Cranfield University

Wilfrid BOURGEOIS

On-line monitoring of wastewater quality using a chemical sensor array

School of Water Sciences

PhD

Cranfield University

School of Water Sciences

Ph. D. Thesis

Academic year 2002-2003

Wilfrid BOURGEOIS

On-line monitoring of wastewater quality using a chemical sensor array

Supervisor: Dr R. M. STUETZ

December 2002

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

... *"sitôt que j'eus achevé tout ce cours d'études, au bout duquel on a coutume d'être reçu au rang des doctes .... je me trouvois embarrassé de tant de doutes et d'erreurs, qu'il me sembloit n'avoir fait autre profit, en tâchant de m'instruire, sinon que j'avois découvert de plus en plus mon ignorance. Et néanmoins j'étois en l'une des plus célèbres écoles de l'Europe, où je pensois qu'il devoit y avoir de savants hommes, s'il y en avoit en aucun endroit de la terre. J'y avois appris tout ce que les autres y apprenoient; et même, ne m'étant pas contenté des sciences qu'on nous enseignoit, j'avois parcouru tous les livres traitant de celles qu'on estime les plus curieuses et les plus rares, qui avoient pu tomber entre mes mains."*

..."*as soon as I had finished the entire course of study, at the close of which it is customary to be admitted into the order of the doctors..... I found myself involved in so many doubts and errors, that I was convinced I had advanced no farther in all my attempts at learning, than the discovery at every turn of my own ignorance. And yet I was studying in one of the most celebrated schools in Europe, in which I thought there must be learned men, if such were anywhere to be found. I had been taught all that others learned there; and not contented with the sciences actually taught us, I had, in addition, read all the books that had fallen into my hands, treating of such branches as are esteemed the most curious and rare."*

René Descartes, 1637.
Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences, Part I.

ABSTRACT

Real-time and on-line monitoring of wastewater quality is a subject of growing concern both in the United Kingdom and internationally. Global organic load parameters are traditionally used to define wastewater characteristics and to demonstrate that a wastewater treatment plant meets statutory law. However these measurements are mainly based on sample collection and retrospective analysis which can be time consuming. Existing on-line instruments remain limited by environmental factors, short lifetimes and fouling problems due to the harsh environment in which they have to be located. The recent availability of commercial sensor array instruments could offer a rapid and relatively simple technique for non-invasive monitoring.

This study explored the use of a commercial sensor array system to continuously monitor wastewater organic load. Preliminary experiments using a 12 CP sensor array (ENose5000, Marconi Applied Technologies) were carried out under laboratory controlled conditions and a sampling system was developed so as to achieve a high sampling frequency. Wastewater samples were continuously pumped through a temperature-controlled flow cell and sparged with nitrogen to generate a headspace gas for sensor array analysis. A study was carried out which quantified the effects and interactions of temperature, gas flow rate and sparger porosity on humidity levels and reproducibility. A sampling procedure was selected based on these observations. Results obtained from recirculated batch wastewater samples showed good reproducibility. PCA and MDA demonstrated that the system could distinguish between different types of wastewater (raw sewage, final effluent, RO water) and between different dilutions of a sample (100%, 50%, 25% RO water). Quantitative analysis using Multiple Linear Regression showed that a good prediction of organic load parameters could be achieved under carefully controlled conditions ($R = 0.86$). A time-dependent relationship was observed and the model rapidly collapsed when applied to unknown samples collected 1 week after those used for training.

A modified system consisting of 8 CP sensors was implemented in the field at Cranfield University's wastewater treatment plant. The array was coupled to an on-line TOC instrument (TOC 4100, Shimatzu, UK) and Racod (USF, UK) in a 12-month monitoring trial of a primary settled effluent. Examination of the sensor array data showed the effect of 'diurnal' and weekly variations, heavy rain, operating anomalies as well as pollution incidents on the profiles. Descriptive statistics of data acquired under normal conditions showed a slight deviation from the normality assumption (positive skewness), and a strong correlation between the sensor responses and RH. Non-linearities were also suggested.

Statistical analysis of the relationship between the measured organic load parameters and sensor responses was carried out. Multivariate studies using both linear (MLR, PLS) and non-linear (Polynomial, Factorial) regression techniques proved unsatisfactory in predicting TOC over long periods of time. The influence of the duration of training proved to be decisive on the long-term validity of the models. By comparison the amount of data used in the training phase had little impact on the outcome of the prediction. Extraction of variance and the use of principal components as independent variables for MLP, Polynomial and Factorial regression did not improve the performances.

The best results were achieved with Artificial Neural Networks, An 8-4-1 MLP using the 8 sensors' $\Delta R/R$ as inputs could successfully predict 91.2% of cases with less than 30% error. The relatively low correlation ($R = 0.49$) was an indication that the predicted values were limited to a narrow range around the mean measured TOC concentration and that the model failed to predict higher and lower values. Pre-processing of the data and exponential smoothing to reduce the effect of noise was shown to be essential and significantly improved the predictions ($R = 0.71$). As a rule, the quality of the data, the time dependent relationship and the high correlation with RH were the 3 most limiting factors. Drift (Sensors 1-4) was also present but did not affect the system's ability to detect changes. On the other hand, Sensors 5, 6, 7 and 8 often failed to return to their original baseline. This poisoning of the sensors had a major impact and significantly increased the prediction error. Reducing acquisition times and increasing the sensor de-purge period was recommended.

Despite the difficulties in establishing a clear relationship between the sensor array data and TOC from field data, this study pointed out the great potential of the technique as an upset early warning device. A model was built which is based on a moving window and simple data mining procedures for the detection and identification of operating anomalies and pollution incidents. With further development the technique could provide a non-invasive alarm system at the inlet of the plant and trigger the sampling of the offending water, bypass the influent or alter the wastewater treatment processes.

ACKNOWLEDGEMENTS

I would like to thank Richard Stuetz for his support and encouragement throughout the course of this three-year study. I am extremely grateful for his trust, friendly guidance and sincerity regarding all aspects of my work within the School of Water Sciences. His open-mindedness and ability to put things into a broader perspective were particularly appreciated.

This work would have not been possible without the help and support of many at Marconi Applied Technologies. I gratefully acknowledge John Warburton, Mark Byfield, and Andrew Pike for their keen interest in my work. A special thank you is due to Gurjit Kang, Pat Casey and Neil Collins for their precious assistance as well as their invariably prompt replies to my most desperate emails.

I would like to thank Chris Jones and Dereck Brown of Northumbrian Water Ltd for their trust and their support of this research. The financial support of EPSRC is also acknowledged.

Thanks must also go to Mathieu Gaugler, Maxime Porterie, Guillaume Gardney, Myriam Servieres and Guillaume Marshall for their help and efforts. I am indebted to them and hope they enjoyed their stay in "sunny" England as much as I have enjoyed their company and good teamwork.

Janice, John and James also deserve a mention for their generosity and kindness. Over the past few years they have become my family on this side of the Channel, and taught me a great deal about the country (not to mention fine ales, single malts and G&T). Thanks to John for letting us use his boat and a warm thank you to Tom and Shirley for putting up with me in Argyll.

Most backing and comfort was given by Juliette Butcher who constantly supported me and kept me sane over these years. This means a great deal to me and I would like to wish you luck with your thesis! Last but foremost I would like to express all my gratitude and my love to my parents and my brother, who have always supported me during my many years of study: Merci Papa, Maman et Gwendal pour votre soutient et votre affection. J'espere que l'éloignement pendant toutes ces années en vallait la peine. Merci également à mes grand-parents et à Pierrick. *Da garan*. This thesis is dedicated to you.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANN	Artificial neural network
ANOVA	Analysis of variance
BAW	Bulk acoustic wave
BOD	Biochemical oxygen demand
BP	Back propagation
CB/CP	Carbon black/composite polymer
CA	Cluster analysis
CC	Canonical correlation
COD	Chemical oxygen demand
CP	Conducting polymer
DFA	Discriminant function analysis
DO	Dissolved oxygen
DV	Dependent variable
ETACS	European testing and assessment of comparability of on-line sensors/analysers
FFANN	Feed forward artificial neural network
HTCO	High temperature catalytic oxidation
IV	Independent variable
MDA	Multiple discriminant analysis
MLP	Multilayer perceptron
MLR	Multiple linear regression
MCERTS	Monitoring certification scheme
MOS	Metal oxide sensor
MOSFET	Metal oxide semiconducting field effect transistor
MVS	Multivariate statistical analysis
NCAS	National compliance assessment sercice
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares
QCM	Quartz crystal microbalance
RH	Relative humidity
RSD	Relative standard deviation
SAW	Surface acoustic wave
SOM	Self organising map
SD	Standard deviation
TOC	Total organic carbon

UK	United Kingdom
VOC	Volatile organic compound
VSS	Volatile suspended solids
WCO	Wet chemical oxidation
WW	Wastewater
WWTP	Wastewater treatment plant

Chapter 1: INTRODUCTION

CHAPTER 1: INTRODUCTION

As public awareness of environmental issues rises and governments take on international commitments to reduce emissions, many industries are now faced with progressively tighter tolerance margins and a rising demand for rigorous quality criteria. As a result of this need to regulate and control levels of pollution, instrument manufacturers have sought to provide suitable environmental monitoring solutions. In this rapidly developing field, the advancement of increasingly sophisticated instrumentation has emerged in areas which, traditionally, had seen little changes (Plumey, 1999). Applications include a broad range of activities. Of particular interest are the potentially polluting industrial installations such as those of the energy, defense, chemical, paper, food, agriculture and waste processing (landfill sites, wastewater treatment plants) industries. This diversity of sectors concerned is a source of many opportunities for the environmental industry. In many cases, this requires costly method development in order to suit application specific requirements

Monitoring of wastewater quality parameters is currently a subject of growing concern both in the United Kingdom and internationally. In the past few years, a number of European regulatory measures and recommendations, such as the 91-271 EEC directive, have put pressure on wastewater treatment plant operators with respect to discharge requirements. In order to comply with these regulations on a permanent basis, and because of the spatial and time dependant variability of wastewater characteristics, on-line monitoring of global organic load parameters is clearly needed.

