ENVIRONMENTAL REFERENCE MATERIALS METHODS AND CASE STUDIES

Karina Schramm-Nielsen

Department of Mathematical Modelling
Ph.D. Thesis No. 58
Lyngby 1998

IMM

ISSN

Printed by IMM/Technical University of Denmark
Bound by Hans Meyer

Preface

The work presented in this thesis has been carried out in the Department of Mathematical Modelling (IMM) at the Technical University of Denmark. The Institute for the Water Environment (VKI), Agern Allé 11, DK-2970 Hørsholm has kindly provided the laboratory facilities for the practical work. The thesis is part of the requirements for acquiring the degree of PhD in engineering.

In addition to the present thesis, two articles have been published on subjects which are not included in the thesis: (Schramm-Nielsen *et al.* 1998a) and (Schramm-Nielsen *et al.* 1998b).

VKI manufactures and conducts research of certified reference materials for environmental analyses. The present work encompasses experimental work in the laboratory and the application of statistical methodology - in theory and in practice, as well as the statistical data analysis and literature research. The experimental work of the thesis has been carried out in the Department of Chemistry in the Division of Environmental Chemistry at VKI.

The main focus of the study has been on the stability of reference materials based on natural liquid matrices[1]. Since data on stability studies

[1]Reference materials for (trace) metals and gasses will not be discussed in this thesis.

for this type of reference materials is rather scarce, data collection has been a fundamental requirement for the realization of this work. Until this point limited attention has been given to the statistical aspects of evaluating stability studies of reference materials. A major object of this thesis has been, on one hand, the investigation of the actual stability of certain types of materials, and on the other - of the statistical modelling of stability data.

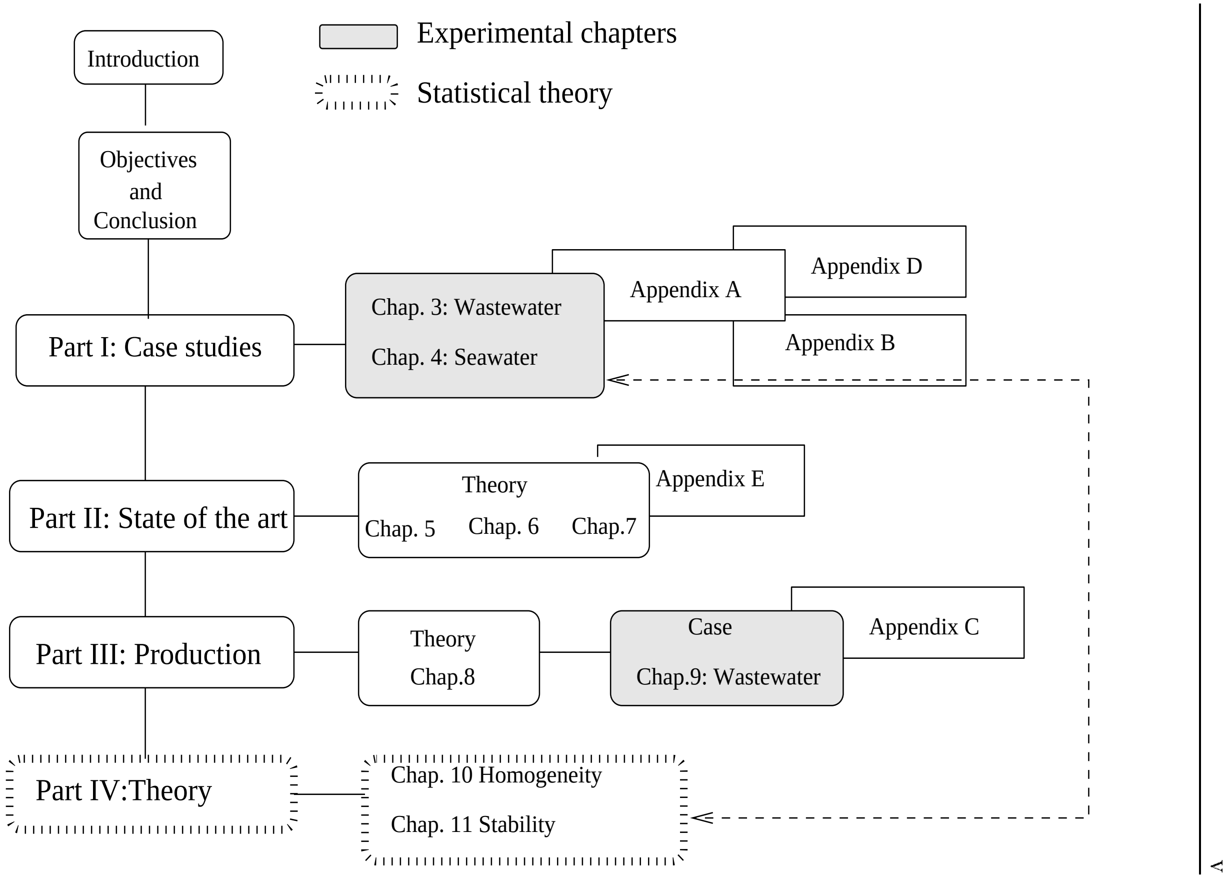
Writing a thesis on an interdisciplinary subject - in this case - about a subject overlapping the fields of applied statistics and analytical chemistry is a challenge. It has been necessary to assume that the reader is familiar with basic statistics and to some extent has a basic knowledge of analytical chemistry.

The thesis consists of four parts: I. Case studies, II. State of the Art, III. Production and IV. Theory. Some parts and chapters are likely to appeal to statisticians and others will be more interesting for chemists.

The structure of the thesis may appear somewhat unusual, since the first part presents the results of the case studies (the experimental work) without presenting the models and the theory in full detail. This approach has been chosen deliberately, in the hope that the reader will stay awake for the main results and that the chemists won't be overwhelmed by the statistics. The second and third part of the thesis *can* be skipped by statisticians, however the issues treated here are intended to explain the context of chemical reference materials and motivate the reader to understand the numerous aspects which the researcher must take into account.

Lyngby, December 15th 1998

Karina Schramm-Nielsen


Experimental chapters
Statistical theory
Introduction
Objectives and Conclusion
Part I: Case studies
Chap. 3: Wastewater
Chap. 4: Seawater
Appendix A
Appendix D
Appendix B
Part II: State of the art
Theory
Chap. 5
Chap. 6
Chap.7
Appendix E
Part III: Production
Theory
Chap.8
Case
Chap.9: Wastewater
Appendix C
Part IV:Theory
Chap. 10 Homogeneity
Chap. 11 Stability

Acknowledgements

I wish to express my gratitude to my supervisor Professor, PhD Henrik Spliid IMM, DTU for his never failing good spirits and sense of humour. He is most likely the only one who will ever get away with calling me an old barge, "Schramm - din gamle pram !" every day and still make me laugh after 4 years.

I also wish to thank Principal chemist at VKI, PhD Jill Merry for her firm belief in my skills and her sparkling mind always full of new ideas and projects. Jill Merry deserves admiration for her tremendous energy and tireless enthusiasm for the subject of reference materials.

A number of people not to be forgotten in this context are

- My office mate Jens Strodl Andersen whose great company, humour and positive attitude towards all challenges I will miss in the future. Quoting a famous last liner backwards I would say - I hope this was only the beginning of a beautiful friendship ! There may be other KK's but the YKK is a singular species !
- My fellow PhD students at IMM Helle M. Sommer, Judith L.Jacobsen, Trine Kvist and Helle Andersen who have contributed with heaps of SAS expertise, qualified comments, lots of moral support and good spirits.
- All of the technical staff at VKI for their kindness and help,

especially Lene Nicolajsen and Rita Muchitsch for their particular support with the (sometimes tricky) equipment in the general chemistry laboratory at VKI.

- Professor Elo H. Hansen, Institut for Kemi, DTU for his advice and guidance at the intiation of the project.
- Bente Nyeland, DMU for good company and support on the conferences abroad.
- Poul Thyregod for always keeping his door ajar.

Last but not least I want to thank my mother Professor at CBS, PhD Jette Schramm-Nielsen and my lion PhD Morten Rostgaard for their love and patience.

Symbolized abbreviations

It has been attempted throughout this thesis to apply abbreviations and symbols which (at least to the author) are as logic as possible in the relevant context. Abbreviations are explained when introduced, but for later reference, the reader may want to return to these pages.

Chapter 11 deals with both time and temperature variables. It is desirable to let the notation indicate the nature of a variable (systematic or random), however with the differentiation between time and temperature there was a conflict. The problem in that chapter has been solved by writing a temperature as t_i° where the top index indicates that t refers to a temperature in degrees C and the lower index is the level. Time is referred to as τ with a few exceptions where a lower index of t is used.

The use of symbols and abbreviations is not extensive and the following tables will hopefully be of assistance whenever needed.

Symbol	Description	Chapter
m	index counter	3
a	index counting main groups	10
a	symbol for a fixed main effect	4
b_i	index counting subgroups within main groups	10
b	intercept	11
b	symbol for a fixed main effect	4
c	symbol for a fixed main effect	4
f	number of degrees of freedom	10
K_p	critical value of Newman Keuls test	3
n	number of observations (total/in group) or number of replicates	
R_{t°	ratio of means at two temperatures	11
t	index for time	3,11
t_i°	temperature variable	11
	° index for temperature in degrees °C, an index number, "$_i$" below ° indicates a temperature level relative to a reference temperature	
p	number of responses/analytes measured on a sample	
p_i	cumulative probability of a standard normal distribution	3
s_r^2	repeatability variance	3
τ	time effect, fixed	3,4
$\mathfrak{T}$	time effect, random	3

Abbreviation	Description
NH_4-N	ammonium-nitrogen
ANOVA	analysis of variance
COD	chemical oxygen demand with potassium dichromate
CRM	certified reference material
CV	coefficient of variation
DS	Danish Standard
GF/A	name for filter/filtration through a 1.6 μm glass microfiber filter
LCL	lower control limit
MEWMA	multivariate exponentially weighted moving average
MS	mean square
NERI	National Environmental Research Institute
NO_{2+3}-N	sum of nitrite- and nitrate-nitrogen
OP	ortho-phosphate
RM	reference material
SAS®	Statistical Analysis System, software package
SI	système international, internationally accepted metric base units
TN	total nitrogen
TP	total phosporous
UCL	upper control limit
UF	ultra filtration (treatment of seawater)
UV	ultra violet (radiation treatment of seawater)
VKI	Insistute for the Water Environment

General glossary

accuracy/accurate Proximity to a true value. "Closeness of agreement between a test result and the accepted reference value". (ISO 5725 1994b)

analyte Chemical compound or element measured in a sample, see also measurand

calibration RM Reference material intended for calibration of an entire chemical analytical method or for the tuning of the measuring equipment.

BCR Community Bureau of Reference - the European Commission, now SMT, Standards Measurement & Testing

digestion - of a sample, treatment which in the analytical process with the purpose of concentrating the analyte or modifying its state so that it is ready for the next step.

GF/A filter 1.6 μm pore size glass microfiber filter used for filtration of natural waters to remove algae, particles etc. prior to chemical analysis.

matrix RM Reference material which matches a certain type of natural sample by resembling its composition, pH, physical state etc.

measurand The selected property of a constituent in a sample which is to be quantified by measurement

matrix In most contexts, matrix is used to describe the media or the state of a sample, e.g. liquid (water), powder, sediment etc.

primary standard “Standard that is designated or widely acknowledged as having the highest metrological qualities and whose value is accepted without reference to other standards of the same quantity” (VIM 1993)

purified water Water purified by either distilling or ion exchange followed by ultrafiltration. The water has very low levels of N (nitrogen) and P (phosphorus) components which makes it suitable for dilution of standards, high concentration level samples etc. without creating background concentration blank values of the components of interest. The water used in the VKI laboratory is ISO grade 2 water (ISO 3696 1987).

precision “Closeness of agreement between independent test results under stipulated conditions”(ISO 5725 1994b), e.g. repeatability conditions.

secondary standard “Standard whose value is assigned by comparison with a primary standard of the same quantity” (VIM 1993)

spike Spiking of a sample in analytical chemistry is the act of enlarging the signal of the analyte by intentionally “polluting” the sample. A known amount of the analyte in question is added to the sample, and the enlarged signal is measured.

traceability (to SI units). There are several definitions (Garner and Rasberry 1993). One is “property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties” (VIM 1993).

Summary

This thesis introduces the reader to the concept of chemical environmental reference materials and their role in traceability and chemical analyses of the environment.

A number of models and principles from the literature are described. Some suggestions are made as to how stability studies can be modelled when the length of the study is unknown.

Experimental data has been collected from two stability studies of aqueous matrices. The first study regards the stability of TN, NO_{2+3}-N and NH_4-N in autoclaved wastewater samples over a period of 22 months. Data was collected specifically for this study with two purposes: 1) to investigate the stability of selected analytes in the chosen matrices and 2) to explore the applicability of various statistical models for the description of stability studies. Three univariate and three multivariate stability models have been applied to these data sets. The methods have been evaluated with regard to their robustness towards variations in the chemical analytical method and with regard to the number of times a significant out of control situation is indicated.

The second study regards the stability of NH_4-N and total phosphorous in autoclaved seawater samples. This study lasted 22 months as well. The

samples were produced and stored according to a 2^3 factorial design. The influences of storage temperature, UV radiation and ultra-filtration on the stability of NH_4-N and total phosphorous have been investigated.

A Youden plot method is suggested for the graphical evaluation of certification data. The plots illustrate consistency between replicate measurements on samples from a batch of reference material, carried out in a number of laboratories according to a staggered nested design.

The development of a reference material is illustrated by a series of experiments with wastewater. The purpose was to improve ortho-phosphate (and total phosphorous) homogeneity. A procedure is suggested which includes freeze-drying and redissolving.

All calculations have been performed in SAS® primarily by means of elementary procedures, analyses of variance procedures, SAS Insight and SAS IML (Interactive Matrix Language).

Resumé (in Danish)

I denne afhandling gives en indføring i koncepter for kemiske referencematerialer til miljøanalyser og deres placering i relation til sporbarhed og miljøkemiske målinger.

En række modeller og principper for stabilitetsstudier fra litteraturen gennemgås sammen med forslag til, hvordan sådanne studier af referencematerialer til miljøanalyser kan modelleres, når længden af studiet ikke kendes på forhånd.

Der er indsamlet eksperimentelle data med to formål: 1) at undersøge stabiliteten af udvalgte analytter i den pågældene matrix samt 2) at undersøge egnetheden af udvalgte statistiske modeller til vurdering af stabiliteten. Data er indsamlet i to stabilitetsstudier af vandige matricer: afløbsspildevand og havvand. I det ene studium er stabiliteten af TN, NO_{2+3}-N og NH_4-N i autoklaverede spildevandsprøver fulgt gennem 22 måneder. Tre univariate og 3 multivariate stabilitetsmodeller er afprøvet på disse data. Metoderne vurderes med hensyn til deres robusthed overfor variation på den kemiske analytiske metode og hvor hyppigt, der er signifikans for, at materialets stabilitet er ude af kontrol.

I det andet studium fulgtes stabiliteten af NH_4-N og total phosphor i autoklaverede havvandsprøver gennem 22 måneder. Prøverne blev

fremstillet og opbevaret efter en 2^3 faktorforsøgsplan. I forsøget undersøgtes indflydelsen af opbevaringstemperatur, UV belysning og ultra-filtrering på NH_4-N og total phosphor.

Til konsistensvurdering af resultater fra certificering af referencematerialer foreslås en grafisk metode - et Youden plot. Metoden foreslås anvendt, når laboratorier udfører gentagne målinger på prøver fra en batch af referencemateriale i et staggered nested design.

Som et eksempel på udvikling af et referencemateriale beskrives en række forsøg udført med det formål at forbedre homogeniteten af phosphor-komponenter i et referencemateriale baseret på naturligt spildevand. Der gives et forslag til en procedure som inkluderer frysetørring og genopløsning.

Alle beregninger er udført i SAS® ved anvendelse af elementære procedurer, variansanalyse og det interaktive modul SAS Insight. De multivariate metoder er programmeret i SAS IML (Interactive Matrix Language).

Contents

Chapter 1

Introduction

The topic of this PhD thesis is a continuation of my master's thesis entitled Quality Assurance of Chemical Reference Materials (Schramm-Nielsen 1995)[1]. The present work is a chemometric project where the main object has been two case studies of stability of natural matrices.

Throughout the thesis the abbreviations CRM and RM are used interchangeably for (certified) reference material. The entire work regards certified RMs - in the sense that a CRM is the outcome. During preparational and experimental steps the material may be regarded as a candidate (C)RM until certified. It should be stressed that in this context the abbreviation RM implies a certified material.

reference material (RM) Material or substance one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials. (ISO Guide 30 1992)

certified reference material (CRM) Reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes its traceability to an accurate realization of the unit in which the property values are expressed and for which

[1] In Danish

each certified value is accompanied by an uncertainty at a stated level of confidence. (ISO Guide 30 1992)

1.1 Traceability of measurements

Traceability and the use of certified reference materials (CRM) are closely linked in a world of increasing demands to provide measurements of internationally acceptable quality. The term "quality of a measurement" in the present context covers the characteristics accuracy and precision - as defined in (ISO 5725 1994b). The relations are illustrated in figure 1.1. The demand for a common platform of quality arises where either

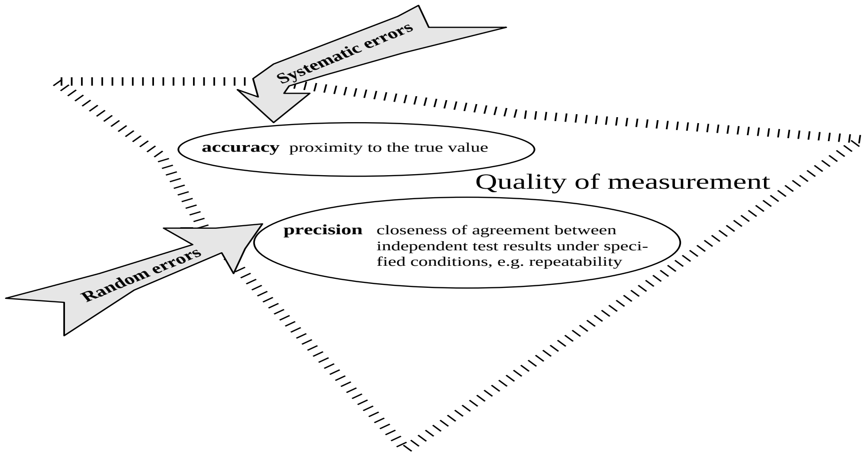


Figure 1.1: The concept of quality in measurements includes accuracy and precision which are influenced by systematic and random errors respectively.

international, national or local comparability of measurements is to be established either because it is dictated by legislation (for food and drug industry, etc.) or required for monitoring programs in national or international frameworks. Standard procedures such as national and international standard methods and the use of certified reference materials (CRM) are tools in achieving this. Comparability of measurement results is a prerequisite for discussing traceability in chemistry.

Traceability to SI units[2] is often discussed in relation to reference material (RM) - in international guidelines (ISO Guide 35 1989) as well as in literature. The question has been raised whether traceability to SI units is relevant for chemical measurements, for instance by (De Bièvre and Lamberty 1995).

A classic example of a physical RM is the meter rod. A measurement of length, say the height of this book could be carried out with a ruler. The reading of the ruler would include uncertainty arising from the precision with which the ruler could be read. The precision of the ruler could be defined by comparing it to the previously official definition of length: the meter rod[3].

The principle of tracing a chemical measurement to the basic SI unit, the mole, is the same.The fundamental principle in establishing traceability is an unbroken chain of comparisons from any metrological level to the one below. A measurement at any level of the traceability hierarchy has a certain uncertainty. The uncertainty is decreasing upwards in the pyramid. The physical traceability hierarchy can be illustrated for instance as in figure 1.2. The second level from the top of the pyramid in figure 1.2 is a national or an international institute which performs measurements with small uncertainties illustrated by the peaked distribution on the left. The end-user is at the bottom. In the physical traceability chain the end user can link physical measurements to the hierarchy of traceability by means of e.g. calibrating scales using weights issued from a national metrological laboratory. The example of a physical

[2]Internationally accepted metric base units.

[3]Presently, the meter definition is based on the wavelength of stabilized lasers which are used directly for measurements of length in the laboratory by means of interferometry (Gærnæs *et al.* 1996). The historical definition is used here only for the purpose of illustration. It is my experience that the comprehension of the concept of chemical reference materials is facilitated by means of this example, when the audience is unfamiliar with the subject

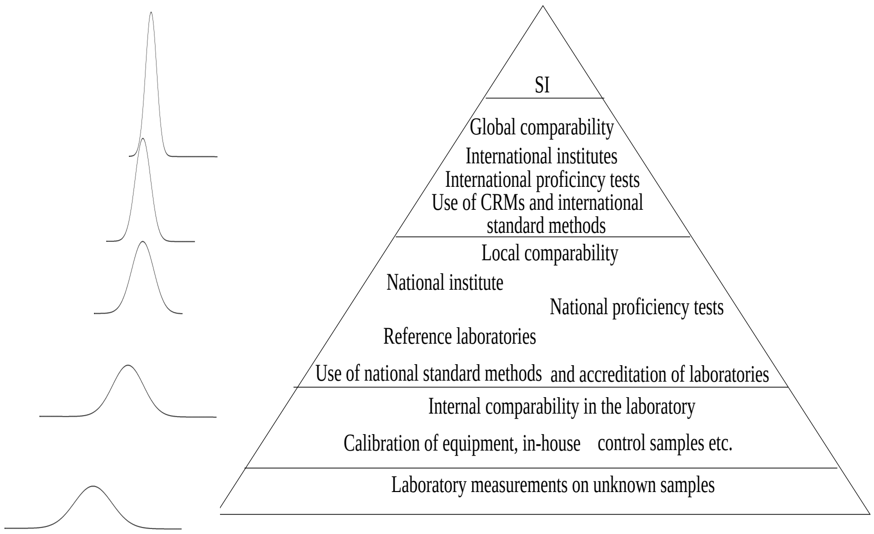


Figure 1.2: Example of traceability hierarchy. Relations between institutions. Concepts and tools which can be used to achieve a certain level.

measurement procedure is an illustration of a calibration hierarchy. The standards used in a calibration hierarchy all have uncertainties. Higher ranking standards have smaller uncertainties. Therefore each step upwards in the pyramid represents a decrease in the uncertainty of measurements (EAL-G12 1995) and at the same time an increase in accuracy.

Some of the tools for achieving traceability in chemical measurements are in-house control samples, participation in proficiency tests, use of CRMs etc. depending on the demand for comparability of the measurement results, as shown in figure 1.2. The employment of a matrix RM certified with a reliable method is a simple method of ensuring traceability in environmental analysis and at the same time verifying the entire analytical procedure (Quevauviller *et al.* 1995). The matrix RM is included in the analytical procedure and treated like a real sample - that is - the matrix

RM undergoes the same digestion[4], extraction steps etc. relevant to the procedure. A calibration RM may be included in the measurement step as suggested by (King 1997).

Every time a sample is modified, be it chemically or physically, the linkage to SI units is disturbed. These disturbances are inevitable when making a chemical analysis on a sample. In the course of a measurement process there are always doubts whether the sample was completely dissolved during pretreatment, whether contamination was avoided or whether the compound of interest was not more or less lost or modified. During the following steps of a measurement process preconcentration, precipitation steps or other modifications of a sample solution increase the possibility of contamination or loss of analyte. Every step of a measurement procedure contributes to the decrease of accuracy.

Because of the above mentioned problems, usually the uncertainty of the whole measurement process is much larger than errors due to the lack of traceable calibrants in the measurement process.

Since, a RM has a certified reference value with a stated accuracy it becomes a tool in the validation of the entire analytical method, particularly when it is combined with the use of e.g. spiked samples, blanks, assessment of selectivity, limit of detection, bias, precision etc. (Acc. and QA in anal. chem. 1994).

De Bièvre et al. point out that the purpose of establishing traceability for chemical measurements is not really a question of actually realizing the comparison of an amount of material to a mole realized as an amount of substance (De Bièvre and Lamberty 1995). A more useful object is to quantify the uncertainty related to a measurement. Instead of focusing on establishing traceability Cameron (Cameron 1975) points out the importance of setting performance requirements on measurements.

Valcárcel and Ríos present a perspective on traceability in analytical chemistry which does not involve reference laboratories at any stage (Valcárcel and Ríos 1995). The authors point out that the traceability

[4] Processing - often involving heating which prepares the analyte for the next step in the analysis.

chain of stoichiometric methods are usually longer than those of e.g. gravimetric methods. With regard to instrumental methods (relative methods) which are frequently used in environmental chemistry, the standardization is based on the establishment through standards of a relationship between the signal and the analyte concentration. The standardization of signal-concentration is linked directly to atomic masses and on to base (SI) standards such as kg or mole through calculation.

Alexandrov and King and stress the point of view that traceability to SI *is* highly relevant for measurements of concentration (Alexandrov 1996) and for RMs (King 1997).

King underlines the unit of the mole as essential to provide a universal common reference point for chemical measurements. Alexandrov represents the point of view that a common primary standard[5] for concentration is impossible to achieve, in the sense that concentration cannot stand alone, it must always be related to a compound: moles of Ca^{2+}, moles of SO_4^{2-} etc. Concentration is always a specific quantity of a particular compound, therefore there are limits to the transferability of the size of the concentration. Instead of the traditional hierarchy, as it is used in the physical measuring field to establish retraceability of measurements, Alexandrov proposes a principle of local traceability of concentration measurements in chemical analysis. For industrial metal and material analysis, and for clinical and biological material analysis, the approach links measurements via local working standard samples and local primary standard samples back to methods of analysis such as gravimetry, volumetry, coulometry etc. and on to SI-base units. For environmental analysis he proposes the same with the exception of the link to primary methods. Instead the chain ends at accuracy methods for the measurement of impurity-sensitive properties.

Quality of measurement results ought to be an element of competition in today's market of environmental analyses. The consumer is conscious of prices and delivery times for required analyses but he should also take into consideration the quality of the delivered results. For an analytical

[5] See vocabulary pp. xiv

laboratory the consequence of such a demand would be a need for documentation and verifiable integration of quality assurance measurements in all steps of the analytical procedure (Acc. and QA in anal. chem. 1994). The use of CRMs as a routine step in all analyses provides this documentation. Systematic errors can be discovered at an early stage, in-house standard calibrants or reference materials can be checked by including the CRM in the analyses. An important aspect to consider in this context is the risk involved with decision making on the basis of data of insufficient quality. Poor quality of measurements could lead to overestimation of the costs involved in establishing protective measures against a potential threat on the environment. At the other extreme the risk could be an environmental disaster as a result of underestimation of a potential risk, an underestimation which was due to measurements of insufficient quality.

The CRMs provide information about comparability between in-house uncertainty and the established uncertainty of the CRM. In other words there is a great need for CRMs in all analytical fields be it food, pharmaceutical or environmental analyses in order to promote better data quality. It should be stressed strongly that an established and well functioning quality assurance system is an absolute pre-requisite for the use of CRMs. Without a QA system the use of RMs is worthless. Uncertainty and traceability are strongly related in a system of traceability.

It can be concluded that there is still a long way to the goal, whether or not one considers full traceability to SI base units to be possible and applicable in chemical measurements. Nevertheless, traceability in chemical measurements is not as early in its childhood as may be expected at a first glance. (King 1997) points out that chemical CRMs are key links in the traceability chain but there is still a gap when it comes to establishing links between countries. There is no doubt that RMs are extremely important at present in attaining as much comparability between measurements as possible, but they will also play an important role in the future for taking the art further than stating measurements traceable "to RM number *nn*".

Lately the discussion on traceability of reference materials among European analytical chemists has turned towards the issue of uncertainty - which is inseparable from the question of of traceability. The Second Eurachem Workshop held in Berlin September 29-30th 1997 on “Measurement Uncertainty in Chemical Analysis” demonstrated that the European analytical chemists are prepared to introduce the concepts of the Guide to Expression of Uncertainty in Measurement (also known as GUM). There was however some concern about how to cope with the realization of uncertainty budgets which are the essence of the guide. Using GUM means the former indication of confidence intervals for certified reference values of CRMs should be replaced by uncertainty statements. The uncertainty statements include type A contributions based on statistical calculations and type B contributions which are to be estimated on the basis of experience, a priori knowledge or by other means. The second type of contributions are sensitive to the choices made by the person calculating the uncertainty budget. Realistically, if uncertainty budgets are going to be a standard requirement, it would be necessary to ensure comparability of these budgets. An obvious beginning would be standardized procedures for the calculation of uncertainty budgets for existing standard methods, national as well as international.

Chapter 2

Objectives and Conclusion

This work considers stability and analysis of stability for candidate RMs to be used for environmental chemical analysis. The main objectives have been

1. To investigate possible statistical methods for analysing stability data of RMs. This includes a survey of methods used in the fields up until this point and an evaluation of the application of various models on data of stable as well as unstable measurands.

2. To investigate the stability of TN, NO_{2+3}-N and NH_4-N in a wastewater candidate RM.

3. To find a method of improving NH_4-N stability in a seawater candidate RM.

4. To investigate possibilities of improving the homogeneity of ortho-phosphate and total phosphate in a wastewater candidate RM.

Conclusions

The area of chemical RMs for environmental analyses is very broad. The perfect manufacturer of RMs would possess an immense knowledge of material chemistry, analytical and environmental chemistry, be a good statistician, an excellent sales person, well informed on environmental legislation, well informed about the trends in environmental politics and on the market of environmental analyses and finally a fortune teller to some extend. This being said, it is with humility and conscience of my limited knowledge in the field that I sign this work. Naturally, a person as the above described does not exist but the constellation of competent employees who posses these qualifications all together would bring the undertaking of a RM development and manufacturing very far.

I have refrained from commenting on the ongoing discussions about traceability of RMs, but instead addressed the limited subjects of statistics in homogeneity and stability testing especially with regard to the case studies performed in this work.

2.1 Case study of wastewater

A case study of a candidate RM based on wastewater from Usserød Wastewater Treatment Plant comprised analyses of TN (total nitrogen), NO_{2+3}-N (sum of nitrite- and nitrate-nitrogen) and NH_4-N (ammonium-nitrogen) over a time interval with the purpose of surveying the stability of these analytes at 4, 20 and 37°C. The study showed that the matrix was stable with regard to TN and NO_{2+3}-N over a period of 22 months. The prospects of making a certified reference material for these two analytes are thus very good. The material was not stable with regard to NH_4-N. A change in concentration was detectable after three months of storage.

2.2 Case study of seawater

A case study of a candidate RM based on seawater of 2.2% salinity was carried out with the purpose of investigating the possibility of improving NH_4-N stability. UV radiation and removal of dissolved organic compounds by ultra-filtration were selected as treatments which might improve material stability. A 2^3 factorial design was set up for the preparation and the stability study of samples when stored at 4 and 37°C was followed over 22 months. Analyses of NH_4-N and TP were performed at intervals. The stability of NH_4-N was not improved by neither UV radiation nor ultra-filtration. The storage temperature was found to be the main contributor to the instability of NH_4-N in seawater whereas the influence of UV radiation was less significant. Contrary to expectations, the ultra-filtration of the sea water seemed to increase the NH_4-N concentration slightly.

2.3 Models for stability studies

Stability studies of RMs can be one of three types:

- ☞ Type I: short term stability study as part of a feasibility study for a RM production
- ☞ Type II: time delimited stability study of longer durability than the above and carried out prior to sale
- ☞ Type III: continuous long term stability monitoring after sales have started with no fixed closure date

At present, primarily two univariate principles have been presented in RM literature for the time delimited studies of types I and II: the first method is a storage procedure according to a grid principle where samples are transferred between varying storage conditions. This approach has been developed into a method of isochronous analysis of all samples at the end of the study (Lamberty *et al.* 1998). The second principle is that of taking the ratio of analytical results at one storage temperature - relative to the results obtained on samples stored at a reference temperature. This method can be applied to type III as well as type I and II studies.

An important scope of the present study was to investigate models and principles applicable to type III studies. Three univariate (reference temperature ratio plot, ANOVA and $\bar{X}$-chart) and three multivariate methods (Hotelling's T^2, MEWMA and M-chart) have been applied to the wastewater stability study data of TN, NO_{2+3}-N and NH_4-N. The procedures have been evaluated with regard to their ability to deal with variation related to the chemical analytical method (robustness).

The univariate methods were applied to one analyte at a time and the multivariate methods were applied on all three analytes simultaneously but for each storage temperature separately.

The reference temperature ratio method was insensitive to variation of the analytical method for the analysis of NO_{2+3}-N (DS 223). The ratio plots of NH_4-N data indicated increasing concentrations with time and temperature. This was in agreement with the analysis of variance on the NH_4-N data. The storage temperature was found to be significant for the ammonium concentration after three months of storage. The $\bar{X}$-chart applied separately for each storage temperature did not indicate an out of control situation for NH_4-N 37°C data until after 10 months. Furthermore, the $\bar{X}$-chart seemed rather sensitive towards variation of the analytical method, whereas the reference temperature ratio method was robust towards this type of variation in the NO_{2+3}-N data, and out-of-control situations occurred numerous times on the $\bar{X}$-chart.

Overall, the reference temperature ratio method agreed with the results of the ANOVA on all three data sets. The poorer performance of the $\bar{X}$-chart is probably due to the fact that the control limits were estimated from the data to be controlled.

Hotelling's T^2 seemed sensitive to variation of the analytical methods and the results were ambiguous for the lowest and the intermediate storage temperatures, but significant for the highest temperature of 37°C. The

M-chart appeared more robust towards measurement method related variation, but the out of control signal appeared considerably later, compared to the ANOVA.

The MEWMA was applied for values of λ (the weight given to the present observation) between 0.1 and 0.5 and the test quantities were significant for all storage temperatures from the first time point and onwards.

The primary limitation of the multivariate methods lies in their inability to identify the components responsible for the detected instability. The application of the multivariate methods is in the present case based on control limits and variances estimated from the data to be controlled which makes it difficult to assess their performance in a more realistic situation. On the other hand, one certain advantage is that, once the necessary information about the variance structure etc. has been gathered, it is relatively easy to process a large number of responses in stability studies of multicomponent RMs.

On the basis of the results in this study the MEWMA method seems to be the best candidate for a multivariate control method of RMs. To document this further it would be recommendable to test the method further on more elaborate data involving e.g. five responses or more and involving several types of analytical methods - some of which could be more or less free of analytical method variation.

2.4 Homogeneity tests

A graphical Youden type method has been suggested and described as a supplement to evaluate the homogeneity of RMs on the basis of certification data. The method has been applied to data from the certification of three types of RMs. The method focuses on inter- and intra-sample consistency of results and identifies observations which are outliers in traditional tests such as Cochran and Grubbs' tests.

2.5 Case study RM development

With the purpose of investigating the possibilities of improving phosphate homogeneity in wastewater RM, a series of freeze-drying experiments has been performed. The residual variance was considerably reduced as improved steps were introduced in the preparation. The residual variance of the ANOVA on the last experiments was comparable to the variation of the analytical method according to VKI control charts. This meant that the denominator of the F-test of variation between samples was small indicating that between-sample variation was not larger than an estimate from the control charts of synthetic samples would have shown.

The experiments demonstrated that homogeneity improvements were achieved, but further experimentation should be done in order to confirm these results.

Chapter 3

Stability of nitrogen compounds in wastewater reference material

This chapter presents the results of the experimental work and a stability study carried out on a wastewater RM.

3.1 Purpose

The purpose of the study on outlet wastewater was to collect data for modelling of stability and to test the stability of three chosen measurands: TN, NO_{2+3}-N and NH_4-N. TN and NO_{2+3}-N were expected to be stable.

A RM for TN in outlet wastewater is especially relevant because it is a legislative compound with threshold values dictated by the Danish Environmental Protection Agency (Nr. 637 1997). Furthermore, the analysis of TN is method dependent. The other compounds are also routine analytes used in environmental monitoring of wastewater.

3.2 Data collection

The wastewater material analysed in this project was prepared from wastewater taken from the outlet of Usserød Wastewater Treatment Plant in July 1996. The water was GF/A filtered and filled into 250 ml. infusion bottles. The entire lot of bottles was autoclaved within 5 hours after the filling at standard conditions 121°C for 20 minutes.

The samples were analysed for TN, NO_{2+3}-N, and NH_4-N immediately after production. The remaining bottles were divided in three lots and stored under the following conditions:

① 4°C

② room temperature 20°C

③ 37°C

Details of the preparation and the methods of analyses are given in appendix A.

During the subsequent 22 months, analyses of TN, NO_{2+3}-N, NH_4-N were performed eight times. At each time point 2 bottles were selected at random from each storage condition and analysed with two replicates per sample as shown in table 3.1. Randomization has been applied in all series of analyses in order to prevent systematic errors. The indices are given as in eq.(3.1).

	Temperature, t_i°		
Time, τ_j	t_1°	t_2°	t_3°
τ_1	sample 1: y_{1111},y_{1112} sample 2: y_{1121},y_{1122}		
. . .			
τ_8			sample 1:y_{8311},y_{8312} sample 2:y_{8321},y_{8322}

Table 3.1: Data structure.

