Populations	Corners	Edges	Middles
Whole populations	-0.6669	-7.4237	-0.2699
Populations purged of outliers	-0.2259	-3.3819	-0.1827

Table B.1: Measure of Skewness

bear in mind that all the populations display a large number of data points (more than 20000 data points for the edge population, more than 35000 for the middle population), therefore the values obtained for the Skewness are to be taken with caution. Indeed, as shown by the graphical representations (see Figure B.1), even if the results seem to be quite good for the middle population for instance, due to the large number of data points, a small departure from the ideal zero (obtained for a Gaussian distribution) can lead to quite a large departure from Normality.

B.1.2 Kurtosis

Populations	Corners	Edges	Middles
Whole populations	0.8019	0.9521	0.5885
Populations purged of outliers	-0.2234	-0.1117	-0.1497

Table B.2: Measure of Kurtosis

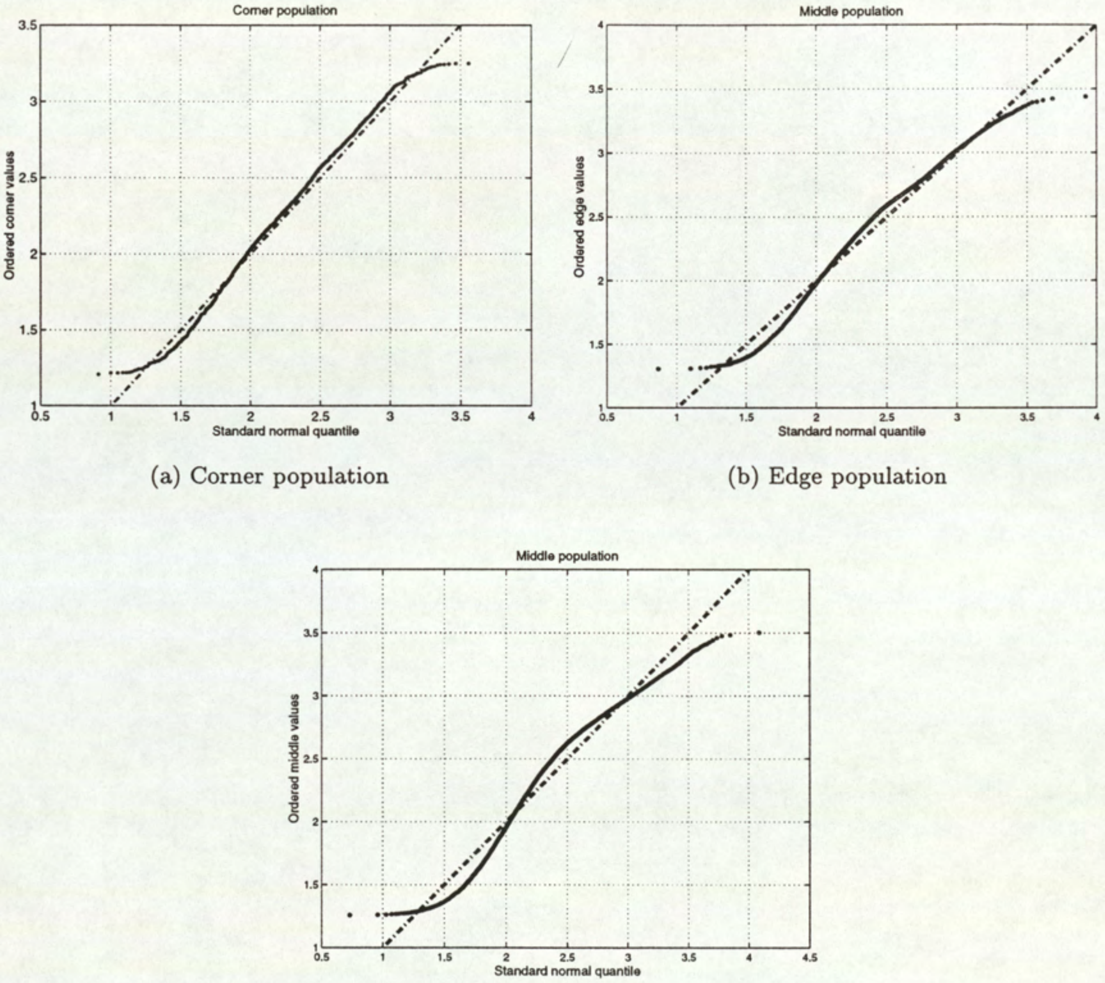
The Kurtosis displays rather good results close to the ideal zero of a Normal distribution. However, as mentioned in the previous section, the large number of data points might hide a significant departure from Normality.

B.2 Graphical representations

To illustrate our point, Figure B.1 represents the populations under investigation using the Normal quantile-quantile plot techniques and gives evidence of a departure from Normality (in particular, as far as edge and middle populations are concerned).

B.3 Goodness-of-fit procedures

The Lilliefors goodness-of-fit test whose results are displayed in Table B.3 confirms the general impression that we had after the preliminary measurements. It gives a numerical evaluation



(a) Corner population

(b) Edge population

(c) Middle population

Figure B.1: Normal quantile-quantile plots: corner, edge and middle populations

of how close (or actually how far) from a Normal distribution, the populations under investigation are.

Lilliefors test	Corners		Edges		Middles	
Critical values	5%	1%	5%	1%	5%	1%
	0.0172	0.0201	0.0072	0.0062	0.0054	0.0046
$D_{Screen\ 12}$	0.0270		0.0629		0.0880	
Conclusion	*Normality rejected*		*Normality rejected*		*Normality rejected*	

Table B.3: Lilliefors goodness-of-fit test for Normality applied to corner, edge and middle populations

B.4 Conclusion

All the methods carried out lead to the same conclusion: they all display some strong evidence for non-Normality.