Biological wastewater treatment plants are a critical component of urban pollution control strategies and must be designed and operated properly in order to effectively

prevent contamination of natural surface waters or to avoid downstream problems in water reuse operations. Wastewater collected from municipalities and communities must ultimately be returned to receiving waters or to the land and must therefore be treated by a series of physical, biological and chemical treatments in order to produce an effluent that complies with the standards set by the regulatory authorities.

The processes commonly used for the treatment of municipal wastewater can be typically divided into four different steps which are described in a number of textbooks (Metcalf and Eddy, 1991; Droste, 1997; Parsons and Stephenson, 2003). These are: preliminary treatment to remove gross solids, primary treatment to remove biological solids and some organic load, secondary treatment to remove organic carbon and nutrients, and tertiary treatment to remove fine solids and nutrients. A classic wastewater treatment scheme is represented in Figure1.1.

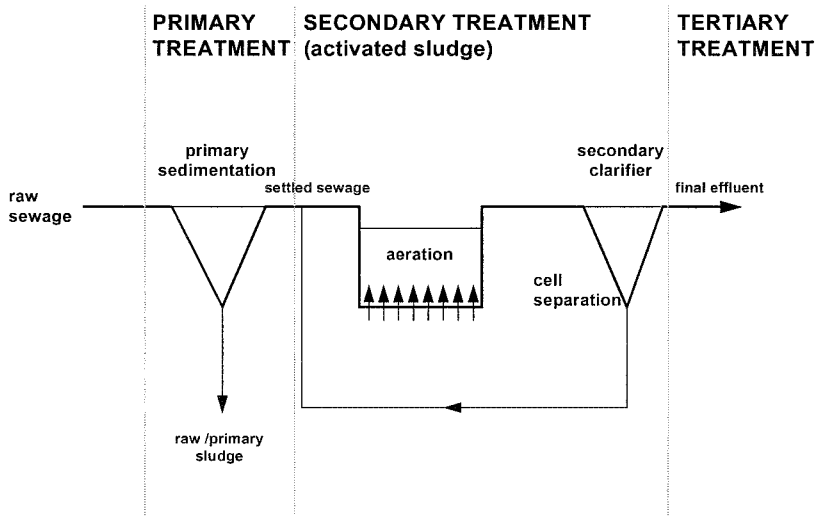


Figure 1.1: Levels of treatment used in municipal wastewater treatment (from Parsons and Stephenson, 2003)

Traditionally, wastewater treatment plants are monitored by determining the quality of the effluent at the outlet of the treatment works using global parameters such as Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Organic Carbon (TOC) and Total Suspended Solids (TSS) (Thomas *et al.*, 1997;

Cecile, 1998; Wacheux, 1998). In addition to providing vital information on the quality of the effluent and treatment efficiency, these procedures demonstrate that a wastewater treatment plant meets statutory discharge requirements (Bourgeois *et al.*, 2001).

At present, however, there is often infrequent monitoring of wastewater quality at most treatment plants and real-time monitoring of wastewater quality is an unresolved problem in sewage treatment. The automation of wastewater systems is not as developed as other process industries, mainly because of the hostile environment in which sensors have to be located (Lynggaard-Jensen, 1999). The lack of sensors suitable for on-line real-time monitoring/control is often due to uncertainties regarding the reproducibility/reliability of existing methods, whereas, more sensitive and standardised laboratory-based techniques are time consuming, and require sample collection and retrospective analysis. Furthermore, straightforward extrapolation of laboratory measurements is not satisfactory, since grab samples (taken on a daily basis in the best case scenario) are unlikely to provide a meaningful and high-resolution picture of the nature of, and variation in wastewater quality. In Europe, the owners and operators of treatment plants, together with the producers of monitoring equipment have expressed an urgent need for new standards and the improvement of comparability, reliability and quality of existing techniques, as well as, to develop new, fast-responding technologies (Collin and Quevauviller, 1998; Pouet *et al.*, 1999). Therefore there is considerable impetus to develop a measurement technique for wastewater monitoring that is reliable, reproducible and non-invasive.

Recent developments in sensor array technology for detecting and characterising odours could offer a rapid and relative simple technique for monitoring for changes in wastewater quality. Odours have long been used by bioprocess operators to identify types and stages of process, as well as detect abnormalities (Namdev *et al.*, 1998). Recent results have shown that a sensor arrays analysis of quiescent wastewater (using a lab-based system) can be used to differentiate between different wastewater types (raw sewage, settled sewage and final effluent) and from different works (Stuetz *et al.*, 1999a). Such studies have shown the potential of this technology in term of the relationships between odour types/ concentrations and

sensor array data, and consequently its wider application to on-line monitoring and control of water and wastewater treatment processes. This demonstration that electronic noses can discriminate between different samples of various nature, as well as monitor their stability and changes in quality with time, was mainly made possible by the recent developments in computing technologies and our better understanding in the field of artificial intelligence. The list of potential applications and the number of system designs reported in the literature is continuously growing, encouraged by the development of new sensor materials and the apparition of smarter, smaller, more versatile and more sensitive devices.

However, despite the progresses reported in the literature, environmental applications, and in particular the study of liquid samples, remain one of the most demanding areas for sensor arrays. Most of the original work is generally limited to bench type laboratory applications and has been carried out with laboratory-based sensor systems originally developed for other applications such as foodstuffs, beverages and perfumes discrimination. With the limited number of trials carried out under realistic condition, there is a case for such system to be developed and tested in the field at an operating wastewater treatment plant. A full investigation of the potential of non-specific sensor arrays must be carried out in order to gain the requisite expertise and understanding that will constitute the basis for future developments.

This study looks at some of the fundamental and practical issues associated with the application of sensor array technology to continuous monitoring of wastewater organic load. The potential of a commercial array of conducting polymer sensors is evaluated with the focus on the development of a sampling strategy and the identification of an appropriate data analysis protocol.

Chapter 2: LITERATURE REVIEW

CHAPTER 2: LITERATURE REVIEW

2.1 CHEMICAL SENSORS AND NON-SPECIFIC SENSOR ARRAYS

2.1.1 Introduction

In the last two decades, there has been an increasing interest in the development of sensor array technology. Although early attempts to use an array of non-specific sensors to discriminate between odours have been made (Moncrieff, 1961; Buck *et al.*, 1965; Dravnieks and Trotter, 1965), the concept of an intelligent chemical sensor array system did not emerge until twenty years later following publications by Persaud and Dodd (1982), Pelosi and Persaud (1988), Ikegami and Kaneyasu (1985) and Kaneyasu *et al.* (1987). Thanks to the recent developments in sensor materials, electronics and computing technologies, the commercialisation of so-called electronic noses which are capable of detecting and recognising complex odours started to take place in the mid 1990's (Gardner and Bartlett, 1999). A definition of an electronic nose has been given by Gardner and Bartlett (1994):

An electronic nose is an instrument which comprises an array of electronic chemical sensors with partial specificity and an appropriate pattern recognition system capable of recognising simple or complex odours.

However this definition limits the technology to those types of sensor array systems that detect odorous compounds as perceptible by biological olfaction. Consequently, the more generic terms "sensor array" or "non-specific chemical sensor array" are often preferred since the sensors respond to both odorous and odourless volatile compounds. These odourless compounds are equally essential to the recognition

process of a sample. Stetter *et al.* (2000) recently redefined an electronic nose in term of structure and function rather than application:

The electronic nose is an instrument comprised of a sampling system, gas/vapour sensor array, and PC for the purpose of quantitative or qualitative analysis.

Although this is still a relatively new technology, sensor arrays have already been used in many field of application, such as foodstuffs, beverages and in the fragrance industry. Commercially available instruments as well as prototypes used and developed by research institutions and universities cover a wide range of chemical sensor principles, system designs and data analysis techniques. An example of a commercial instrument with auto-sampling system and PC for data analysis is shown in Figure 2.1. In every case, the fundamental principle is that arrays of broadly tuned sensors produce patterns of response that can be used as descriptors or fingerprints to characterise a sample (Figure 2.2). This strategy provides a unique way to measure complex odours as well as to detect individual species where traditional selective chemical sensing methods may have difficulties.

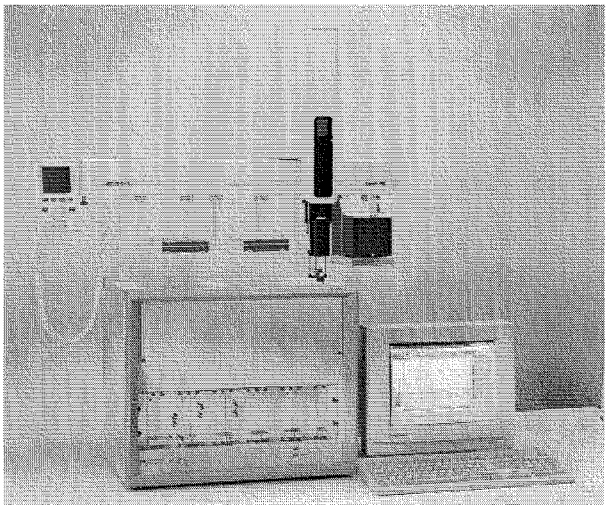


Figure 2.1: Commercially available sensor array: eNose 5000, Marconi Applied Technologies, UK.