The purpose of the study was to reveal temperature related instability in the material and to apply statistical methods capable of indicating possible changes. The temperatures were selected as representative of precautious (4°C), ordinary (room temperature) and worst case (37°C) storage conditions (conditions during shipping were not considered). Instability is expected to be temperature dependent in the sense that samples stored at high temperature will be prone to accelerated decay. Samples stored at low temperature are expected to develop in the same manner only at a lower rate.

3.3 Results

The results of the stability study on N compounds are presented in figures 3.1, 3.2 and 3.3. Results are given in mg/l for TN and NO_{2+3}-N, and in μg/l for NH_4-N.

The sample means are plotted against a numerical time variable which measures the time in months from preparation to analysis.

For the purpose of illustration, the mean values of the measurands at time τ_0 i.e. right after preparation have been used as reference values in plots and calculations (indicated by the symbol $\bar{X}(\tau_0)$). The reference lines in the plots (figures 3.1, 3.2 and 3.3) are the mean values for the measurands at τ_0.

All of the following statistical analyses and graphical methods which include a temperature effect etc. are based on the data set where $\tau > 0$ since the data obtained at the starting point had not yet been exposed to varying storage conditions.

It was expected from previous experiences at VKI with other natural water matrices that NH_4-N levels might increase with time. The plot of the raw data in figure 3.3 gives a first indication that this is also the case with the wastewater samples in this experiment.

In addition to the plot of sample means shown in this chapter, plots of means calculated by temperature and time are shown in appendix A.

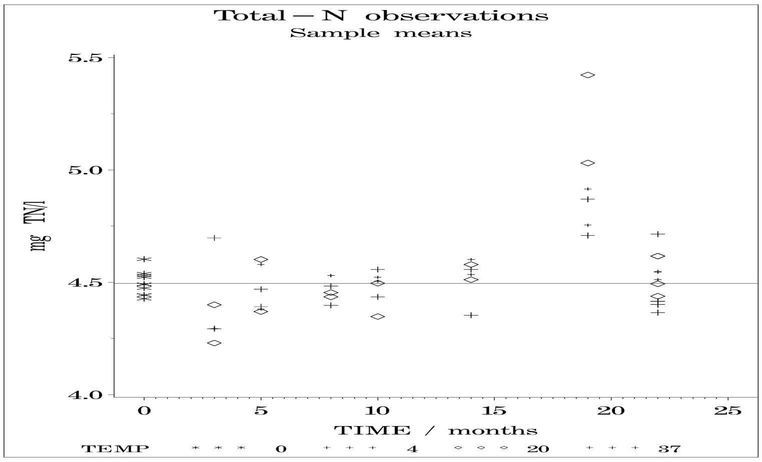


Figure 3.1: TN raw data: sample means. Reference line at $\bar{X}(\tau_0)$.

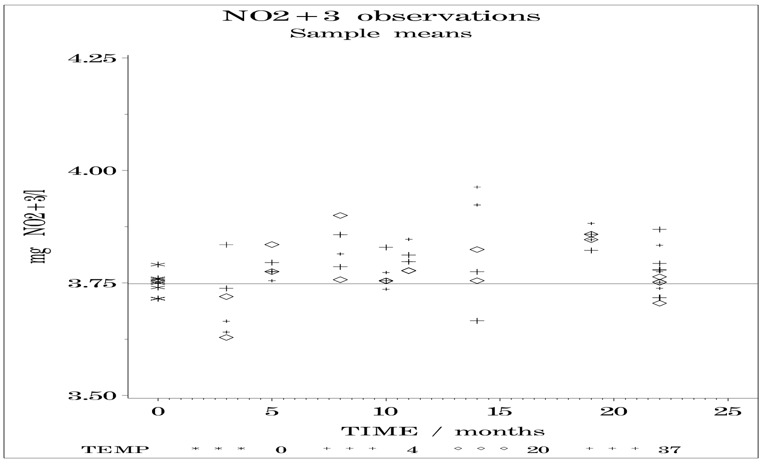


Figure 3.2: NO_{2+3}-N raw data: Sample means. Reference line at $\bar{X}(\tau_0)$

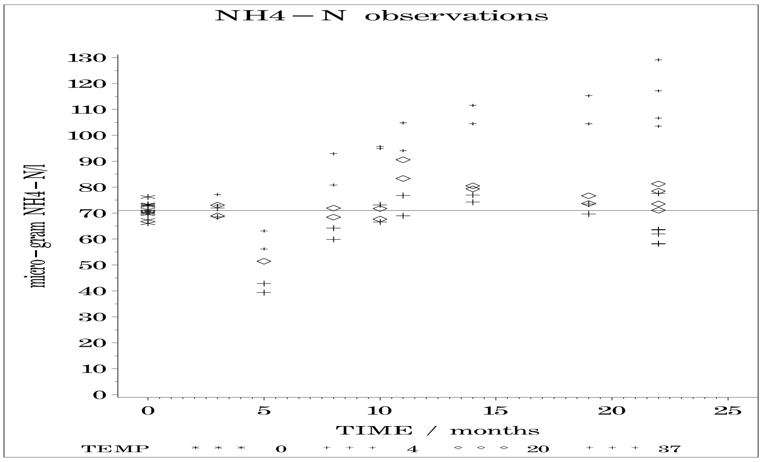


Figure 3.3: NH_4-N raw data: Sample means. Reference line at $\bar{X}(\tau_0)$.

3.4 Statistical analyses

3.4.1 Model adequacy checking

The raw data for each measurand have been evaluated by means of quantile-quantile plots (QQ-plots) of the difference between 2 replicate analyses on one sample and of the difference between two bottle means from one storage temperature measured at time point τ_i (see table 3.1). The difference between replicates analyses on one bottle is:

$$Y_{i,j,k,1} - Y_{i,j,k,2} \tag{3.1}$$

where $i = time$, $j = temperature$, $k = sample$ and $r = 1, 2$ is the replicate. The difference between bottle means within the combined conditions of storage temperature and time is:

$$\bar{Y}_{i,j,1} - \bar{Y}_{i,j,2} \tag{3.2}$$

where $i = time$, $j = temperature$, $k = 1, 2$ is the bottle number.

For each of the measurands, the normality assumptions are verified for both sample replicates and difference between bottle means by use of the Anderson-Darling test (Encyclopedia of Statistical Sciences 1982), (Stephens 1974). The QQ-plots are shown in appendix A, section A.2.1.

The Anderson-Darling test[1] is an empirical distribution function (EDF) statistics which measures the discrepancy between the empirical distribution function and a theoretical distribution function (the normal distribution in this case). The null hypothesis that X is normal distributed is rejected for large values of the test statistic.

The test for the normal distribution with two estimated parameters is based on

$$w_i = \frac{x_{(i)} - \bar{x}}{s} \tag{3.3}$$

where $\bar{x}$ and s are the usual estimates of population mean and variance and $x_{(i)}$ is the i'th ordered observation in the range $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. The cumulative probability of a standard normal distribution for the value w_i is labelled p_i and entered (in ascending order) into the formula for the test statistic:

$$A^2 = -n - \{\sum_{i=1}^{n}(2i-1)[\ln(\mathrm{p_i}) + \ln(1 - \mathrm{p_{n+1-i}})]\}/\mathrm{n} \tag{3.4}$$

The null hypothesis of normal distribution characteristics of the data is accepted for the intra- and inter-sample differences (eq.3.1,3.2) at $\alpha = 0.05$.

3.5 Univariate methods

3.5.1 Reference temperature ratio plots

A common problem in long term stability studies of RMs is that maintaining tight control of the analytical method can be difficult. The

[1] The Anderson-Darling test is available in the SAS® procedure "PROC CAPABILITY".

applied analytical methods used for the analysis of TN, NO_{2+3}-N and NH_4-N (see appendix A) in this study are examples of analytical methods with measurable day to day variance. Straight forward illustration of the results by plotting data versus time can demonstrate changes which are in fact caused by variances in the analytical method and not related to changes in the RM. Analytical variations can be corrected for, by expressing the mean values of samples stored at each chosen temperature level relative to the mean value of samples stored at a reference temperature, when all samples are measured at the same time. The method has been presented by (Faure and Wagstaffe 1993). Figures 3.4-3.6 show the reference temperature ratio plots for TN, NO_{2+3}-N and NH_4-N. At each time point the mean values of samples stored at 20°C and 37°C are shown as ratios of the 4°C mean value at the same time point. The expression is

$$R_{t^\circ} = \frac{\bar{X}_{t_i^\circ}}{\bar{X}_{t_{ref}^\circ}}$$

t° is an index which describes a temperature variable at level i or at the reference temperature level. The reference temperature ratio method is discussed in further detail in chapter 11 and an illustration of the principle is shown in figure 11.2. Table 3.2 shows the uncertainties for the ratio method, U_t, calculated according to eq.(11.5).

Measurand	Time	U_t $[\mu g/L]$ Temperature 20°C	37°C
TN	τ_1	$7.83 \cdot 10^{-3}$	$5.79 \cdot 10^{-3}$
	τ_2	$1.03 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$
	τ_3	$3.83 \cdot 10^{-3}$	$3.21 \cdot 10^{-3}$
	τ_4	$3.30 \cdot 10^{-3}$	$2.76 \cdot 10^{-3}$
	τ_6	$2.89 \cdot 10^{-3}$	$3.23 \cdot 10^{-3}$
	τ_7	$6.27 \cdot 10^{-3}$	$9.72 \cdot 10^{-3}$
	τ_8	$4.90 \cdot 10^{-3}$	$4.475 \cdot 10^{-3}$
NO_{2+3}-N	τ_1	$7.00 \cdot 10^{-3}$	$6.11 \cdot 10^{-3}$
	τ_2	$1.67 \cdot 10^{-3}$	$7.48 \cdot 10^{-3}$
	τ_3	$3.88 \cdot 10^{-3}$	$4 \cdot 10^{-3}$
	τ_4	$2.10 \cdot 10^{-3}$	$2.10 \cdot 10^{-3}$
	τ_5	$9.35 \cdot 10^{-3}$	$1.37 \cdot 10^{-3}$
	τ_6	$3.95 \cdot 10^{-3}$	$8.88 \cdot 10^{-3}$
	τ_7	$9.41 \cdot 10^{-4}$	$9.92 \cdot 10^{-4}$
	τ_8	$2.73 \cdot 10^{-3}$	$2.68 \cdot 10^{-3}$
NH_4-N	τ_1	$4.39 \cdot 10^{-2}$	$4.56 \cdot 10^{-3}$
	τ_2	$6.04 \cdot 10^{-2}$	0.107
	τ_3	$5.6 \cdot 10^{-2}$	0.127
	τ_4	$6.64 \cdot 10^{-2}$	$7.95 \cdot 10^{-2}$
	τ_5	$9.79 \cdot 10^{-2}$	0.128
	τ_6	$2.63 \cdot 10^{-2}$	$6.31 \cdot 10^{-2}$
	τ_7	$4.08 \cdot 10^{-2}$	$9.95 \cdot 10^{-2}$
	τ_8	0.15	0.27

Table 3.2: Uncertainties for the reference temperature ratio method, eq.(11.5)

Figures 3.4 and 3.5 do not show systematic changes for TN and NO_{2+3}-N, whereas figure 3.6 shows increasing ratios for NH_4-N. The non systematic variations for TN and NO_{2+3}-N are at this point assumed to be related to variations in the analytical method.

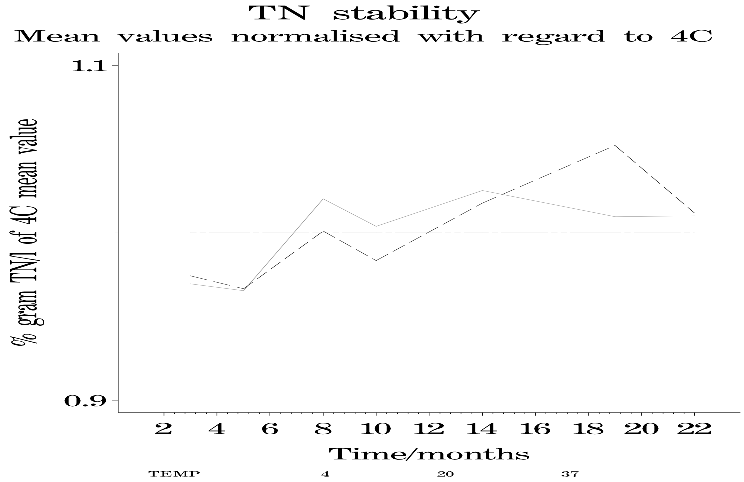


Figure 3.4: TN: mean temperature values at each time point $\bar{X}_{\tau,t_i^\circ}$ normalised to $\bar{X}_{\tau,t_4^\circ}$.

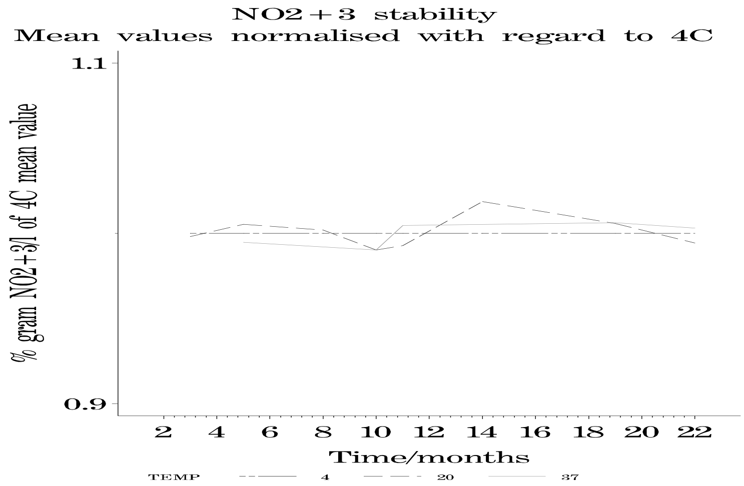


Figure 3.5: NO_{2+3}-N: mean temperature values at each time point $\bar{X}_{\tau,t_i^\circ}$ normalised to $\bar{X}_{\tau,t_4^\circ}$.

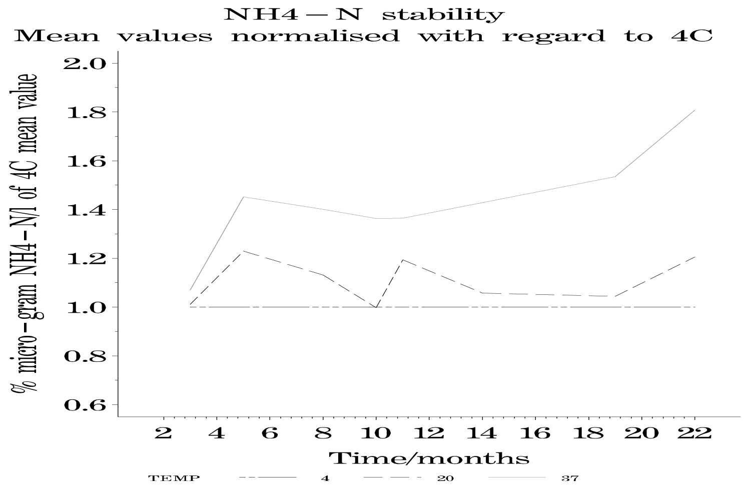


Figure 3.6: NH_4-N: mean temperature values at each time point $\bar{X}_{\tau,t_i^\circ}$ normalised to $\bar{X}_{\tau,t_4^\circ}$.

3.5.2 ANOVA

In addition to the temperature ratio plots, analysis of variance (ANOVA) has been applied according to the following general linear model where capitals represent random effects and small letters represent fixed effects.

$$Y_{ijkl} = \mu + t^\circ_i + \tau_j + t^\circ\tau_{ij} + S(t^\circ\tau)_{k(ij)} + E(t^\circ\tau S)_{l(ijk)} \qquad (3.5)$$

The response Y is modelled by μ - the overall mean, a temperature effect, t°, and a time effect τ. Time and temperature effects and the interaction between them, $(t^\circ\tau)$ are considered fixed. The number of storage conditions (i.e. number of levels of the temperature effect) runs from 1 to n and the subscript of τ indicates the length of the study running from 1 to m.

The random variation between samples is nested within all other factors because a sample has been subject to a particular storage temperature and it is used only once, i.e. at the time of analysis. The time effect describes a possible trend in the concentration as illustrated in figure 11.2.

The residual is considered normal distributed, i.e.

$$E(t^\circ\tau S)_{l(ijk)} \in N(0, \sigma^2_{E(t^\circ\tau S)})$$

and furthermore

$$\sum t^\circ_i = 0 \qquad i = 1...n \qquad \text{temperature} \qquad (3.6)$$

$$\sum \tau_j = 0 \qquad j = 1...m \qquad \text{time} \qquad (3.7)$$

$$\sum t^\circ\tau_{ij} = 0 \qquad \text{interaction} \qquad (3.8)$$

$$S(t^\circ\tau)_{k(ij)} \in N(0, \sigma^2_{S(t^\circ\tau)}) \qquad k = 1, 2 \qquad \text{sample} \qquad (3.9)$$

The least square normal equations which are the result of minimizing the residual sum of squares are linearly dependent, therefore a unique solution depends on the constraints given above.

The time effect - a special problem

In practice the time effect is confounded with the random variation of the analytical method. An alternative approach is to model the time effect as a random variable $\mathfrak{T}$ with a linear trend, $b \neq 0$, and a random contribution of normal distributed white noise with zero mean, ω. Suppose the stability check is performed at time t with the random contribution $\mathfrak{T}$ representing the confoundation between time and the analytical method.

The time effect, $\mathfrak{T}$, with index t in model (3.5) could be described by the regression equation:

$$\mathfrak{T}_t = b \cdot t + \epsilon_t \tag{3.10}$$

The expanded form includes a correlation between time points controlled by α:

$$\mathfrak{T}_t = b \cdot t + \alpha \mathfrak{T}_{i-1} + \omega_t \tag{3.11}$$

The problem is now converted into a time series with an autocorrelated error structure. A model of this form allows a trend superposed by random noise. (Given the fact that observations are sampled discretely, taking the differences between observations in time, may be appropriate.)

A model such as 3.11 requires quite substantial data - at least more than the 7-8 time points observed in the present study. For this reason it has not been attempted to apply the extended model to data. However, with sufficient data, this formulation of the problem could mean that a very early prediction of instability becomes realistic by means of a time series based type of control chart. A second order constant dynamic linear model is described in (Shoemaker and Tsui 1993). The model is suggested for the early detection of linear trends for instance in the case of degradation phenomena with small shifts in the mean and it might be applicable in this situation when sufficient data is available.

The application

The model (3.5) has been applied stepwise including data from $\tau \leq \tau_2$ up until τ_m. Any trends in the stability of the material is expected to manifest itself in the time and temperature effects which are considered fixed. It would be natural to expect a significant interaction effect between time and temperature if there were a stability problem. The confounding between the time effect and the variation of the chemical method of analysis must be taken into consideration in the interpretation of the ANOVA results. For τ_1 only temperature and sample variation could be tested. All tests were considered at $\alpha = 0.05$. The overall results of the analyses were:

Time point	τ_1	$\leq \tau_2$	$\leq \tau_3$	$\leq \tau_4$	$\leq \tau_5$	$\leq \tau_6$	$\leq \tau_7$	$\leq \tau_8$
Elapsed time in months	3	5	8	10	11	14	19	22
Measurand	Significant variance contribution at $\alpha = 0.05$							
TN								
$t°$					■			
τ	%				■			✗
$t°\tau$	%				■			
$S(t°\tau)$					■			
NO_{2+3}-N								
$t°$								
τ	%	✗	✗	✗	✗	✗	✗	✗
$t°\tau$	%					✗	✗	✗
$S(t°\tau)$								
NH_4-N								
$t°$		✗	✗	✗	✗	✗	✗	✗
τ	%	✗	✗	✗	✗	✗	✗	✗
$t°\tau$	%	✗	✗	✗	✗	✗	✗	✗
$S(t°\tau)$	✗	✗	✗	✗	✗	✗	✗	✗

Table 3.3: Results of ANOVA according to eq.(3.5) applied to wastewater data tested at $\alpha = 0.05$. ✗ effect significant at $\alpha = 0.05$. ■ no observations. % effect not included in model.

Reproducibility of the analytical methods
$s^2_{analytical}$ - the analytical uncertainty (of the chemical method) is calculated as according to (ISO 5725 1994c). The reproducibility variance, s^2_R of ISO 5725 consists of repeatability variance and an intermediate measure of precision, $s^2_{L_i}$ - the variance between laboratories in a proficiency test. $s^2_{L_i}$ does not apply to the design used here, since all measurements were performed in the same laboratory. However, since there were several months between the measurements, standard solutions etc. had been replaced, and therefore the variance between laboratories is replaced by the variance between time points. Assume that the variance between time points is of the same order as the variance between laboratories, then - using the terminology of ISO 5725:

$$s \backsimeq s_{analytical,t_i^\circ} = \sqrt{s^2_{r_i} + s^2_{L_i}} \tag{3.12}$$

where $s^2_{r_i}$, the repeatability variance for temperature t_i° is based on the cell standard deviations s_{ij} as shown in table 3.4. $s^2_{L_i}$ expresses the between-time variance. Equations are given in appendix D.

	Temperature, t_i°		
Time, τ_j	t_1°	t_2°	t_3°
τ_1			
τ_2	s_{ij}		
τ_3			
. . .			
τ_8			

Table 3.4: Spread within cells

Information about the method reproducibility on natural samples can be useful for the estimation of control limits for non-synthetic RMs (see section 3.5.3). Also, all of the chemical analyses carried out to obtain the present data sets have been carried out by the author (and are thus biased by nature). The comparison of the experimental analytical reproducibilities to the that of VKI control charts (which include within-and between-days variation) is also a comparison of the author's laboratory performance to the performance of trained laboratory technicians. In tables 3.5,3.6 and 3.7 the reproducibilities have been calculated separately for each storage temperatures. The residual variance from the ANOVA is also included, this is the overall repeatability estimate according to (D.1) with all temperatures included.

Range test procedure

Newman-Keul's range test procedure (Montgomery 1997b) has been applied in the cases where the time effect was significant in the ANOVA. Provided that the presence of a systematic time trend causes the concentration of analyte to either increase or decrease, it would be expected that the groups identified by the range test procedure would exhibit a chronological pattern. The Newman-Keul's range test statistic is based on the studentized range

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{\sqrt{MS_E/n}} \tag{3.13}$$

where $\bar{y}_{max}$ and $\bar{y}_{min}$ are the maximum and minimum sample means out of a groups of p sample means. $q_\alpha(p,f)$ is the upper α percentage point of the studentized range for p group of means with f degrees of freedom. Taking the standard error of each group average, the critical value of the test is

$$K_p = q_\alpha(p,f)\sqrt{MS_E/n} \tag{3.14}$$

Group means are compared two at a time, thus $a(a-1)/2$ pairs of means are considered. The null hypothesis of no difference between a set of means is rejected for differences between group means larger than the critical value K_p.

TN

The ANOVA of TN shows no significant temperature effects when new time points are included stepwise. From previous experiences with analyses of TN on wastewater material from the same origin it was expected that sample variation would not be significant (Schramm-Nielsen 1995). The time effect is significant in the ANOVA including $\tau \leq \tau_8$ (marked by ✗ in table 3.3). This could be due to method variation confounded with the time effect. The Newman-Keul range test statistics for the full data sets identifies two groups. One group consists of the observations from time-point τ_7 only, all other data sets do not differ significantly from each other and therefore belong to the same group. Based on the result of the range test it is a reasonable conclusion that the significance of the time effect for the full data set is due to variation in the chemical analytical method and does not express a trend in concentration.

The material is stable and homogeneous with regard to TN within the 22 month period of observation.

The reproducibility of the chemical analytical method is calculated on the basis of the experimental data and separately for each temperature according to (3.12) as illustrated in figure 3.7. The experimental reproducibility calculated from 20°C data is about twice the method standard deviation of the analytical method derived from the VKI control chart of synthetic samples using the analytical method according to DS 221 (see table 3.5). The experimental reproducibilities calculated from 4°C and 37°C data are of the same order as that of the VKI control chart. Data is given in table 3.5.

Total experimental uncertainty	
4°C	$s_{analytical,t_1^\circ}$ =16.24 μg/l
20°C	$s_{analytical,t_2^\circ}$ =29.81μg/l
37°C	$s_{analytical,t_3^\circ}$ =15.99μg/l
VKI control chart QC Type WW3	$s_{synthetic}$ = 13.86 μg/l
ANOVA (Overall exp.repeatability)*	$\sqrt{MS_E}$=14.25

Table 3.5: DS 221 TN: Comparison of analytical uncertainty according to (3.12). * Eq.(D.1) appendix D applied across temperature levels.

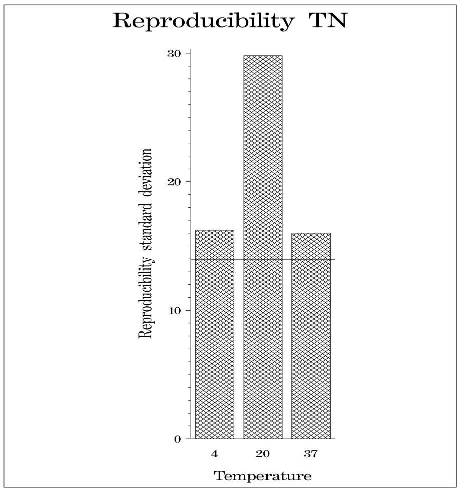


Figure 3.7: TN experimental reproducibility, calculated separately for each storage temperature according to (3.12). VKI control material uncertainty marked for comparison.

NO_{2+3}-N

In the ANOVA's of the sample NO_{2+3}-N content, the time effect is found to be significant for $t \geq \tau_2$. The interaction effect between time and temperature is significant at $t \geq \tau_6$. Newman-Keuls test applied stepwise to time means indicates that groups can be identified. The groups overlap in some cases and they do not exhibit a time related systematic trend. The significance of the time effect is therefore explained as variations in the chemical analytical method. The material is stable and homogeneous with regard to NO_{2+3}-N. The analytical uncertainty is evaluated in table 3.6. The experimental analytical uncertainty is on the same level as the uncertainty for VKI synthetic control material, QC Type WW2.2 using an auto-analyzer adapted version of DS 223. The reproducibility calculated separately for each temperature according to (3.12) is illustrated in figure 3.8.

Total experimental uncertainty	
4°C	$s_{analytical,t_1^\circ}$ =5.35μg/l
20°C	$s_{analytical,t_2^\circ}$ =6.36 μg/l
37°C	$s_{analytical,t_3^\circ}$ =7.64 μg/l
VKI control chart QC Type WW2.2	$s_{synthetic}$ = 6.40 μg/l
ANOVA(Overall exp.repeatability)*	$\sqrt{MS_E}$ = 5.12

Table 3.6: DS 223 NO_{2+3}-N: comparison of analytical uncertainty according to (3.12). * Eq.(D.1) appendix D applied across temperature levels.

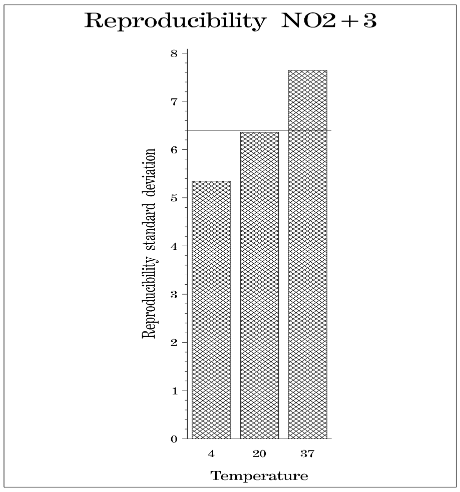


Figure 3.8: NO_{2+3}-N experimental reproducibility, calculated separately for each storage temperature according to (3.12). VKI control material uncertainty marked for comparison.

NH_4-N

From τ_2 and onwards all of the effects in model (3.5) are significant. The material is not stable with regard to NH_4-N. Temperature and time effects are significant as well as the interaction effect, and the material is inhomogeneous from $\tau \geq \tau_1$. (Aminot and Kérouel 1995) reported increased levels of ammonium in autoclaved seawater samples[2] stored for 12 months and longer. Aminot explains the concentration changes (in the range of 0.05-0.07 μmol/l per year) by diffusion of atmospheric air through polypropylene caps and demonstrated this by experiments. At VKI seawater samples in glass bottles were seen to show increased levels of ammonium after 14 months of storage. Since the concentration changes seen here are considerably larger than those reported by Aminot, the most

[2]North Atlantic surface water

likely explanation is a chemical decomposition of organic N compounds into NH_4-N. The concentration changes seen for NH_4-N is not reflected in the TN concentrations. A probable explanation is that while the distribution between organic N and ammonium changes, the sum is constant. Furthermore, the ammonium concentrations are in the range of 60-100 μg/l whereas the TN concentrations are about 4.5 mg/l. A concentration change in the μ area will not affect a total N concentration of 4.5 mg/l.

Figure 3.9 shows the mean temperature values at each time point. The standard deviations of the mean are indicated with bars. It is seen that from time τ_3 the means of samples stored at 37°C begin to move out of control. The centre line on the plot is the assumed reference value, i.e. the mean value at time τ_0. Note that in this plot the ammonium concentrations are to scale and not expressed as ratios.

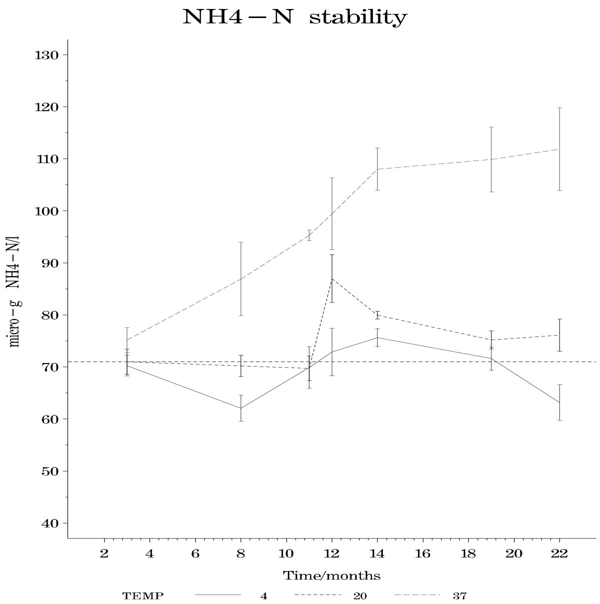


Figure 3.9: NH_4-N: Mean temperature values $\bar{X}_{\tau,t_i^\circ}$ at each time point with standard deviations of the mean indicated

Total experimental uncertainty	
4°C	$s_{analytical,t_1^\circ}$ =11.16 μg/l
20°C	$s_{analytical,t_2^\circ}$ =8.76 μg/l
37°C	$s_{analytical,t_3^\circ}$ =20.03 μg/l
VKI control chart QC Type RW1	$s_{synthetic}$ = 2.69 μg/l
ANOVA(Overall exp.repeatability)*	$\sqrt{MS_E}$ = 3.79

Table 3.7: DS 224 NH_4-N: comparison of analytical uncertainty according to (3.12). * Eq.(D.1) appendix D applied across temperature levels.

The analytical uncertainty is evaluated in table 3.7. The experimental analytical uncertainty is about 4-10 times the uncertainty for VKI synthetic control material, QC Type RW1 using the analytical method DS 224. The reproducibility calculated separately for each temperature according to (3.12) is illustrated in figure 3.10. The reproducibility standard deviation increases for the highest temperature, which indicates that differences between samples also increase with the temperature when the material turns unstable. Figure 3.11 shows the standard deviations when eq.(3.12) is applied for each time point instead of for each temperature. This corresponds to reversing columns and rows in table 3.4. s_{ij} is still the cell standard deviation but s_{Li} expresses the deviation between storage conditions.

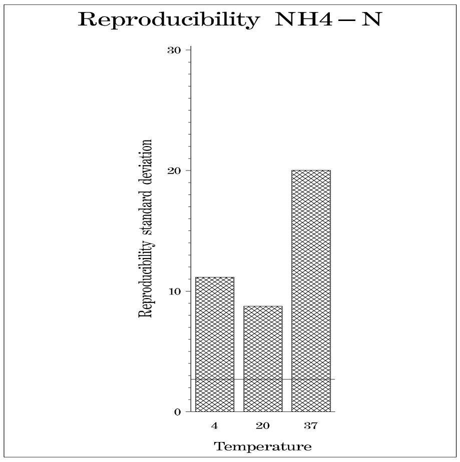


Figure 3.10: NH_4-N experimental reproducibility, calculated separately for each storage temperature according to (3.12). VKI control material uncertainty marked for comparison.

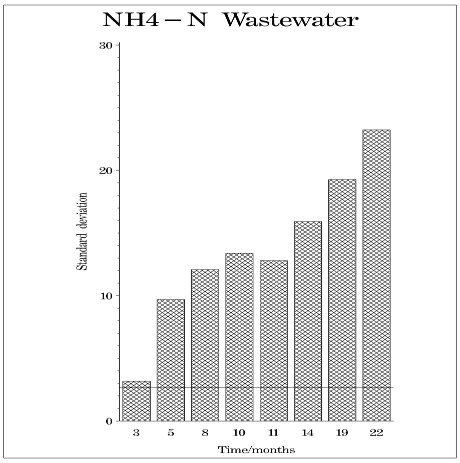


Figure 3.11: Standard deviation according to (3.12) calculated separately for each time point.

3.5.3 $\bar{X}$ control charts

The reproducibilities and repeatabilities calculated according to (ISO 5725 1994c) have been used for the estimation of upper and lower control limits for Shewart $\bar{X}$ control charts. A separate chart is constructed for each storage condition and the input is the mean of m samples measured at each τ.

The mean values at τ_0 have been used as target values in the place of real certified reference values. Upper (UCL) and Lower (LCL) Control Limits have been calculated according to

$$LCL = \mu_0 - 3\hat{\sigma}/\sqrt{m_i} \quad \text{where} \quad \mu_0 = \text{mean value at } \tau_0 \tag{3.15}$$

$$UCL = \mu_0 + 3\hat{\sigma}/\sqrt{m_i} \quad \text{and} \quad \hat{\sigma} = \sqrt{V(\bar{X}_{t^\circ_{ref}})} \tag{3.16}$$

$$m_i = \text{no. of samples in a cell (table 3.1)} \tag{3.17}$$

$$V(\bar{X}_{t^\circ_{ref}}) = \hat{\sigma}^2_{between\ time\ points} + \frac{\hat{\sigma}^2_{repeatability}}{n_h} \tag{3.18}$$

$$n_h = \text{harmonic mean of number of samples/cells in table 3.1, see eq.(D.3) in appendix D} \tag{3.19}$$

$V(\bar{X}_{t^\circ_{ref}})$ - the variance of the mean values at 4°C has been chosen for the control limits. The reason for this is that the lowest temperature represent stable conditions under which it is presumed certain that the material will be stable. The estimates $\hat{\sigma}^2_{between\ time\ points}$ and $\hat{\sigma}^2_{repeatability}$ are the variance measures calculated according to (3.12). Since changes may take place at the higher storage temperatures and the aim is that the material should not differ from the 4°C level, the limits based on 4°C reproducibility have been applied to the charts for higher storage temperatures.

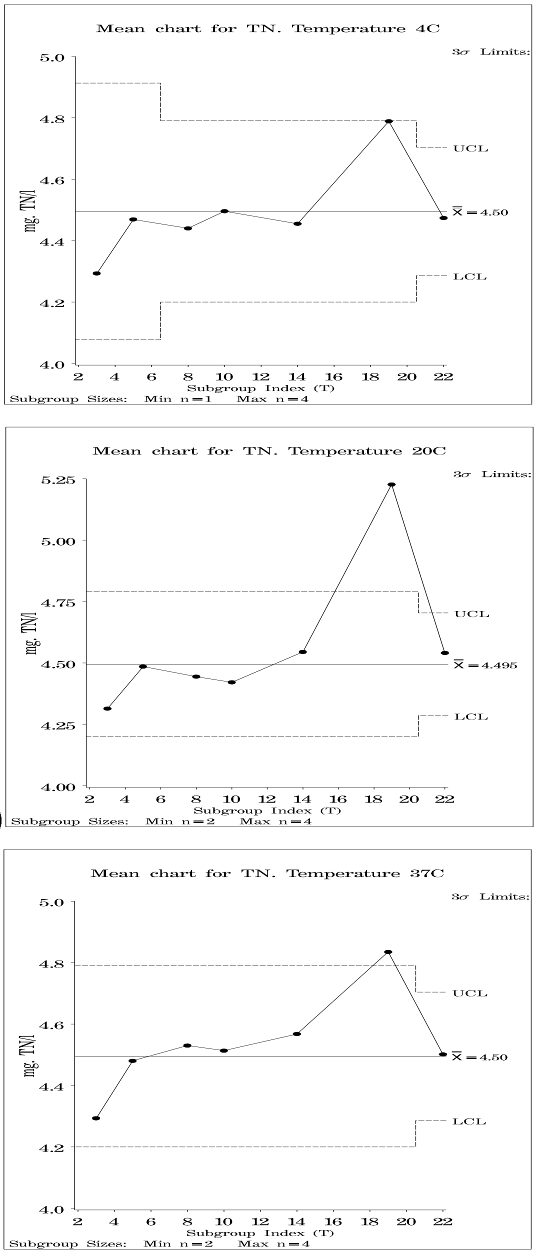
The control limits calculated in this case example are of course subject to the conditions that

☞ μ_0 is not a real certified reference value, and therefore less certain as a reference level.