Appendix C

Harmonic interpolation

C.1 A note on interpolation

The following figure gives part of the tabulation of the function $f(x)$ with a step h: This

$$\begin{array}{ccccc}
x_0 & f_0 & & & \\
 & & \delta'_{1/2} & & \\
x_1 & f_1 & & \delta''_1 & \\
 & & \delta'_{1,1/2} & & \delta'''_{1,1/2} \\
x_2 & f_2 & & \delta''_2 & \\
 & & \delta'_{2,1/2} & & \\
x_3 & f_3 & & &
\end{array}$$

means that $f_i = f(x_i)$ with $x_{i+1} = x_i + h$.

Interpolation of $f(x)$ at values other than those tabulated are given by the entries in the last three columns obtained by differences between the elements in the column just before. Thus,

$$\begin{aligned}
\delta'_{1/2} &= f_1 - f_0 \\
\delta''_1 &= \delta'_{1,1/2} - \delta'_{1/2} \\
\delta'''_{1,1/2} &= \delta''_2 - \delta''_1 \\
&\dots
\end{aligned}$$

Linear interpolation between x_1 and x_2 approximates therefore the function f by

$$f(x) = f_1 + p\delta'_{1,1/2}$$

with $p = (x - x_1)/h$. This is the most commonly used form of interpolation. But occasionally *harmonic* interpolation is advisable.

C.2 Harmonic interpolation

Harmonic interpolation is nothing but a variant of the general linear interpolation described previously, where the argument x is replaced by $1/x$.

Let's consider a practical example. In Section 4.1.2 for instance, one of the problems is to determine the critical value F of the F-distribution for $\alpha = 0.01$, $\nu_1 = 524$, $\nu_2 = 291$ (which corresponds to the edge-middle population for screen 1). Since both ν_1 and ν_2 are too large to be given by tables, a double harmonic interpolation has to be performed.

First, an harmonic interpolation is conducted in ν_2 (for ν_1 fixed, $\nu_1 = 24$):

ν_2	$1/\nu_2$	F[a]			
∞	0	1.791			
120	1/120	1.950	0.159		
60	2/120	2.115	0.165	0.006	
40	3/120	2.288	0.173	0.008	0.002

[a] obtained from tables for $\nu_1 = 24$

Then we compute an harmonic interpolation between $\nu_2 = \infty$ and $\nu_2 = 120$.

For $\nu_2 = 291$, $p = (1/\nu_2 - 0)/(2/120 - 1/120)$, *i.e.* $p = 0.4124$.

Then $F(0.01, 24, 291) = 1.791 + p(1.950 - 1.791)$, *i.e.* $F(0.01, 24, 291) = 1.8566$.

The same computation is conducted for $\nu_1 = \infty$, 12, 8. The results obtained are respectively, 1.1571, 2.2473, 2.5737. The second step consists of conducting an harmonic interpolation in ν_1 (for ν_2 fixed, $\nu_2 = 291$):

ν_1	$1/\nu 1$	F[a]			
∞	0	1.1571			
24	1/24	1.8566	0.6995		
12	2/24	2.2473	0.3907	-0.3088	
8	3/24	2.5737	0.3264	-0.0643	0.2445

[a] obtained from the previous harmonic interpolation: $\nu_2 = 291$

Then we compute an harmonic interpolation between $\nu_1 = \infty$ and $\nu_1 = 24$.

For $\nu_1 = 524$, $p = (1/\nu 1 - 0)/(2/24 - 1/24)$, *i.e.* $p = 0.0458$.

Then $F(0.01, 524, 291) = 1.1571 + p(1.8566 - 1.1571)$.

We finally obtain: $F(0.01, 524, 291) = 1.1891$.

Bibliography

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical data.* Wiley.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, pages 60–73. Oxford Univ. Press.

Bojanic, D., Keighley, W. W., Russell, M. J., and Wood, T. P. (1997). Factors for the successfull integration of assays, equipment, robotics and software for high throughput screening. In Devlin, editor, *High Throughput Screening - The Discovery of Bioactive Substances*, pages 493–508, New York. Marcel Dekker.

Devore, J. L. (1991a). *Probability and Statistics for Engineering and the Sciences.* Duxbury Press, third edition.

Devore, J. L. (1991b). *Probability and Statistics for Engineering and the Sciences*, chapter 10, pages 390–392. In (Devore, 1991a), third edition.

Fouquart, P. F. (1997). Quality control of high throughput screening. Master's thesis, Aston University.

Kanji, G. K. (1993). *100 Statistical Tests*, pages 29, 35–37. SAGE Publications.

Lindley, D. and Scott, W. (1984). *New Cambridge Elementary Statistical Tables.* Cambridge University Press.

Mason, R. L., Gunst, R. F., and Hess, J. L. (1989). *Statistical Design & Analysis of Experiments*, pages 31–40, 58–61, 78–86. Wiley.

Miller, R. G. and Ruppert, D. (1986). *Beyond ANOVA : the Basics of Applied Statistics*, pages 5–7, 71–76, 259–261. Wiley, New York.

Neave, H. (1978). *Statistics tables* for mathematicians, engineers, economists and the behavioural and management sciences. George Allen & Unwin Hyman Ltd.

Neave, H. and Worthington, P. (1988). *Distribution-Free tests*, pages 89–98, 100–103, 149–157. Unwin Hyman Ltd.