Each sensor may respond differently to a volatile or gas mixture. The intensity of the sensor response will depend on the concentration of the analytes present as well as on the affinities of these chemical species towards individual sensors. Because of their overlaying selectivity, the relative signals from the individual sensors represent a pattern that is unique to the gas/headspace sample being measured. These signatures are extremely rich in information and are usually interpreted using multivariate statistical pattern recognition techniques and artificial neural networks. Some of the most commonly used pattern recognition techniques are discussed in section 2.1.4.

In principle such instruments can be applied to any product that gives off volatiles with or without a smell provided that this occurs within the sensitivity range of the sensors (Haugen and Kvaal, 1998). A schematic of a typical sensor array system and its different components is given in Figure 2.2.

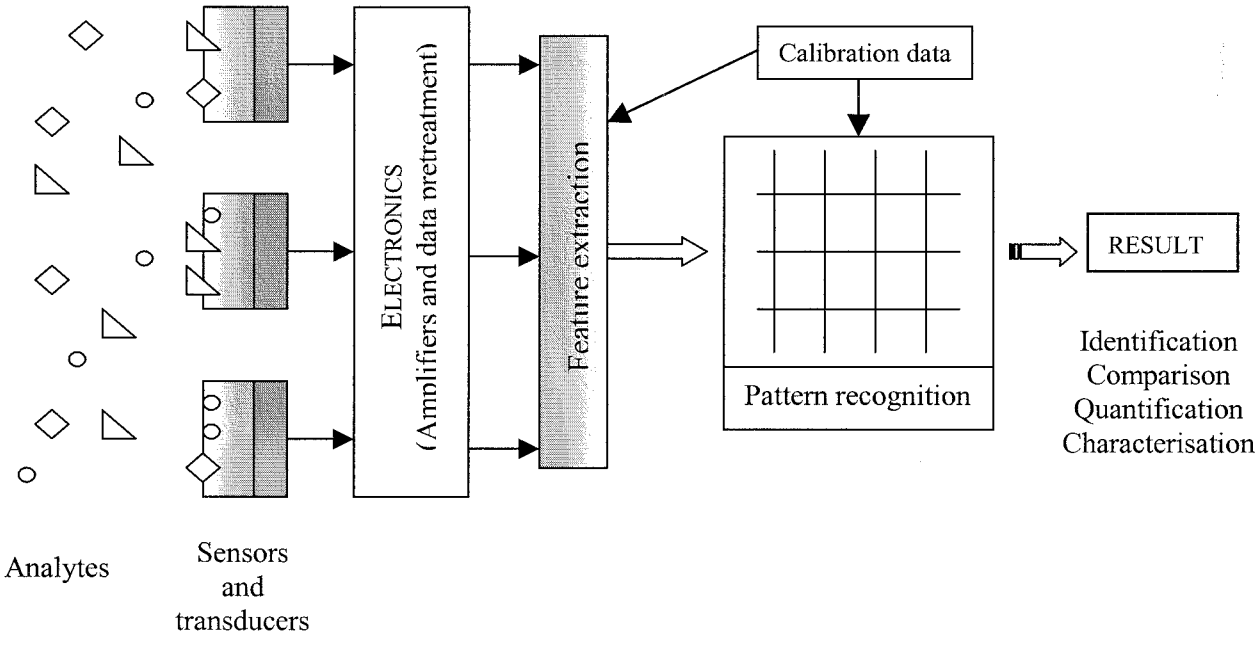


Figure 2.2: Schematic set-up of a non-specific sensor array for gas and odour recognition (adapted from Gopel, 1998)

2.1.2 Sensor types overview

The types of sensors that can be used in sensor arrays are quite varied. A definition of a chemical sensor is given by the IUPAC (International Union of Pure and Applied Chemistry):

A chemical sensor is a device that transforms chemical information, ranging from the concentration of a specific component to total composition analysis, into an analytically useful signal. As opposed to dosimeters which are non-continuously operating devices exhibiting irreversible characteristics.

In a recent review, D'Amico and DiNatale (2000) give some basic definitions of sensor properties such as sensor response curve, sensitivity, selectivity, noise, drift and resolution. The authors also pointed out how the continuous progress in microelectronic technologies, together with the developments in material sciences have been largely responsible for the birth and rapid growth of the sensor discipline over the past few years. Typically, chemical sensors comprise an appropriate chemically sensitive material interfaced to a transducer (Figure 2.3) and can be classified into four separate categories according their transduction principles (Hierlemann, 2002):

1) Mass sensitive sensors ("chemo-mechanical" sensors) that respond to mass, viscosity changes due to bulk absorption
2) Thermal sensors (thermocouples) that measure a change in temperature due to chemical interaction of the analyte with the sensor
3) Optical sensors which measure a change in light intensity
4) Electrochemicals sensors which measure a change of potential or resistance through charge transfer.

These reactions can take place on the surface or in the bulk of the sensor material (Haugen and Kvaal, 1998). Numerous implementations of these principles have emerged. However the most frequently used sensors in commercial instruments are

the metal oxide semi-conductors (MOS, MOSFET[1]) and organic polymers. The structure and working principle of a conducting polymer (CP) chemoresistor is shown in Figure 2.4. Sensor arrays based on surface acoustic waves (SAW), bulk acoustic waves (BAW) and quartz crystal microbalances (QCM) are also commercially available (Figure 2.3). Thus the measured quantity in an individual detector can be the frequency shift of a resonating crystal in a QCM or SAW configurations, change in the optical absorption or emission properties of a die that has been impregnated into a polymer, or a change in the resistance of a CP or carbon-black polymer composite (CB/PC) film (Matzger *et al.*, 2000).

As a guide, a summary of the performances of the different techniques is given in Table 2.1. It should be noted that some of the performances given in this table may be disputed due to constant developments in sensor technologies.

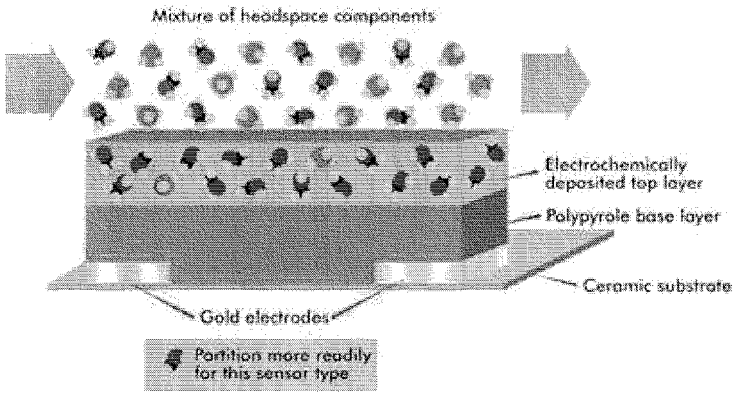


Figure 2.3: General structure and working principle of a conducting polymer sensor (from http://osmetech.plc.uk/technology/polymers.htlm)

[1] MOSFET: Metal Oxide Semiconducting Field Effect Transistor

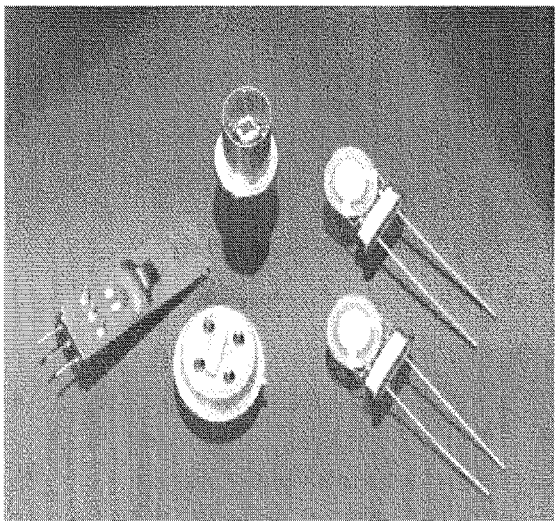


Figure 2.4: Most commonly used sensors in commercial sensor array. Clockwise from top; MOS, BAW (x2), SAW and CP.

Table 2.1: Comparative properties and performance of most frequently used gas sensors in electronic nose instruments (from Haugen and Kvaal, 1998).

Performance	MOS	MOSFET	CP	QMB	SAW
Selectivity	Poor	Moderate	Moderate	High	High
Sensitivity	>0.1ppm	>0.1ppm	0.01ppm	>0.1ppm	Ppb
Reproducibility	Poor	Good	Good	Moderate	High
Temperature dependence	Low	Low	High	Moderate	High
Humidity dependence	Low	Moderate	High	Low	Low
Carrier gas	Synthetic air (O_2)	Synthetic air (O_2)	Inert/Synthetic air (O_2)	Inert/Synthetic air (O_2)	Inert/Synthetic air (O_2)
Operating temperature (^oC)	300-400	100-200	Ambient	Ambient	Ambient
Response time (sec)	0.5-5	0.5-5	20-50	20-50	20-50
Recovery time	Fast	Fast	Slow	Slow	Slow
Lifetime (years)	3-5	1-4	1-2	<2	<2

The principles and applications of different types of sensors have been extensively reviewed in numerous studies (Persaud *et al.*, 1996; Persaud, 1997; Gopel, 1998; Gardner and Bartlett, 1999; Albert *et al.*, 2000; Lee and Lee, 2001; Wolff *et al.*, 2001; Vig, 2001; Wilson *et al.*, 2001) and are also presented in a number of universities' and manufacturers' web sites (Table 2.2). In the next section we concentrate on the particular type of sensors that have been used in this study, i.e. conducting polymer sensors

Table 2.2: Universities and Manufacturers web sites for further information on chemical sensors

URL's:
http://www.nose-network.org (short course lectures notes, members only)
http://www.ipc.uni-tuebingen.de/weimar/research/maintopics/
http://www.inapg.inra.fr/ens-rech/siab/astec/elba/sommelen.htm
http://www.osmetech.plc.uk/technology/
http://www.appliedsensor.com/
http://www.cyranosciences.com
http://www.alpha-mos.com
http://lennartz-electronic.de

2.1.3 Conducting polymer sensors

Because of their unique electrical properties, organic electrically conducting polymers present a great interest as chemical sensors materials. Since their discovery in the late 1970's by Shirakawa, H, MacDiarmid A.G and Heeger, A.J. (www.nobel.se/chemistry/laureates/2000/public.html) they have been the subject of considerable research and development. Conducting polymers have first been used in prototypes sensor arrays by Persaud and Pelosi, 1985 and later in the first commercial electronic nose (OdourMapper Ltd, 1993).