- ☞ The repeatabilities and reproducibilities are calculated on the basis of the limited data from this study
- ☞ -and the estimates of these measures of variance are likely to be biased considering the fact that all measurements were carried out by one person.

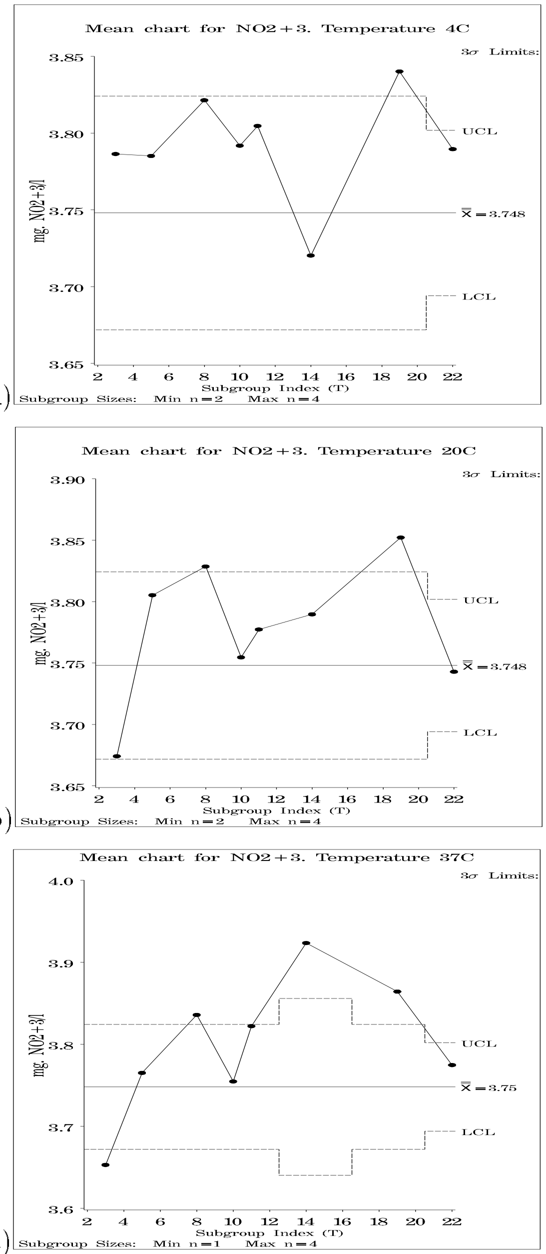
Figure 3.14 demonstrates the trend for NH_4-N at 37 °C already seen in the ANOVA. The charts for TN and NO_{2+3}-N are less successful in the sense that several out of control situations occur, which are apparently not trend related and are not in agreement with the ANOVA results.

The $\bar{X}$ chart procedure has been included for demonstration of the possible use of variance measures of the analytical method. An indisputable weakness of the present charts is that the reproducibilities and repeatabilities were not estimated on beforehand but on the basis of the data plotted in the charts. Further comments on the problems related to simultaneous control charts for analytes in the same matrix are found in sec.11.3.



a) b) c)

Figure 3.12: TN $\bar{X}$ control chart. a) 4°C b) 20°C c) 37°C.



a) b) c)

Figure 3.13: NO_{2+3}-N $\bar{X}$ control chart. a) 4°C b) 20°C c) 37°C.

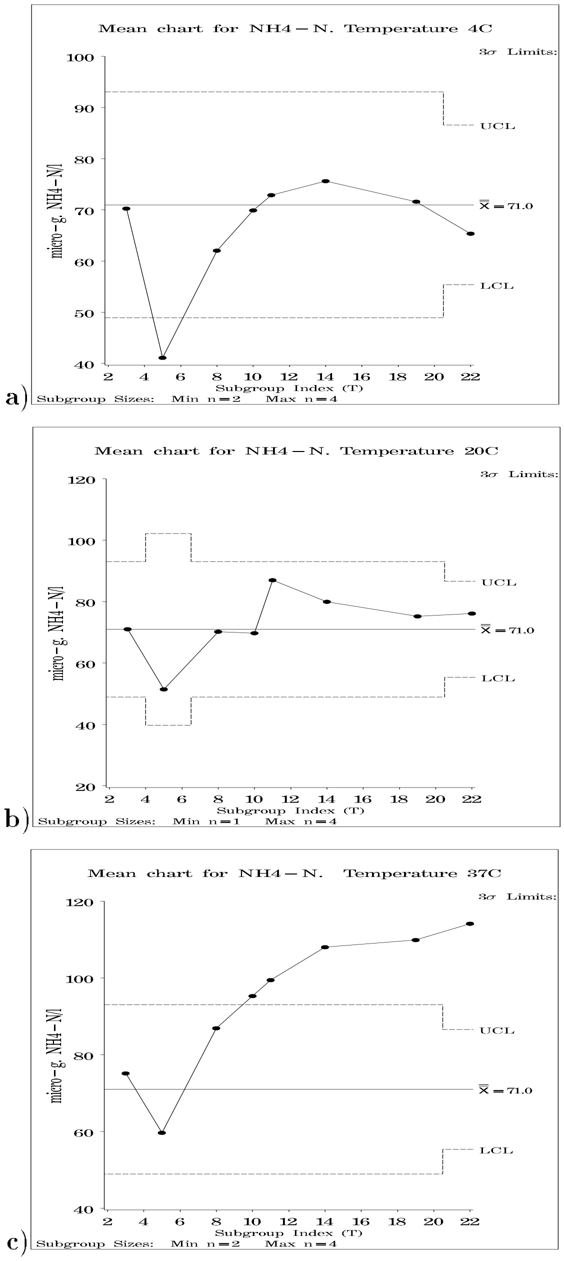


Figure 3.14: NH_4-N $\bar{X}$ control chart. a) 4°C b) 20°C c) 37°C.

There are two remarkable indications on the charts which need further comments:

- at $\tau = 19$ months very high concentrations are observed on all of the charts for TN and NO_{2+3}-N
- a decline in the NH_4-N concentration is seen at $\tau = 5$ months

There is no apparent explanation for the high concentrations of TN and NO_{2+3}-N to be found in data. The results of the synthetic control samples are within the control limits and the reduction capacity of the Cd column was satisfactory. A possible explanation for the high levels of NO_{2+3}-N could be that there was a difference in temperature between the control samples and the wastewater samples in spite of the precautions taken to bring all samples to room temperature. This however could not be the case for TN since *all* samples were autoclaved with peroxodisulphate as part of the analytical procedure. The measurement step did not take place on the same day as the autoclaving which meant that all samples were certain to have the same temperature.

It is more likely that an explanation should be sought for in the standard deviations at time $\tau = 19$ months. Eq. (3.12) has been applied to each time point as explained on pp. 33.

Figures 3.15 and 3.16 show the results for TN and NO_{2+3}-N. The TN variation is very large at $\tau = 19$ months and the NO_{2+3}-N variation is rather small compared to the other time points (see also figures 3.1 and 3.2). The coincidence that this occurs at the same time point could be part of the explanation.

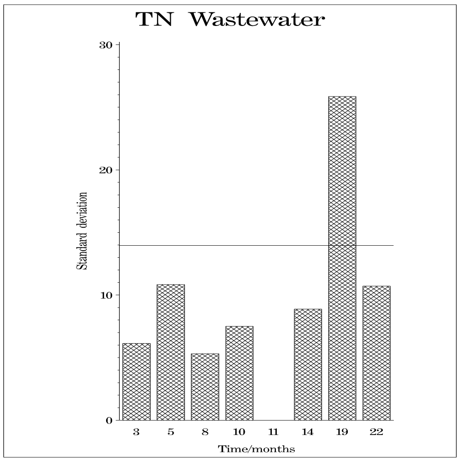


Figure 3.15: TN: standard deviation according to (3.12) calculated separately for each time point.

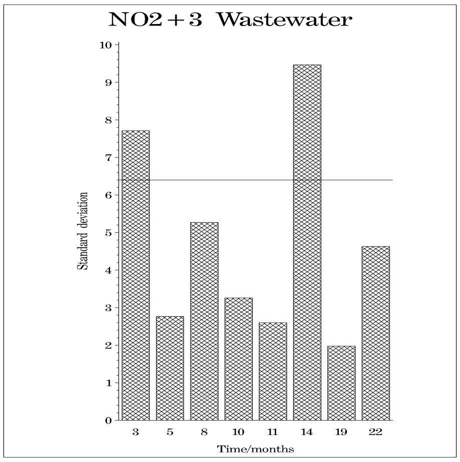


Figure 3.16: NO_{2+3}-N: standard deviation according to (3.12) calculated separately for each time point.

3.6 Multivariate methods

The following sections present three multivariate methods: Hotelling's T^2 (Montgomery 1997b), a so-called M-chart procedure (Chan and Li 1994) and a multivariate exponentially moving average control chart (MEWMA) (Lowry *et al.* 1992). The tests are multiple and compare several responses at the same time. The time mean values of TN, NH_4-N and NO_{2+3}-N calculated separately at each of the storage temperature are used as input, i.e. each of the three methods is applied to each of the storage temperatures. The results of the analyses of the wastewater data are presented here and discussed. A detailed explanation of the methods and their equations is found later in chapter 11.

3.6.1 Hotelling's T^2

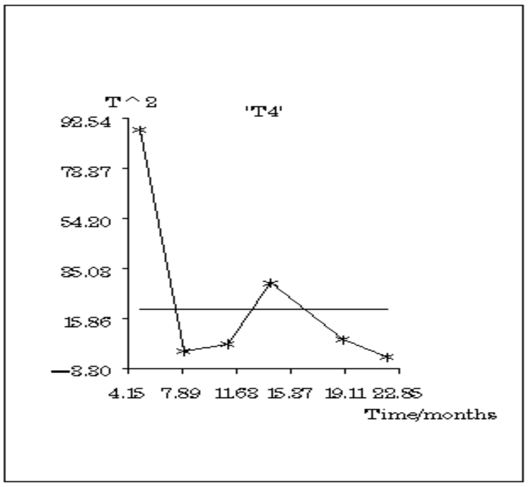


Figure 3.17: Result of Hotelling's T^2 for 4°C

The theory of the Hotelling's T^2 test is described in chapter 11. It is used for a Shewart-type control chart which employs information about the current sample only. The multivariate test which dates from 1947 is designed to give a signal when a statistically significant shift has taken place in the multivariate mean value. In the present case the mean values at time τ_0 have been used.

The covariance matrix for the responses TN, NH_4-N and NO_{2+3}-N has been estimated on the basis of the entire data set. The upper control limit of 19.47 (indicated by a horizontal line in the plots) is based on a sample size of 2 (which is the size of the subsamples taken at each temperature level at each time point) and a confidence level of $\alpha = 0.05$ (see eq.11.16). The method of Hotelling's T^2 does not correct for variation in the analytical method. The calculation of the test statistic is based on the basic model which assumes that the overall mean follows a p-dimensional normal distribution. Figures 3.17 - 3.19 show the results of the Hotelling T^2 test for all three temperatures. T^2 is significant at times τ_2 and τ_5 for samples stored at 4°C.

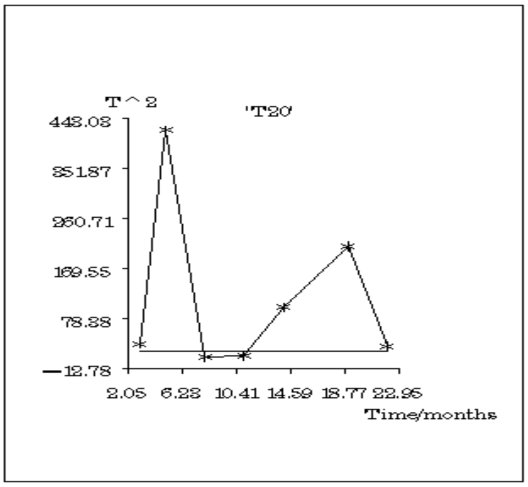


Figure 3.18: Result of Hotelling's T^2 for 20°C

For samples stored at ambient temperature, the T^2 test statistic is significant at all time points except at times τ_3 and τ_4. T^2 is significant at all time points from τ_1 to τ_7 for samples stored at 37 °C.

The tendency of the 37°C data to show significance in almost the entire period of the time study is seen for other methods as well; the ANOVA of the NH_4-N concentrations and the M-chart in section 3.6.2.

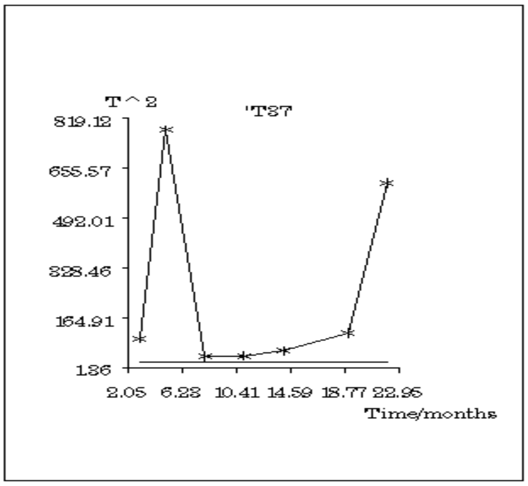


Figure 3.19: Result of Hotelling's T^2 for 37°C

There is no apparent explanation for the two significant test statistics in figure 3.17, nor can it be explained why all but two test statistics for the 20° C data fall outside the control limit. The outcome of the test is ambiguous for 4 and 20° C.

3.6.2 M-chart procedure

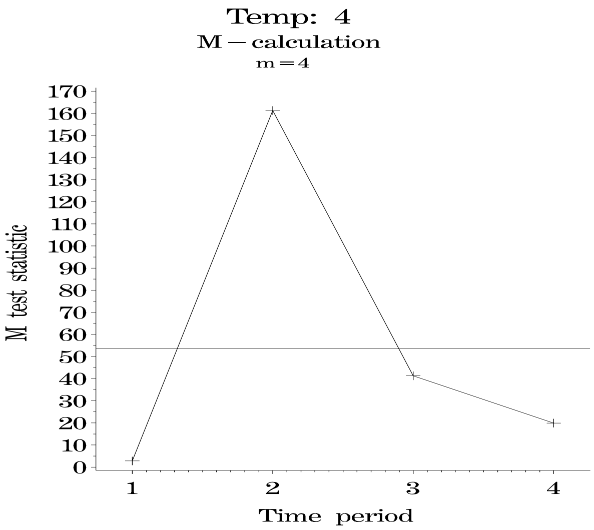


Figure 3.20: M-test statistic for 4°C and m=4

The M-chart procedure described in (Chan and Li 1994) is a dimension reducing control chart method for detecting linear trends in a process mean (in this case the process means of three N-analytes in wastewater). A test statistic M is calculated for a sub-vector of the vector of all the time points in a stability study. The length of the sub-vector must be greater than the number of measured signals, p. The sub-vector includes m time periods within the full time span of the stability study ($m \leq p+1$). The first sub-vector includes the time points $\{\tau_1, \tau_2, \tau_3, \tau_4\}$, the second sub vector moves one step forward and includes the time points $\{\tau_2, \tau_3, \tau_4, \tau_5\}$ etc.

The method is described in detail in chapter 11. It is suggested that the method may be applied to multicomponent RMs because it enables an easily understood graphic representation of data which cannot otherwise be illustrated by plots.

For the purpose of illustration the method is applied to the wastewater data which has the dimension $p = 3$ (for the three measured signals TN, NO_{2+3}-N and NH_4^+). Because of the limited number of time points in the data set the M-chart procedure has been applied with $m = 4$ as the length of the time sub-vector. The method is applied separately to the data sets at each of the storage temperatures.

The test statistic is calculated for the first four time points and plotted as shown in figure 3.20. Since the test statistic is smaller than the F-distribution percentile[3] $F(p, m-p)_{1-\alpha}$ (marked by a reference line in the plot) the time sub vector is moved one step forward and the test statistic is calculated anew.

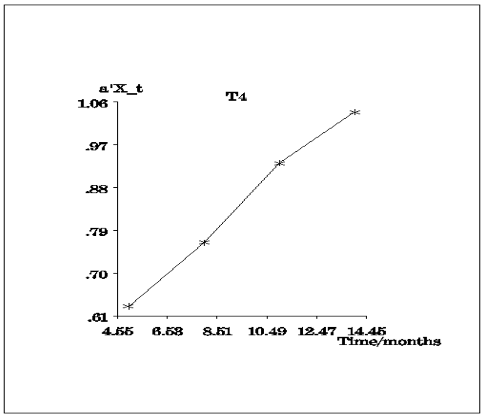


Figure 3.21: Temp. 4 °C: plot of $a^{'}X_{\tau}$ for the second time period.

At 4°C the test statistic falls outside the limit in the second time period. When the test statistic is significant, a projection of the data belonging to the significant time period is determined.

An example of the projection of measured data points versus time is shown in figure 3.21. The plot is linear and indicates that in the time period from $\tau = 5$ months to $\tau = 14$ months $(\tau_2, \tau_3 \tau_4, \tau_5)$ there is a linear trend in measured mean values of the wastewater RM stored at 4 °C.

Figures 3.22 - 3.24 illustrate the same for the 20 and 37 ° C data.

[3] A significance level of $\alpha = 0.10$ has been allowed here, see chapter 11 for comments.

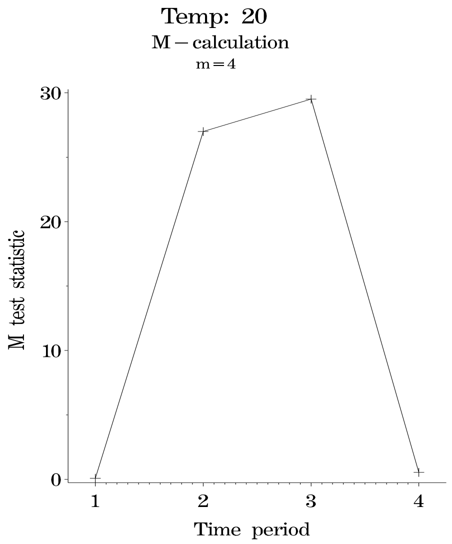


Figure 3.22: M-test statistic for 20°C and m=4

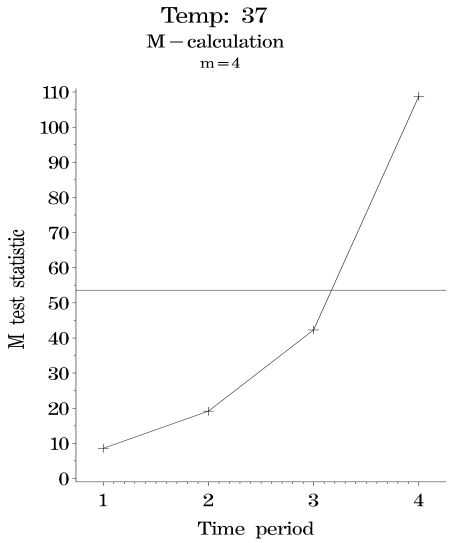


Figure 3.23: M-test statistic for 37° C and m=4

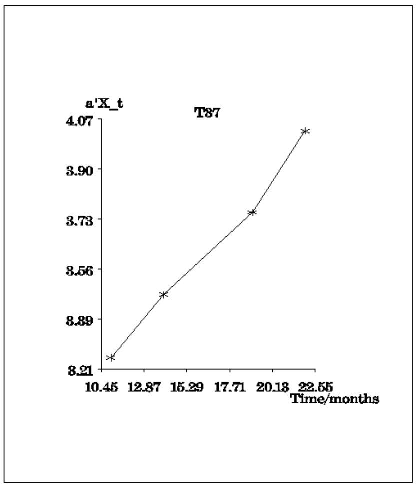


Figure 3.24: Temp. 37 °C: plot of $a^{'}X_{\tau}$ for the fourth time period

3.6.3 MEWMA

The multivariate exponentially weighted moving average (MEWMA) control chart is one of numerous multivariate control chart methods. The multivariate version of the EWMA was developed by (Lowry *et al.* 1992). (The equations are given in chapter 11.) Like the M-chart method, the MEWMA control chart uses more than just the current sample in the calculations. The current sample is given the weight λ and the previous test portion is included (remembered) with a weight of $(1-\lambda)$. (Lowry *et al.* 1992) used $\lambda = 0.1$ and found it to be an effective choice for detecting small shifts in the mean vector. The method does not correct explicitly for method variation, but the weight given to the previous test portion could be interpreted as a "memory" of the method variation included in previous results.

The control limit is related to the average run length (ARL) of the chart and is given in (Lowry *et al.* 1992) for $\lambda = 0.1$ and dimension $p = 3$. With ARLs of 50, 100 and 200 the control limits are 7.08, 9.00 and 10.97 respectively. As the following tables will show, the test statistics are much larger than the control limits in the majority of the cases - which indicates out-of-control situations for all three storage temperatures.

Temperature: 4°C			
λ	Time	Test statistic T_i^2 eq.(11.29)	Significance*
$\lambda = 0.1$	$\tau \leq 5$	$16{\cdot}10^4$	yes
	$\tau \leq 8$	$73{\cdot}10^3$	yes
	$\tau \leq 11$	$41{\cdot}10^3$	yes
	$\tau \leq 14$	$24{\cdot}10^3$	yes
	$\tau \leq 19$	$15{\cdot}10^3$	yes
	$\tau \leq 22$	$11{\cdot}10^3$	yes
$\lambda = 0.2$	$\tau \leq 5$	$40{\cdot}10^3$	yes
	$\tau \leq 8$	$16{\cdot}10^3$	yes
	$\tau \leq 11$	$7{\cdot}10^3$	yes
	$\tau \leq 14$	$2.7{\cdot}10^3$	yes
	$t \leq 19$	10^3	yes
	$\tau \leq 22$	589	yes
$\lambda = 0.3$	$\tau \leq 5$	$18{\cdot}10^3$	yes
	$\tau \leq 8$	$6{\cdot}10^3$	yes
	$\tau \leq 11$	$2{\cdot}10^3$	yes
	$\tau \leq 14$	262	yes
	$\tau \leq 19$	7.28	yes **
	$\tau \leq 22$	0.29	no
$\lambda = 0.4$	$\tau \leq 5$	10^4	yes
	$\tau \leq 8$	$3{\cdot}10^3$	yes
	$\tau \leq 11$	567	yes
	$\tau \leq 14$	2.0	no
	$\tau \leq 19$	105	yes
	$\tau \leq 22$	58	yes
$\lambda = 0.5$	$\tau \leq 5$	$6{\cdot}10^3$	yes
	$\tau \leq 8$	$1.5{\cdot}10^3$	yes
	$\tau \leq 11$	121	yes
	$\tau \leq 14$	126	yes
	$t \leq 19$	248	yes
	$\tau \leq 22$	90	yes

Table 3.8: MEWMA test statistic for 4°C data. * Critical value for $\lambda = 0.1$
** When ARL=50

Temperature: 20°C			
λ	Time	Test statistic T_i^2 eq.(11.29)	Significance*
$\lambda = 0.1$	$\tau \leq 3$	17	yes
	$\tau \leq 5$	601	yes
	$\tau \leq 8$	390	yes
	$\tau \leq 11$	301	yes
	$\tau \leq 14$	43	yes
	$t \leq 19$	6.28	no
	$\tau \leq 22$	5.22	no
$\lambda = 0.2$	$\tau \leq 3$	4.22	no
	$\tau \leq 5$	709	yes
	$\tau \leq 8$	399	yes
	$\tau \leq 11$	271	yes
	$\tau \leq 14$	10.0	yes
	$t \leq 19$	1.69	no
	$t \leq 22$	40	yes
$\lambda = 0.3$	$\tau \leq 3$	1.88	no
	$\tau \leq 5$	798	yes
	$\tau \leq 8$	375	yes
	$\tau \leq 11$	216	yes
	$\tau \leq 14$	0.8	no
	$t \leq 19$	24	yes
	$\tau \leq 22$	91	yes
$\lambda = 0.4$	$\tau \leq 3$	1.06	no
	$\tau \leq 5$	885	yes
	$\tau \leq 8$	329	yes
	$\tau \leq 11$	155	yes
	$\tau \leq 14$	22	yes
	$t \leq 19$	58	yes
	$\tau \leq 22$	131	yes
$\lambda = 0.5$	$\tau \leq 3$	0.68	no
	$\tau \leq 5$	969	yes
	$\tau \leq 8$	267	yes
	$\tau \leq 11$	100	yes
	$\tau \leq 14$	66	yes
	$t \leq 19$	86	yes
	$\tau \leq 22$	145	yes

Table 3.9: MEWMA test statistic for 20°C data. * Critical value for $\lambda = 0.1$

Temperature: 37°C			
λ	Time	Test statistic T_i^2 eq.(11.29)	Significance*
$\lambda = 0.1$	$\tau \leq 3$	$11{\cdot}10^4$	yes
	$\tau \leq 5$	$74{\cdot}10^3$	yes
	$\tau \leq 8$	$48{\cdot}10^3$	yes
	$\tau \leq 11$	$32{\cdot}10^3$	yes
	$\tau \leq 14$	$19{\cdot}10^3$	yes
	$t \leq 19$	$11{\cdot}10^3$	yes
	$\tau \leq 22$	$6{\cdot}10^3$	yes
$\lambda = 0.2$	$\tau \leq 3$	$29{\cdot}10^3$	yes
	$\tau \leq 5$	$24{\cdot}10^3$	yes
	$\tau \leq 8$	$14{\cdot}10^3$	yes
	$\tau \leq 11$	$8{\cdot}10^3$	yes
	$\tau \leq 14$	$3{\cdot}10^3$	yes
	$t \leq 19$	816	yes
	$\tau \leq 22$	11.05	yes
$\lambda = 0.3$	$\tau \leq 3$	$13{\cdot}10^3$	yes
	$\tau \leq 5$	$13{\cdot}10^3$	yes
	$\tau \leq 8$	$7{\cdot}10^3$	yes
	$\tau \leq 11$	$3{\cdot}10^3$	yes
	$\tau \leq 14$	574	yes
	$t \leq 19$	0.77	no
	$\tau \leq 22$	532	yes
$\lambda = 0.4$	$\tau \leq 3$	$7{\cdot}10^3$	yes
	$\tau \leq 5$	$9{\cdot}10^3$	yes
	$\tau \leq 8$	$4{\cdot}10^3$	yes
	$\tau \leq 11$	$1.5{\cdot}10^3$	yes
	$\tau \leq 14$	35	yes
	$t \leq 19$	256	yes
	$\tau \leq 22$	$1.2{\cdot}10^3$	yes
$\lambda = 0.5$	$\tau \leq 3$	$4.6{\cdot}10^3$	yes
	$\tau \leq 5$	$7{\cdot}10^3$	yes
	$\tau \leq 8$	$3{\cdot}10^3$	yes
	$\tau \leq 11$	713	yes
	$\tau \leq 14$	32	yes
	$t \leq 19$	572	yes
	$\tau \leq 22$	$1.5{\cdot}10^3$	yes

Table 3.10: MEWMA test statistic for 37°C data. * Critical value for $\lambda = 0.1$

3.7 Discussion

In the present case three analytes were measured on the wastewater RM. Table 3.11 summarizes the results of the methods applied to data.

Univariate methods
The univariate methods both indicated instability for NH_4-N. The reference temperature method works well as a diagnostic tool for visual inspection, but it is not capable of identifying sources of impact. For analytical methods with no day to day variance, there is a possibility of performing regression analysis on the data points in the reference temperature ratio plots for potentially unstable analytes. An alternative is the time series formulation of the problem presented in section 3.5.2, provided that substantial data is available. The ANOVA is capable of identifying sources of impact. The ANOVA confirmed the suspicion from the reference temperature ratio plots that NH_4-N was unstable and the analysis furthermore revealed variance contributions for NO_{2+3}-N which were related to the variations of the chemical analytical method.

The $\bar{X}$ control chart with control limits based on the experimental analytical variance indicated out of control situations for TN at $\tau = 19$ months for 20 and 37°C and several times for NO_{2+3}-N. These results did not reflect the results of the ANOVA. For NH_4-N the control chart showed that the material was out of control from $\tau = 10$ months and onwards when stored at 37°C. The sign of instability is significant at a later time with the control chart than with the ANOVA.

The reference temperature ratio plot method and the $\bar{X}$ chart require very little work from a computational point of view. As diagnostic tools they can be used by a technician in the lab equipped with a calculator, paper and a ruler. The ANOVA requires a little more effort and preferably some kind of statistical software, preferably capable of handling unbalanced data as well as random and fixed effects in the same model statement.

Method	Comment
Univariate methods	
Reference temp. ratio	TN: no indication
	NO_{2+3}-N: no indication
	NH_4-N: indication of instability
ANOVA	TN: sample homogeneity, material stability
	NO_{2+3}-N: sample homogeneity, material stability
	NH_4-N: instability from $\tau = \tau_2$ and onwards, sample inhomogeneity
$\bar{X}$ chart	TN: one out of control situations for 20°C and one for 37°C, both at τ_7.
	NO_{2+3}-N: several out of control situations, but no signs of trend.
	NH_4-N: all points in control for 4°C and 20°C. Obvious trend for 37°C which is out of control from τ_4 and onwards.
Multivariate methods	
Hotelling's T^2	4°C: two out of control test statistics for $\tau = 5$ and $\tau = 14$
	20°C: five out of control test statistics for $\tau = \{3, 5, 14, 19, 22\}$
	37°C: all test statistics are out of control
M-chart	4°C: one out of control statistic for the time span $\tau = 5$ to $\tau = 14$
	20°C: no out of control statistics
	37°C: one out of control statistic in the last time span $\tau = 11$ to $\tau = 22$
MEWMA:	large majority of out of control statistics for all data sets with λ values ranging between 0.1-0.5

Table 3.11: Outline of statistical analyses on wastewater stability data.

Multivariate methods

When the dimension of the data matrix is small it is fairly easy to analyse data one response at a time (or one temperature at a time) and identify which component (or temperature level) is responsible for changes. However, if the number of measured responses is large and possible trends less obvious, the methods of plotting and analysing one component at a time, one temperature at a time etc. becomes a tedious and time consuming task.

The multivariate methods can handle all responses in the same calculation. The statistics have been calculated separately for each temperature level, which reduces the total number of calculations by a factor of p - the number of responses. The results of Hotelling's T^2 were ambiguous for 4°C and 20°C data, whereas all of the calculated test statistics for 37°C data were out of control. The M-chart method identified one out of control situation for 4°C. This was in agreement with the Hotelling's T^2 statistics since $\tau = 5$ months and $\tau = 14$ months gave significance for both methods.

The M-chart method accepted all 20°C data as being in control, and found the 37°C data to be out of control from $\tau = 11$ months and onwards. The conclusion about instability of the 37°C data, agreed with the ANOVA of the NH_4-N data.

The M-chart method includes sub vectors of information from several time points, but with equal weight to all observations included. The MEWMA includes increasing amount of information and assigns weights to the current sample and the information statistic of the previous observations. The impact of an observation from time τ_m is reduced with λ with each increment of time. If the time points between observations are far spaced in real time, then the influence of observation τ_m will last longer than if the time points are fairly close. The MEWMA may also be considered a conservative method because of its memory of past test statistics and this may not be an advantage with chemical methods of analyses which are subject to day to day variation. With the present data sets MEWMA identifies almost all observations to be out of control. If this is due to the influence of the NH_4-N data, the response of the MEWMA method is very prompt.

Assuming that the conclusion about NH_4-N instability is true, then it would be a fair expectation that the multivariate methods signal out of control situations at some point in the stability study. The MEWMA and Hotelling's T^2 appear to be very sensitive (possibly to method variation) and both methods give signals more often than the univariate methods. Of the multivariate methods the MEWMA and Hotelling's T^2 appear to agree most with the univariate ANOVA, whereas the M-chart and the $\bar{X}$ chart react considerably later than these two methods.

The advantage of the multivariate methods is mostly due to their diagnostic nature and their capacity for handling multiple responses. For a large number of responses, the somewhat heavier computational work should outweigh the burden of plotting numerous univariate control charts and trying to draw a final conclusion on the state of the material. As mentioned earlier there is another aspect of the multicomponent reference materials which must be taken into account when working with stability studies. Plotting one response at a time may involve a serious risk of drawing wrong conclusion, because of the probability distortion: the probability that all p responses of a perfectly stable material will be plotted simultaneously inside their control limits when plotted separately is $(1-\alpha)^p$ (α being the type I error probability) as discussed on page 166 in chapter 11.

The multivariate methods provide information about trends, but they cannot identify neither which response(s) is (are) the source of the change, nor can these methods differentiate between inhomogeneity, method variation or instability. In cases where the number of responses is larger than three, a suggestion could be to recalculate the multivariate statistics excluding individual responses for the purpose of identifying components possibly responsible for instability. The ANOVA applied to each response is specific because of its univariate approach, but it is also more time consuming if the number of responses p is large.

Conclusions on the application of the univariate and multivariate methods chosen for the present data are

- The reference temperature ratio plot confirmed the expectation that NH_4-N was unstable.
- The ANOVA found that NH_4-N is unstable and that the instability is significant from τ_2 and onwards.
- The $\bar{X}$ chart identified instability for 37°C at τ_4, one step earlier than the M-chart but later than the ANOVA.
- Hotelling's T^2 gave ambiguous results for lower temperatures, but indicated general out of control for 37°C data from the very beginning of the stability study.
- The M-chart identified instability at a later time point (τ_5) than the ANOVA for NH_4-N data (indication as early as τ_2). There was some ambiguity about the M-chart result for 4°C.
- The MEWMA method signaled overall out of control situations for all three temperatures.

Clearly, the multivariate methods cannot do the job alone. Provided that a good type of multivariate method is chosen for the RM in question there is a possibility of greatly reducing the statistical work in all of the *in-control* situations.

All of the multivariate methods presented here require more computational effort than the univariate methods. A software package which can handle matrix calculations is required. The user must have basic programming skills and some statistical knowledge in order to interpret the results.

Chapter 4

Stability of ammonium in seawater RM

This chapter presents the results of the experimental work and a stability study carried out on a seawater RM.

4.1 Purpose

Problems with increasing ammonium concentration in a seawater RM had been experienced previously at VKI. The purpose of the study on seawater was to investigate the possibilities of improving the ammonium stability. As already mentioned on page 33, the reason for increased ammonium concentration on seawater previously observed at VKI could be that dissolved organic compounds were transformed into ammonium. Two types of treatment were chosen - ultra-filtration (UF) and ultra violet (UV) radiation treatment.

The purpose of the ultra-filtration was to reduce the problem of ammonium instability in seawater. The ultra-filtration was supposed to remove some if not all of the potentially unstable compounds from the water, so that ammonium levels would be lower and possibly more stable in the ultra-filtrated water.

Another approach was to enhance the possible transformation into ammonium by UV radiation. The UV radiation transfers energy to the material which means that chemical bonds could be broken thereby enhancing the ammonium production. In this case it would be expected that ammonium levels would be higher in samples treated this way than in samples not exposed to UV radiation. The ammonium concentration in the exposed samples would then be expected not to alter over time.

4.2 Data collection

The seawater RM analysed in this project was prepared from a batch of seawater sampled in the Sound (the narrow body of water between Sweden and Denmark). The salinity of the water was measured to 22 ‰. All standards and control solutions used in the analyses were based on synthetic seawater of the same salinity.

A 2^3 factorial design was set up for the preparation of the sea water. All of the seawater was pre-filtered using a GF/A filter (glass microfiber filter, see General Glossary). Part of it was passed through an additional filter used for blood treatment analysis (ultra-filtration). A UV radiation treatment was also applied to a portion of the water. For the purpose of the experimental design two storage temperatures were selected: 4 and 37°C. The selected factors were:

Factor symbol	Treatment type (abbr.)	Level of treatment
A:	Radiation (UV)	no UV radiation
		UV radiation
B:	Ultra-filtration (UF)	no filtration
		filtration
C:	Storage	4°C
		37°C

Table 4.1: Treatments used in the seawater experiment.

The treated water was filled into 250 ml. infusion bottles. The entire lot was autoclaved within 3 hours at standard conditions 121°C for 20 minutes and stored. Details of the preparation and the methods of analyses are given in appendix B. During the following 22 months, analyses of TP (total phosphorous) and NH_4-N were performed at intervals. At each time point 1 sample was selected at random from each combination of treatment and storage conditions. The samples were analysed with two replicates per sample. The primary aim was to identify main effects which influence ammonium levels.

4.3 Results

The results of the stability study on ammonium are presented in the following two figures 4.1 and 4.2. The TP stability study data is presented in figures 4.3 and 4.4. All results are given in micro grams.

Since the material was prepared especially for this study there are no certified reference values available. For the purpose of illustration, the mean values of the measurands at time τ_0 i.e. right after preparation have been used as reference values in plots and calculations. The mean values have been estimated on the results of samples which had not been exposed to UV radiation because it was suspected that contamination of some of the selected samples may have taken place during analyses. (The presence of ammonium in the air could have caused the problem.)

The central reference lines in the plots (figures 4.3 and 4.4) are the mean values for the measurands at τ_0.

All of the following statistical analyses, graphical methods etc. are based on the data set where $\tau > \tau_0$ since the data obtained at the starting point had not yet been exposed to varying storage conditions.