The most commonly used conducting polymers for gas sensing applications are derived from aromatic and heterocyclic compounds. Typically, the fundamental structural unit of a CP is a linear backbone composed of repeating conjugated monomers such as pyrrole, thiophene or aniline as represented in Figure 2.5.

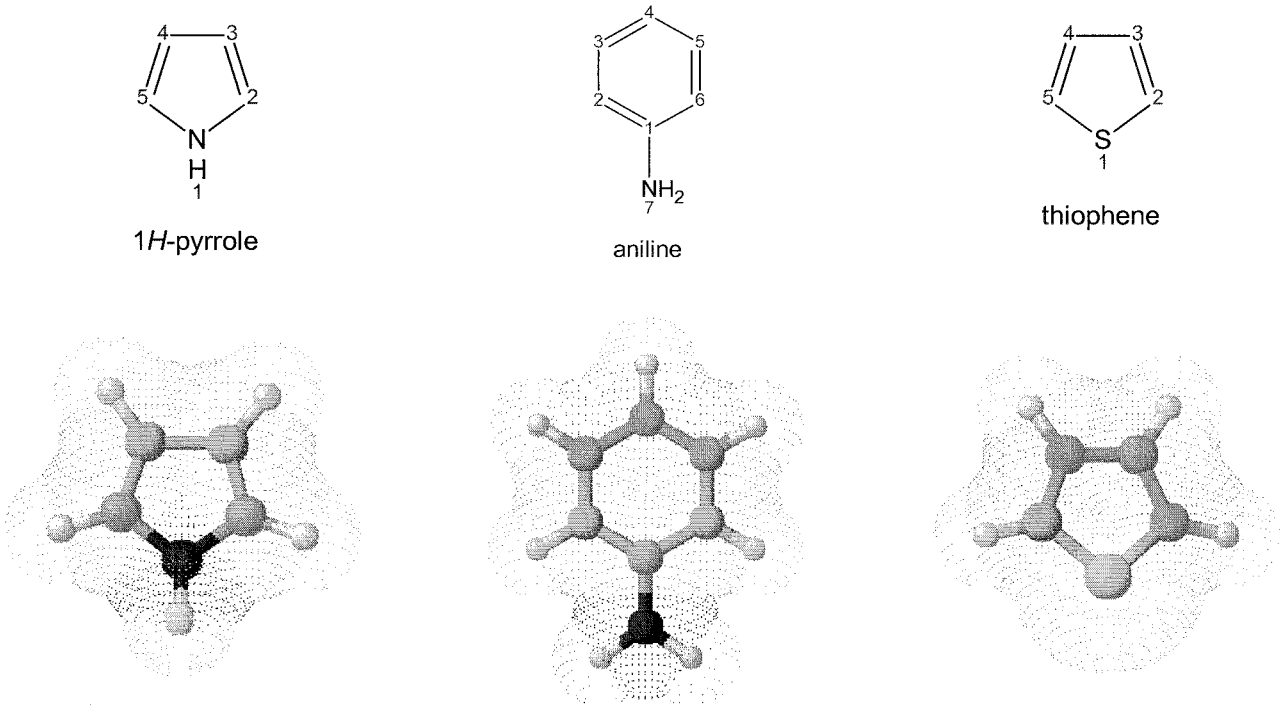
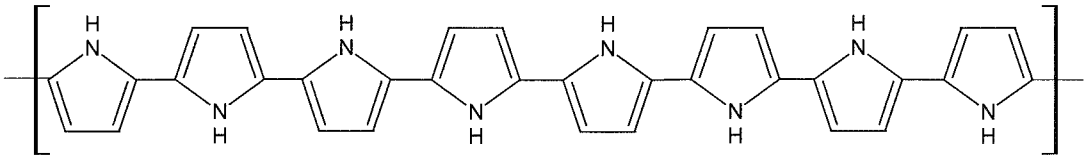
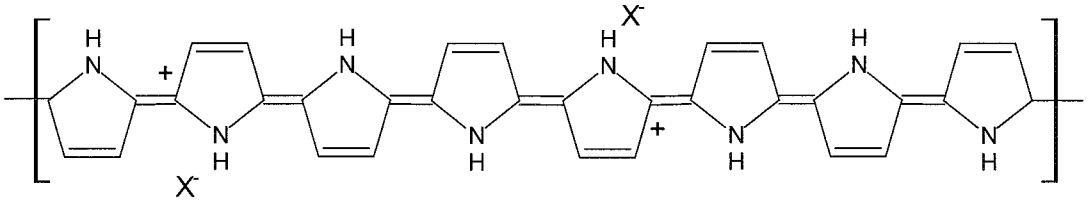


Figure 2.5: Most commonly used monomers to make conducting polymer sensors

Of most interest to us is the case of pyrrole and its derivatives, which by coupling in the alpha (2) position produce a polymer with extended conjugation as seen in Figure 2.6. This polypyrrole, however will only acquire its semi-conductive properties after partial oxidation of the polymer chain. The lonely anions (polaron and bipolarons) can then diffuse between the chains due to the relatively weak interchain binding of the polypyrrole, whilst the cations form part of the chain (Hodgins, 1995; www.nobel.se/laureates/2000/public.html).



A) Reduced state



B) Bipolaron

Figure 2.6: The structure of polypyrrole before oxidation (a) and fully doped (B)

These chemical sensors are responsive to the presence of a wide range of analytes in the vapour phase. Vapours adsorbed by a CP sensor reversibly alter the electronic properties of the polymer, causing measurable changes in its electrical resistance. Generally this is measured as a percentage change of the original resistance ($\Delta R/R$). These changes have been studied by a number of researchers. In particular, Gustaffson and Lundstrom (1987) and Miasik *et al.* (1986) studied the response of CP's to gases such as ammonia and hydrogen sulphide. Other early studies (Pelosi and Persaud, 1988; Persaud and Pelosi, 1985, 1992; Bartlett *et al.*, 1989; Bartlett and Ling Chung, 1989) investigated the response of polypyrrole films to a wide range of organic vapours. The observed conductivity changes may, or may not, be linearly dependant on the concentration of the analyte presented to the sensor, depending on the particular transduction mechanisms involved in the CP of concern (Albert *et al.*, 2000). However, the way in which these organic vapours affect the conductivity of the polymer is still poorly understood (Persaud, 1997). Some of the mechanisms by which a volatile may affect the behaviour of a conducting polymer chemiresistor have been suggested (Topart and Josowicz, 1992; Janata, 1992; Bartlett and Gardner, 1992)

and are also reviewed in Gardner and Bartlett (1999). Experimental and theoretical behaviours of intrinsically conducting polymers have been discussed in several reviews (Roncali, 1992; Skotheim *et al.*, 1998; Reddinger and Reynolds, 1999).

Despite the signal transduction mechanisms remaining unclear, CP-based sensors are of great interest since their responses are generally rapid (a steady state response can be reached in as little as 20 sec) and reversible at room temperature. Advantages over other sensor technologies include the great diversity of material that can be synthesised, their broad selectivity, good reproducibility, high sensitivity, low power consumption (few microwatts), long sensor life (few years) as well as their resilience to poisoning. Another attractive feature of CP's is their ease of preparation, which readily allows for miniaturisation and mass production of sensors. A diverse range of polymers can be prepared by chemical or electrochemical oxidation (doping) of the appropriate monomer. Electrochemical polymerisation is more widely used. It allows the process to be controlled through the choice of applied potential or current, and can be monitored, thus allowing for reproducible deposition of the polymer onto the device (Gardner and Bartlett, 1999), whilst chemical oxidation generally results in a more random organisation of the polymer (e.g. amorphous polypyrrole). However, the properties of the resulting polymer do not just depend on the choice of the monomer. The sensitivity of a sensor element can be readily altered by changing the polymerisation conditions (voltage, solvents) and the type and concentration of the counteranions that are used to balance the charges on the polymer chain during the growth stage. These charges compensating counterions often play a great role in determining the properties of the material since they can make up for a significant portion of the polymer. The oxidation state of the conducting polymer can also be altered after deposition in order to tune the sensor to the analyte of interest (Albert *et al.*, 2000; Pearce *et al.*, 1993). Finally, by attaching a carboxylic group or an alkyl chain in the beta (3) position it is possible to make a hydrophilic or hydrophobic film respectively (Gardner and Bartlett, 1999). Bio-materials such as enzymes, antibodies and cells may also be incorporated into such structures which shows the potentially limitless range of material and properties achievable.

The major drawbacks of CP's are their sensitivity to changes in temperature and humidity and long term drift. Drift is a common problem that affects the majority of

chemical sensors and their reproducibility over time. This general slow change in sensitivity of the sensors can be due to ageing effects, slow morphological changes in the sensor material and other long term effects (Gardner and Bartlett, 1999). Sensitivity to changes in humidity, temperature and gas flow rate are also important characteristics to most sensors. In the case of conducting polymer chemoresistors, the baseline temperature-dependence can generally be described by a Mott variable range-hopping model (Meikap *et al.*, 1993). Figure 2.7 and Figure 2.8 (a,b) illustrate how these changes in behaviour to both temperature and humidity depend upon the type of monomer and counterions of the conducting polymer. Although these can be limiting factors, CP sensors still have many advantages over metal oxide sensors, BAW, SAW and other devices (chiefly their diversity, rapid and reversible response and low power consumption). Consequently, they remain popular and are often a first choice for sensor arrays or electronic noses in a number of applications.