In addition to the plot of sample means shown in this chapter, plots of means calculated by each combination of treatment and storage temperature are shown in appendix B.

Figure 4.1 shows NH_4-N 4 °C degree data with the GF/A filtered samples represented by the upper figure and the ultra-filtrated samples shown in the lower. Data is divided in the same manner in figure 4.2 which shows the NH_4-N 37°C data.

Figures 4.3 and 4.4 illustrate the TP data at 4 and 37 °C respectively. Judged by these figures, the TP concentration in the seawater samples appear to remain unchanged within the time period of this study.

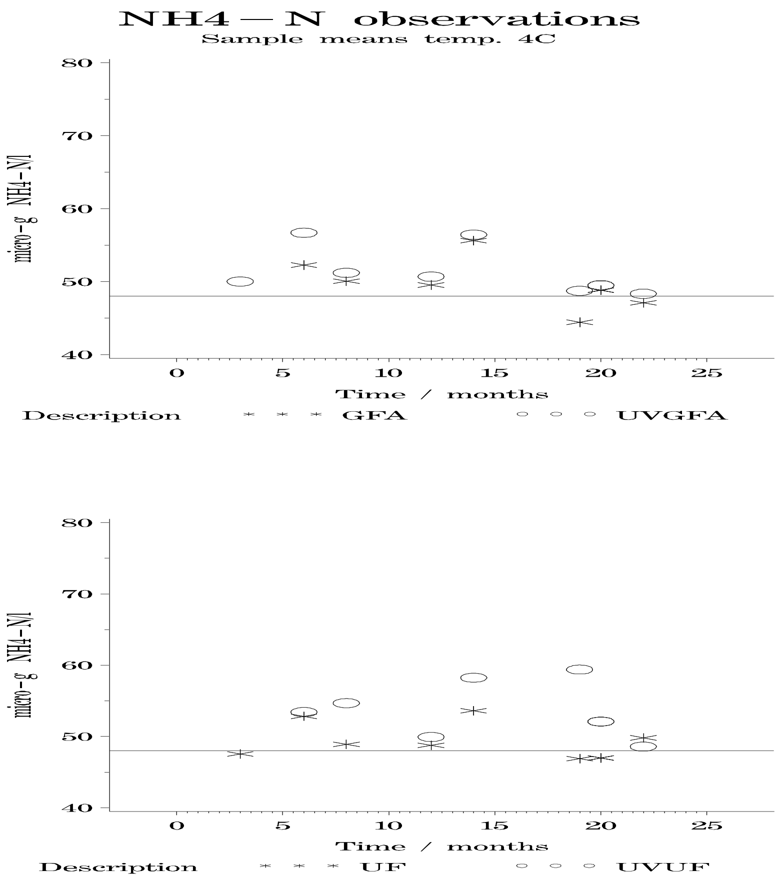
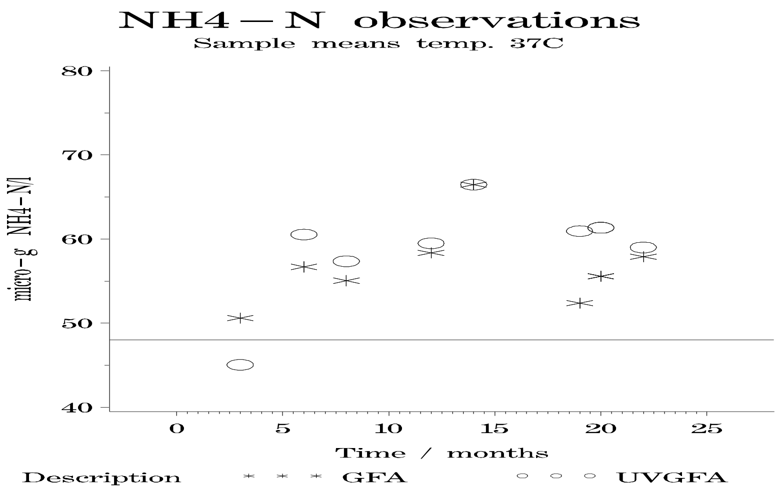


Figure 4.1: NH_4-N 4°C raw data: sample means. Reference line at $\bar{X}(\tau_0)$.



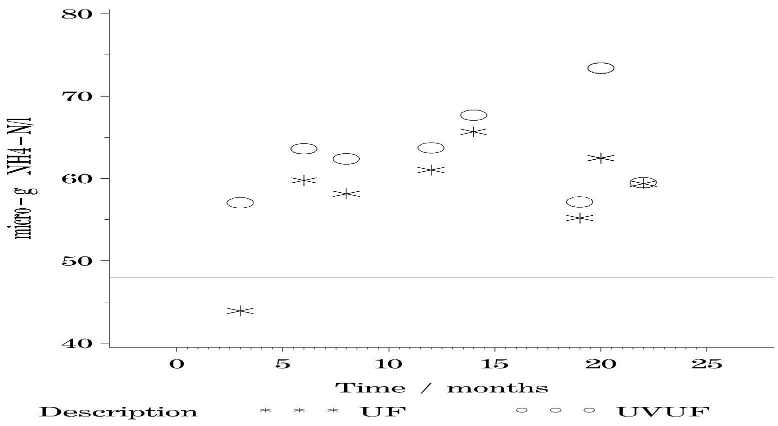


Figure 4.2: NH_4-N 37°C raw data: sample means. Reference line at $\bar{X}(\tau_0)$.

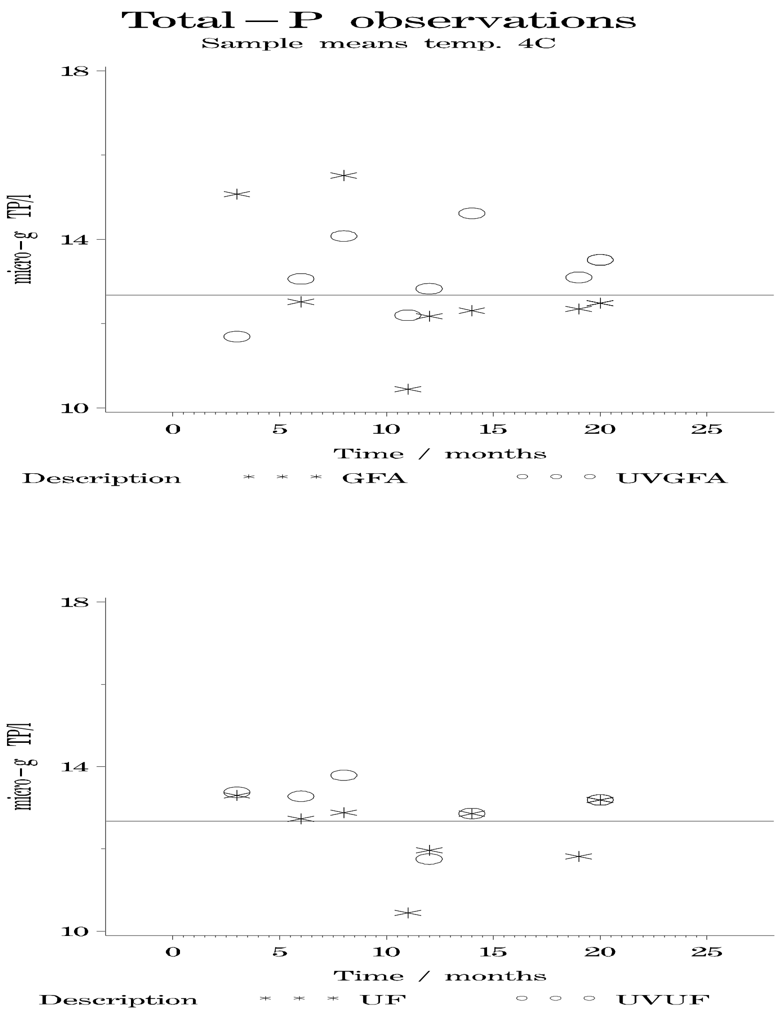


Figure 4.3: TP 4°C raw data: sample means. Reference line at $\bar{X}(\tau_0)$.

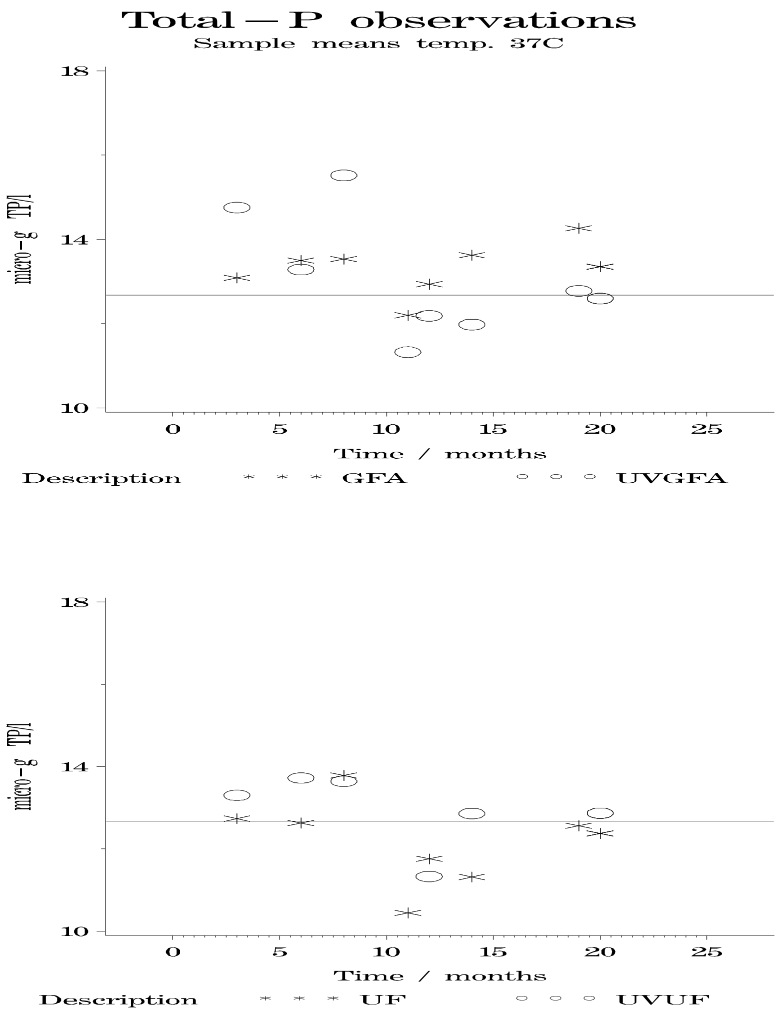


Figure 4.4: TP 37°C raw data: sample means. Reference line at $\bar{X}(\tau_0)$.

4.4 Statistical analyses

4.4.1 Model adequacy checking

The raw data for each measurand have been evaluated by means of quantile-quantile plots (QQ-plots) of the difference between replicates on a bottle as described in section 3.4. i.e. for each measurand:

$$Y_{i,j,k,1} - Y_{i,j,k,2} \tag{4.1}$$

where $i = time$, $j = temperature$, $k = treatment$ and $r = 1, 2$ is the replicate.

The plots are shown in appendix B, section B.0.5. For each of the measurands, the normality assumptions are verified by use of the Anderson-Darling test (Encyclopedia of Statistical Sciences 1982), (Stephens 1974) as described earlier in chapter 3.

4.4.2 Reference temperature ratio plots

The reference temperature ratio plots were introduced in chapter 3. The method is also applied to the seawater data. The mean values of samples stored at 37 °C relative to the mean value of samples stored at the reference temperature (4°C) are shown in figures 4.5 and 4.6.

Figure 4.5 indicates that NH_4-N concentration levels at 37°C increase with time. This is in agreement with previous experience, and the same trend was seen for wastewater in the preceding chapter.

Figure 4.6 shows no time dependent trend for TP concentration levels at 37°C. On the basis of this type of plot, it is supposed that the small fluctuations are due to variation of the chemical analytical method.

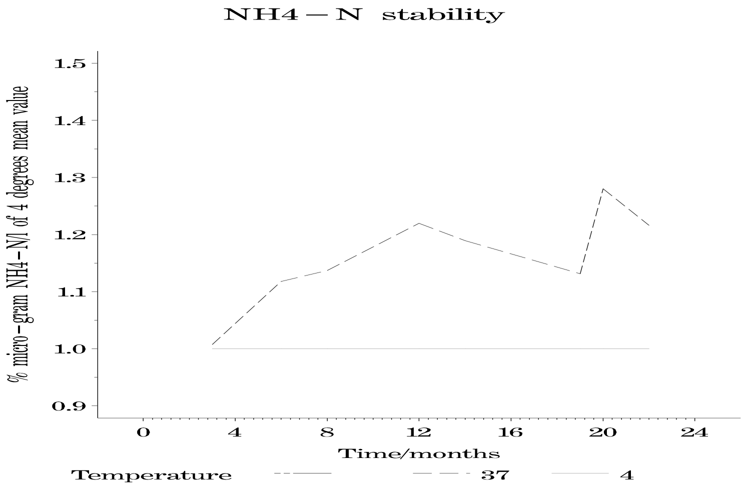


Figure 4.5: NH_4-N: mean temperature values at each time point $\bar{X}_{time,temp}$ normalised to $\bar{X}_{time,4}$.

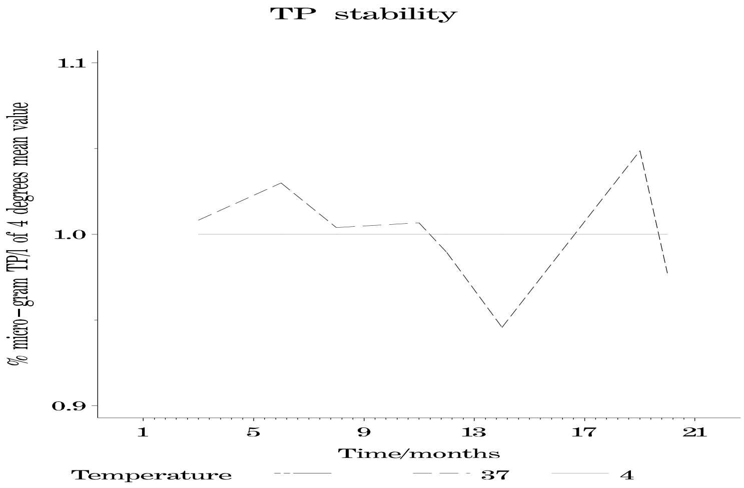


Figure 4.6: TP: mean temperature values at each time point $\bar{X}_{time,temp}$ normalised to $\bar{X}_{time,4}$.

4.4.3 ANOVA

The ANOVA model applied is that of a 2^3 factorial design for the effects listed in table 4.1. The labels a for radiation, b for filtration etc. are in small caps in the model to indicate that these effects are considered fixed. The factorial experiment is carried out once and a sufficiently large number of samples were produced for the further experiments. The development of ammonium and TP in the samples is monitored over time and for that reason the time effect is included in the model as a fixed effect. The problems related to the time factor in this design are identical to those discussed in the preceding chapter. The response is described by the following model:

$$Y_{ijkl} = \mu + a_i + b_j + ab_{ij} + c_k + ac_{ik} + bc_{jk} + abc_{ijk} + \tau_l + E_{ijkl} \tag{4.2}$$

It is a slightly modified form of the model for a 2^3 factorial since τ_l is an additive term in (4.2). The constraints on the model regard the fixed effects as argued in chapter 3. The residual is considered normal distributed, i.e.

$$E_{ijkl} \in N(0, \sigma_E^2)$$

and furthermore

$$\begin{aligned}
&\sum a_i = 0 && i = 1,2 && \text{radiation} && (4.3)\\
&\sum b_j = 0 && j = 1,2 && \text{filtration} && (4.4)\\
&\sum c_k = 0 && k = 1,2 && \text{storage condition} && (4.5)\\
&\sum ab_{ij} = 0 \quad \sum ac_{ik} = 0....etc. && && \text{interaction} && (4.6)\\
&\sum \tau_l = 0 && l = 1...m && \text{time} && (4.7)\\
& && && && (4.8)
\end{aligned}$$

The model has been applied stepwise including data from $\tau \leq \tau_2$ up until τ_m. For τ_1 the time effect was eliminated from the model. The overall results of the analyses were:

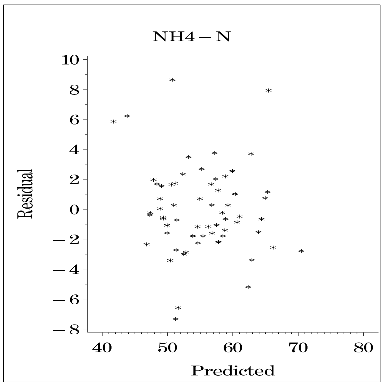
Time point	τ_1	$\leq \tau_2$	$\leq \tau_3$	$\leq \tau_4$	$\leq \tau_5$	$\leq \tau_6$	$\leq \tau_7$	$\leq \tau_8$	$\leq \tau_9$
Elapsed time in months	3	6	8	11	12	14	19	20	22
Measurand	Significant variance contribution at $\alpha = 0.05$								
NH_4-N									
a			✗	■	✗	✗	✗	✗	✗
b				■				✗	✗
c			✗	■	✗	✗	✗	✗	✗
T	%	✗	✗	■	✗	✗	✗	✗	✗
TP									
a									
b					✗	✗	✗	✗	✗
c									
τ	%				✗	✗	✗	✗	✗

Table 4.2: Results of ANOVA according to eq. (4.2) applied to seawater data. a = radiation treatment, b = filtration type, c = storage temperature, τ = time. ✗ The effect is significant at $\alpha = 0.05$. ■ no observations. % effect not included in model.

NH_4-N

The ANOVA of NH_4-N shows that the time effect is significant from τ_2 and onwards. (The time effect cannot be tested at τ_1.) The three-way interaction effect abc was found to be significant on one occasion. Since no other interaction effects were found to be significant, it is assumed that abc could be related to method variation and the term has therefore not been included in table 4.2.

The temperature and the radiation main effects are found to be significant



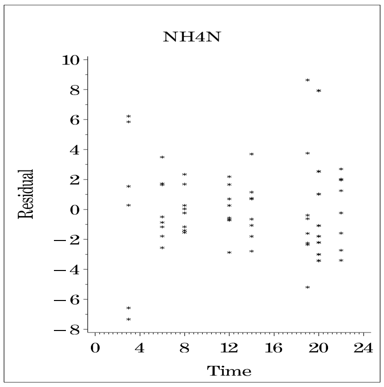


Figure 4.7: NH_4-N: residuals.

from τ_3 and onwards. The filtration main effect is significant at from τ_8 and onwards.

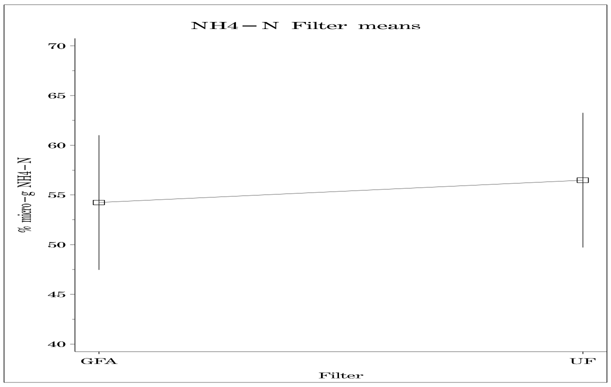
The influence of the main effects are illustrated on figures 4.8 a)-c). The temperature effect is the most influential. The effects of radiation and filtration are smaller. Since no interaction effects between the storage temperature and the treatments 'radiation' or 'filtration' are found to be significant, it cannot be concluded that radiation or filtration treatments have contributed to the stability of the material.

The UV radiation did indeed increase the NH_4-N concentration but the process was not sufficiently enhanced to stabilize the ammonium concentration and prevent further NH_4-N increase.

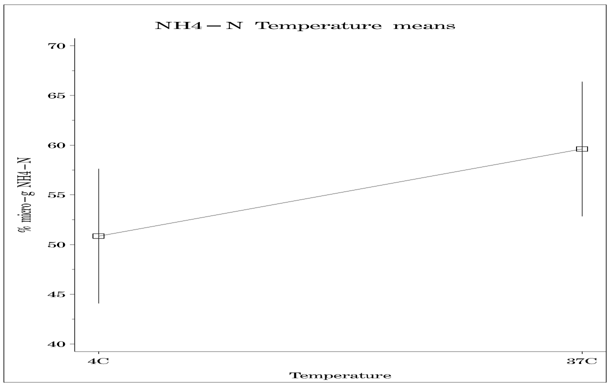
Contrary to the expectations, the ammonium levels for the filtration main effect were found to be higher for the ultra-filtrated samples[1] than for the GF/A filtrated.

No particular patterns are observed when plotting residuals versus predicted values or time, figure 4.7 a) + b).

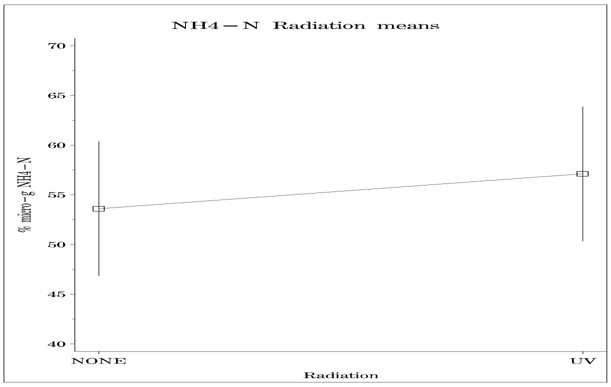
[1] N.B.: the ultra-filtrated samples were also GF/A filtrated prior to the extra filtration.



a)



b)



c)

Figure 4.8: NH_4-N: main effects. Bars indicate the 95 %confidence interval.

TP

The ANOVA of TP shows that the time effect and the filtration main effect are significant from $\tau \geq \tau_5$ and onwards. The main effect of filter which is significant for $\tau \geq \tau_5$ is shown in figure 4.9. The plot of residuals

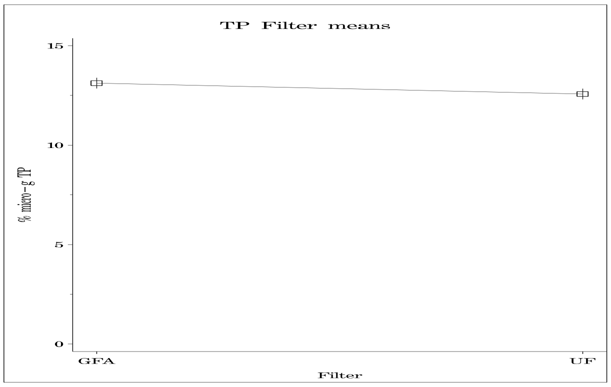


Figure 4.9: TP: main effect of filter. Bars indicate the 95 %confidence interval.

versus predicted values shows increasing variance of the residuals with increasing predicted values (shown in figure 4.10). Considering the fact that the TP concentration levels are very low in seawater, the ANOVA has been verified by a repeated analysis of the log-transformed TP data. This changes nothing with regard to the significance of the filter main effect nor the time effect. The effects are significant at the same time points in the ANOVA of the transformed data and the log transformation does not improve the residual plot.

a) b)

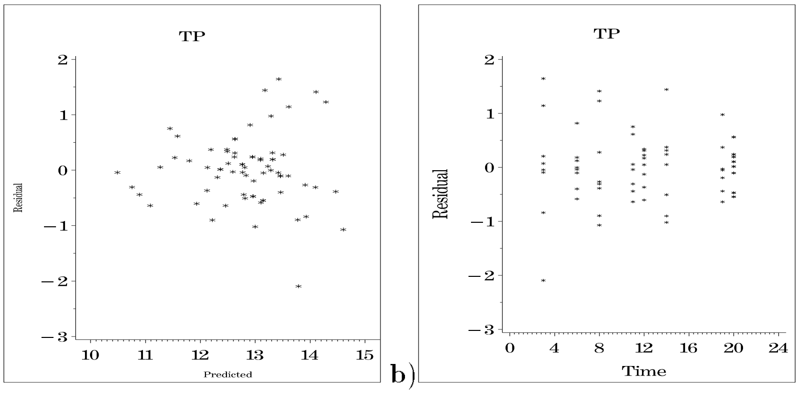


Figure 4.10: TP: residuals

4.5 Discussion

In the present case two analytes were measured on the seawater RM. The purpose of this experiment was screening for main effects which could have an impact on NH_4-N stability. The effect of the treatments was also investigated with regard to TP concentrations. All tests were performed at $\alpha = 0.05$.

NH_4-N
The statistical analyses of the univariate responses confirmed the expected ammonium instability. As expected the storage temperature was found to be a significant source of variation. Samples stored at 37°C showed increasing ammonium concentrations over time, the samples stored at 4°C exhibited lower concentrations and a less pronounced increase. The concentration increase seen in a similar RM at VKI was comparable to the changes seen here. Since the change is larger than those Aminot (Aminot and Kérouel 1995) ascribed to diffusion from the atmosphere, it is assumed that chemical decomposition of organic N is responsible for the changes.

The UV radiation treatment did result in increased ammonium concentration levels, but the treatment effect did not stabilize the material sufficiently to prevent the increasing concentration over time. Contrary to the initial expectation, the ultra-filtration treatment resulted in higher ammonium levels than the GF/A filtration alone. No interaction effects were found to be significant.

TP

The samples exhibited no temperature dependence over time with regard to TP concentration. A time effect - related to analytical method variation was observed for $\tau \geq \tau_5$. A filtration effect was observed for $\tau \geq \tau_5$ with slightly increased TP concentrations for ultra-filtrated samples. The same type of increase was observed with regard to ammonium concentration.

The conclusions on the application of the univariate methods for the seawater data are

- The reference temperature ratio plot confirmed the expectation that NH_4-N was unstable.
- The ANOVA identified this to be the case and found the instability to be significant from $\tau > 8$ months and onwards.
- Slightly increased NH_4-N concentrations were seen in samples treated by UV radiation regardless of storage temperature and filtration type.
- Slightly increased NH_4-N concentrations were seen in ultra-filtrated samples regardless of storage temperature and radiation type.
- Radiation and filtration have not improved the stability of NH_4-N in seawater.

The ultra-filtration and the radiation treatments did have an impact on ammonium concentration levels in autoclaved seawater samples but these treatments did not result in stabilization of ammonium concentration levels. The TP concentration levels were influenced positively by the ultra-filtration treatment. No signs of instability were found for this measurand.

Chapter 5

Innovation and traditions

This chapter gives a short presentation of the history of RMs and the issues which are debated in the field.

The field of chemical RMs has its roots about 30 years back in the mid sixties with a number of biological RMs primarily for inorganic constituents. The next two decades brought about some botanical, food related and clinical standards and some animal tissue materials from large standardization bodies such as NIST[1] in the USA (Iyengar and Wolf 1998)

The struggle for obtaining comparability of measurements on natural samples by means of standardized procedures has been among the many tasks of international and national bodies of standardization. The Environmental Protection Agency in the USA distributed some RMs in the late 1970s without charge as part of its work on method development and quality assurance of analysis of natural samples (Jenks and Boekholt 1998).

Still, at this point most of the materials were single substance materials and only a few of them natural matrix materials. Around this time most of the interest for RMs lay in academic research environments and in government agencies and the materials were not really a commercial trade product. Since then analytical chemistry laboratories in many countries

[1] National Institute of Standards and Technology, at the time National Bureau of Standards.

have taken over tasks of environmental analyses that used to be the responsibility of the governments. This change brought with it the visualization of the need for laboratories to demonstrate quality of analysis in order to achieve knowledge of the validity of data. The need for matrix RMs as support in the analysis of environmental samples is still growing and as a consequence BCR[2] and NIST are no longer alone on the market: other private companies see the potential in producing for this market.

Along the way the BERM symposia[3] have served as an assembly place for the discussions of all the aspects of RMs: preparation, data evaluation for certification, natural matrices, spiked standards for various concentration levels in the same matrix, statistical aspects of homogeneity, trace elements, measurement accuracy, laboratory accreditation, proficiency testing, proper use of RMs, calculation of certified values, specimen banking and the never ending discussion of true values and traceability.

The BERM symposia demonstrate that RM development is an ever changing and live art. New analytical methods such as micro-analytical techniques will renew the demands for RM homogeneity with small sample sizes and sharpened attention on emissions to the environment results in legislative outlet demands/limits, monitoring programs for analytes not previously surveyed and consequently new demands for RMs. An interesting observation was presented at the latest BERM meeting (Antwerp, 1997) by (Jorhem 1998) that many scientific journals (the examples were taken in food analysis, but probably the same holds for the environmental journals) publish papers which provide no proper validation of results. The demand for such documentation prior to publication of results is a vision for the future, but looking back over how things have developed it is probably not a far fetched one. General trends in the RMs available on the market seems to be:

- ✓ more materials based on natural matrices
- ✓ more focus on the uncertainty of measurement procedures in general,

[2] Bureau Communautaire de Référence, now SMT, Standards Measurement & Testing
[3] Biological and Environmental Reference Materials

but also on the uncertainty of certified values.

(Rasberry 1998) predicts growth in usage of RMs for the next 20 years in general, but with a lot of detailed arguments for and against different directions which may be taken. Just to mention a few, Rasberry foresees

- ✓ increased analytical diversity increasing the demand for RMs, but at the same time increased equipment and method robustness
- ✓ increased legislative regulation and increasing number of field laboratories
- ✓ increased trade combined with a growing lack of trust in commerce

-all of which may increase the demands for RMs. However, speculations about the possibilities of the web, of linking scientists at different institutions via computers or instruments that calibrate themselves by national calibration services at a metrological institute may overrule the above premonitions and change the scenery of the RM market completely.

Chapter 6

Some requirements for chemical RMs

This chapter outlines some of the formal requirements for a RM.

The very first and foremost object of a RM producer is to provide the market with a material which is "adequate for purpose". The statement is loose in its formulation, but at the same time broad enough to cover all the aspects that a producer must take into consideration. Directions for RM producers can be found in (ISO Guide 31 1988),(ISO Guide 34 1996),(ISO Guide 35 1989), and (Doc. BCR/48/93 1994). In this context it has been chosen to focus on the questions of homogeneity and stability.

6.1 Homogeneity

This section will focus on within-unit homogeneity. A discussion of between unit homogeneity follows in chapter 10. The purpose of the homogeneity requirement is obvious and should need no further argumentation. What should be given some attention is that the difficulty of obtaining suitable homogeneity depends on the material.

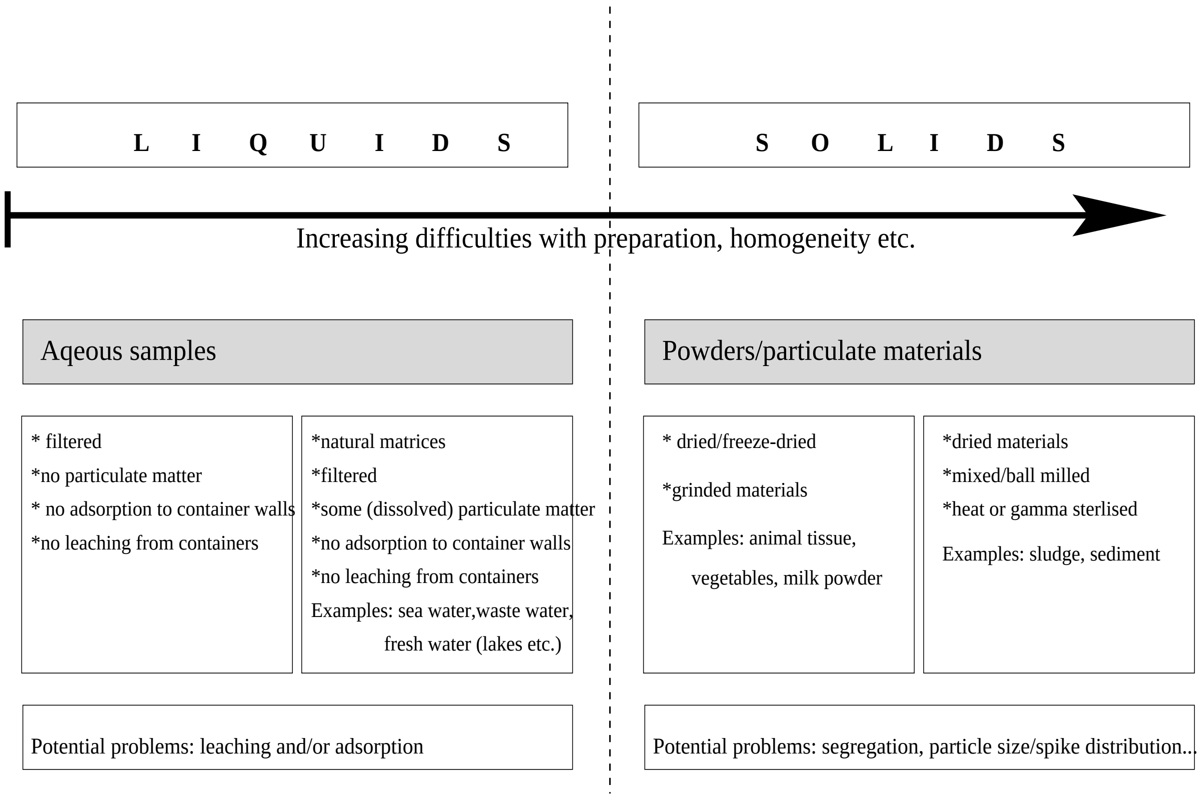


Figure 6.1: Illustration of two broad categories of RMs and some of the potential problems attached to them.

On a qualitative scale liquid materials are easy to handle in terms of homogeneity whereas solids such as sludge or sediment (spiked or natural) materials can cause substantial problems.

A lot of material found to be homogeneous by the supplier's investigations may be found inhomogeneous when the size of the subsample is reduced to smaller and smaller amounts. Therefore it must be underlined that "homogeneous" is not homogeneous at all conditions, but at terms found "adequate for purpose" and regards the size of subsample as stated by the producer. This aspect is naturally more critical for the solid than the aqueous materials. It is recommendable that the homogeneity be checked on a sample size which is e.g. half the sample size used in the standard methods of analysis for which the RM is certified. A material can be accepted as homogeneous even if the F-test for the null hypothesis of no sample variance is greater than the critical value for $p = 0.05$. The condition is that the estimated sample variance s_s^2 is less than 0.3σ where σ is the target standard deviation for the certification (Lawn *et al.* 1997).

6.2 Stability

The stability of a reference material is the ability of the material " when stored under specified conditions, to maintain a stated property value with specified limits for a specified period of time" (ISO Guide 30 1992).

Since stability studies are not always carried out to completion before a RM is put on the market but may continue beyond release, an expiration data given by the manufacturer can only be based on his best knowledge and experiences obtained during the production of the material.

The expected stability of a RM should ideally match the life time of the stock for two reasons. If a material turns unstable, any remaining stock should be re-certified or it will be wasted and lead to fiscal losses for the manufacturer. On the other hand if the material remains stable for years after production with no traceable increase or decrease in the certified compounds, it is in the interest of the manufacturer and the user to maximize the relation of mass production benefits (large batch sizes) to storage expenses. Since liquid RMs for environmental chemical analysis rarely exceed sizes of 25 cm^3 per unit, the storage cost will usually not be critical when no specific climatic conditions are required.

6.3 Application

According to (ISO Guide 30 1992) most CRMs should be considered secondary standards (see pp. xiv), which means that their values depend on a primary reference which is one metrological level above the secondary. This places them accordingly in the measurement hierarchy. Nevertheless, CRMs of a suitable matrix matching composition are to be preferred over primary standards (see pp. xiv) of pure solutions, when it comes to matching a test sample.

(ISO Guide 33 1989) states the advantages of CRMs with a relevant matrix match:

- ✔ assessment of trueness and precision of the measurement method

- ✔ establishment of metrological traceability of results

The above achievements are conditioned on the correct use of the RM in which the identical treatment procedure for test sample and RM is essential. It should also be noted that all information given in certificates have reference to amounts of sample analysed and method(s) of measurement applied as stated in the certificate. If the RM is used for other purposes than the intended use and perhaps analysed with significantly different methods of measurement, the information on uncertainty given in the certificate may not apply.

6.4 Types of reference materials

Solid and liquid chemical RMs can be classified according to various characteristics

☞ the field of application which can be e.g. one of the following

- environmental analyses
- biological analyses
- food industry
- other industries such as pharmaceutical, oil and mining industries etc.

☞ physical state - as one of the following two

- liquids
- solids

☞ chemical composition

- element or compound
- inorganic or organic
- matrix complexity
- natural sample or artificial sample

☞ properties

- degree of homogeneity
- degree of stability

Figure 6.2 is a qualitative illustration of the degree of stability and homogeneity of some types of RMs.

Without quantifying the placement of the examples given in figure 6.2, the approach of visualising where a potential new RM would be placed, might give the manufacturer some idea of the batch size - stability relationship.

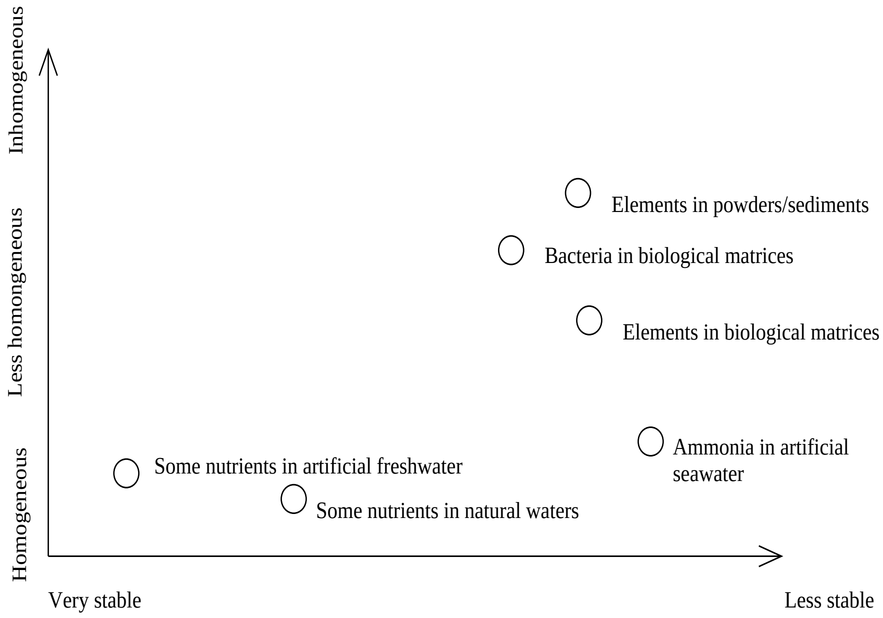


Figure 6.2: Examples qualitatively illustrating the degree of homogeneity and stability in various materials.

Stability and homogeneity experiences for materials in the vicinity of a potential new RM might be included in the research step to evaluate possible performance of a candidate material.

Chapter 7

Aspects of chemical environmental analyses

The present chapter discusses the general concepts, uncertainties and problems related to the chemical analysis of a real sample.

7.1 Introduction

An outline of the sources of uncertainty in environmental analyses could be as shown in figure 7.1. The uncertainties introduced in the course of these steps can be divided into two main groups: contributions from the sampling site and contributions from technical conditions. For a real sample of the environment the background uncertainty comprises

$\sigma^2_{bgd} = \sigma^2_{site} + \sigma^2_{tech}$

where σ^2_{tech} includes

$$= (\sigma^2_{sampling} + \sigma^2_{analytical/n})/m \tag{7.1}$$

for n replicate analyses per sample and m samples taken per site. The example is taken from the field of geochemistry (Thompson and Ramsey 1995) and describes the overall conditions of a chemical environmental analyses.

Figure 7.1 basically reflects the concept used in the International Measurement Evaluation programme to describe a quantitative measurement - shown by the fully drawn line (-). The dotted line (· · ··) shows the steps performed by a laboratory participating in an interlaboratory exercise, collaborative trial or certification procedure. The dashed line (- - -) illustrate roughly the steps carried out by the producer of a natural matrix CRM. Some steps between sampling and storage conditions have been left out for clarity.

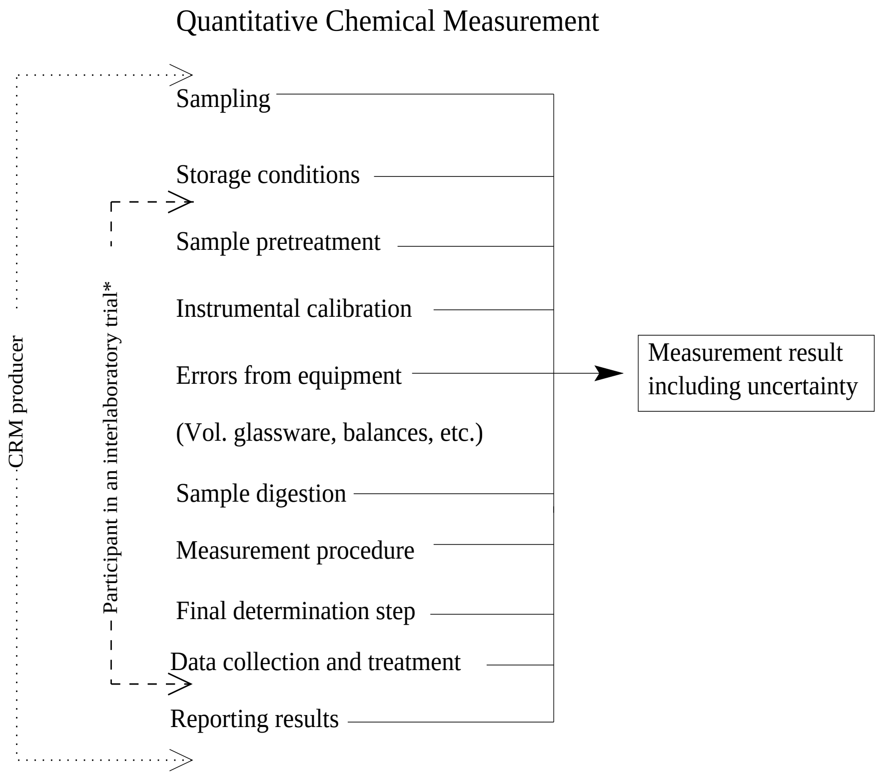


Figure 7.1: Schematic view of factors contributing to uncertainty in the result of a chemical measurement. See text for detailed explanation.

The measurement on a RM is a subject to a subset of all these contributions with some extra twists. Figure 7.2 is an attempt to illustrate the relation between the fresh unknown sample taken from the environment and a CRM based on a natural matrix. It should be noted that in reality the site of sampling and the site from which the CRM originates are NOT identical as shown in the sketch. The sketch should give some idea of the complex uncertainty relations for CRMs and real samples analysed in environmental laboratories. CRMs have the extra uncertainty contribution from the certification exercise, but the background variation is less important for the CRM. The CRM does not attempt to deliver true values of the sampling cite, but serves as a tool for internal quality control in each analytical laboratory. The task of the analytical laboratory is however to deliver trustworthy results describing the state of the sampling site, therefore the background variation is important in the local uncertainty budget.

Sampling errors occur not only in the actual sampling process where a portion of material is taken from its natural environment, but can also be introduced in the sampling of blanks for the measurement process. Furthermore, sampling errors may also be committed by users when sampling reference material for analyses. Homogeneity or heterogeneity are characteristics important to the units of the RM batch as well as to the bulk material from which the RM is sampled.

RMs are accepted as homogeneous when material variability is negligible in relation to measurement uncertainty. The RM may be accepted as homogeneous when the material variability is of the same magnitude as the measurement uncertainty. In such cases the material variability is included as an uncertainty component (Cali and al. 1975).

All of the components of uncertainty given in figure 7.1 apply to the analysis of real unknown samples, but they also form the conditions under which a candidate RM is treated and analysed for homogeneity check, stability monitoring and certification.

There are other types of error which are caused by chemical or environmental conditions such as contamination of the samples during

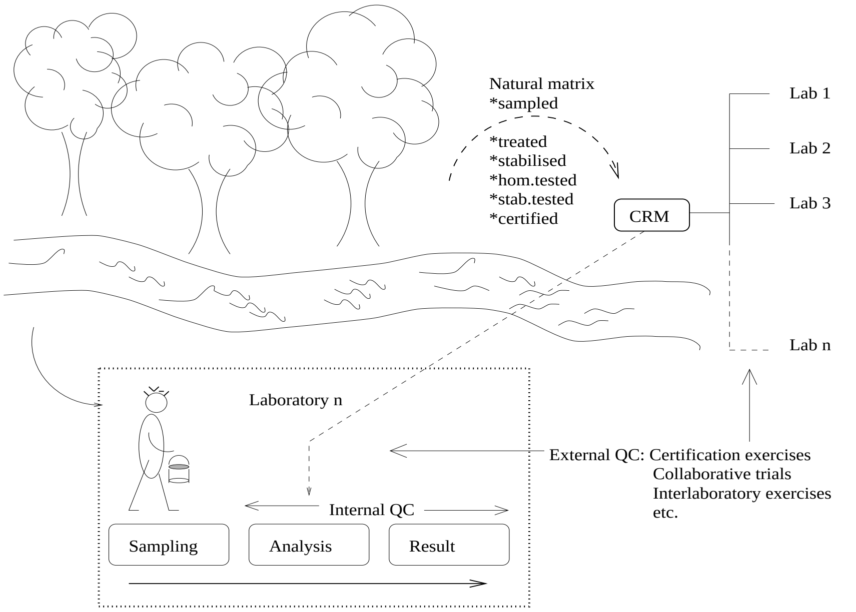


Figure 7.2: Illustration of some of the relations between CRMs and real samples and the uncertainty conditions for both.

collection, contamination of the samples during handling in the laboratory[1] and errors in the actual measurement process. The latter type of error may be of various sorts depending on the type of measurement, e.g. incomplete sample digestion, matrix interferences, calibration problems etc.

Much attention has been paid in literature to the uncertainty regarding all of the steps (see figure 7.1) from storage conditions to the reporting of results (e.g. (Buzoianu and A.-E. 1997), (Ellison and Barwick 1998), (van der Veen and Alink 1998) (Williams 1998), (QA in environmental monitoring 1994), (Funk *et al.* 1995), (Grant and Pelton 1972),(Thompson 1998), (Thompson and Ramsey 1995), (Ramsey *et al.* 995),(Argyraki *et al.*

[1]Contamination has been discussed in section E.2

1995), (Maier 1991) (Quevauviller 1993a), (Reed 1995) and (GUM 1993).) One step - the actual sampling is also vitiated by errors, but this aspect has often been given less attention than the others. A further discussion of this problem is given in section 7.3.

7.2 Uncertainty in measurement

A recent trend in analytical chemistry is the focus on uncertainty in measurement. This was demonstrated e.g. at the Second Eurachem Workshop on "Measurement uncertainty in chemical analysis" held in Berlin (Sept. 29-30 1997). It is the concept of "uncertainty in measurement" which marks a change from earlier terminology which -also in statistics tends to discuss "errors" rather than uncertainty. Whereas the term "error" implies something which may be corrected or avoided, the term "uncertainty" in connection with chemical measurements indicates the nature of the procedure - that there is uncertainty from the beginning.

As mentioned earlier CRMs are used to help achieve traceability in the entire measurement procedure. An aspect which remains uncovered is how to assess the uncertainty of the CRMs; it might not be possible, yet indeed unsatisfactory. In nuclear physics the scientist reaches a boundary limit which is the smallest particle that can be found. Beyond this point he may suspect smaller units, but cannot prove it. In the same manner it could be stated that the chemist will have to satisfy himself with the fact that no CRM for chemical analysis can exist without some unknown uncertainty. The unknown uncertainty of RMs includes of e.g. sample inhomogeneity and measurement uncertainty.

7.3 Uncertainty in sampling

At a first glance it seems obvious that sampling water should be much easier than sampling powders or solids. One condition is common for liquids and solids: the importance of keeping all equipment uncontaminated with the analyte of interest. In the cases where particulate material is not present in the solution, sampling of liquid materials is easier with regard to homogeneity. It should be considered that when sampling e.g. wastewater or seawater, hourly or seasonal fluctuations may influence the concentration of analytes in the bulk material. If care is taken, subsequent dilution or spiking to achieve target levels for the analytes of interest might be avoided.

7.3.1 Powder/solid sampling

With powders and solids the sampling step and collection of bulk material are subject to considerable sources of error. It is desirable to predict the size of this error in advance because such knowledge would greatly improve the estimate of the overall measurement uncertainty. Understanding the nature of these errors may also help to reduce them.

Not only type of material (solid, fluid etc.), but also physical location of sampling site and species (e.g., biological RMs) are important to the error of the sampling step.

In the examples given in appendix E sampling of sediment material with a grab was mentioned several times. This type of sampling implies that the experimenter assumes the material to be sufficiently homogeneous not to interfere with the accuracy of the experimental results. Especially when bulk material is sampled, transported and treated in various ways, segregation may take place according to particle size, shape and density (Grant and Pelton 1972).

Segregation of particles is important in the sampling of bulk material. When it comes to subdividing samples for analyses in the laboratory, random variations may exercise a greater influence. The total variance of general sampling of bulk material has been described by (Grant and Pelton 1972).

$$S^2 = A/W + B/N \tag{7.2}$$

where

S^2	total variance of the system
A	sampling constant; a component of the random variance
B	sampling constant; a component of the segregation variance
W	size of the gross sample
N	number of samples collected

A is estimated from a series of small samples dominated by random effects and with only small segregation. B is estimated from a series of large samples where the segregation effect dominate. W can be substituted by N·w, where w is the sample weight.

7.3.2 Sampling exercises

The problem of sampling has been treated recently by van der Veen, Thompson, Ramsey and Argyraki among others, (van der Veen and Alink 1998), (Thompson and Ramsey 1995), (Ramsey *et al.* 995), (Argyraki *et al.* 1995)(Thompson 1998). The essence of the discussion in these articles is that proficiency testing in sampling can be carried out in much the same manner as ordinary proficiency testing. The same comparison applies to collaborative trials and the analogous collaborative trial in sampling. The object of the sampling exercises is to establish the magnitude of the sampling uncertainty which contributes to the technical uncertainty together with the analytical uncertainty. The collaborative trial and the proficiency testing in sampling differ from the traditional schemes in that the exercise only require that participants send technicians to take samples at a designated site according to a fixed protocol. The samples are analysed centrally by one laboratory in order to separate the sampling variations (between samplers) from the analytical variations between laboratories.

Composition variations between samples are caused by heterogeneity of the target and inadequacies in the sampling procedure. Samples usually differ from each other and from the average composition. This means that the differences between samplers is greater than the performance variations of each one in the same manner that interlaboratory variations are often seen to be greater than within laboratory variations in interlaboratory comparisons of chemical analysis.

(Thompson 1998) proposes a design for estimation of sampling precision which resembles the traditional nested designs used for homogeneity testing of RMs. Based on mean squares (MS) between (B) and within (W) n samples analysed according to a nested design, the sampling variance is estimated as

$$\hat{\sigma}_s^2 = (MS_B - MS_W)/n \tag{7.3}$$

which is the same type of estimate as for between samples variation in a homogeneity test. The important difference is the focus on the repeated application of a sampling protocol.

$\hat{\sigma}_s^2$ is estimated on the basis of the afore mentioned design. (Thompson 1998) argues that sampling uncertainty estimates must be based on designs especially made for that purpose, given the fact that ordinary measurements have both sampling and analytical uncertainty. The estimate of the sampling variance from ANOVA requires that the analytical variance is less than one third of the sampling variance. Another noteworthy aspect of figure 7.2 is that if the site variation constitute the majority of the total uncertainty of the analysis result, little is gained by improving the technical uncertainty in the laboratory.

7.3.3 Reference sampling targets

Reference sampling targets (RST) could be a new field for RM producers. (Thompson and Ramsey 1995) suggest RSTs for research of the sampling influence in analytical chemistry and for validating sampling protocols. This could bring an end to the general hesitation among analytical chemists with regard to ignoring the sampling uncertainty. The chemists are well aware that it is there, but few have any idea of how to cope with it. (Thompson and Ramsey 1995) suggest RSTs certified in the same manner as regular CRMs. Several sites such as natural waters, the atmosphere and even sediments disqualify as RSTs because of the dynamics involved (flow etc.). However, unused polluted land where pollution has ceased quite a long time ago, could serve the purpose. In contrast to CRMs, heterogeneity of such a site is desirable. Thompson admits that this type of undertaking would be rather costly, but nonetheless not too excessive if the needs of large industries are involved, especially seen in the light that contaminated land is often unsuitable for other purposes.

Chapter 8

Methods of production

This chapter contains a description of the steps to be taken when starting a RM production, the questions to be answered etc. Solids and liquid materials are considered.

8.1 Motivation for starting a production

A recently produced RM for pesticide analysis in ground water produced by VKI and NERI is an example of what may initiate a RM project.

The Danish ground water monitoring program of the action plan for the aquatic environment prescribes that a number of inorganic and organic compounds are to be monitored in the ground water. Pesticide analyses are complicated and many of the Danish laboratories have relatively little experience in analysing certain of these pesticides. In order to assure high analytical quality of these analyses (which will in the years to come probably comprise a list of several hundreds of compounds) a RM was needed. Particularly in the fields of environmental analyses and in food analyses the demands for official control and monitoring are changing rapidly.

As mentioned in the example above, legislation is probably one of the primary initiators for RM production. At VKI the procedure of settling the needs for a new RM is often by means of questionnaires to laboratories working with the particular type of analysis of interest.

To summarize in a few words: starting a production of a new RM is normally based upon

- a demand for a certain type of analysis to be performed
- requirements for the quality of analytical results
- the resulting need for a reference material which is not an "in-house" standard

8.2 Selection of batch raw materials

For environmental RMs it is desirable for the user that they are based on natural matrices, because the matrix similarity with real samples is one of the great advantages of RMs over in-house standards of pure compounds. This is not always possible, however desirable it may be. The starting material may therefore be either

- natural batch material:
  - seawater
  - ground water
  - sludge
  - sediment etc
- synthetic material: solutions based on purified water water for instance

The natural batch material may be taken as it is or spiked[1] to achieve certain target levels for the compounds of interest. Synthetic materials based on concentrates may be preferred for testing parts of the analytical process and identify irregularities related to the calibration procedures, the instruments or extraction and sample digestion. The combination of

[1] See General Glossary for explanation

spiking the unknown sample and using a synthetic RM based on a concentrate may be an advantage because the matrices of environmental samples can differ considerably depending on local conditions.

In some cases either the producers or the users of RMs will have experience with samples from a number of locations through daily work in the laboratory. An experienced chemist would therefore have one or several ideas about where to look for candidate batch raw material.

The aspects needed to consider at this point depend heavily on the type of material to be produced and its purpose as it has been identified through questionnaires and the like. Taking wastewater as a case example, the issues to consider at this point could be:

- ☞ Should the water be taken from a plant in an area which is mainly industrial, urban or rural ?
- ☞ Should it be taken in the inlet or the outlet from the plant ?
- ☞ Do the locations under consideration have compound concentrations in the target area ? Will spiking be needed ? Is it desirable to spike ?
- ☞ What is the sample volume needed per analysis
- ☞ Should there be particulate matter in the water ?
- ☞ Are there any experiences with analytical difficulties (interferences or matrix effects) for one or more of the locations under consideration (which may then be ruled out) ?

Another example could be a sediment for trace metal analyses in which case the producer would need to consider points such as

- ☞ Matrix composition: what should the contents of sand, clay, organic matter etc. be ?
- ☞ Which locations are easily accessible and is special authorization needed?
- ☞ Are the possible locations for sampling equally representative of the analytical problems to be solved by means of the RM ? I.e. is the matrix relevant and are the concentration levels suitable ?

☞ Are there (trace) metals present in the candidate materials ?

☞ Will spiking, if needed, be feasible with regard to homogeneity and stability ?

☞ Is the particle size distribution in the candidate material(s) close to what is required for analytical purposes or is much grinding and sieving needed ?

☞ How can sample homogeneity be achieved ?

The above mentioned examples are not meant to be exhaustive in their areas, but only to illustrate that specialist knowledge and experience is necessary to make these decisions. Unless there is absolute certainty about the optimal qualities of one particular location it is recommendable to include several of different materials in the initial experimental phases. Difficulties may show up unexpectedly. If no specialist knowledge is available, exhaustive and expensive experimentation will be needed to test numerous points like those just mentioned.

8.3 Containers

The choice of container type depends on

- sample matrix (liquid or powder)
- stabilization procedure (lyophilizing, autoclaving, chemical preservation)
- possible reactions between the sample and the container material

etc. In the following two alternatives, glass and plastic, are discussed in relation to liquid RMs. For powders there will be other relevant containers such as metal cans.

8.3.1 Glass

The question of inertness of the container walls is difficult to answer satisfactorily, because the term "glass" is not specific and because of the many variations in liquid matrices and analytes that may be filled into bottles. In the present study it has been a general impression that little is known about how commercially available containers interact with various aqueous solutions. The composition of a certain material used for containers can only be known to a certain degree because, the manufacturers will not or cannot provide detailed information as glass is a complex material produced from natural sources. This leaves a tremendous gap in the field of knowledge about container-fluid interactions and container stability. Information of this is very hard to get in print, nor do experts in glass manufacturing for instance know much about glass stability. In the present context, glass stability is particularly interesting with regard to phosphorus in seawater.

The physical chemistry of glass surfaces exposed to liquids is worth a study in itself. The following is not an attempt to describe the subject fully. Any durable glass surface almost always develops a surface film with a different concentration than the bulk material (Hench 1977). This happens immediately after manufacturing. The composition and qualities of this surface film is important with regard to the durability of the glass when exposed to different types of liquid. Depending on the physical conditions (such as pressure, temperature and pH) and the chemical conditions (presence of ions) diffusion processes may take place at the surface film. One of the aspects of glass behaviour of relevance to the present subject is that surface reaction processes of glass are altered during autoclaving.

(Hench 1977) summarizes five different types of glass which he characterizes according to the state of the surface film. The regarded characteristics are: the thickness of the surface film, its concentration of silica and its resistance towards attack all determine the durability of the glass.

(Hench 1977) describes durability tests of glass exposed to autoclaving and notes that the surface reaction processes are altered during autoclave testing. The reactions during acceleration of glass durability tests by autoclaving do not correspond to the surface changes at low temperature aqueous exposure of glass. However, the observations are relevant for the present purpose.

The autoclave test was performed on commercial container glasses autoclaved at standard conditions for 1 hour[2]. At room temperature aqueous exposure leads to dealkalization and a hydrated silica surface[3] During the autoclave treatment the hydrated silica goes into solution along with Na^+ and Ca^{2+} species[4]. Hydrated silica has low solubility in neutral solutions at 37°C, but silica is found in solution at autoclaving temperature levels (121°C), (Hench 1977).

(Simmingsköld 1958) describes experiments on glass powders exposed to water attack at 97.5°C. (These conditions are probably somewhat comparable to stability tests in glassware at elevated temperature, whereas the afore mentioned reference, (Hench 1977), is more relevant for the RM preparation stage involving autoclaving stabilization.) The flow experiments described, show that at a pH value of 9.6 SiO_2, Na_2O and CaO are released. The Si release is the largest and the calcium release is the lowest measured over a 120 minute period. The same results are seen for all tested glass types also at higher pH values.

At extreme pH values (pH 12-13) ion exchange of Na^+ at the glass surface may take place followed by Ca^{2+} exchange. None of the samples studied here are in this extreme pH range and would probably not attain such levels even during autoclaving.

If an aqueous RM is intended for any of these ions, it may be advisable (depending on the pH of the solution) to consider possible effects of glass-water interaction and autoclaving. Some of the glass powders used in the experiments of (Simmingsköld 1958) yielded from 5-15 mg SiO_2/l up

[2] 1 atmosphere, 121°C

[3] The process is called a stage 1 reaction which is dominant for solutions of pH < 9. An exchange of H^+ ions with alkali ions takes place at the surface.

[4] The network dissolution is called stage 2 and dominates at pH > 10. Si-O-Si bonds are attacked by OH^-.

to 35-50 mg SiO_2/l with increasing pH values above 9.5. (Blomqvist *et al.* 1993) report complicated interferences from silicate (fluoride and arsenate) in the antimonylmolybdenum blue method also used by Danish Standard for phosphate analysis. Silicate leaching may not be a problem in general, but on the other hand may be important for certain types of aqueous RMs.

(Aminot and Kérouel 1995) also notes dissolution of silicate from glass ware (100-150 μmoles/l) and a positive effect on phosphate during autoclaving. The silicate is accompanied by phosphate leaching. Aminot states that phosphate dissolution from glass walls influences the stability of phosphate in autoclaved marine samples. The phosphate leaching is governed by pH.

(Aminot *et al.* 1992) state that in connection with determination of silicate, glass is attacked by natural waters and by seawater in particular. Traces of phosphorus in the glass container may therefore be of interest.

Of the glass types tested by Aminot and co-workers, borosilicate dissolves slower than ordinary soda-lime glass. The seawater dissolution rate of glass bottle walls is temperature controlled, and at the same time some phosphorus from the glass is leached into the solution. The primary reaction is silicate leaching and the phosphorus leaching is an accompanying reaction. One type of bottles used in the experiments described in (Aminot *et al.* 1992) was a soda-lime glass bottle. These bottles had been used for seawater storage for 2 years prior to the experiments. The silicate and phosphate leaching attained the highest levels in these bottles of all the glass types tested (up to 1000 μmol/l silicate and up to 0.2 μmol/l phosphate). Aminot and co-workers also found that if crystals formed during the experiments, precipitate phosphate was found. The precipitate was assumed to be co-precipitation of calcium carbonate and phosphate.

With regard to long term storage of marine phosphate samples, (Koroleff 1983) discourages the use of glass containers because the potential uptake of phosphorus by the glass. (Koroleff 1983) does however, recommend glass for nitrite, ammonia and urea.

The bottles used in the present study (for both seawater and wastewater) are made of bor-silicate. The manufacturer had treated the surface with ammonium sulphate during production to improve the hydrolytic quality of the glass.

8.3.2 Plastic

Containers of plastic for aqueous RM are at present not a realistic alternative to glass because of the potential difficulties of ensuring

- ✘ transportation safety
- ✘ sterilization
- ✘ air tightness

Ampoules of plastic like the type used for artificial tears for contact lenses could be an interesting alternative, since units of plastic weigh less and are easy to open. Such an approach may call for batch sterilization which presents a number of problems. It is e.g. difficult to autoclave large amounts of liquid with standard autoclaves. Autoclaving on an industrial scale is a possibility if economically feasible, but the following filling of plastic ampoules or the like would require strict precautions against contamination. In all, such a procedure may not be economically nor practically feasible.

(Koroleff 1983) recommends intermediate storage of seawater in glass bottles when phosphate analyses are to be performed. Otherwise (Koroleff 1983) advises preservation with sulphuric acid and storage in polyethylene containers. Deep freezing in polyethylene bottles is also recommended .

The conclusion of the study of containers is that with regard to the compounds in focus in the present work, glass is adequate for liquids containing nitrogenous compounds, whereas plastic must be preferred for all liquids containing phosphorus compounds. These conclusions agree with the practical experience at VKI.

8.4 Stabilization

Stabilization is a necessary measure when working with natural matrices. Wastewater and sludge may constitute a health risk if not treated. For these types of matrices and for marine matrices as well, biological and chemical activity may cause stability problems. Therefore measures must be taken to stop all biological and chemical activity of the batch material in order to produce homogeneous and stable reference samples. In the following a number of different preservation methods are discussed.

8.4.1 Autoclaving

Autoclaving is an extremely efficient bacteria killing process for liquids. The autoclaving process causes exponential reduction of the number for bacteria. The treatment is time and temperature dependent. With the standard treatment of 121°C for 20 minutes, sterile conditions should be obtained in most cases. However, since the reduction is exponential it might, in some cases, be considered to autoclave several times to obtain a sufficiently stabilized material.

In addition to killing bacteria, the autoclaving may change enzymes which were present in the aqueous media from the beginning or may have occurred as a result of cell rupture during filtration. (Aminot and Kérouel 1995) summarize experiments with autoclaving of purified water, and spiked seawater. The waters were taken from various locations. The experiments showed that nitrate concentration up to 50 μmole were not affected by autoclaving, whereas a minor effect was seen for nitrite. Ammonia contents increased between 0.1 and 0.3 μmole/l at all tested concentration levels. The rise in ammonia is attributed to dissolved organic matter content in the water. Successive autoclaving was performed. Purified water remained uninfluenced, while ammonia levels as a result of hydrolysis of dissolved organic nitrogen increased in what appeared to be a kinetic-controlled process. Aminot also notes significant positive effects on phosphate levels. Internal processes such as hydrolysis of complex forms and precipitation of calcium phosphate combined with leaching of phosphorus from the bottle walls are offered as explanations. The processes are thought to be competitive.

The fact that autoclaving may affect concentration levels is not crucial with regard to RM production where the issues are homogeneity between sampling units and stability, rather than the exact original concentration in the environment. Autoclaving appears to have a number of advantages with regard to aqueous material because this stabilization method prevents further enzymatic activity in the material - activity which might otherwise cause hydrolysis of organic nitrogen or phosphorus compounds.

8.4.2 Irradiation

Radiation is often used in food industry especially for killing organisms in spices where neither chemical preservation nor heating or freezing treatment will ensure satisfactory biological inactivation. An example of Co-radiation of mussel tissue is given in appendix E.4.5. Radiation of bottled or ampouled aqueous material suffers the difficulty of ensuring equal dosage for all single units.

Aminot and Kérouel note unsatisfactory results using irradiation with accelerated electrons and state that since enzymes can be very resistant to irradiation, they do not consider this method recommendable for materials where enzymatic activity could be suspected of changing the concentration of compounds of interest (Aminot and Kérouel 1995).

Irradiation may be appropriate for some particulate materials for environmental or food reference materials, but appears to be a less frequently used method. (See also appendix E.) One reason for this could be that irradiation processes are accompanied by the formation of free radicals which can enhance unwanted chemical reactions in the material. Another reason for this might be import restrictions on irradiated material and consequently a concern for sales figures with the RM producers.

8.4.3 Drying

Drying of various sorts is frequently used in RM production (Appendix E) -especially for food and biological RMs. Methods applied regularly are

- ✘ oven drying
- ✘ spray drying
- ✘ freeze-drying (lyophilizing)

Biological material is often oven dried and ground, whereas RM based on liquids (milk powder e.g.) is spray dried. The method of drying must be selected according to the sensitivity of the material in question.

Particularly in cases where organic compounds are of interest, drying at elevated temperatures may cause loss of compounds of interest or of dry matter in general (organic carbon). In such cases freeze-drying where water evaporates at low pressure and temperatures of about -50°C may be a recommendable and gentle method. It appears to be a frequently used technique for food RMs (appendix E).

(Bortlisz and Velten 1994) show an example of naphtalene determination in sludge from biological clearing tanks, where the loss is almost 100 % when samples are dried at 105°C, whereas freeze-drying preserved the naphtalene content of the samples. Also, loss of dry mass was shown to be larger for heat dried than freeze-dried samples.

The suitability of freeze-drying is obvious when organic compounds and dry matter content of the material are of interest, however three aspects should be kept in mind:

- Volatile compounds will still be at risk of evaporating.
- The material is much more hygroscopic when freeze-dried as compared to heat dried, because of the extremely large surface.
- It is not a sterilization method, which means that contamination/health risks persist in the case of sludge, waste water etc. Subsequent heat treatment of some form (at less than 100 °C) may be called for.

Spray drying has not been considered any further since it has not been used in VKI production of RMs up until now. One of the reasons for this is that most spray drying equipment in Denmark is used for foods and related products. The production of environmental RMs often involve natural matrices which are not considered safe in a bacteriological respect, therefore the owners of spray drying equipment are likely to be rather reluctant to accept a task of spray drying this type of materials.

8.4.4 Chemical preservation methods

Whereas some of the above mentioned methods of stabilization modify the matrix physically to some extent, chemical modification does not necessarily influence the physical characteristics of a sample (pH being one exception). Chemical preservation should be applied only with the greatest consideration, since obviously there are risks of

- contamination
- matrix modification
- analytical inference

along with the risk of unwanted precipitation. Furthermore, chemical preservation might not stabilize the material immediately or may not have the full desired effect on bacteria etc. In other cases chemical preservation is a must to preserve the compound of interest. The preservation can be applied either to provide a specific pH environment or bind compounds or elements in specific compounds adequate for the purpose of the material. Chemical preservation may be combined with the methods mentioned above to promote/enhance homogeneity and stability.

Particularly with regard to nutrients in marine waters (Aminot 1994) adds the following to the previously mentioned risks connected with chemical preservation: with marine waters, acidification may cause loss of nitrite in the form of dinitrogen because the former reacts with ammonia or primary amines. There is also a risk of hydrolysis of condensed or organic forms of phosphorus. (Aminot 1994) recommends acidification only for filtered samples intended for silicate determination. Mercuric chloride may be used for low turbidity marine waters, but represents potential problems with analytical interferences in some methods for ammonia and nitrate determination.

Preserving samples in glass containers with acid or base can lead to hydrolysis processes if the final pH level is far from the neutral region of pH 7. This should be considered before choosing a combination of chemical preservation and glass bottles/ampoules.

8.4.5 Freezing at -18 to -23°C

The paper by (Dore *et al.* 1996) is focused on the determination of in situ values of nitrite, nitrate, soluble reactive phosphate and silicate with and without preserving samples by freezing. (Dore *et al.* 1996) conclude that freezing is an adequate method of preservation of unfiltered samples for all of the mentioned measurands. (As mentioned in section 8.3 (Koroleff 1983) also recommends freezing of phosphate samples in polyethylene containers.)

A few deviations are found in (Dore *et al.* 1996)'s research: in the comparison between fresh and frozen samples of low level soluble reactive P (phosphate), a slight increase was seen in the frozen samples. The effect, however was negligibly small with standard manual or automated colorimetric methods. For low-particulate waters it is concluded that no prefiltration was necessary and no changes greater than two times the analytical variability for N, P and Si were detected over a one year storage period.

The conclusions are restricted to low-particulate oligotrophic oceanic seawaters. The authors note that nutrients in other types of waters may encounter difficulties such as precipitate formation and silicate polymerization when frozen.

Freezing of aqueous samples is questionable, because of the severe risk of insoluble salts. (Aminot 1994) mentions that phosphate can decrease when samples are frozen and that silicate polymerizes increasingly with dropping salinity levels. There may also be rare biological processes which do not stop at low temperatures and volatile organic compounds may also be lost at these conditions. Synthetic samples of exact known composition may be frozen with less concern, whereas freezing of natural samples especially the aqueous types should be considered carefully. If the risk of precipitation etc. is minimal, freezing can be an attractive alternative to e.g. chemical preservation, because the matrix is maintained unaltered without any risk of introducing contaminants by the preservation chemicals. As a means of preserving RMs, freezing constitutes a serious problem with regard to transportation. It would probably be a rarely used means of stabilization.

8.5 Initial experimental phase

The purpose of the initial experimental phase is of course to answer all the questions that rose as s result of the considerations stated above. Desk work will rule out some solutions, but at some point the action must be moved into the laboratory in order to narrow down the possibilities.

8.5.1 Statistical approach

A fractional block design is an efficient and fast way of identifying important factors. Further investigations of interactions can be done in a second stage. Suppose the desk work has narrowed down the factors of interest to be

- 3 possible locations for sampling batch material
- 3 types of stabilization
- 2 types of containers
- 2 options for spiking

A reasonable approach in this case would be to split the investigation in two parts. Part I would investigate batch materials and stabilization methods and Part II would explore the possibilities for containers and spiking, based upon the results from Part I.

Part I is a 3^2 full factorial design which in most cases can be split into blocks of three for analytical comfort.

Factor	Description	Levels
A	type of stabilization	0,1,2
B	batch of material*	0,1,2

Table 8.1: *Batch raw material is considered a factor in the main part of this section because it is assumed that the origin of the batches is well known and in that sense reproducible. Furthermore, it is assumed that the batches are selected on presumption that they differ with regard to characteristics and qualities.

Each type of batch material is combined with each of the stabilization methods. The actual preparation will in most cases have to be carried out according to available man power and equipment capacity.

Assuming that any two way interaction between A and B is of minor interest at this stage, the confounding could be chosen by the defining alias relation I=AB^2 which results in the following experimental design.

		A		
		Method 0	Method 1	Method 2
B	Batch 2	b^2 (I)	ab^2 (II)	a^2b^2 (0)
	Batch 1	b (II)	ab (0)	a^2b (I)
	Batch 0	1 (0)	a (I)	a^2 (II)

Table 8.2: 0, I, II indicate in which block the combination belongs.

The experiment indicated with "ab^2 (II)" in the second cell of the first row is the experiment where the level of factor A is the intermediate, a, for method 1 and factor B has the level 2, b^2, which is batch two. The combination ab^2 of batch 2 and method 1 is performed in block II. The power of the design will increase noticeably if more than one sample is analysed in each cell and with replicate analyses.

However, it must be stressed that at this point failed experiments are very likely to occur. Suppose that one or more of the stabilization methods is a chemical preservation step. If the batch material is some type of natural water, precipitation may occur and thus render some of the cells empty.

A number of methods exist for estimating missing values. (Montgomery 1997a) proposes an approximate analysis for the estimation of one or more missing observations by minimizing the error mean square. The risk of this method is that it gives too many significant results because the mean square for methods (regarding batch raw material as blocks) is biased. An alternative is the exact analysis also suggested by Montgomery, which makes treatment comparisons possible, however with precisions that are not the same for all combinations. Whichever approach is chosen, great care must be taken in the evaluation of the statistical tests since the empty cells represent combinations of batch raw material and stabilization method which have not been realized.

Part II of the experiments is a 2^2 full factorial consisting of minimum 4 samples to be analysed with regard to the relevant analyte(s). Again, the power of the design is improved by preparation of several samples for each cell and by performing replicate analyses. Thereby, it is also possible to get an early estimate of homogeneity between units.

The approach given in this section for the example of 2x3 + 2x2 factors (on three and two levels) is hypothetical. It is difficult to make good experimental designs for combinations of 2 and 3 levels. Descriptions of such experiments can be found in (Kempthorne 1952).

The realistic approach

As already mentioned above moving from the desk into the laboratory with a premade experimental design may be a wakening to a very real world, where liquids precipitate, analytes are destroyed by stabilization methods or adsorbed to container walls. In short all that is unforeseen. In spite of these expected empty cells in the data schemes, the systematic statistical approach is recommendable.