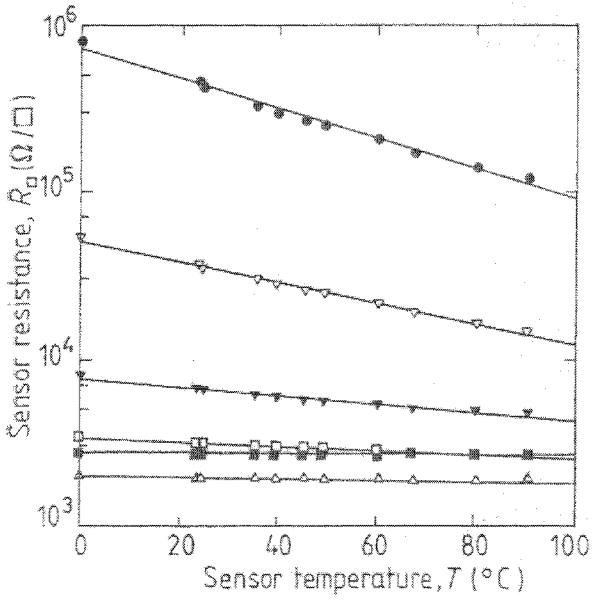


Figure 2.7: Effect of temperature on the baseline (sheet) resistance of polypyrrole chemoresistors with different counterions: BSA (●); PSA (▽); HxSA (▼); HpSA (■); OSA (□); and NSA (Δ). (from Gardner and Bartlett, 1999)

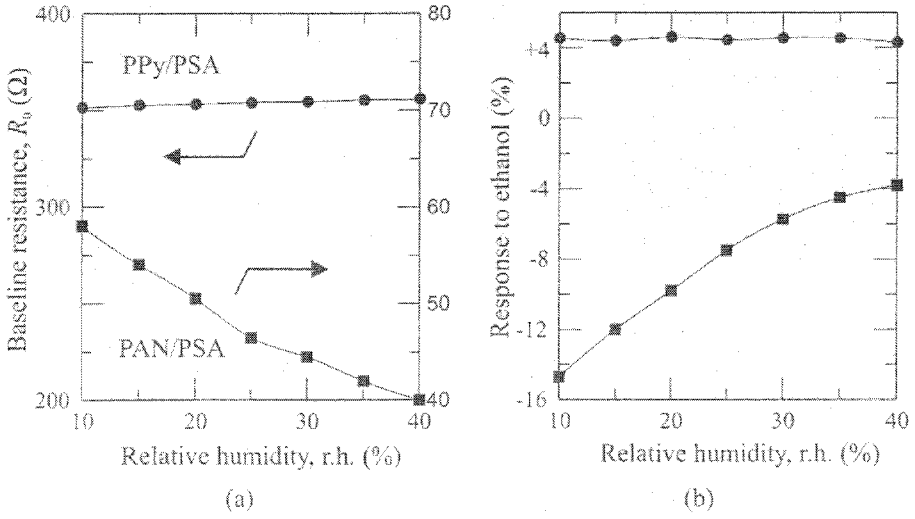


Figure 2.8: Effect of relative humidity on (a) the baseline R_0 and (b) the steady state response $\Delta R/R_0$ to ethanol vapor of polypyrrole (●) and polyaniline (■) chemoresistors in air at 20°C (from Gardner and Bartlett, 1999).

2.1.4 Data analysis and pattern recognition

In the strongly multidisciplinary field of sensor array analysis, the use of appropriate data analysis protocols is essential. Shaffer *et al.* (1999) remarked how pattern recognition algorithms have become a critical component in the successful implementation of chemical sensor arrays and electronic noses. Pattern recognition techniques are used in a wide range of application areas such as speech recognition, medical diagnosis, financial forecasting, bioinformatics and industrial process control (Pardo and Sbeveglieri, 2002). Pattern recognition can be defined as the transformation of an input data set, such as sensor responses to an output set of attributes, such as the type of sample or a concentration (http://www.appliedsensor.com/). Since a wide range of pre-processing, dimension reduction, and learning techniques can be combined, there can be a great number of possible approaches to any particular problem. Sensor arrays can be used to generate a

great deal of data in a very short time and a significant challenge lies in finding ways to extract useful information from this data in order to solve the problem at hand. Furthermore, the data generated from a sensor array have a high dimensionality. Thus, it can be represented in a multidimensional space whose dimension corresponds to the number of variables measured, generally the number of sensors in the array. As a result of this multidimensionality, simple graphical analysis of the raw data is generally not possible and methods to reduce the dimensionality of the data sets are needed. In the vast majority of cases Multivariate Statistics (MVS) and Artificial Intelligence are normally used.

Multivariate analysis is an ever-expanding set of techniques for data analysis (Hair *et al.*, 1998). It provides a methodology to extract qualitative and quantitative information from large amount of data that consist of several measurements (variables) on the same set of cases. Such data analysis techniques can be used to determine which cases can be grouped together (cluster analysis), or belong to a predetermined group (discriminant analysis), or to reduce the dimensionality of the data by forming linear combinations of the existing variables (principal component analysis, factor analysis, canonical correlations). The derived configurations will represent most of the variation in the original data with a smaller number of variables thus enabling the analyst to describe the data in a more straightforward manner by graphical or other statistical methods (Unistat 4.5, 1997). Quantitative methods such as multiple linear regression (MLR) and Partial Least Square (PLS) are well known in analytical chemistry and in the field of gas sensing, and can also be used to analyse simple mixtures of odourant components (Gardner and Bartlett, 1999).

Typically, multivariate analysis can be divided into supervised and unsupervised techniques (Table 2.3). In an unsupervised technique the objective is to discriminate between unknown samples by enhancing the differences between their associated input vectors, while in a supervised technique it is to classify unknown samples as known ones used in an initial calibration, learning or training stage (Gardner and Bartlett, 1999). Alternatively it is also possible to distinguish between the traditional statistical methods and the relatively new emerging techniques such as artificial neural networks that seek to solve multivariate problems in a manner similar to the

human cognitive process. Figure 2.9 shows some of the most commonly used techniques and their classification.

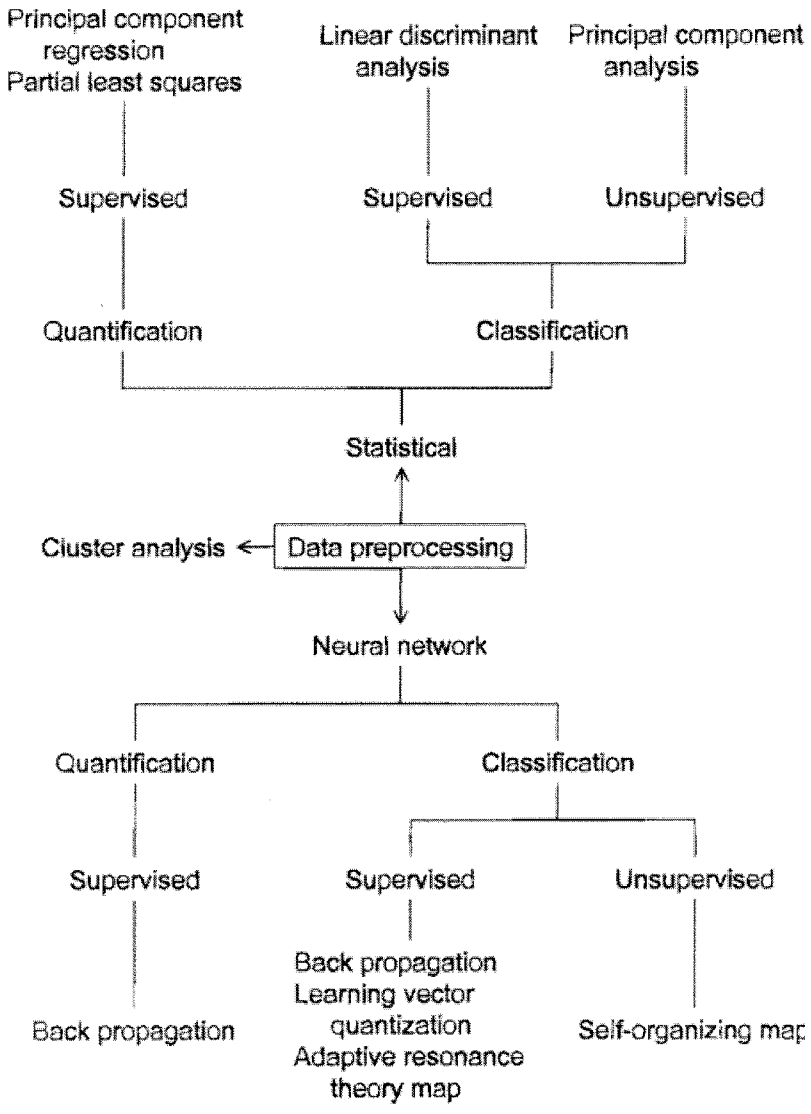


Figure 2.9: Most commonly used data analysis techniques for sensor array applications (from Jurs *et al.*, 2000)

Table 2.3: Summary of multivariate techniques commonly used to analyse sensor array data

Data Analysis Method	*Linear*	*Supervised*
STATISTICAL		
Principal Components Analysis	*Yes*	*No*
Cluster Analysis	*Yes*	*No*
Transformed Cluster Analysis	*Yes*	*No*
Linear averaging / interpolation / recalibration	*Yes*	*No*
Canonical Discriminant	*Yes*	*No*
Discriminant Analysis	*Yes*	*Yes*
Feature Weighting	*Yes*	*Yes*
Canonical Correlation	*Yes*	*Yes*
Regression methods (PCR, PLS, MLR)	*Yes*	*Yes*
ARTIFICIAL INTELLIGENCE		
Back-Propagation	*No*	*Yes*
Fuzzy Neural Network	*No*	*Yes*
Radial Basis Function	*No*	*Yes*
Probabilistic Neural Network	*No*	*Yes*
Genetic Algorithm	*No*	*Yes*
New Fuzzy C-Means Algorithm	*No*	*Yes*
Fuzzy C-Means Algorithm	*No*	*No*
Self Organizing Map	*No*	*No*

Indeed, these techniques have a very diverse character and in the absence of a methodological framework, comparison of methods and results reported in the literature can be difficult.