In the 3^2 example given in the previous section of investigating stabilization methods and batch raw material the actual chemical analysis of the analyte(s) may not be carried out. The design table may simply serve as a tick box of "feasible" or "not feasible" for the various combinations.

The realistic approach to the initial experimental phase consists of equal portions of systematic planning, preparedness for the unexpected and readiness to change course according to results.

Chapter 9

Case: lyophilized wastewater RM

This chapter presents the results of the experimental work with freeze-drying of wastewater.

The attempt to produce freeze-dried RM reported in this thesis is an example of stepwise experimental development. As such it is an example of how work could proceed in the research of RM. The starting point is an idea or a problem which needs to be solved. Experiences are made along the way and changes the course of the experiments and the preparation procedure.

Detailed description of the experiments can be found in appendix C. This chapter will focus on the overall course and outcome of the experiments.

9.1 Motivation

It is a general experience from preparation of samples for interlaboratory studies that aqueous samples in glass containers turn inhomogeneous with regard to P (phosphorous) compounds. The reason for this is likely to be adhesion of phosphorous compounds to the container walls or some kind of interaction between these compounds and the surface of the container. In a production of RM based on natural wastewater, inhomogeneity of P compounds is a potential problem from the beginning due to particulate matter.

The purpose of the experiments was to investigate the possibility of preventing the inhomogeneity problems caused by adhesion to the container walls by changing the state of the wastewater. Freeze-drying the solid material out of the water might produce a material which could be redissolved and analysed immediately. In this manner the material should theoretically be unable to change during the time from production to use.

9.2 Experiments

The freeze-drying was performed on wastewater from Usserød Wastewater Plant. The water had been stored at 4°C for some months before freeze-drying. Not all experiments have been performed with repeated measurements.

A time scale of the experiments is shown in figure 9.1. The first experiments were performed with the object of evaluating the process of re-dissolving the dried material. It was also of interest to know the phosphorous and nitrogen levels in the samples with regard to dilution prior to analyses. Some analyses were performed of nitrate and ammonium, but the study was focused on ortho-phosphate and TP since homogeneous samples for nitrogen compounds can be produced as aqueous samples by traditional methods. The following sections describe the motivation for and details of the steps taken in the experiments. Readers interested in a detailed description of the experiments are referred to appendix C which also contains the output results of the statistical analyses and a number of data plots.

9.2.1 Selection of raw material

It was decided to work with water from Usserød Wastewater Treatment Plant, since the aqueous samples studied in this work originated from the same location.

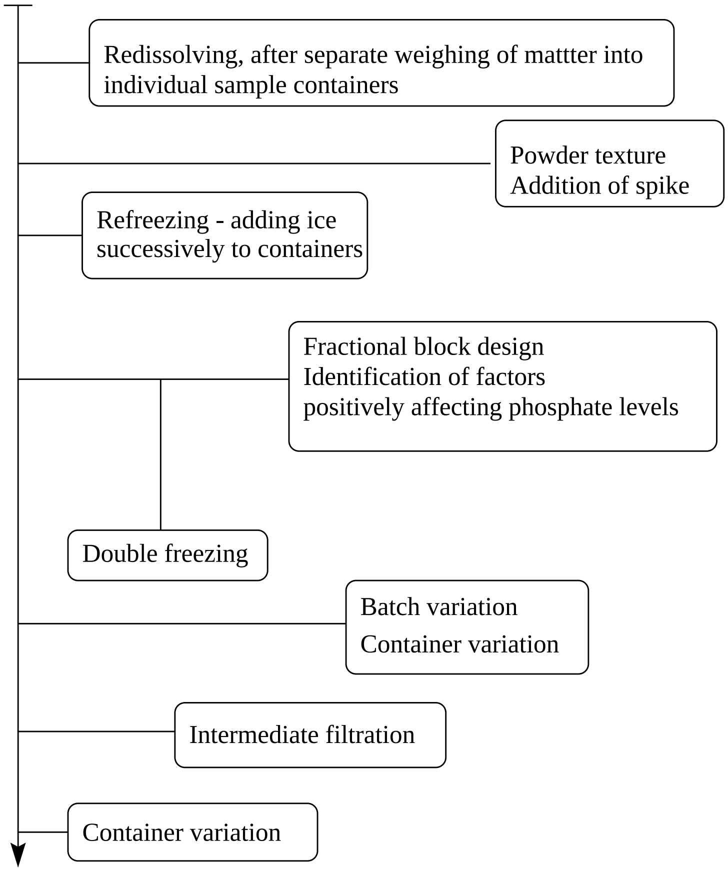


Figure 9.1: Time perspective of freeze-drying experiments.

9.2.2 Freeze-drying

Freeze-drying was carried out at VKI using existing equipment, which is, from a practical view point, an on-off device. The samples are placed in glass dishes under a plexi-glass globe and a vacuum pump reduces the pressure thereby reducing the temperature to about -50°C. Under normal circumstances VKI procedures prescribe that samples should be placed at -70°C for 30 minutes prior to freeze-drying. For practical reasons and due to VKI contamination security procedures, this step was omitted since the experiments involved wastewater. The samples were instead placed in a regular freezer at -18°C prior to freeze-drying.

9.2.3 Powder texture

Sometimes, it is necessary to spike a batch of RM. In the case of aqueous RMs made artificially, spiking is the principle of production. It is not unlikely that natural waters may need spiking for one or more parameters, because the criteria for selecting a certain type of material may collide with the desire to keep a certain concentration level for one or more analytes.

The addition of spikes could change the characteristics of the freeze dried material in one or several ways.

An ammonium spike was added to the wastewater prior to freeze-drying. It was noted that the consistency of the powder changed from being very 'fluffy' and 'fly away' when no spike was added to something a little more manageable and dense with the added spike. Other than the physical appearance of the powder, no other effects were seen with regard to ortho-phosphate homogeneity. Therefore the conclusion can be drawn that a spike addition can improve the physical quality of the powder without analytical side effects - provided of course that the compound for the spike does not have any known matrix effect with regard to ortho-phosphate/total phosphorous (TP) (or the measurand of interest). In fact the addition of a non interfering spike or some "blank" re-dissolvable additive, which will increase the amount of dry matter resulting from the freeze drying, could make the material easier to handle.

9.2.4 Re-freezing

Instead of removing dried material from the glass dishes, consecutive freeze drying cycles were run, where new ice was added on top of already dried powder. This is described further in appendix C.

9.2.5 Fractional block design

A fractional block design was carried out to identify preparational steps which had a positive effect on orto-phosphates/. The following factors listed in table 9.1 were included: spiking, consecutive freeze-drying - filling new ice into freezing containers without removing already dried powder, dissolving dry sample in sulphuric acid then purified water, and ultra-sound treatment. The low levels of these factors were: no spike, emptying of containers before new ice was to be dried, dissolving sample in purified water then preserving with sulphuric acid, and shaking of sample instead of ultra-sound treatment.

Factor	Description
A	Addition of spike or not prior to freeze-drying
B	Consecutive freeze-drying cycles without removing dry sample or removal of dried sample before new cycle
C	Addition of sulphuric acid and Mili-Q water in the dissolution step or dissolution only in Mili-Q water
D	Ultrasound treatment or shaking of sample
E	Glass or plastic container for dissolved sample

Table 9.1: Outline of 2^5 factorial experiment.

The full experiments consits of $2^5 = 32$ experiments in which n samples should be analysed with repeated measurements. The immediate purpose of the experiment was to determine the main effects of the above mentioned factors. Thus the full factorial design was reduced to $2^{-2} \cdot 2^5 = 8$ experiments. In each of the eight experiments 1 sample was analysed with 3 replicate measurements. Having chosen the principal fraction of the full design, the following experiments were carried out:

	Description				
Name	Spike	Re-freezing	Dissolving in H_2SO_4	Sample prep. before analysis	Container
1	-	-	-	shaking	glass
cde	-	-	yes	ultrasound	plastic
bde	-	yes	-	ultrasound	plastic
ad	yes	-	-	ultrasound	glass
abe	yes	yes	-	shaking	plastic
ace	yes	-	yes	shaking	plastic
bd	-	yes	yes	shaking	glass
abcd	yes	yes	yes	ultrasound	glass

Table 9.2: Principal block of fractional block design.

Statistical analysis of the experiment was carried out according to the following model where all indices run from 1 to 2 except $n = 1, 2, 3$. Letter codes are as indicated in table 9.1.

$$Y_{ijklmn} = \mu + A_i + B_j + C_k + D_l + E_m + AB_{ij} + AC_{ak} + E_{ijklmn} \quad (9.1)$$

The output of the statistical analysis is shown in appendix C, section C.1. The factors A (spiking), B (re-freezing) and D (shaking) were found to be significant with regard to ortho-phosphate concentration. These levels of the factors influenced phosphate levels in a positive direction. Ultrasound had a negative effect on the ortho-phosphate concentration level and was dismissed immediately.

For the remaining experiments, spiked, re-frozen powder was used and samples were shaken prior to analysis. It was decided to adopt the procedure of dissolving in sulphuric acid before adding purified water to samples although ortho-phosphate concentration levels were not influenced.

9.2.6 "Double freezing"

On the basis of the initial experiments it was clearly seen that too much uncertainty was connected with weighing of dried material into each sample container. It was impossible to reconstruct aqueous samples with suitable homogeneity in this manner and another approach was taken. Instead of weighing matter in individual portions into each sample container, a concentrate was made. An amount of sample was weighed to attain an approximate target level of 100 μg ortho-phosphate/l. The sample was dissolved, filtered and the concentrate transferred by pipette into the individual sample containers. The sample containers with the concentrate were placed in the freeze-dryer and the concentrate was dried again. After freeze-drying the samples could be re-dissolved with purified water and analysed as desired directly in the containers.

9.2.7 Batch and container variation

A two-factor factorial experiment was carried out to investigate whether there would be a difference between freeze-drying cycles run on different days. A preliminary test of glass versus plastic for the redissolved samples was included here to construct the two-factor factorial.

The experiment involved two batches of material freeze-dried under the same conditions, but produced with a 4 week interval. Blue-cap bottles washed in diluted hydrochloric acid and ordinary plastic containers used for ortho-phosphate/TP analysis were used. The output of the statistical analysis is shown in appendix C section C.1.

The experiment showed that there was no difference between batches produced at different times. A clear distinction could not be made between glass and plastic containers.

Further experiments testing various types of containers were carried out in search of reproducible results which could show satisfactory homogeneity between samples. In some cases the statistical analysis would reveal homogeneity between samples and differences between types of container for ortho-phosphate/TP, but the reverse conclusions had to be drawn for TP. In the cases where homogeneity was achieved the statistical models fitted data poorly (R^2=0.5-0.75) and the noise contribution (MSE) was large.

9.2.8 Filtering

Since filtering is not normally part of the procedure in the VKI ortho-phosphate/TP analyses it was decided that filtering should not be performed after the final dissolving of dry matter. However, it was found that an intermediate filtration step of a concentrated solution improved homogeneity in the final step. This is mentioned in section 9.2.6 about "double freezing" of material.

9.3 Choice of containers

The fact that natural samples present homogeneity problems with regard to ortho-phosphate/TPs when kept in glass containers, raises the question whether plastic vials could solve the problem for the aqueous samples.

Part of this study has included preparation of dry samples in plastic containers, but no attempt has been made to produce aqueous samples in plastic. The experiences made in these experiments regard redissolved powder samples. It is therefore uncertain whether the observations made from these experiments can be expected to hold for fresh samples in plastic vials.

In the freeze-drying experiments both glass bottles rinsed in diluted HCl and plastic vials were used in order to investigate whether there was a significant effect on the redissolved sample.

9.4 Statistical Analyses

The outlines of the experimental designs are summarized in appendix C and a summary of the analysis of variance (ANOVA) of the results is found in table C.2. The full output of the statistical analysis is shown in appendix C section C.1.

9.4.1 Outliers

Grubb's and Cochran's outlier tests were applied (ISO 5725 1994a). One Grubb's and one Cochran's outlier were found in two experiments on container type. The outliers are marked in the plots on figures C.9 and C.11.

9.4.2 Homogeneity of lyophilized wastewater

Table C.2 in appendix C shows that MSE (Mean Square Error) of the ANOVA has been reduced considerably during the course of the experiments. At the same time the proportion of the variability explained by the analysis of variance model has decreased. In all of the statistical analyses presented in this table, the variation between samples has been tested against the residual variance of the model.

Method	QC Type	$s_{synthetic}$	Experiment	MSE
DS 291 Orthophosphate	WW2.1	1.01	15005-1997	1.39^2
			2006-1997	2.00^2
			2508-1997	0.89^2
			2201-1998	0.61^2
DS 292 Total phosphorous	WW3	1.66	2006-1997	0.39^2

Table 9.3: Uncertainty analytical methods (P analytes) and MSEs of the ANOVAs. $s_{synthetic}$ from VKI control charts, it includes within and between day variation. All units are in μg/l.

An alternative approach to this could be to compare the sample mean square to the variance of VKI quality control charts for synthetic samples of total phosphate and ortho-phosphate. As shown in table 9.3, the mean square errors of the ANOVA's of the most successful experiments are comparable to the uncertainty of the analytical methods. The comparison of the sample mean squares to the shown MSEs are fair in the sense that the order of the model error corresponds to the uncertainty of the analytical method ($\sqrt{MS_E} \sim s_{synthetic}$).

Sample homogeneity with regard to ortho-phosphate/TP was accepted for the experiments 2006-1997 and 2508-1997. The outcome of the experiments have lead to the following recommended procedure for the production of freeze-dried RM for ortho-phosphate/TP analyses:

- ☞ GF/A filtering of wastewater which has been stabilized (stored at 4°C for some time).
- ☞ Spiking if desired to attain target concentration levels for one or more analytes, or for the purpose of stabilizing the texture of the powder (in the present case, wastewater was spiked with NH_4-N to a concentration level of 250 μg/l.
- ❄ Freezing in small portions at -18°C, e.g. in ice cube bags.
- ❄ Freeze-drying in successive cycles, adding new ice (cubes) on dried material
- ❋ Storage of the powder in a desiccator.
- ☞ Dissolution of a weighed amount of powder in 4M H_2SO_4, followed by purified water to attain an appropriate concentration of orthophosphate/TP.
- ☞ Shaking for minimum 30 minutes.
- ☞ GF/A filtering of concentrate.
- ☞ Transferring of concentrate by pipette to sample plastic vials, which can be used for the subsequent chemical analysis.
- ❄ Freeze-drying of samples.
- ☞ Redissolution of dried samples by the addition of 4M H_2SO_4 (for the preservation of samples according to the analytical method prescription) followed by purified water to attain sufficient sample volumes for replicate analysis according to DS 291 or DS 292.
- ☞ Chemical analysis.

The question of whether to prefer plastic containers over glass has not been answered satisfactorily in the experiments. Furthermore it would have been desirable to have obtained several replicates of the successful experiments in which the null hypothesis of sample homogeneity could be accepted. Due to the limited time resources in this project it has not been possible to follow up with further experiments. The course of the work has shown that the residual variance was considerably reduced, as the steps of the above outlined procedure were included in the sample preparation. Furthermore it has been shown that the addition of a spike improved manageability of the powder. Pipetting a concentrate into individual sample containers in which the subsequent analysis can be conducted directly, greatly improved inter-sample homogeneity. On the basis of these experiments it can be concluded that there are possibilities of producing a homogeneous samples for ortho-phosphate/TP on the basis of natural wastewater, but some further experimentation will be needed to confirm this. The economical aspect of such an undertaking will not be discussed here.

Chapter 10

Homogeneity testing

This chapter discusses the statistical aspects of investigating the homogeneity of a RM.

The term homogeneity has a slightly different meaning depending on what it refers to (see figure 10.1). This chapter is concerned with the statistical aspect of homogeneity. The purpose of testing statistical homogeneity is to verify that all sampling units from a batch of RM can be considered identical, or if a heterogeneity (in terms of difference between sampling units) is detected - to state its contribution to the overall variability of the material. Physical homogeneity is of importance for certain types of RMs, but is not dealt with any further in the present context.

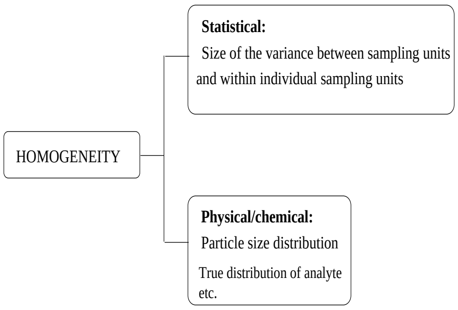


Figure 10.1: An example of the contextual meaning of homogeneity.

10.1 The levels of homogeneity testing

The problem of statistical homogeneity can be viewed from two angles: The micro approach which regards homogeneity within sampling units and the macro approach concerned with between sampling units homogeneity.

Aqueous low particulate samples are at the macro level, whereas intra-unit homogeneity is more critical for dry materials. Because of the heterogeneous physical nature of these types of materials, the purpose of homogeneity testing is to measure the degree of statistical homogeneity. With regard to liquid RMs, especially those prepared as solutions, the main purpose of homogeneity testing is to reveal impurities which may interfere with the analytical procedure, because in most cases homogeneity of an analyte, for example, can be expected based on physical grounds (ISO Guide 35 1989). Within-unit variation must be investigated if each sampling unit of RM is intended for sub-sampling. In such cases part of the purpose of homogeneity testing is to establish the minimum sample size which guarantees a certain required homogeneity between subsamples from a sampling unit. Recent developments in chemical analytical techniques has changed the conditions for homogeneity testing of RMs. Improved methods imply that formerly undetectable small differences between sampling units become detectable today. Homogeneity testing can be carried out in two steps:

- ☞ Homogeneity established or examined by a manufacturing laboratory.
- ☞ Homogeneity evaluated on the basis of the certification procedure when a nested experimental design is used.

In many cases data from the certification procedure will be used as an alternative confirmation of homogeneity. There are, however, problems in relation to the use of certification data as will be discussed further in section 10.1.2.

Sampling for homogeneity testing performed by the manufacturer may be performed in one of three ways:

- ☞ Random sampling if no particular circumstances prevail

☞ Systematic sampling, e.g. from a filling line, if time variations can be expected

☞ Stratified sampling for cases where there is a natural grouping of sampling units, e.g. batch preparation, autoclaving in portions smaller than the entire batch size etc.

10.1.1 Homogeneity testing by a manufacturing laboratory

The question of homogeneity testing performed by the manufacturer has been contested recently by (Lamberty *et al.* 1998). The authors hold the point of view that the degree of heterogeneity seen in relation to the size of the certified uncertainty interval is more important than the outcome of a hypothesis test of homogeneity based on MS-ratios. In other words - the rejection or acceptance of the homogeneity hypothesis is not the central point, but the relative size of the material variability in relation to the certified uncertainty interval is. This point of view stresses the **pooled** residual of all participants in a certification exercise as the most important: the pooled residual is used for constructing the certified uncertainty interval, and the homogeneity test involves the comparison of the inter-sample mean square to this interval. The subject was touched upon earlier in sec. 6.1, page 83 where the proficiency testing criteria of the estimated sample variance was described: s_s^2 being less than $0.3\sigma_{certification}$.

(Lamberty *et al.* 1998) point out that between- and within-sample homogeneity should be tested by a method with high repeatability. Furthermore the experimental uncertainty which includes between-sample uncertainty and method uncertainty should be corrected for within sample homogeneity.

10.1.2 Homogeneity testing on the basis of certification data

As specified in (ISO Guide 35 1989) certification data can be used as a second verification of material homogeneity. However, laboratories participating in certifications, proficiency testing etc. do have different repeatabilities: within laboratory performance is not the same. This creates a problem in the homogeneity test.

When a homogeneity test is performed on certification data using ANOVA, the residual variance is the pooled residual variance (repeatability) of all participants and the F-test for homogeneity is based on this combined residual. This means that small repeatabilities from competent laboratories can be inflated by the results of laboratories with poorer performance. The competent laboratories may be able to detect inhomogeneities in the test material if a certain statistical test was based on their results alone. Pooling results of different quality can thus be problematic. Two alternatives have been considered to the ANOVA F-test:

- ✘ A method based on individual F-type tests for each laboratory.
- ✘ A distribution free rank test method

At present VKI uses a staggered nested design for certification of RMs. Two examples of staggered nested designs used recently at VKI are shown in figure 10.3. At VKI the above mentioned alternatives are to be applied to data sets containing a maximum of 5 results in 2-3 series (analysed on 2-3 different days) per participant. This conflicts with the advantage of the staggered design over the fully nested design, namely that the degrees of freedom are evenly (or almost evenly) assigned to all the factors in the staggered design and fewer observations are needed compared to the fully balanced nested design.

(The table in figure 10.2 shows the formulas for calculating degrees of freedom in staggered nested designs including all laboratories.)

Therefore, an individual F-test procedure of the homogeneity hypothesis for each laboratory will lead to a test with low power with respect to drawing safe conclusions. The two-day design yields an $F(2,1)_{(1-\alpha)}$ test and the three-day design a $F(1,1)_{(1-\alpha)}$ test for $H_0 : \sigma^2_{samples} = 0$ in one laboratory.

One approach to a distribution free test for homogeneity in a two factor design with no interaction is the Friedman's test which in its basic form is a rank test for the main effects. The test for the case with more than one observation per cell has been described by (Mack and Skillings 1980). (Brits and Lemmer 1990) have extended the theory to include an adjusted Friedman test for the main effect in nested designs. It is, however, difficult to construct a test for the lowest level in the nested design (which is of interest here) because of the problem with assigning ranks when different cells are considered.

Source of variation	Degrees of freedom
Days	$f_1 = a - 1$
Samples	$f_1 = \sum_{i=1}^{a} b_i - a$
Residual	$f_3 = N - \sum_{i=1}^{a} b_i$

Figure 10.2: Formulas for the calculation of degrees of freedom in nested three-factor designs (Cummings and Gaylor 1974).

For certification studies where the object is to have a large number of participants and where only relatively few results are supplied by each participant, individual F-testing is not a feasible possibility.

Since a simple alternative to the F-test procedure of homogeneity testing in a staggered nested design cannot be suggested, a graphical method is proposed to help in the data evaluation. Section 10.2 describes a plot method and a distance test designed to match the designs described in figure 10.3. The theoretical description is followed by examples of application to certification data.

10.2 Youden type consistency plots

The (ISO 5725 1994c) for Accuracy of measurement methods and results uses Mandel's h and k plots to evaluate the laboratories and the variability of the measurement method. Mandel's h is a between-laboratory consistency statistic and Mandel's k is a within-laboratory consistency measure. Both are calculated for each laboratory and the results are displayed graphically. The present thesis proposes a supplement to Mandel's h and k plots for the illustration of differences between replicate analyses on sampling units and between sampling units within and across days. Data is assumed to be data from a certification exercise.

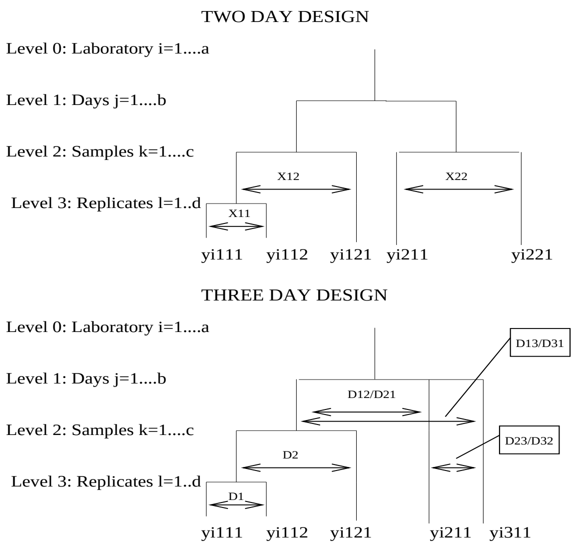


Figure 10.3: Staggered nested design covering two to three days used for certification exercises.

The results within each laboratory are referred to as labelled in figure 10.3. On the basis of a three factor staggered nested design, differences and means are calculated according to the general idea of similar quantities used in (ISO 5725 1994c). The calculated quantities in tables 10.1 and 10.2 represent weighted means and numerical differences within the branches of the tree in figure 10.3. The calculated means and numerical differences are subtracted, thereby forming quantities which describe differences across the branches as marked in figure 10.3. E.g. the intra-sample (marked by X_{11} in the figure), the inter-sample differences (marked by X_{12} and D_2) or the inter-day difference in means.

For the nested designs the following means are calculated:

Mean	Description
Two day design	
$z_{0i} = \sum_{l=1}^{d} \frac{y_{i11l}}{n_{i11l}}$	Mean of results on sample 1 on day 1
$z_{1i} = \frac{1}{n_{i1\cdot}} \sum_{k=1}^{c} \frac{y_{i1k\cdot}}{n_{i1k}}$	Mean of all results on day 1
$z_{2i} = \frac{1}{n_{i2\cdot}} \sum_{k=1}^{c} \frac{y_{i2k\cdot}}{n_{i2k}}$	Mean of all results on day 2
Three day design	
$z_{4i} = \frac{1}{n_{i3\cdot}} \sum_{k=1}^{c} \frac{y_{i3k\cdot}}{n_{i3k}}$	Mean of all results on day 3

Table 10.1: Means for Youden type plots

The index i has been omitted in the following table 10.2. For each laboratory i the above quantities are used to calculate the following differences.

The X-variables are used for the two-day design and the D-variables are used for the three-day design. Although, they are not conventional intermediate measures of reproducibility and repeatability, these quantities give information about the degree of precision achieved by the laboratories in the analyses (mostly the D quantities) and about the intra- and inter-sample consistency of the results (mostly the X quantities and D_1, D_2).

Difference	Description
	Two day design
$X_{11} = y_{i111} - y_{i112}$	Difference between replicates on sample 1 on day 1.
$X_{12} = z_0 - y_{i121}$	Difference between the mean of sample 1 and the result on sample 2, on day 1 i.e. inter-sample difference on day 1.
$X_{22} = y_{i211} - y_{i221}$	Inter-sample difference on day 2.
	Three day design
$D_1 = y_{i111} - y_{i112}$	Difference between replicates on sample 1 on day 1.
$D_2 = z_0 - y_{i121}$	Difference between the mean of sample 1 and the result on sample 2, on day 1 i.e. inter-sample difference on day 1.
$D_{12} = z_1 - z_2$	Difference between mean of day 1 and mean of day 2.
$D_{13} = z_1 - z_3$	Difference between mean of day 1 and mean of day 3.
$D_{23} = z_2 - z_3$	Difference between mean of day 2 and mean of day 3.
$D_{21} = z_2 - z_1$	Difference between mean of day 2 and mean of day 1.
$D_{31} = z_3 - z_1$	Difference between mean of day 3 and mean of day 1.
$D_{32} = z_3 - z_2$	Difference between mean of day 3 and mean of day 2.

Table 10.2: Differences for Youden type plots

The plots of (X_{i11}, X_{i12}) for the two day design and (D_1, D_{12}) for the three-day design show the intra-sample versus the inter-sample difference within day 1. The plot of (X_{21}, X_{22}) shows the inter-sample difference within day 1 versus the inter-sample difference within day 2 in the two-day design. The plots of (D_{12}, D_{13}) for the three-day design illustrate the difference between mean of day 1 and mean of day 2 versus the difference between the mean of day 1 and mean of day 3. The plots of (D_{23}, D_{21}) and (D_{31}, D_{32}) represent the same type of differences with day 2 as a reference mean, and day 3 as a reference mean respectively. The information in the three plots overlap, but it was chosen for completeness to plot data in this manner.

If, generally, two measures of deviance, called x_i and y_i are considered, we define $\mathbf{w}_i = (x_i, y_i)$ as a two-dimensional set of scores for laboratory i. Data are assumed to follow a bivariate normal distribution. A numerical F-test quantity (analogous to the Hotelling's T^2) for the distance between $\mathbf{w}_i$ and the population mean $\boldsymbol{\mu} = (\mu_x, \mu_y)$ is:

$$T^2 = \frac{1}{2}\left[(\mathbf{w}_i - \hat{\boldsymbol{\mu}})\prime\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{w}_i - \hat{\boldsymbol{\mu}})\right] \in F(2, n-1)_{(1-\alpha)} \tag{10.1}$$

where $\hat{\boldsymbol{\mu}}$ is the empirical mean vector and $\hat{\boldsymbol{\Sigma}}$ is the empirical (2x2) variance-covariance matrix. The above expression is approximate, however, the expression is assumed to be satisfactory for the present purpose.

The test has been applied to pairs of the quantities calculated according to the equations in table 10.2. The test is sensitive to laboratory results which are accurate (i.e. have correct mean), but imprecise or both inaccurate and imprecise. It was seen during the statistical analysis that not all Cochran and Grubb outliers turn out as significant in the distance test. (Cochran and Grubb outlier tests are applied to certification data as part of the statistical routine at VKI.) Part of the explanation for this is that the Cochran and Grubb outlier tests use all of the observations from each laboratory for the calculation of the test statistic. T^2 is applied to subsets of data for each laboratory, and thus it treats intermediate deviations which generally differ from the overall deviations diagnosed by the Grubb's and Cochran's tests.

The expression (10.1) can also be used to construct the joint $(1-\alpha)$ confidence region for the observations. Since all the plotted quantities represent differences from zero, the confidence ellipses should in theory be centered at the origin, provided that samples were completely identical, i.e. homogeneous. Deviations from this pattern, slanted or oblong shapes etc. can be interpreted as signs of inhomogeneity and/or problems with analytical performance. A small cross in the plots marks the location of the two plotted variables and the length of the arms indicate the standard deviation of each variable. The cross is not visible if observations are positioned close to the mean. Reference lines for the origin have been added to the plots.

10.2.1 Examples of Youden type plots and distance testing

The following figures show the results of the Youden type plot and the distance test applied to certification data of four RMs produced at VKI: a pesticide RM mentioned earlier in chapter 8[1], WW4 and WW4a intended for COD (COD - chemical oxygen demand), and LL3 for mercury analyses respectively. 50, 95 and 99 % confidence ellipses for the observations have been added to the plots.

The first example shown here is for the pesticide simazine. The original design for this certification was a fully nested design, but data were far from complete. For the purpose of illustration the data has been reduced to match the staggered nested design. The data from this certification were characterized by a general tendency of having larger between-days than between-laboratory variances.

[1]Produced in cooperation with the National Environmental Research Institute, NERI.

In figure 10.4 a) the outlier on the left-hand side of the plot represents a laboratory whose results deviate excessively both with regard to within-sample and between-sample differences within one day. This data point influences the shape of the confidence ellipse considerably. The removal of this point (see figure 10.4 b)) only moves the centre of the ellipse slightly, but the shape of the ellipse changes considerably. The right-hand extreme point on figure 10.4 b) represents a laboratory which differs mainly with respect to within-sample compared to the rest of the results. Apart from observation of laboratory no. 11, the rest of the observations are distributed in a pattern around the origin in figure 10.4 as

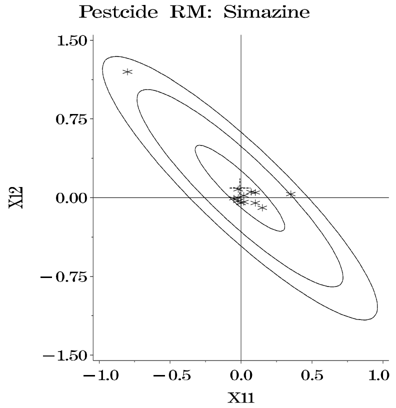


Figure a) One outlier: lab 11

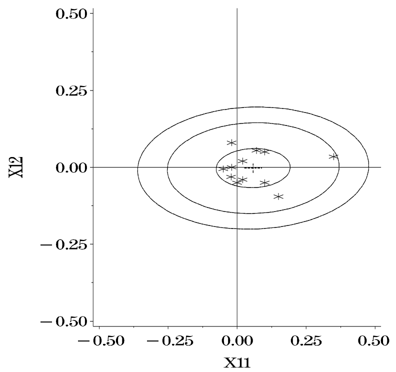


Figure b) Lab 11 removed

Figure 10.4: Youden type plot: pesticide RM. One outlier is observed in figure a). The calculations are repeated excluding this observation, plot shown in figure b).

could be expected for a homogeneous test material (RM).

The plot of differences between samples within day 1 versus the differences between samples on day 2 in figure 10.5 shows one observation far from the others. Laboratory no. 11 is responsible for this observation. The rest of the observations are positioned in the area of the origin.

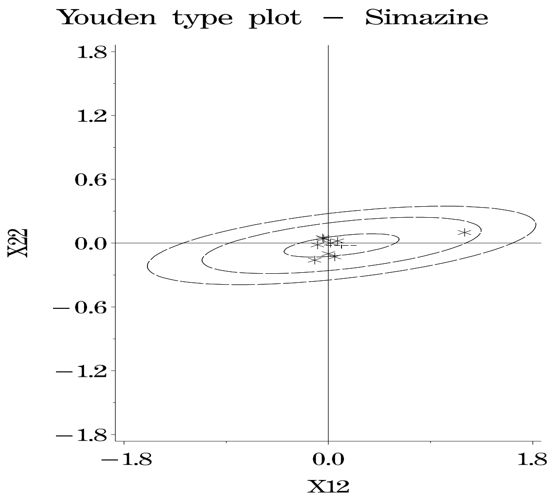


Figure 10.5: Youden type plot: pesticide RM. Differences between ampoules between days. See also figure 10.3 and table (10.2.

Figures 10.6, 10.7, 10.8 and 10.9 show certification data on a VKI RM for COD. The material is produced in a high-level version, WW4 and a low level version WW4a. The raw data is trimmed according to the Grubbs and Cochran statistics and according the the result of the distance test (T^2) applied at $\alpha = 0.05$. The first plot, a) in each figure shows the raw data. The second plot, b) shows the trimmed data where distance testing has been repeated until all observations were within the 95 % confidence ellipse.

The oval shaped vertically positioned confidence ellipses in figure 10.6 indicate slightly larger differences in X_{12} (difference between ampoules on day 1) than in X_{11}. A probable explanation for this could be that inter-sample differences are larger than the differences between replicate analyses on a single sampling unit. Figure 10.7 of the differences between ampoules on different days also shows confidence ellipses which are slightly

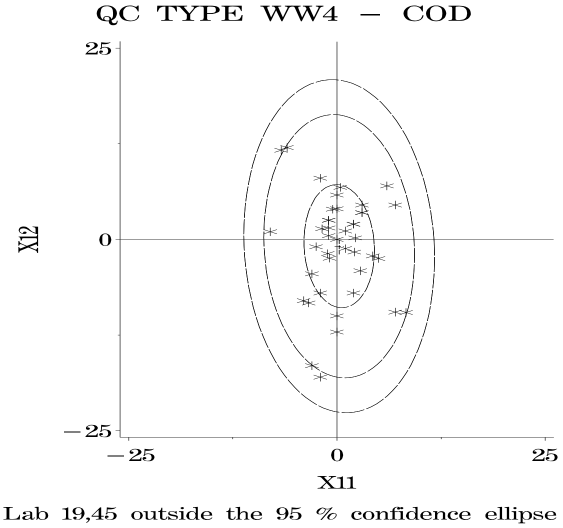


Figure 10.6: Plot of high-level certification data for QC Type WW4 COD. Differences between and within ampoules within a day.

slanted to the left and oval shaped along the X_{22}-axis. This is caused by the outlying laboratory no. 6. Once this result is removed, it is seen that there are no problems with the homogeneity of the samples.

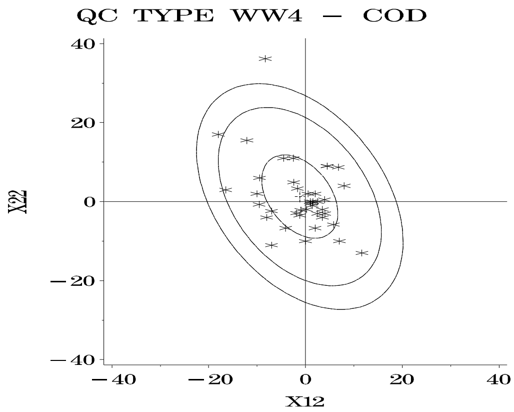


Figure a) Lab 6 outside the 99 % confidence ellipse

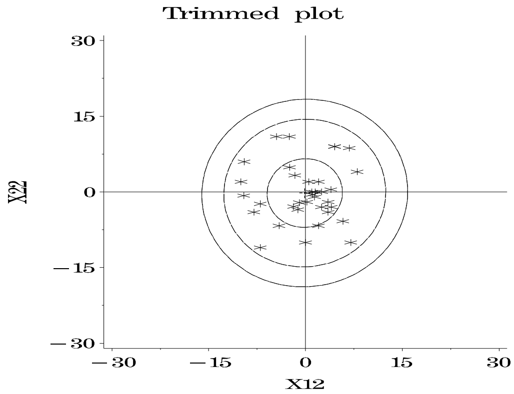


Figure b) 5 observations trimmed by T−test

Figure 10.7: Plot of high-level certification data for QC Type WW4 COD. Differences between ampoules between days.