When considering the application of multivariate statistical analysis the first question to be asked is: can the data variable be divided into independent and dependent classification (Hair *et al.*, 1998). The analysis of sensor array data aims to determine the underlying relationships between one set of independent variables (i.e. sensor array output) and another set of dependent variables (i.e. odour-class and component concentration) (Gardner and Bartlett, 1999). A decision tree to assist the researcher in his choice of a technique is given in Figure 2.10. As a broad rule, dependence techniques are mostly used for prediction of the dependent variable (DV) from

independent variables (IV's) while with interdependence techniques, all the variables are analysed simultaneously, in an effort to find an underlying structure to the entire set of variables.

Therefore the choice of an appropriate data analysis method is highly dependent on the nature of the data and the type of application. However, there is no universal approach which will be right in all situations. The following sections contain a brief description of the techniques that have been commonly used by researchers with sensor array data. More detailed information on multivariate analysis techniques can be found in many text books (Dunteman, 1984; Fausset, 1994; Bishop, 1995; Tabachnick & Fidel, 1996; Hair *et al.*, 1998; Gardner & Bartlett, 1999), reviews (Jurs *et al.*, 2000) and statistical software manuals (Unistat; Statistica) as well as relevant web sites (http://www.nose-network.com; http://www.galatic.com/algorithm; http://www.statsoft.com).

2.1.4.1 Classification Methods

Cluster Analysis

Cluster Analysis (CA), also referred to as Numerical Taxonomy, is an analytical technique used to determine inherent or natural structure in the data, or provide a convenient summary of the data into a given number of groups. The term CA actually encompasses a number of different classification algorithms which can be used to organise data into meaningful structures or taxonomies, typically as part of exploratory data analysis. The aim of CA is to group a sample of entities into clusters, so that objects in the same clusters are more similar to one another than they are to objects in other clusters. In CA, the groups are not predefined and the technique has been used in a number of research areas for group identification such as biological taxonomy, target marketing, psychiatric profiling (Hair *et al.*, 1998) as well as in electronic nose applications (for aroma identification). Specifically the objective is to maximize the homogeneity of objects within the clusters while also maximizing the heterogeneity between the clusters.

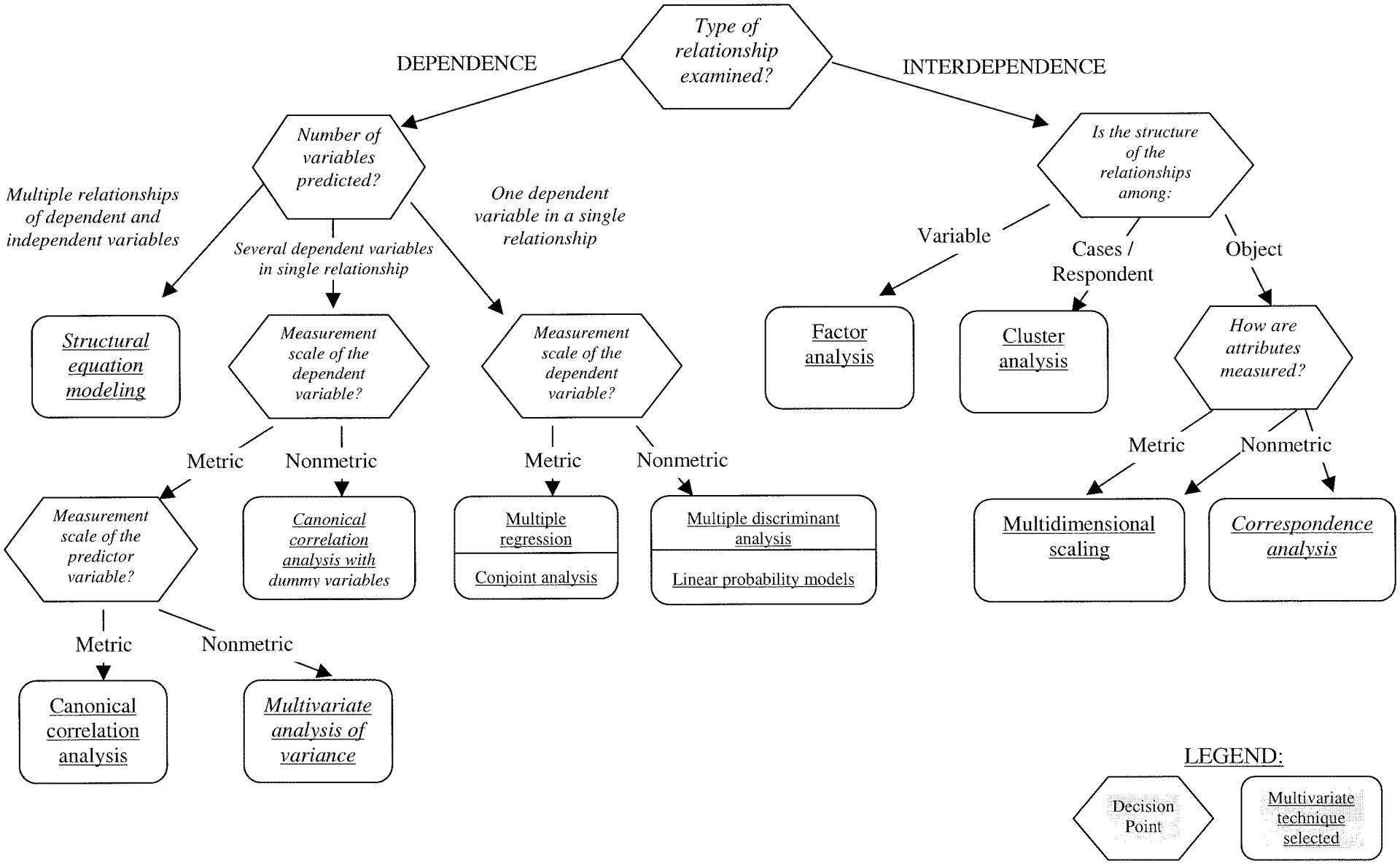


Figure 2.10: Selecting a multivariate technique (from Hair *et al.*, 1998)

Figure 2.11 shows the two main categories of CA: hierarchical and non-hierarchical. The major difference between these two methods is that while the hierarchical method starts sequentially; i.e. starting with the most similar pair and forming higher clusters step by step, the non-hierarchical methods evaluate overall distribution of pairs and then classify them into a given number of groups. In hierarchical CA, clusters are formed by either a process of agglomeration or division. As described in Hair *et al.* (1998), CA usually involves 3 steps. The first is the measurement of some form of similarity among the entities to determine how many groups really exist in the sample. The second step is the actual clustering process, whereby entities are separated into clusters. The final step is to profile the variables to determine their composition. In many cases this may be accomplished by applying discriminant analysis to the groups identified by the cluster technique.

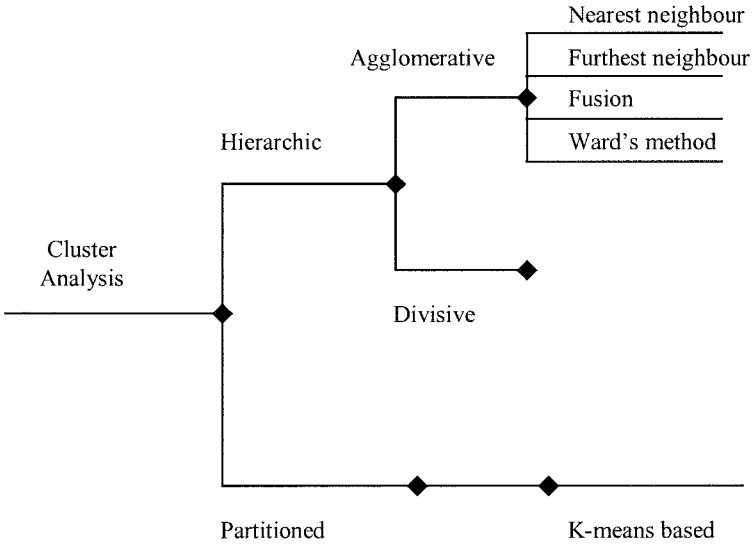


Figure 2.11: Family of clustering algorithms commonly employed in multivariate analysis (from Gardner and Bartlett, 1999)

CA is an easy technique to use that rapidly provides the user with pertinent information. However it has no statistical basis upon which to draw statistical inferences from a sample to a population, and it is therefore used primarily as an exploratory technique. The nature of sensor array data is such that it is often desirable to use a more powerful pattern analysis method (Gardner and Hines, 1997).

Discriminant function analysis

In research where groups have already been identified, the emphasis is frequently on predicting group membership from a set of variables. The primary purpose of discriminant analysis (DA) is to classify samples into well-defined groups or categories based on a training set of similar samples. The ultimate aim being to unambiguously determine the identity or quality of an unknown sample (www.galactic.com). DA assumes that the IV's are metric and normally distributed. It is applicable in situations in which the total sample can be divided into groups based on a non-metric dependent variable characterizing several known classes. For instance, DA can be used to test whether a particular clustering of cases obtained from a cluster analysis is a likely one as suggested previously. It will report whether the group assignment is true or false, as well as the probability of the sample belonging to a particular group.

Computationally, Discriminant Function Analysis (DFA) is very similar to the analysis of variance (ANOVA). The basic idea underlying DFA is to determine whether groups differ with regard to the mean of a variable, and then to use that variable to predict group membership of new cases. If the means for a variable are significantly different in different groups then we can say that this variable discriminates between the groups (Statistica Help). In the case of a single variable, the significance test which indicates whether or not a variable discriminates between groups, is the F test. F is essentially computed as the ratio of the between-groups variance in the data over the pooled (average) within-group variance. In DFA, each discriminant function (DF) is calculated so that the F ratio on the analysis of the variance is maximised.

In the case of sensor array analysis, several variables are included in the study. As a result we have a matrix of total variances and co-variances; and likewise a matrix of - pooled within-groups variances and co-variances. The two matrices are compared in the same way, via the multivariate F test in order to determine if the differences (if any) between groups are significant. Multiple discriminant analysis (MDA) is the appropriate multivariate if the single DV is dichotomous (e.g. male-female) or multichotomous (e.g. high-medium-low) and therefore non-metric. As Gardner & Bartlett (1999) noted: it is common practice to use part of the dataset (i.e. training set) to calculate the coefficients of the discriminant functions and then use the rest of the data (test set) to cross-validate the classification process. Stepwise DFA is a variation of this procedure where variables are added to the DF one by one until the addition of a new variable does not significantly improve the discrimination. Amongst others, Shaffer *et al.* (1999), Romain *et al,* (2000), Nicolas *et al.* (2000), Nicolas *et al.* (2001) and Delpha *et al.* (2001) have successfully used DFA for gas sensing using sensor array systems.

Principal Component Analysis

Principal component analysis (PCA) is the most fundamental and one of the most popular MVS-based monitoring method (Rosen & Lennox, 2000). The technique has found widespread use in many scientific areas and can be considered as the core multivariate analysis procedure. All other MVS methods (except for cluster analysis) can be regarded as variations of PCA (Unistat 4.5). The basic idea behind PCA is that the dimensionality of the variable space (sensor space) is reduced by introducing a number of new variables (principal components) while at the same time extracting maximum variance from the data.

Bishop (1995) defined PCA as: "A linear reduction technique, which identifies orthogonal direction of maximum variance in the original data, and projects the data into a lower dimensionality space formed of a subset of the highest-variance components". In other words, PCA can be described as a method to fit a line (or component) in the direction of greatest variability of the measured variable space. Next, a line is fitted in the second greatest direction of variability, orthogonal to the

first line and thus, a plane is obtained. The following line is fitted in the third direction of greatest variability, orthogonal to the plane. This is continued until it is established that no systematic variability remains (Figure 2.12). The coefficients which transfer the original data into the new coordinates are called eigenvectors (or loading vectors).

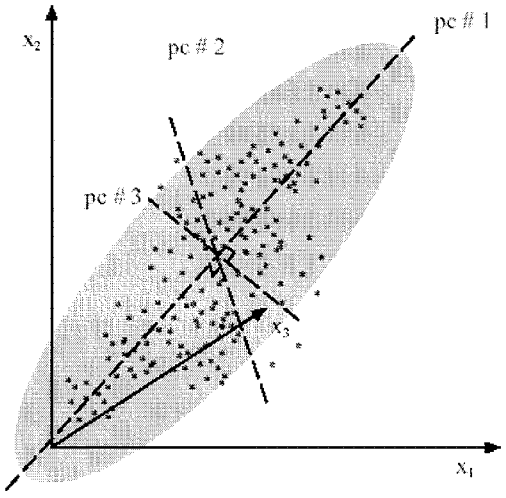


Figure 2.12: PCA as a successive fitting of lines (or components) in the directions of greatest variability (from Rosen *et al.*, 2002)

The application of PCA to a dataset provides two main quantities: the factor loadings and the scores (Principal Components). The factor loadings are the correlations between the original variables (sensor responses) and the factors. These are key to understanding the nature of a particular factor and allow for eigenvalues to be determined. In mathematical terms eigenvalues (also called latent root) are the column sum of squared loadings for a factor. The eigenvalues represent the variance in the data and in the case of sensor array analysis give an evaluation of the contribution of each variable to the total information contained in the dataset. It is therefore often used as a tool to design and optimise a sensor array for a particular application.

The Scores are the transformed variables obtained by multiplying the original data matrix with the matrix of eigenvectors (Unistat 4.5). In other words the scores are the values of the measurement points projected onto the PC's. The Score Plot or plot of principal components gives a graphic representation of the relationship between observation points and allows us to visualise similarities/differences between samples, as well as the presence of outliers or possible trends such as drift in the data. The method essentially removes any sensor collinearity and is therefore particularly well adapted to the analysis of sensor array data when sensor responses are correlated. Consequently, PCA has been extensively used in sensor array studies (Persaud *et al.,* 1996; Persaud *et al.,* 1999; Shaffer *et al.,* 1999; Baby *et al.,* 2000; McEntergart *et al.,* 2000; Nicolas *et al.,* 2000; Wilson *et al.*, 2000; Capone *et al.,* 2001; Delpha *et al.,* 2001; Llobet *et al.*, 2001 and Nicolas *et al.,* 2001), as well as in wastewater monitoring and other process monitoring applications (Gurden *et al.*, 1998; Rosen and Olsson, 1998; Rosen and Lennox, 2001 and Rosen and Yuan, 2001). An example of a typical PCA plot is given in Figure 2.13. The figure illustrates how the dimensionality of data obtained from an array of 5 tin-oxyde sensors has been reduced, and shows how the variance has been extracted to discriminate between samples.

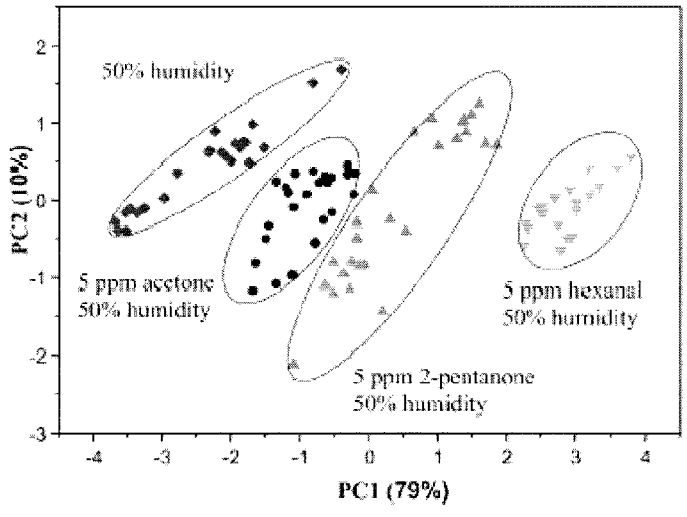


Figure 2.13: PCA plot for 4 analysed compounds using a 5 tin-oxyde sensor array system (from Capone *et al.,* 2001)

2.1.4.2 Regression Methods

Principal Component Regression

Principal Component Regression (PCR) uses the output from a PCA to create a quantitative model for complex applications (e.g. calibration of a multi-sensor array). In a PCR, the scores are used as input to a linear regression model. A schematic of the two-step process associated with PCR is given in Figure 2.14. The PC's used are generally these accounting for the largest amount of variances in the artificial dataset, i.e. the first few PC's. The number of PC's used in the model influences the prediction accuracy. If too many PC's are used, the model will contain information that is unnecessary for a good generalisation and may degrade prediction for data not included in the calibration process as a result of overfitting (Jurs *et al.*, 2000). On the other hand, a model with too few PC's will not be able to capture the relevant information in the original variables, also resulting in poor predictions.

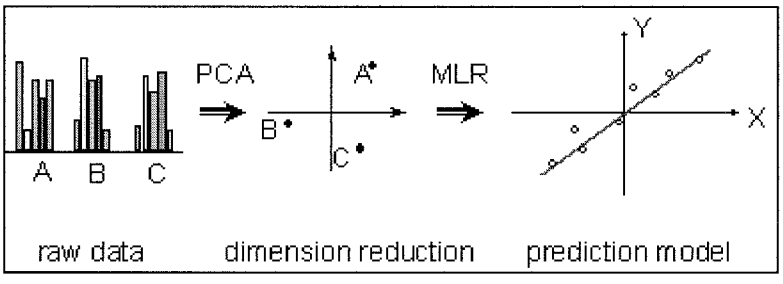


Figure 2.14: PCR as a two-step process (from www.appliedsensors.com)

Figure 2.15 shows the influence of the number of PC's on the Root Mean Squared Error of Prediction (RMSEP) of *p*-aminophenol (*p*-AP) and *p*-phenilenediamin (*p*-PDA) using spectrophotometer data. Such plots are useful in selecting the optimal number of PC's when developing a model.