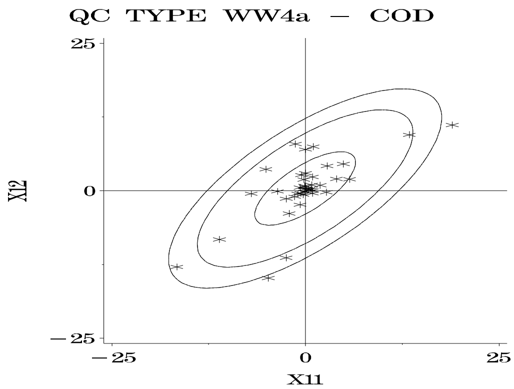


Figure a) Lab 38,44 outside the 95 % confidence ellipse
Lab 28,43 outside the 99 % confidence ellipse

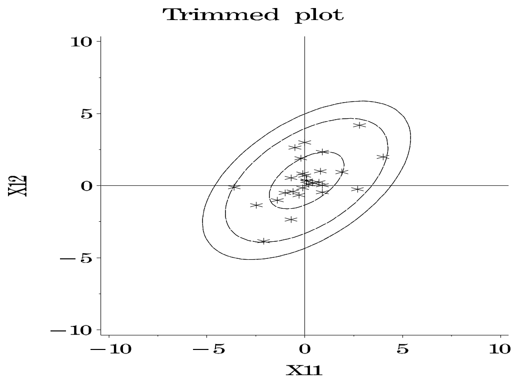


Figure b) 5 Cochran outliers removed
Iteratively trimming of significant T−outliers

Figure 10.8: Plot of low-level certification data for QC Type WW4a COD. Differences between and within ampoules within a day.

The low-level version of the COD RM is the QC Type WW4a. The four plots in figures 10.8 and 10.9 show the Youden type plots for these observations. Figure 10.8 a) is the plot of raw data. Figure 10.8 b) is the result after trimming of outliers - Cochran,Grubbs and distance outliers.

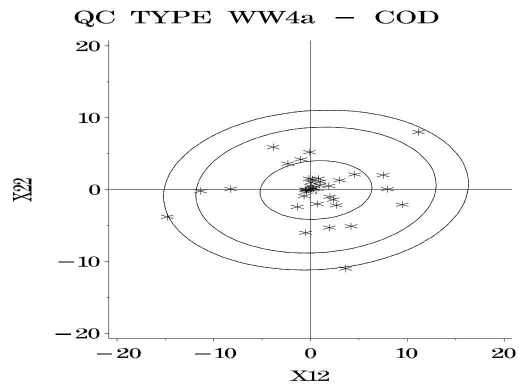


Figure a) Lab 5,28 (Cochran) outside the 99 % confiden—ce ellipse. Lab 43 (Cochran) outside the 95 % confidence ellipse

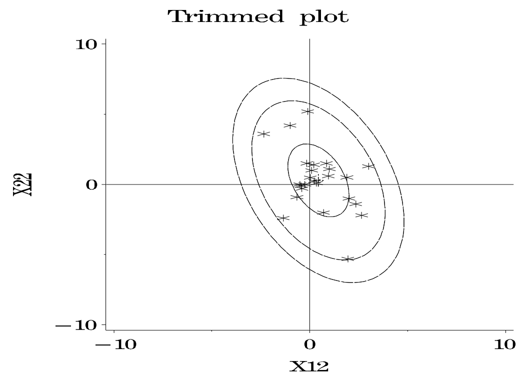


Figure b) Six outliers removed (5 Cochrans)
Iteratively trimming of significant T—outliers

Figure 10.9: Plot of low-level certification data for QC Type WW4 COD. Differences between ampoules between days.

About half of the trimmed observations could be identified as Cochran's outliers (and/or distance outliers) i.e. observations which deviate too much in variance from the set. It is noted that the observations which deviate the most on the X_{12}-axis fall out as Cochran outliers. This is an indication that the laboratories in question have too large standard deviations compared to the rest of the set. The laboratories in question may have a less tight control of the analytical method and its repeatability than the rest of the laboratories.

By comparing figures 10.6 b) and 10.8 b), it is seen that the numerical spread of observations along the X_{12}-axis is smaller and the observations are clustered closer to the origin for the low-level WW4a than for the high-level WW4. Because of the difference in concentration levels between WW4 and and WW4a, the plotted differences show relatively larger values for the high level material. A similar tendency is seen when figures 10.7 b) and 10.9 b) are compared.

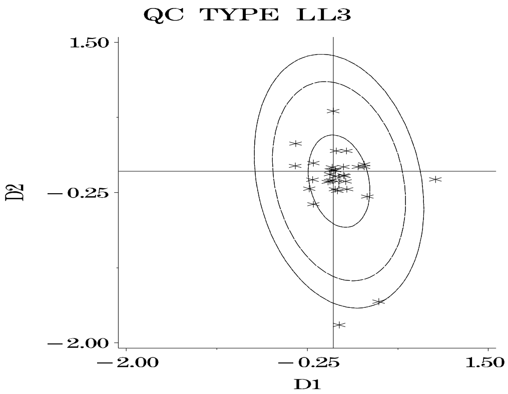


Figure a) Lab 3 outside the 99 %, lab 23 outside the 95 % confidence ellipse. Both Cochran outliers

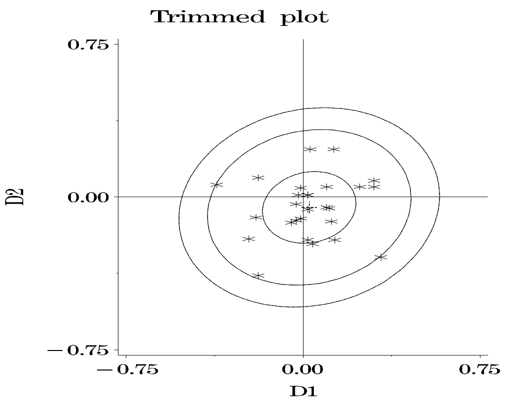


Figure b) Four Cochran outliers removed

Figure 10.10: Plot of certification data for QC Type LL3. Difference between and within ampoules within a day.

Certification raw data of the VKI RM LL3 for Hg analysis is shown in figures 10.10, 10.11, 10.12 and 10.13. The certification was carried out using the three-day design shown on the lower part of figure 10.3. Figure

10.10 of the difference between ampoules on day 1 resembles the patterns seen in the same (trimmed) plots for WW4 and WW4a. The plots in figures 10.11-10.13 overlap in their demonstration of data because of the following relations (also listed in table 10.2):

$$D_{12} = -D_{21} \quad D_{13} = -D_{31} \quad D_{23} = -D_{32}$$

The correlation between the variables D_{12} and D_{13} in figure 10.11 is close to 1 since replicate measurements on ampoule 1 on day 1 are used in the calculations of both variables. Two observations are outside the 95 % confidence ellipse in figure 10.11 b), but they are not significant as neither Grubb's nor Cochran outliers. However, they are significant in the T-test, eq.(10.1) at $\alpha = 0.05$.

The correlations between variables on the trimmed plots in figures 10.12 and 10.13 are not as close to 1 as in figure 10.11. The differences plotted in figure 10.12 are more spread out along the D_{21}-axis than in the horizontal direction. The same two labs as observed in figure 10.11 b) are outside the 99 % confidence interval in figure 10.12 b) and in figure 10.13 b).

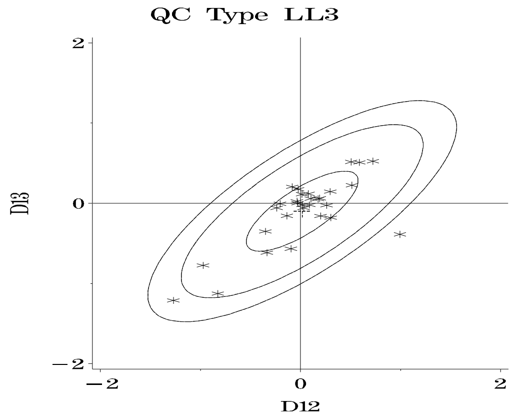


Figure a) Lab 24 outside the 99 % confidence ellipse

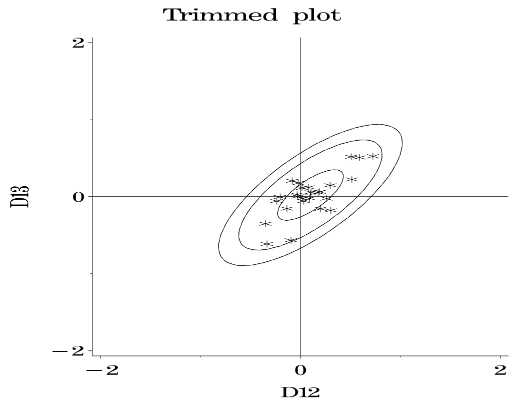


Figure b) Two Cochran outliers removed

Figure 10.11: Plot of certification data for QC Type LL3. Difference between mean of results on day 1 and the results of day 2 (x-axis) and difference between mean of results on day 1 and results on day 3 (y-axis). See lower part of figure 10.3 and table (10.2).

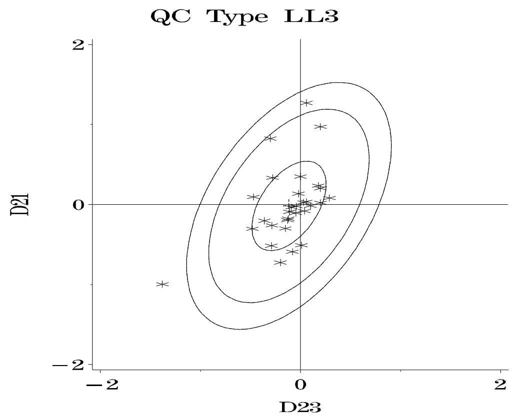


Figure a) Lab 24 outside the 99 % confidence ellipse

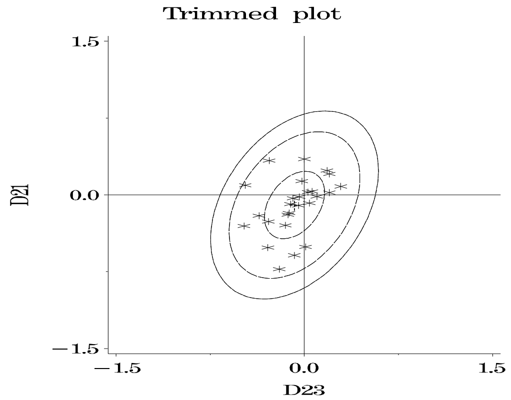


Figure b) Two Cochran outliers removed

Figure 10.12: Plot of certification data for QC Type LL3. Difference between day 2 and day 1 (x-axis), difference between day 2 and day 3 (y-axis). See lower part of figure 10.3 and table (10.2).

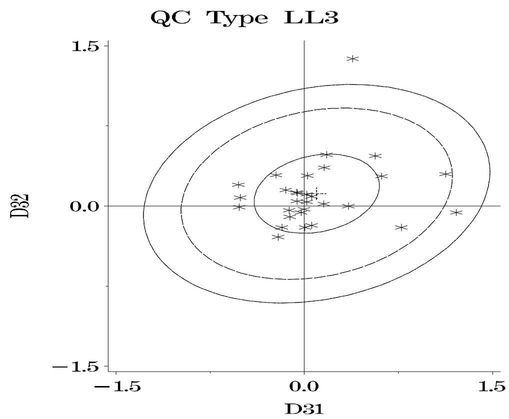


Figure a) Lab 24 outside the 99 % confidence ellipse

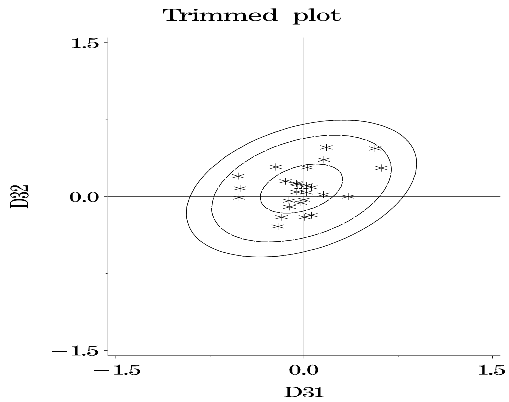


Figure b) Two Cochran outliers removed.

Figure 10.13: Plot of certification data for QC Type LL3. Difference between mean of results on day 3 and the results of day 1 (x-axis) and difference between mean of results on day 3 and results on day 2 (y-axis). See lower part of figure 10.3 and table (10.2).

10.3 Data trimming

Chemists are in general very reluctant to exclude data from certification and proficiency testing data. The argument is that observations should never be excluded if there is no technical reason for suspicion. The statistician however looks differently at the data: observations originate from one or more underlying distributions. If the purpose of the data analyses is to achieve consensus values which represent analytical results achievable by a majority of laboratories, exclusion of data which deviate in one way or another is inevitable. The statistician is probably less scared of excluding influential observations because he has confidence in what calculated statistics and various graphic plots tell him. An observation which shows strong signs of not belonging to the parent distribution of the set, cannot be retained if the assumption of one underlying distribution is to be maintained.

The dilemma of wearing statistician's shoes and a chemists laboratory coat is that on one hand the chemist's concern of accidentally excluding the one laboratory which has in fact achieved the only correct results. On the other hand the statistical analyses tell that these (possibly correct, possibly erroneous for unknown reasons) results are very influential on the statistical analyses and therefore they should be excluded -not on the grounds that they are erroneous but because they are not likely to originate from the same underlying distribution as the other results. The alternative is to carry out the analyses separately for the influential observations and compare the results with the results of the analyses of, either the full data set (including all results), or the reduced data set (where the influential and suspicious observations have been excluded.)

Chapter 11

Stability Studies

11.1 Introduction

In spite of the many international guidelines regarding RMs, limited attention has been paid to stability studies, although stability is a primary requirement and of great concern for both the producer and the user. The guidelines do mention the stability testing requirement, but explicit methodology has not yet been worked out.

ISO Guide 34 states in a few lines that stability of property values should be checked at intervals over a range of practical storage conditions. Because of the great number of RMs and their variety it is impossible to give precise instructions for numbers of samples, intervals etc. Stability monitoring of RMs can be categorized in either one of the following

☞ Type I: short term stability study as part of a feasibility study for a RM production

☞ Type II: time delimited stability study of longer durability than the above and carried out prior to sale

☞ Type III: continuous stability monitoring after sales have started with no fixed closure date

A principle for stability monitoring must take the following into account

- ☞ RM based on natural raw material will exhibit larger between sample variation than a purely synthetic material. The variation between units may be significant.
- ☞ Some analytical methods have measurable day-to-day variance
- ☞ As opposed to pharmaceutical products a RM production concerns one single large batch, not continuously produced series of batches
- ☞ The certified value with its stated uncertainty (and confidence level) must be included in predictions about expected shelf life of the material. (This corresponds to the label claim for pharmaceutical products.)

In the following a summary is given of monitoring principles applicable to one or more of the three categories mentioned above.

11.1.1 Traditions in pharmaceutics

The pharmaceutical industry has a long and extremely regulated tradition for stability studies or shelf life estimation.

A survey of literature on the topic has uncovered two main issues.

- size of experimental design
- choice of model for expiry estimation

Both subjects are of course closely linked to the large costs of developing drugs and both of them apply with some moderation to the field of chemical reference materials. (This will be discussed in section 11.1.3.)

The object of a stability study for a drug is to fix an expiration date within the limits of which the potency of the drug is maintained so that quality requirements and safety of the patient are fulfilled. For most drugs there are a number of factors to be taken into consideration:

- Differences between batches of product

- Differences between concentrations/strengths of active ingredient
- Differences between types of package

All cross-combinations of these factors need examination with regard to the question of whether stability changes are the same. Obviously, the full factorial design for such an investigation is far too expensive. Fractional designs have several advantages and can be chosen so that only high and low levels of a factor are considered, leaving out the intermediate levels as in the example given by (Ju and Chow 1995).

In this example, the samples representing intermediate levels are not analysed in the stability study, but kept aside as a back-up in the case a stability loss shows up between the highest and lowest levels. (Ju and Chow 1995) also suggest designs with partial time points where e.g. two package types are tested at different time points: package type 1 is tested at $\tau_{p1}=\{0, \tau_3, \tau_9, \tau_{24}\}$ and package type 2 is tested at $\tau_{p2} = \{0, \tau_6, \tau_{12}, \tau_{24}\}$ so that the number of total assays is reduced. The principle is illustrated in figure 11.1. The indices represent time in units of months. (Nordbrock

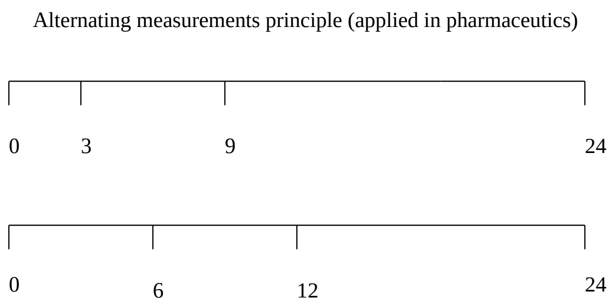


Figure 11.1: Time limited stability study design used in pharmaceutics, see text for explanation.

1992) chooses the power of detecting a significant slope difference of stability loss over time as the criterion for design selection. A fixed sample size is set and the design with the highest power of detecting slope differences is selected.

(Ju and Chow 1995) argue that Nordbrock's approach doesn't take into consideration the reliability and accuracy of estimated shelf life. (Ju and Chow 1995) suggest using the precision of the shelf life estimation as a design criterion.

In (Rahman 1992) two expiration dating methods are presented. Method 1 follows the FDA[1] guidelines: an estimated degradation line fitted to stability data and a 95 % lower confidence bound on the true degradation line is fitted. The expiration data is the time point τ' where the expression of the 95 % lower confidence bound equals the lower specification of the drug potency.

The approach in method 2 is to calculate a 95 % lower confidence limit, b_1 of the slope, b of the true degradation line. A new line is constructed using b_1 as its slope and the expiration date is determined as the time point where the new line equals the lower specification of the drug potency.

In Rahman's detailed discussion it is demonstrated that in most practical cases the FDA method (method 1) leads to greater expiration dates than method 2. It seems that there is a tradition for using linear regression estimation (Lin 1990), (Morris 1992), (Rahman 1992).

According to Morris, degradation functions often follow a simple first-order exponential model. (Morris 1992) shows that within a range of 10-15 % maximum allowable degradation, the difference between linear and exponential modelling is of negligible importance. If however, the stability study is very long, the assay variability is small or there is a lower specification limit of less than 90 % of label claim, then exponential methods are more advantageous.

11.1.2 Traditions in food industry

With regard to RMs for food there seems to be a far less standardized approach to stability monitoring. Whereas the pharmaceutic world appears to have a systematic approach to assay types and sizes and to the

[1] Federal Drug Administration, USA

modelling of degradation, the fields of reference material for foods (and environmental samples) can be characterized as less stringent.

None of the literature on food RMs surveyed for this study has presented attempts to apply models to check for instability. It may be added that the reason for this could be that much attention and effort in the field of food RMs is devoted to the question of homogeneity and sampling constants, especially as new analytical methods make use of very small sample sizes compared to the past.

(Maier *et al.* 1993) for instance, mention influence factors on stability as well as specific statistical tests for Poison distribution of bacteria in the investigated material for food microbiology.

The researchers in this field make a point of storing the samples at minimum two temperatures for short term studies, and up to 4 different temperatures for long term studies. One argument for selecting numerous temperature levels is that the storage conditions should also reflect extreme exposures which might occur during transport. Another reason is the desire to investigate accelerated disintegration at extreme conditions. This is based on the assumption that decay will follow the same course at all temperatures, but at different speeds, so that any change seen at an elevated temperature will eventually occur at lower temperature levels. Food RMs are mostly powder materials and are often stored at -20°C and -40°C for stability studies. Freezing is usually avoided for aqueous environmental RMs because of the risks of irreversible matrix modification.

A t-test is often applied to stability data (Vercoutere and Cornelis 1988), but in many cases it is just mentioned that the samples did not change significantly (Vercoutere *et al.* 1995) or that stability improved (Maier *et al.* 1993) without stating the statistical approach in any further detail.

(Hollmann *et al.* 1993) state the within laboratory coefficient of variation (CV) as a criterion for evaluating stability. In addition to statistical tests, food RMs are observed qualitatively, and changes in color or smell of the product may be taken as indications of instability.

11.1.3 Application to aqueous reference materials

(Hollmann *et al.* 1993) make an important observation with regard to the stability studies of RM: it is often difficult to distinguish long-term repeatability of the analytical method from instability variability.

It is noted that the papers on pharmaceutical products did not discuss the difficulty of the variability of the analytical method involved. Either, it has been considered and not found to be worth mentioning, or it is supposed that the decay of pharmaceutical products is indisputable and certain to manifest itself through any blurring caused by analytical uncertainty. (Rahman 1992) states two models neither of which includes a random uncertainty term in their initial formulations.

(Hollmann *et al.* 1993)'s point about long-term repeatability versus instability variances strikes a central area of the problems with stability studies of RMs in general. In spite of its extreme importance, it is an issue which has been discussed little so far. It would seem probable that some of the systematic approaches to stability studies from the pharmaceutical fields could be implemented with many advantages to environmental RMs. The reluctance to do so may be founded on economical aspects as touched upon earlier.

The RM end-user is not endangered on his health if the actual content of a RM has increased or decreased significantly since the date of certification. There is no strict legislative regulation of the validity period for reference materials, but it is certainly in the interest of the manufacturer to have the stability under control. Selling instable material is of no real use and causes confusion in the customer's QA (Quality Assurance) systems and can thus damage the reputation of the manufacturer and hence have economical consequences. From a profit point of view it is therefore worthwhile giving the stability monitoring attention.

The approaches in (Ju and Chow 1995) and (Nordbrock 1992) should be considered for environmental RMs i.e. either optimizing assays with regard to the power of detecting a significant slope difference over time. First of all a precise model for product degradation is needed. Assume that the response of the compound A in a RM is measured at time $\tau = t$

$$Y_t = \mu' + \epsilon_t \tag{11.1}$$

where
μ' = reference value
ϵ_t = analytical uncertainty (short term repeatability)

and $\epsilon_\tau \in N(0, \sigma_E^2)$. If several samples are measured repeatedly then at time τ:

$$Y_{ij,t} = \mu' + S_{i,t} + \epsilon_{ij,t} \tag{11.2}$$

i = sample number
j = replicate measurement number
and $S_{i,t} \in N(0, \sigma_S^2)$. In case of degradation of the product it is assumed, for a start, that the change will follow a linear model with either an increase or decrease in the concentration of compound A

$$Y_{ij,t} = \mu' + b\tau + S_{ij,t} + \epsilon_{ij,t} \tag{11.3}$$

μ' the reference value with its stated (certified) uncertainty represents the intercept, τ is time, and either $b > 0$ (increasing concentration) or $b < 0$ (decreasing) or $b = 0$ (stable compound concentration). Eqs. (11.1) - (11.3) do not include the uncertainty of the reference value. Other models like for example exponential decrease, can be treated analogously.

11.2 Modelling long-term stability for a reference material

(Lamberty *et al.* 1998) give a review of the latest method development with regard to stability studies of RMs. In many articles about environmental (or food) RMs, stability studies are mentioned in a brief section stating that "*Samples were stored at temperatures* t_1°, t_2° *and* t_3° *and checked after time* τ_1 *and* τ_2 *etc. No signs of instability were found.*" or similar phrasings as in e.g. (Hollman *et al.* 1993), (Hollmann *et al.*

1993), (Vercoutere and Cornelis 1988), (Maier *et al.* 1993) and (Dahl *et al.* 1990). There is no contradiction between this approach and the statements in ISO Guide 34, but there seems to be basis for a more informative and reassuring approach.

A recent publication on the issue by (Pauwels *et al.* 1998) adopts the methodology of the pharmaceutical field for the estimation of expected shelf life.

11.2.1 Continuous monitoring - graphical representation methods

Time zero ratios

Around 1989 ratio versus time zero (τ_0) in stability studies is introduced by Griepink (Lamberty *et al.* 1998). The results at any time point $\tau > \tau_0$ are divided by the mean at τ_0 and the stability is evaluated on the basis of these ratios.(Lamberty *et al.* 1998) note that this method is only satisfactory for analytical methods with very reliable long-term reproducibility.

Reference temperature ratios

For less reproducible methods which are likely to exhibit drift over time, stability measurements at time τ are better expressed as ratios to measurements at a reference temperature also at time τ. A low temperature, t°_{ref}, is selected to be used for comparison with samples stored at higher temperatures. The ratio of means at two temperatures, R_{t° is calculated at each time point:

$$R_{t^\circ} = \frac{\bar{X}_{t^\circ_i}}{\bar{X}_{t^\circ_{ref}}} \tag{11.4}$$

The uncertainty is based on variance coefficients of n measurements at

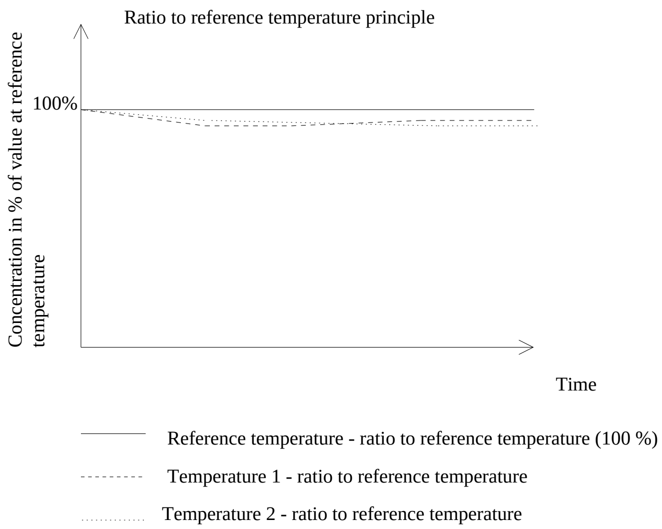


Figure 11.2: Illustration of continued stability study by ratios to a reference temperature, see text for explanation.

each temperature level:

$$\begin{aligned} U_{t_i^\circ} &\simeq (CV_{t_i^\circ}^2 + CV_{t_{ref}^\circ}^2)^{1/2} \cdot R_{t^\circ} \\ &\simeq \left(\frac{\hat{\sigma}_{t_i^\circ}^2}{\bar{X}_{t_i^\circ}^2} + \frac{\hat{\sigma}_{t_{ref}^\circ}^2}{\bar{X}_{t_{ref}^\circ}^2}\right)^{1/2} \cdot \frac{\bar{X}_{t_i^\circ}}{\bar{X}_{t_{ref}^\circ}} \end{aligned} \tag{11.5}$$

where $U_{t_i^\circ}$ is the overall uncertainty of the ratio R_{t° between the mean concentration measured at time τ of samples stored at temperature i and samples store at a reference temperature. This approach has been used by several authors in the European RM community (e.g. (Quevauviller *et al.* 1993) and (Reijnders *et al.* 1994)). (Lamberty *et al.* 1998) add that the relative ratios of concentrations should be regressed versus time and the 95 % confidence interval of the line should be compared to the uncertainty range of the reference value. The temperature ratio approach has the advantage of cancelling out analytical variations which are time related.

11.2.2 Time delimited studies - storage systems

Grid principle

(Faure and Wagstaffe 1993) introduced a scheme for short term stability studies where samples showing no instability when stored for a few weeks at a low temperature, were transferred to a higher temperature level and checked after another few weeks (see figure 11.3). The design seemed suited for revealing instability which could arise as a consequence of extreme temperatures during transport, but is not intended for long-term studies where the reference of a low temperature is needed to confirm stability at higher temperatures through (relative) comparisons.

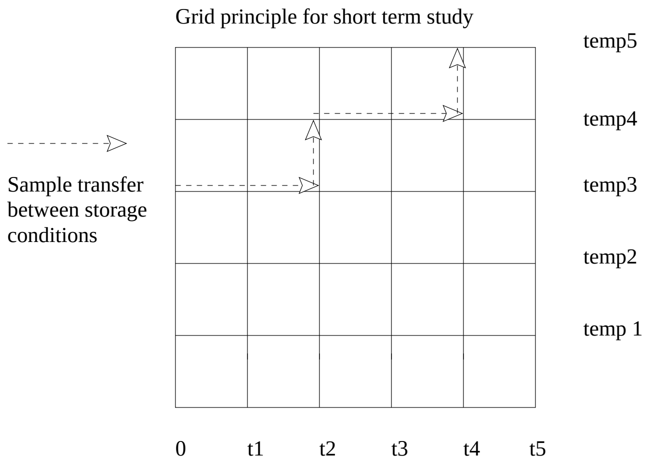


Figure 11.3: Illustration of (Faure and Wagstaffe 1993)'s grid principle for short stability studies, see text for explanation. Samples are measured at each time point in the grid.

Isochronous measurement principle

(Lamberty *et al.* 1998) suggested a design to deal with this problem based on the principle of transferring samples to a higher temperature level as in (Faure and Wagstaffe 1993). The difference from (Faure and Wagstaffe 1993)'s design is that in (Lamberty *et al.* 1998)'s isochronous scheme, analyses are performed only once: at the end of the study and then on all samples. The principle is illustrated in figure 11.4.

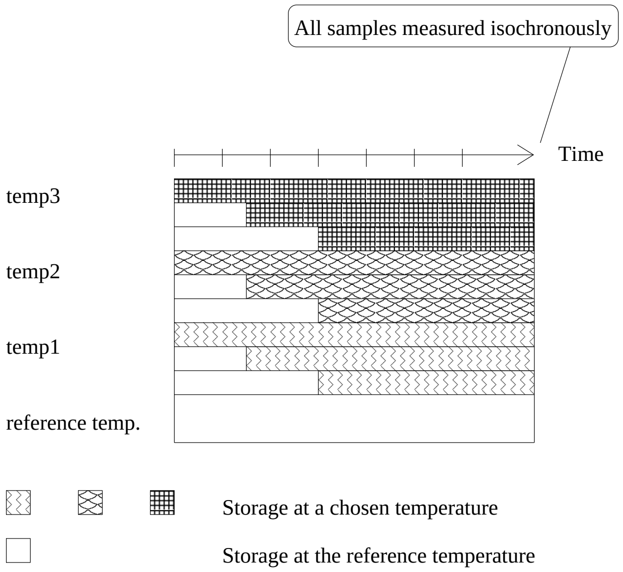


Figure 11.4: Illustration of stability studies by isochronous measurement, see text for explanation.

The advantage of this approach is that the problem of long-term method variation or drift is eliminated because all samples are analysed at the same time. However, it does require sufficient laboratory capacity to deal with a large number of samples within a short period of time. The design applies to a stability study of a pre-determined length such as it would be in a pre-certification stability study from the time of production to the time of release.

The obvious disadvantage is that any instability that shows up after the first short term stability study has been completed, will not be revealed until the analysis of all ampoules/units at the end of the long-term study. Another thing is that the design obviously does not apply to the stability monitoring after product release (the authors do not claim it to be intended for that purpose).

The calculation method of relative ratios as it is presented for the isochronous designs (and in earlier publications) ignores the sample variation. With the combined uncertainty expressed as $U_{t_i^\circ}$ there is no analysis of sample variations at different temperature levels. This may be relevant because a change in sample variation may be an early indication of instability.

11.2.3 Variance component modelling

The variance component model offers several advantages with regard to identifying influential sources of variation in a stability study. The model statement can be adapted to suit any type of stability study. In general it is of interest to identity variation due to storage conditions. Depending on the number of samples analysed at each stability check, there is a possibility of surveying sample homogeneity over time as well. The power of the statistical tests of course depends on the size of the design - i.e. the number of samples which the producer is willing to use in each stability check.

Three types of scenarios are given here for a stability study. The assumptions are:

- The residual is assumed normal distributed $E \in N(0, \sigma_E^2)$
- Fixed effects are subject to the restriction $\sum \tau_i = 0$ for example
- Random effects are assumed $R \in N(0, \sigma_R^2)$, for example

The following symbols are used:

Source of variation	Nature	Symbol
Time	Random	$\mathfrak{T}$
Time	Fixed	τ
Storage temperature	Fixed	t°
Humidity	Fixed	h
Sample	Random	S
Error	Random	E

Table 11.1: Symbols used in variance component models for stability.

Case 1

① The applied analytical method is not influenced by day to day variation.

② Changes in sample homogeneity is not of interest/not expected.

③ Stability is monitored at various storage temperatures.

This situation only requires a model which includes storage temperature and time as a fixed effects apart from the overall mean:

$$Y_{ijk} = \mu + \tau_i + t^{\circ}_j + \tau t^{\circ}_{ij} + E_{k(ij)} \qquad (11.6)$$

If degradation takes place, the decay is assumed to follow a first order degradation as illustrated in figure 11.2. The fixed time effect in the model corresponds to the term $b\tau$ in eq.(11.3). Note, however that the method of estimating the τ_i effects is different from estimating the regression coefficient for the decay rate.

Case 2

① The applied analytical method has significant day to day variation.

② Changes in sample homogeneity is not of interest.

③ Stability is monitored at various storage temperatures.

$$Y_{ijk} = \mu + \mathfrak{T}_i + t^\circ_j + \mathfrak{T}t^\circ_{ij} + E(\mathfrak{T}t^\circ)_{k(ij)} \tag{11.7}$$

In this situation the time effect is treated as a random effect expressing variation related to the analytical method. If a decay trend is present, this effect will be confounded with the method variation unless the design is expanded with measurements on two or three different days at each stability check. Another possible way of revealing a systematic time trend could be to include information about method variation from an existing QA system.

Case 3

① The applied analytical method has significant day to day variation.

② Changes in sample homogeneity are of interest.

③ Stability is monitored at various storage temperatures and different humidity conditions.

$$Y_{ijklm} = \mu + \mathfrak{T}_i + t^\circ_j + \mathfrak{T}t^\circ_{ij} + h_k + ht^\circ + S(\mathfrak{T}t^\circ)h_{l(ijk)} + E(\mathfrak{T}t^\circ hS)_{m(ijkl)} \tag{11.8}$$

This design will demand a large number of samples to be reserved for the stability study in order to sufficient power of the test statistic from the beginning of the study. Such an extensive model is not very likely for aqueous samples, but it may be relevant for solid materials (powders).

A mixed model such as eq.(11.8) can be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{11.9}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices and $\mathbf{e}$ is the residual. β is a vector of parameters for fixed effects, e.g. storage and humidity in Case 3. $\mathbf{u}$ is a matrix of variance components describing random effects with variance

$var(\mathbf{u}) = \mathbf{G}$, shown here for Case 3:

$$\boldsymbol{\beta} = \begin{bmatrix} t_1^\circ \\ t_2^\circ \\ .. \\ h_1 \\ h_2 \\ .. \end{bmatrix} \qquad Var(\mathbf{u}) = \begin{bmatrix} \sigma^2_{\mathfrak{T}} & & \\ & \sigma^2_{\mathfrak{T}t^\circ} & \\ & & \sigma^2_{S(\mathfrak{T}t^\circ h)} \end{bmatrix}$$

The mixed models equations containing fixed as well as random effects can be written as

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{11.10}$$

In case 1 the model contains only fixed effects and the u-matrix is in the model. Normality is assumed for the random parameters and furthermore

$$\mathrm{E}\begin{bmatrix} \mathrm{u} \\ \mathrm{e} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathrm{Var}\begin{bmatrix} \mathrm{u} \\ \mathrm{e} \end{bmatrix} = \begin{bmatrix} \mathrm{G} & 0 \\ 0 & \mathrm{R} \end{bmatrix}$$

11.3 Modelling several parameters in the same matrix

Multicomponent RMs present an additional aspect which should be taken into account: the tediousness of processing p characteristics one at a time at each time point. Moreover there is the risk of being mislead when estimating several confidence limits if a method of individual control charts is used.

Suppose a multicomponent RM of p components is monitored by use of some type form of control chart applied individually for each of the p measurands. α is the probability of committing a type I error on each individual chart. The probability of all p measurands plotting simultaneously inside their control limits when assuming independence between components and measurements of components, is

$$P\{\text{all p measurands in control}\} = (1-\alpha)^p$$

If this is applied to a 3-sigma $\bar{X}$-charts for a multicomponent RM of dimension 3 then, theoretically,

$$P\{\text{all 3 measurands plot in control}\} = (1-0.0027)^3 = 0.991$$

when the material **is** stable. For characteristics which are not independent there it is more complicated to compute this probability.

The joint probability of committing a type I error for a RM consisting of three components is

$$\tau = 1-(1-\alpha)^p = 1-(1-0.0027)^3 = 0.008$$

for the 3-sigma $\bar{X}$ control chart.

For individual measurands, i.e. single component RMs the best choice of control charts is probably Shewart type charts, because the interval between samples is large (weeks to months) and because variability (noise from the measurement method) often dominates changes in the mean.

11.3.1 Multivariate control charts

Suppose measurements on a RM are carried out in the following manner: at time τ a set of measurements of p characteristics is carried out on a sample of n units. In total this procedure is repeated m times. An observation is written as $x_{ij\tau}$ where $i = 1, ..n$ describes the sample number, $j = 1, ..p$ indicates the measured characteristic, and $\tau = 1, ..m$ is the sampling time point. The p measured characteristics $\mathbf{x_{ij}}_\tau = (x_{i1\tau}, ..., x_{ip\tau})'$

are assumed normally distributed as $\mathrm{N}_p(\mu, \Sigma)$. The set observations are:

$$\left(\begin{bmatrix} x_{111} \\ x_{121} \\ x_{131} \\ . \\ . \\ x_{1p1} \end{bmatrix} \quad .... \quad \begin{bmatrix} x_{n11} \\ x_{n21} \\ x_{n31} \\ . \\ . \\ x_{np1} \end{bmatrix} \right) .... \left(\begin{bmatrix} x_{11m} \\ x_{12m} \\ x_{13m} \\ . \\ . \\ x_{1pm} \end{bmatrix} \quad ..... \quad \begin{bmatrix} x_{n1m} \\ x_{n2m} \\ x_{n31} \\ . \\ . \\ x_{npm} \end{bmatrix} \right)$$

Hotelling's T^2 control chart

The Hotelling's T^2 control chart is an extension of the chi-square control chart for two quality characteristics to control p characteristics jointly. (A description can be found in (Montgomery 1997b).) The Hotelling's T^2 control chart method applied to N-components in wastewater reference material was shown in chapter 3, section 3.6.1. The test statistic is based on the set of averages represented by the $p \times 1$ vector. For each time point $\tau = 1, ..m$ the vector

$$\bar{x}_\tau = \begin{bmatrix} \bar{x}_{.1\tau} \\ \bar{x}_{.2\tau} \\ . \\ . \\ . \\ \bar{x}_{.p\tau} \end{bmatrix}$$

is calculated. If μ is the mean and Σ is the covariance matrix for a sample $\{x_1, x_2, .., x_p\}$ then the quantity

$$\chi_0^2 = n(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu) \tag{11.11}$$

follows a χ^2-distribution with p degrees of freedom. $\mu = [\mu_1, \mu_2 ... \mu_p]'$ could be taken as the vector of certified values. If certified values are not available, μ can be replaced by the estimate $\bar{\bar{x}}_{...}$. Σ is the covariance matrix. The control is based on an estimate for Σ, as follows.

Let indices j and h count the characteristics and index τ count the sample number (the sample is taken at time τ).

$$\bar{x}_{.j\tau} = \frac{1}{n}\sum_{i=1}^{n} x_{ij\tau} \begin{cases} j = 1, 2, ..., p \\ \tau = 1, 2, ..., m \end{cases} \tag{11.12}$$

$$S_{jh\tau} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij\tau} - \bar{x}_{j\tau})(x_{ih\tau} - \bar{x}_{h\tau}) \begin{cases} \tau = 1, 2, ...m \\ j = 1, 2, ...p \\ h = 1, 2, ...p \end{cases} \tag{11.13}$$

The statistics $\bar{x}_{j\tau}$, $S^2_{j\tau}$ and $S_{jh\tau}$ are averaged

$$\bar{S}^2_j = \frac{1}{m}\sum_{\tau=1}^{m} S^2_{j\tau} \qquad \bar{S}_{jh} = \frac{1}{m}\sum_{\tau=1}^{m} S_{jh\tau}$$

and from the vector $\bar{\bar{x}}$ containing the elements $\bar{\bar{x}}_j$ and the $p \times p$ average of sample covariances

$$\mathbf{S} = \begin{bmatrix} \bar{S}_{11} & \bar{S}_{12} & \bar{S}_{13} & . & . & . & \bar{S}_{1p} \\ & \bar{S}_{22} & \bar{S}_{23} & . & . & . & \bar{S}_{2p} \\ & & & . & & & \\ & & & & . & & \\ & & & & & . & \\ & & & & & & \bar{S}_{pp} \end{bmatrix} \tag{11.14}$$

The final test statistic is

$$T^2 = n(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) \tag{11.15}$$

which under the null hypothesis follows an F-distribution with $(p, mn - m - p + 1)$ degrees of freedom.

Suppose the T^2 chart is used in a preliminary stability study - as a phase 1 case where the interest is to test whether the material is in control - in other words to obtain an in-control set of observations. For the phase 1 case the lower control limit is zero and the upper limit is

$$UCL = \frac{p(m-1)(n-1)}{mn - m - p + 1}F(p, mn - m - p + 1)_{1-\alpha} \tag{11.16}$$

For further monitoring e.g. after certification has taken place the upper control limit is

$$UCL = \frac{p(m+1)(n-1)}{mn-m-p+1} F(p, mn-m-p+1)_{1-\alpha} \tag{11.17}$$

M-chart method

In (Chan and Li 1994) a method for multivariate control charts is presented. The example is based on a manufacturing process in which the quality of a product is characterized by several characteristics. The situation can be applied to multicomponent reference materials - that is - reference material intended for quality control of several measurands in the same matrix. The method in (Chan and Li 1994) aims at efficiency in detecting linear trends as a sign of an out of control situation. The M-chart proposed in the paper is a dimension-reducing method for detecting linear trends of the process mean with time. The M-chart method applied to N-components in wastewater reference material was shown in chapter 3, section 3.6.2. The method is based on the following:

Again, p measured characteristics $\mathbf{x} = (x_1, ..., x_p)'$ are assumed normally distributed as $N_p(\mu, \Sigma)$. At each time point τ_i, $i = 1, ..., m$, a sample of n units is selected. At the time points τ_1, τ_2,.. τ_m the subgroup means are $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$,...$\bar{\mathbf{x}}_m$ respectively as shown in the expression on page 167, i.e. $\bar{\mathbf{x}}_2$ is the p-dimensional mean vector of the measurements made on samples at time τ_2.

For a given timeperiod of length m, the method provides a test for detecting a linear trend. The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = ... = \mu_m = \mu_0$$

which is tested against the alternative

$$H_1 : \mu_1, \mu_2, ..., \mu_m \text{ depend linearly on } \tau_1, \tau_2, ..., \tau_m$$

through a linear relationship $\alpha'\mu_1......\alpha'\mu_m$. μ_i is the population mean of the i'th subgroup and α is a vector.

A linear trend test statistic M_i is calculated for the subgroup means $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_{i+1}$,...,$\bar{\mathbf{x}}_{i+m-1}$, where m is a predetermined number of consecutive process means. m is chosen so that $m > p$ for μ known. Let the time period mean be

$$\bar{\tau} = \sum_{j=1}^{m} \tau_j / m \tag{11.18}$$

$$\text{and} \tag{11.19}$$

$$\mathbf{X} = (\bar{x}_1 - \mu_0, ...\bar{x}_m - \mu_0)' \tag{11.20}$$

$$\text{where } \ \mu_0 \ \text{ is the target value} \tag{11.21}$$

$$\mathbf{w} = (\tau_1 - \bar{\tau}, \tau_2 - \bar{\tau}, ...\tau_m - \bar{\tau})' \tag{11.22}$$

$$\mathbf{z} = \mathbf{X}'\mathbf{w}/|\mathbf{w}| \tag{11.23}$$

$$|\mathbf{w}|^2 = \mathbf{w}'\mathbf{w} \tag{11.24}$$

$\mathbf{X}$ is a m$\times$p matrix and $\mathbf{w}$ is an m-vector. The sum of squares due to regression, SSR, of $\mathbf{X}\tilde{\mathbf{a}} = \tilde{\mathbf{a}}'(\bar{\mathbf{x}}_1 - \mu_0), ...\tilde{\mathbf{a}}'(\bar{\mathbf{x}}_m - \mu_0)$ on $\mathbf{w}$ and the residual sum of squares, SSE form the ratio F-test, $M(\tilde{\mathbf{a}})$ of the null hypothesis. The $\mathbf{a}$ which maximizes $M(\tilde{\mathbf{a}})$ is given by

$$\mathbf{a} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}}{|(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}|} \tag{11.25}$$

and

$$M_i = \frac{q_i}{1 - q_i} \tag{11.26}$$

$$q_i = \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}\right]' \mathbf{z} \tag{11.27}$$

If any of the points $(1, M_1), (2, M_2), ..., (\tau - m + 1, M_{\tau - m+1})$ fall outside the upper control limit $F(p, m-p)_{1-\alpha}$, the $(1-\alpha)$ quantile of the F-distribution, then reject H_0. The plot of $(\tau_i, \mathbf{a}'\bar{\mathbf{x}}_1), ..., \tau_m, \mathbf{a}'\bar{\mathbf{x}}_m)$ is considered to be linear when M_i is above the upper control limit. The covariance matrix Σ is not involved directly in the calculations and μ can also be estimated if it is unknown. The certified reference values can be used as μ when the stability of a multicomponent RM is inspected over time. If the method is used in a pre-certification stability study, μ_0 can be replaced by the grand sample mean over the period $\bar{\bar{\mathbf{x}}} = \bar{\mathbf{x}}_1 + ... + \bar{\mathbf{x}}_m$. The degrees of freedom for the quantile of the F-distribution are reduced to $F(p, m-p-1)_{1-\alpha}$.

MEWMA

(Lowry *et al.* 1992) have developed a multivariate version of the exponentially weighted moving average control chart (MEWMA). The MEWMA is described by the following equations

$$Z_i = \lambda x_i + (1-\lambda)Z_{i-1}, \quad i = 1, ..., m \tag{11.28}$$

where $0 \leq \lambda \leq 1$ is the reduction in weight given to the previous observation and Z_0 can be chosen equal to a reference value.

$$T_i^2 = Z_i' \Sigma^{-1} Z_i > h_4 \tag{11.29}$$

is the quantity plotted on the control chart, Σ is the covariance matrix of Z_i. Σ has to be estimated. The upper control limit, h_4 (Lowry *et al.* 1992) is chosen to suit a specified in control average run length (ARL). The covariance matrix is in principle estimated according to equation 11.14. If it is standardized by subtracting the mean, the non-centrality parameter will be zero and control limits i.e. values of h_4 are given in (Lowry *et al.* 1992).

11.4 Discussion

The following discussion summarizes the models and methods presented in the previous sections. Table 11.2 gives an outline of the applicability of the models and methods to the three types of stability studies presented in section 11.1. The table contains the symbol ✔ if it is *possible* to apply a certain model/method to a stability study. (Adequacy for purpose is not evaluated in the table.) The symbol τ is used when the model/method is capable of detecting trends in concentration (stability trends). $\Im$ is used if a model/method can identify time related random sources of variation.

Principle		Stability Study Type I	Type II	Type III
Graphical represent.	Time zero ratio	✔	✔	✔
	Reference temp. ratio	✔ τ	✔ τ	✔ τ
Storage systems	Grid method	✔	✔	
	Isochronous method	✔ τ	✔ τ	
Univariate	ANOVA	✔ τ $\Im$	✔ τ $\Im$	✔ τ $\Im$
	$\bar{X}$-chart		✔	✔
Multivariate modelling	Hotelling's T^2	(✔)τ	✔τ	✔τ
	M-chart (multicomp. RM)	(✔)	✔ τ	✔ τ
	MEWMA	(✔)	✔	✔

Table 11.2: Monitoring principles - evaluation summary. ✔ applicability to stability studies. τ trend indicating properties. $\Im$ variance source identifying properties. (✔) less suited because of limited number of observations. Type I is a short term stability study, type II is a time delimited stability study, and type III is a continuous stability monitoring, see also page 151.

The ratio methods in table 11.2 can be applied to time delimited as well as long term stability studies. The advantage of the ratio methods is primarily the form - which is the graphical representation. The plot requires only a limited number of calculations and can be produced and are easily interpreted by a laboratory technician.

As mentioned in section 11.2.1, the reference temperature ratio method avoids the problem of analytical variations, whereas the time zero ratio is best suited for methods such as e.g. isotope dilution mass spectrometry and the like which are highly repeatable and reproducible methods. If the analytical variation cannot be kept in sufficiently tight control, the continuous methods suffer risk of confounding the effects of time and

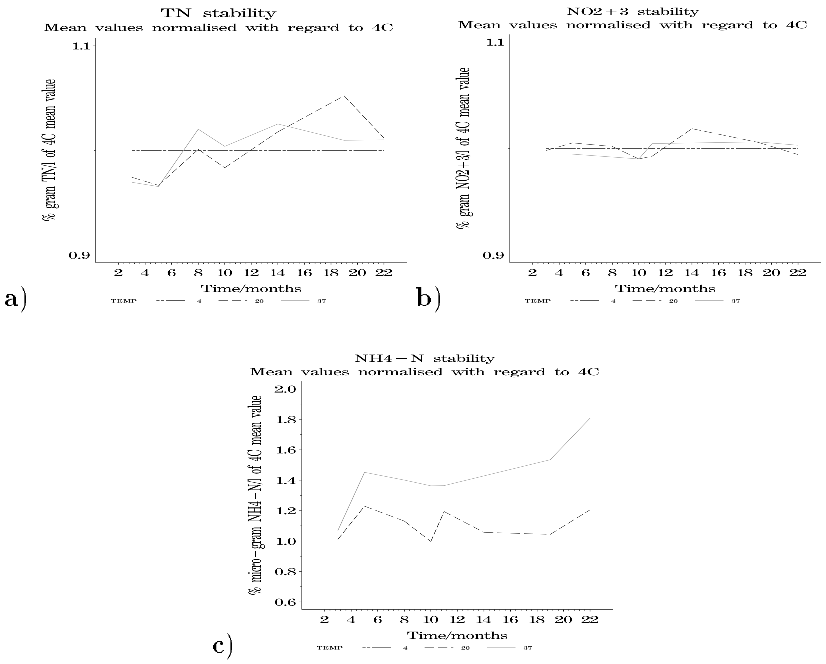


Figure 11.5: Reference temperature ratio plots - shown earlier in chapter 3 a) TN b) NO_{2+3}-N c) NH_4^+-N.

method variation. Neither of the ratio methods can discriminate between time effect and analytical variation.

However, as shown in chapters 3 and 4 the indicating capability of the reference temperature ratio method showed a trend most distinctly for NH_4-N. (See the small prints in figure 11.5.) This was confirmed by the ANOVA. For TN and NO_{2+3}-N the graphical representation did not indicate the same type of indisputable trend, and the ANOVA of these data did not identify temperature as a significant source of variation for either of these data sets. In fact method variation over time was significant in the ANOVA of the NO_{2+3}-N data (the effect $\mathfrak{T}$ in model 3.5, summarized in table 3.3), but the reference temperature ratio plot did not react to this. Based on the present data sets of stable analytes, TN and NO_{2+3}-N it is concluded that the reference temperature ratio plot method is robust towards the variations of the applied analytical methods.

The grid method is applicable only to time delimited studies. The advantage of the grid method is its flexibility in identifying relevant storage temperatures, shipping conditions and temperatures relevant for accelerated decay studies. If e.g. instability shows up when samples are transferred to temperature level 4, conclusions can be drawn from this that storage must take place at lower temperature levels, that there may be an instability problem which needs further investigation, or that shipping conditions must be carefully controlled. The design also offers the possibility of reducing the number of test items at this point, because samples are transferred to higher storage temperatures in stead of storing samples at all of the storage conditions from the beginning. Data collected in a grid storage study design could be combined with either the ratio methods or ANOVA.

The isochronous method effectively overcomes the problem of significant analytical method variation. However, with this method there is no knowledge available of how the material develops until at the very end of the pre-determined time period. It can also be more difficult to identify the time of change because it may have taken place at any number of time units after samples were transferred to a higher storage temperature. The design could be improved by increasing the number of times for transfer of samples from the reference temperature to $t°$, but such a step also increases the amount of analytical work at the end of the study. If the number of samples is too large to be analysed in one series, care must be taken in the design of the analyses to make sure that analytical variation between days can be distinguished. An ANOVA model would be the best suited choice for detailed data analysis. The ratio methods can be applied for graphical inspection of the results, if chemical analysis can be carried out under repeatability conditions.

The ANOVA model (3.5) applied to the wastewater data in chapter 3, identified temperature dependent concentration changes for NH_4^+-N, method related variation for NO_{2+3}-N and sample variation for TN.

Hotelling's T^2 is most effective in detecting shifts from in-control mean levels when the shift is fairly large. As illustrated in figure 11.2, a change in concentration is expected to take place gradually. It must be taken into account that this type of method may be sensitive to analytical method variations. As seen in chapter 3, table 3.11, the number of significant statistics increased with the storage temperature. For the intermediate storage temperature all of the statistics were significant except for two intermediate time points, and similar incoherent observations were made for the lowest storage temperature where the statistics of two non adjacent time points were significant. With the present wastewater data sets, the conclusion was unambiguous for the 37°C data, but it was harder to have complete confidence in the results seen for the low and intermediate storage temperatures. Two of the time points $\tau=5$ months and $\tau=14$ months stand out for the Hotelling's T^2 and also for the M-chart statistic on the 4°C data.

One advantage of the M-chart is that it is effective for detecting linear trends as a sign of out of control situations. A linear trend is exactly what is expected in a situation of change (fig. 11.2). A second advantage is that it does not involve the covariance matrix directly and μ can be estimated if necessary (in pre-certification studies). Whereas the Hotelling's T^2 was out of control for all time points at 37°C wastewater data, the M-chart is more reluctant and does not signal out of control until $\tau=11$ months and onwards for the highest storage temperature, 37°C.

Summing up some major points from the discussion it may be concluded on the basis of the investigations in chapters 3 and 4 that

① Ratio methods are indicative, but suffer risk of confounding effects when analytical variation control is not sufficiently tight.

② The reference temperature ratio method exhibited robustness towards measurement method variation on the NO_{2+3}-N wastewater data analysed in chapter 3. (The method related variation was indicated by the ANOVA.)

③ Isochronous measurement only applies to time delimited studies and demand a relatively large laboratory capacity for handling many analyses under repeatability conditions.

④ ANOVA models can identify sources of variation.

⑤ Data measured on multicomponent RMs are tedious to treat by one signal at a time methods.

⑥ M-chart procedure may have too long retention times until a signal is given, depending on the number of unstable components in relation to the total number of measured components.

⑦ The problem of separating time effect and analytical variation is not easily solved at low cost.

⑧ Variation between units is larger for natural samples than for synthetic samples, the power of any type of test will necessarily be influenced negatively by this fact.

Appendix A

Production of wastewater reference materials

A.1 Preparation

The samples were diluted 5 times prior to the phosphate analyses and 10 times prior to the nutrient analyses with the exception of the ammonium analysis which was performed without dilution. NO_{2+3}, ortho-phosphate and ammonium must analysed on the day the bottles are opened. Two samples from each storage temperature are measured with replicate measurements in randomized order in all analyses.

07.06.96	Water was taken from the outlet at Usserød Wastewater Treatment Plant and stored at 4°C.
12.06.96	The wastewater was passed over a GF/A filter into two containers. The filter paper was changed once during the procedure. The water was filled into 250 ml infusion bottles. This step lasted about 1.5 hours. Because of a troublesome cap closing device the procedure of closing the bottles was not finished until 5 hours after the completion of the filling. Half an hour later, the bottles were autoclaved at 121°C for 20 minutes. In total there were 185 bottles.
25.07.96	60 bottles were stored at 20°C, another 60 at 37°C, and the rest were kept at 4°C.

A.1.1 Analytical results

The samples were analysed using VKI equipment and materials. The applied methods were Danish Standard (DS) methods - in some cases adapted to VKI equipment.

Analyte	Method	Comments
TN	DS 221 *Determination of nitrogen content after oxydation by peroxodisulphate* 1. udg. December 1975	Auto Analyzer
NO_{2+3}-N	DS 223 *Water Analysis - Determination of the sum of nitrite- and nitrate-nitrogen* 1. udg. December 1975	Auto Analyzer
NH_4-N	DS 224 *Water Analysis. Determination of ammonia-nitrogen* 1. udg. December 1975	Manual
OP	DS 291 *Water Analysis - Phosphor Photometric method* 2.udg. Marts 1995	Manual
TP	DS 292 *Water Analysis - Total Phosphor Photometric method* 2.udg. Marts 1995	Manual

Table A.1: Analytical methods used for wastewater analyses.

A.2 Statistics and data plots.

Sample no. 1 has been excluded as a Grubbs' outlier. The Grubbs' test was performed on the mean of replicate measurements of ortho-phosphate and total phosphorous.

A.2.1 QQ-plots of raw data

The following figures are the plots of differences between replicates on samples and the differences between sample means within storage temperature and time. The calculations are as described in equations (3.1) and (3.2) in chapter 3. In figure A.1 the tails of plot a) and b) are somewhat irregular. The observations responsible for this are the last observations in the data sets, at $\tau = 22$. At this point a colloid precipitate had formed in the samples, which is probably responsible for the enlarged intra-sample differences. The Anderson-Darling test was performed for intra-sample as well as between-sample differences.

The Anderson-Darling test was performed separately for each storage temperature on intra-sample differences. The QQ-plots were also verified separately for each storage temperature, but these plots are not included here.

For the inter-sample differences (calculated within storage temperature and time), the Anderson-Darling test was performed for the full data sets across storage conditions and time. The normality assumption was readily accepted at $\alpha = 0.05$ for TN, NO_{2+3}-N, and NH_4-N.

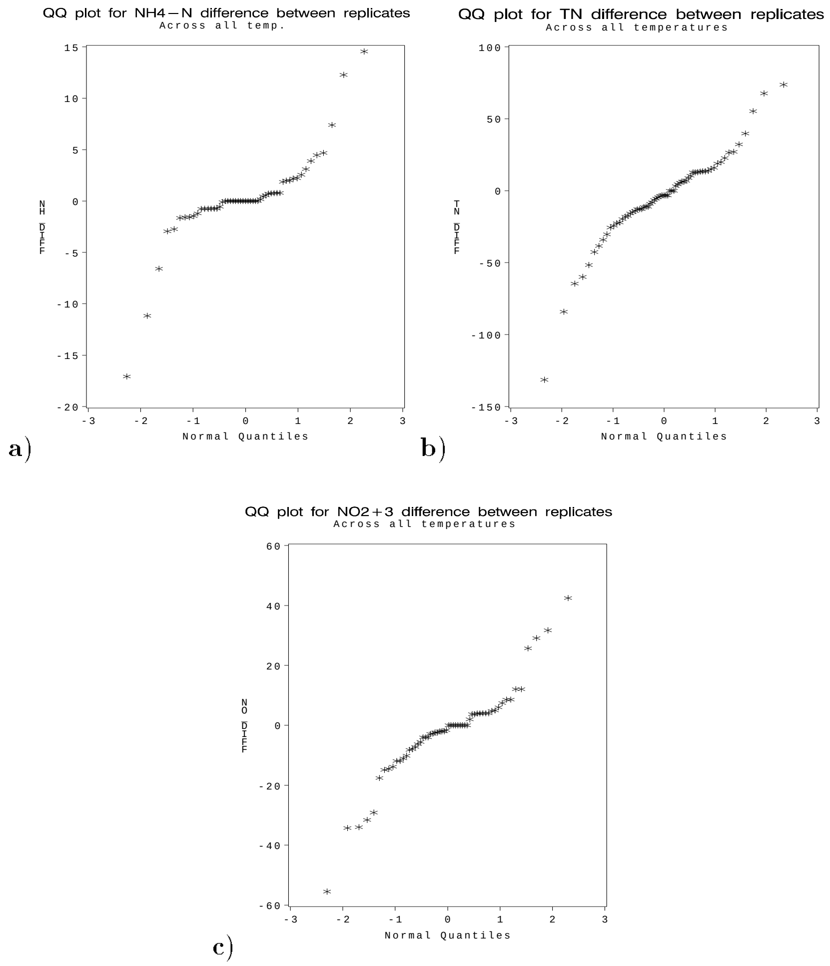


Figure A.1: Wastewater: QQ-plot for difference between replicate measurements. a) NH_4-N raw data, b) TN c) NO_{2+3}-N.

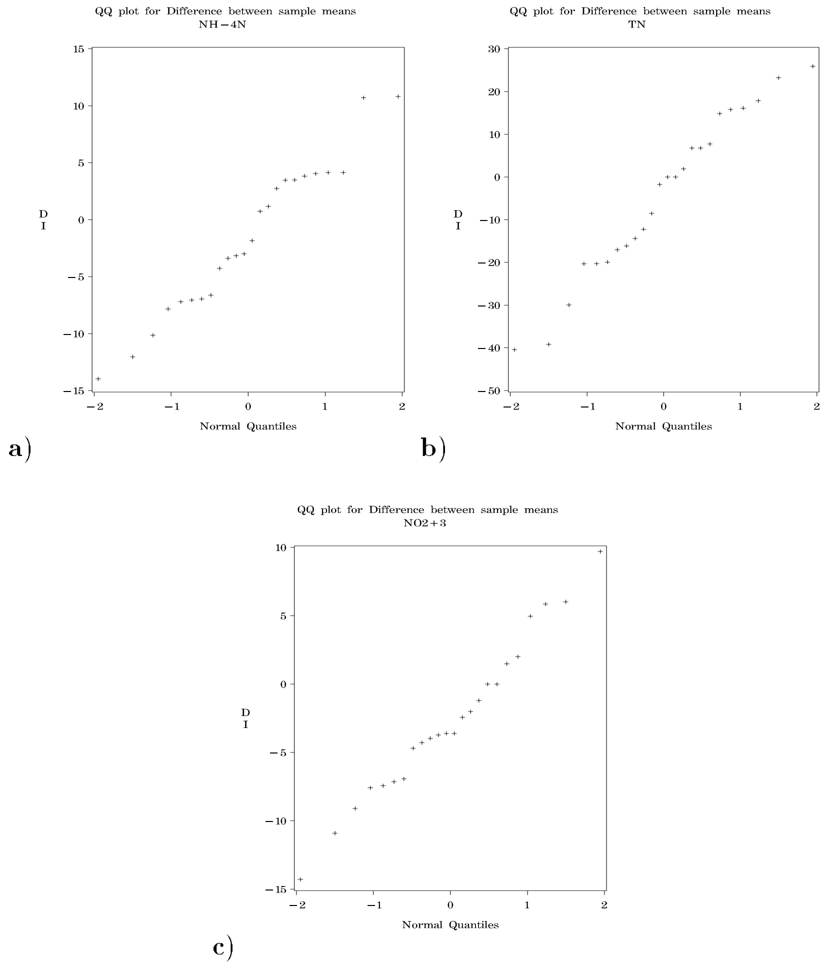


Figure A.2: Wastewater: QQ-plot for difference between sample means within storage condition and time. a) NH_4-N raw data, b) TN c) NO_{2+3}-N.

A.2.2 Plot of means by temperature and time

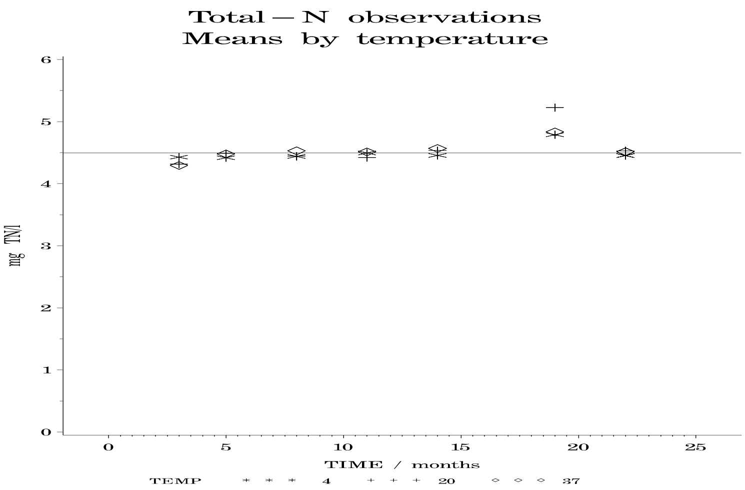


Figure A.3: TN data. Means calculated within each storage temperature, plotted versus time.

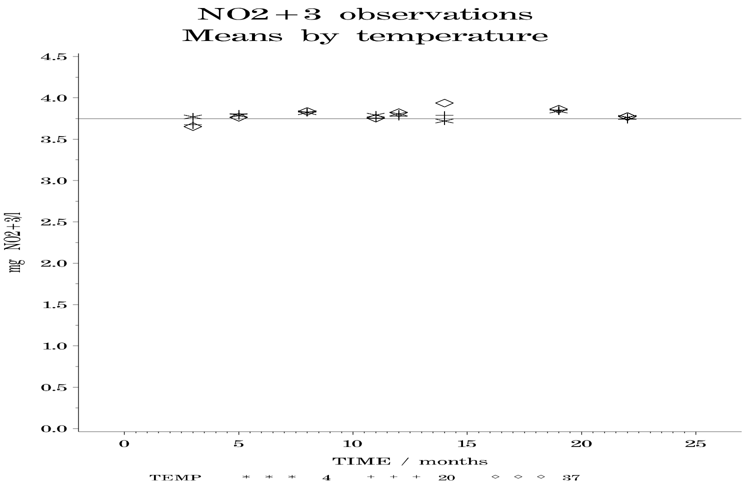


Figure A.4: NO_{2+3} data. Means calculated within each storage temperature, plotted versus time.

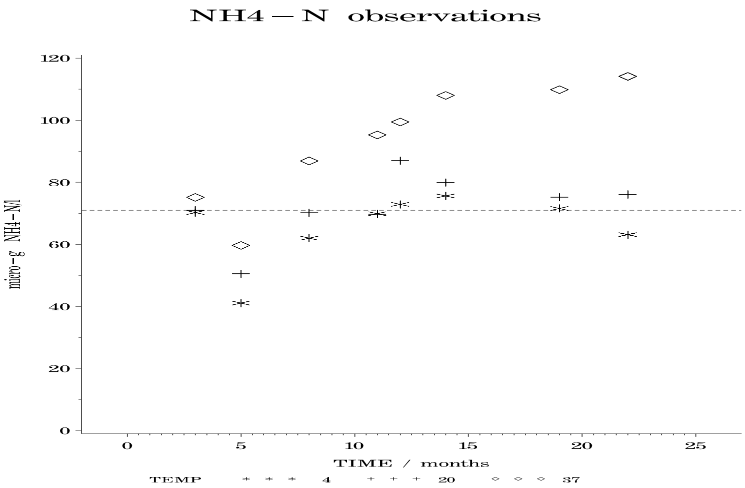


Figure A.5: NH_4-N data. Means calculated within each storage temperature, plotted versus time.

Appendix B

Production of seawater reference materials

Seawater needs pre-filtering to remove algae and the like. The ultra-filtration referred to as an actual filtration step in the following, is filtration through a filter used for blood treatment: dialysis. A final autoclaving step in sealed bottles was common for all samples. For these experiments only two temperatures were chosen: 4°C and 37°C.

The material was prepared and stored according to the following table.

Data	UV radiation	Filtering	Storage
(1)	No	No	4°C
a	Yes	No	4°C
b	No	Yes	4°C
ab	Yes	Yes	4°C
c	No	No	37°C
ac	Yes	No	37°C
bc	No	Yes	37°C
abc	Yes	Yes	37°C

Table B.1: Preparation and storage of seawater RM.

Eight samples are analysed at each point in time with 2 replicates per sample. Samples were analysed in randomized order. The sample mean of two determinations was entered in the ANOVA. The analysis of variance gives when applied separately at each time point:

Variation	Degrees of freedom
Main effect of UV radiation	1
Main effect of filtration	1
Interaction UV*filtration	1
Main effect of storage temp.	1
Interaction UV*storage temp.	1
Interaction filtration*storage temp.	1
Interaction UV*filtration*storage temp.	1
Residual	8

Table B.2: Effects in ANOVA of seawater experiments.

B.0.3 Preparation

The sketch in figure B.1 illustrates the arrangement for the UV treatment. The lamps were suspended over the water surface by means of a special type of rack used for providing light to biological experiments at VKI.

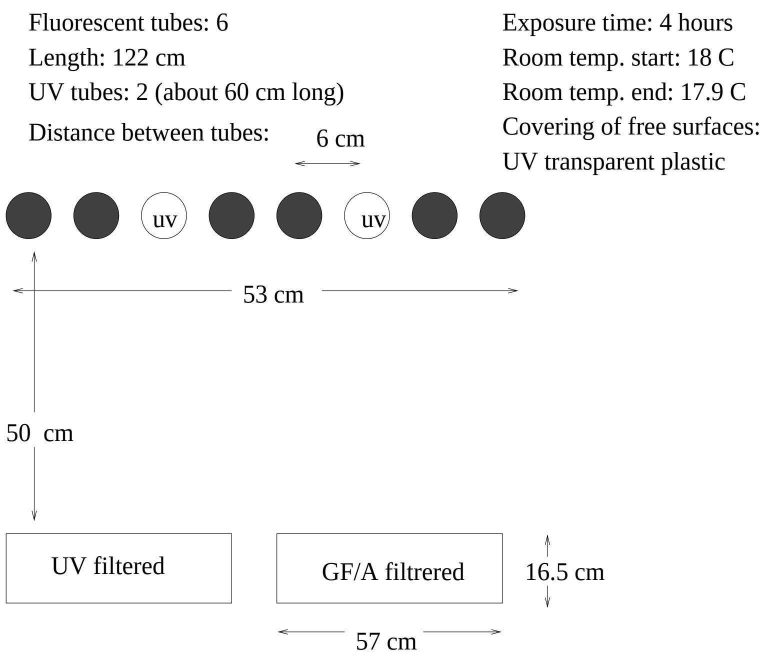


Figure B.1: Illustration of set up for UV radiation of seawater

17.06.96	Two containers of seawater were taken from Øresund. Stored at 4°C until further treatment.
02.07.96	The water in the two containers differed slightly in salinity. The water from both containers was pumped simultaneously through a GF/A filter into a third container by means of a T-tube. Thorough violent shaking followed to ensure mixing. The salinity was measured and found to be 2.24 %. All containers used for filtration had been washed in hydrochloric acid and rinsed twice with the water passed through the filter. The filtration process lasted about 1 hour and took place at 15°C. The ventilation was shut off during that period to minimize risk of contamination. The inner tube of the pump used for filtration was new and had not been used for any other type of water. All other tubes used were clean tubes from the chemistry laboratory at VKI.
03.07.96	Half of the water was poured into clean, open containers and covered the best way possible by UV transparent plastic. The water was UV radiated for four hours and the ventilation was shut off during that time. A white back ground light was also turned on in the room during the UV radiation period. After UV radiation half of the water was pumped GF/A filtered into a clean container washed in hydrochloric acid. The other half of the water was passed through a GF/A filter and an ultra-filter into a clean container. The water was stored at 4°C.
04.07.96	The four types of water were bottled. Bottling lasted about 3 hours and took place in the same room as the UV radiation. The ventilation was shut off during bottling. As much of the handling as possible was carried out wearing gloves and a fully covering paper suit. The bottles were cooled at 4°C for 2 hours until autoclaving. Autoclaving conditions: 121°C for 20 min. Bottles were stored at 4°C after autoclaving. Batch: 112 bottles
25.07.96	Half of the bottles were moved to storage facilities at 37°C. The other half remained at 4°C.

B.0.4 Analytical results

Seawater samples were analysed for phosphates and ammonium right after production. The methods used were DS 224 and DS 292 as listed in table A.1. It was suspected that he results labelled * in the table below were caused by contamination of samples during the analyses. Levels of these magnitudes were not seen in any later analyses of the material. The results were excluded from the statistical analyses.

Sample	Ortho-phos.	Total phos.	N	NH_4-N
	μg/l			
UF	10.75	12.40	226.23	31.78
UF	7.78	11.09	233.18	31.23
GFA	8.52	12.84	188.75	27.83
GFA	7.22	14.37	187.61	28.29
UV-UF	8.52	100.45	509.17*	72.70*
UV-UF	8.52	100.23	512.97*	74.69*
UV-GFA	8.71	34.03	260.41*	47.58*
UV-GFA	14.27	34.03	268.01*	52.54*

Table B.3: Analytical results for seawater RM at τ_0. Units in μg/l. * Suspicion of pollution.

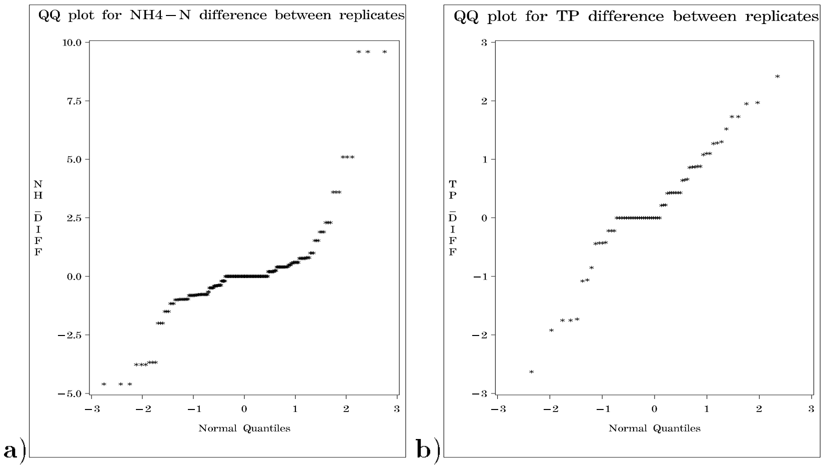


Figure B.2: Wastewater: QQ-plot for difference between replicate measurements. a) NH_4-N raw data. b) TP raw data.

B.0.5 QQ-plots of raw data

The following figure B.2 shows the plots of differences between replicates on samples. The calculations are as described in eq.(4.1) in chapter 4. The somewhat heavy tails in figure a) is due to observations late in the study at the highest storage temperature. The instability of ammonium in the samples is responsible for the increasing intra-sample differences. The Anderson-Darling test has been performed on the full data sets of differences between replicates and in sub sets of data divided according to storage condition and combination of filtering and light treatment. All of the tests were accepted at $\alpha = 0.05$.