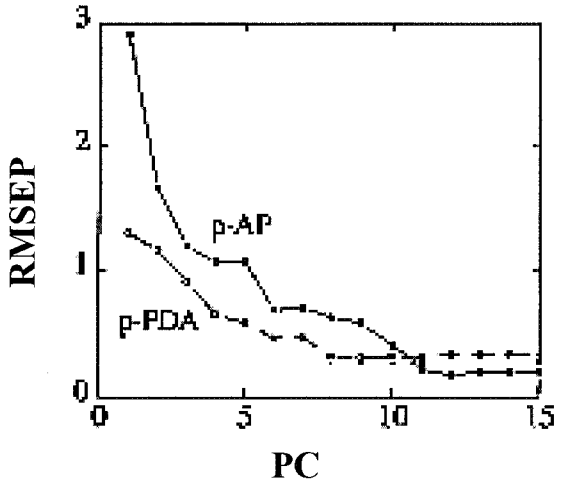


Figure 2.15: Selection of the number of PC's for the PCR modeling of *p*-AP and *p*-PDA. (from Lopez-Cueto *et al.*, 2000)

Multiple Regression

One of the keys to multivariate quantitative analysis is the assumption that the concentrations of the components of interest are somehow related to the data from the measurement technique used for analysing the samples. Multiple Regression is the appropriate method of analysis when the research problem involves a single metric dependent variable (e.g. sample concentration) presumed to be related to two or more metric independent variables (e.g. sensor responses) (Hair *et al.*, 1998). The objective of multiple regression is to predict the scores on the DV in response to changes in the IV's. This is most often achieved using the least squares regression technique which is one of the simplest methods: in essence, a line is fitted through the data points so that the squared deviation of the observed points from that line (predicted points) are minimised. There are a variety of fundamental equations that can be used for multiple regression.

Multiple linear regression

In essence, in MLR the least square fit is calculated for a straight line represented by an equation of the form:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

where $b_1 ... b_n$ are the regression coefficients and a the intercept.

While this method is relatively rapid and easy to understand, MLR suffers from a few major conceptual limitations. In addition to its sensitivity to noise, MLR assumes that the relationship between the IV's and the DV is linear in nature and that the data is normally distributed. Furthermore, MLR will not produce accurate results if the IV's are highly correlated. This so-called collinearity problem is often true for most sensor-array systems. Consequently more sophisticated mathematical techniques are often preferred. Principal Component Regression (PCR) is a typical example where a dimension reduction is made with a PCA, followed by a MLR regression model on the reduced dataset (www.appliedsensors.com).

Polynomial regression

Polynomial regression is a common non-linear estimation model that can in theory estimate any kind of relationship between a DV and a series of IV's. Here the relationship between DV and IV's can be expressed in a regression equation of the form:

$$y = b + c_1 x + c_2 x^2 + ... + c_i x^i$$

where b and $c_1 .. c_i$ are constants and i is the degree of the polynomial

Technically speaking, the nature of this model is actually linear and very similar to MLR in which the straight line of best fit is nothing else than a first degree

polynomial. In the equation above, b represents the intercept and $c_1 \ldots c_i$ are the regression coefficients. The non-linearity of the model lies in the terms $x^2 \ldots x^i$

As described in the Statistica user manual, polynomial regression designs contain main effects and higher-order effects for the continuous IV's but do not include interaction effects between IV's. For example, the polynomial regression design to degree 2 for three continuous predictor (independent) variables P, Q, and R would include the main effects (i.e., the first-order effects) of P, Q, and R and their quadratic (i.e., second-order) effects, but not the 2-way interaction effects or the P by Q by R 3-way interaction effect.

$$y = b_0 + b_1 P + b_2 P^2 + b_3 Q + b_4 Q^2 + b_5 R + b_6 R^2$$

Polynomial regression designs do not necessarily contain all effects up to the same degree for every IV. For example, main, quadratic, and cubic effects could be included in the design for some predictor variables, and effects up the fourth degree could be included in the design for other IV's.

Factorial regression

In factorial regression designs, there may be many more such possible combinations of distinct levels for the continuous IV's than there are cases in the data set. To simplify matters, full-factorial regression designs are defined as designs in which all possible products of the continuous IV's are represented in the design. For example, the full-factorial regression design for two continuous IV's P and Q would include the main effects (i.e., the first-order effects) of P and Q and their 2-way P by Q interaction effect, which is represented by the product of P and Q scores for each case. The regression equation would be

$$y = b_0 + b_1 P + b_2 Q + b_3 P * Q$$

Factorial regression designs can also be fractional, that is, higher-order effects can be omitted from the design. A fractional factorial design to degree 2 for 3 continuous

predictor variables P, Q, and R would include the main effects and all 2-way interactions between the predictor variables, but not the 3-way P by Q by R interaction:

$$y = b_0 + b_1 P + b_2 Q + b_3 R + b_4 P*Q + b_5 P*R + b_6 Q*R$$

Similarly a fractional design to degree 3 would include all main effects, 2-way interactions and 3 way interactions, and so on.

Partial Least Square

Partial Least Squares regression or Projection to Latent Structures (PLS) is an extension of the multiple linear regression model. It is probably the least restrictive of the various multivariate extensions of the multiple linear regression model which allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than IV's. Furthermore, PLS can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression. (Statistica user manual)
PLS is a tool extensively used for quantitative analysis, in chemometrics and in multi-sensor applications (Di Natale *et al.*, 1997). It is, in many aspects closely related to PCR since it is based on the combined properties of MLR and PCA (Gardner and Bartlett, 1999). However, in PLS, the decomposition is performed in a slightly different fashion. Unlike PCR, PLS is a one step process, i.e. there is no separate regression step, and PLS performs the decomposition on both the sensor response and concentration data simultaneously (Figure 2.16). Instead of first decomposing the sensor response matrix into a set of eigenvectors and scores, and then regressing them against the concentrations as a separate step, PLS actually uses the concentration information during the decomposition process. In effect this generates two sets of vectors and two sets of corresponding scores (which are different from those of PCR): one for the sensor data and one for the sample concentration. Since the two sets of scores can be correlated to each other, a calibration model is constructed (galactic.com). The main idea of PLS is to get as much concentration information as possible in the first few loading vectors and PLS

generally requires fewer PC's than PCR to achieve similar levels of accuracy. As Gardner and Bartlett (1999) pointed out, PLS is generally preferred to MLR for the analysis of gas mixtures, because it accepts collinear data, separates out sample noise and makes meaningful linear combinations in the dependent concentration matrix. This is unlike PCR where the relation between the sensor responses and the individual gases of interest is ignored until the final regression step.

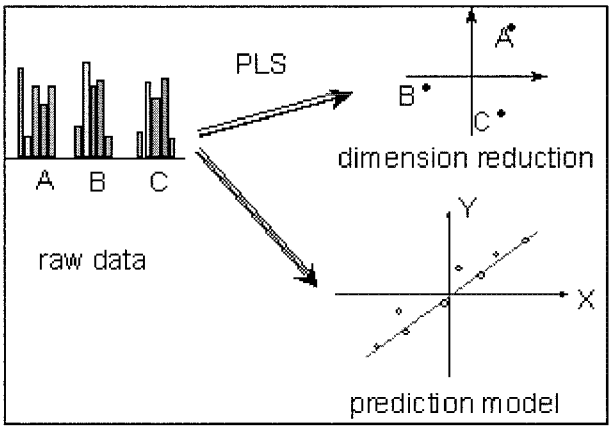


Figure 2.16: PLS as a one-step process (from www.appliedsensors.com)

Canonical Correlation

In a previous section we discussed Multiple Regression analysis which can predict the value of a single (metric) variable from a linear function of a set of independent variables. Canonical Correlation (CC) analysis can be viewed as a logical extension of MLR. However, with CC analysis, there are several continuous dependent variables as well as several continuous independent variables. Whereas multiple correlation involves a single dependent variable, CC involves multiple dependent variables and the aim is to correlate simultaneously these variables. The underlying principle is to develop a linear combination of each set of variables to maximise the correlation between the two sets. Although other techniques may be presented in a more interpretable manner, CC places the fewest restrictions on the type of data on which it operates and represents a useful tool for multivariate analysis, particularly

since interest has spread to considering multiple dependent variables (Hair *et al.*, 1998). However it appears from the most recent literature that the use of CC for the analysis of sensor array data remains relatively limited. Indeed there is a growing tendency to use alternative new techniques such as artificial neural networks where CC could be potentially used.

2.1.4.3 Artificial Neural Networks

Artificial Neural Networks (ANN) are a totally different approach to data analysis from any of the classical multivariate techniques reviewed above. Instead of conceptualising the problem as a mathematical one, ANN aims to mimic the processes of learning in the cognitive system and the neurological functions of the brain in order to develop a learning strategy and predict new observations. Historically, ANN grew out of research in the field of neurobiology in the late 1950's and only became commonly used in the 1980's when some practically useful algorithms were developed and the calculation power of personal computers became sufficiently large (http://www.appliedsensor.com/).

Linear regression methods, such as these described in the previous sections, cannot model complex non-linear patterns such as those often found in sensor array data. Consequently new flexible analysis techniques that can perform both relationship identification (as in MLR or DFA) or data reduction and structure analysis (as in PCA or CA), while being able to cope with non-linear, non-parametric data have become increasingly popular. This is particularly true for sensor array data processing when exploration and prediction, but not explanation, are the focal points of the research.

Neural networks normally have great potential for parallelism, since the computations of the components are largely independent of each other. Some people regard massive parallelism and high connectivity to be defining characteristics of NNs. However there is no universal definition of an ANN. Here we give two definitions found in the literature.

"...a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes." (DARPA, 1988).

"...a neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Inter-neurone connection strengths known as synaptic weights are used to store the knowledge". (Haykin, 1994).

ANN design

ANN are incredibly varied and with new ones (or at least variations of old ones) being developed every week, there can be as many different models as there are users and types of applications. There are two main types of ANN: Feedback (recurrent) and Feedforward. Feedback ANN can be quite difficult to train and are mostly of interest to researchers in ANN. So far, Feedforward ANN (FFANN) such as Multilayer Perceptron (MLP) have proved most useful in solving real problems and these are the most commonly used NN techniques for sensor array applications. Feedforward ANN are the type of NN described here. The basic unit of an Artificial Neural Network is the neurone. Figure 2.17 shows how each neurone receives a number of inputs, multiplies the inputs by individual weights, sums the weighted inputs, and passes the sum though a transfer function (also known as the activation function) which can be stepwise, linear or sigmoid.

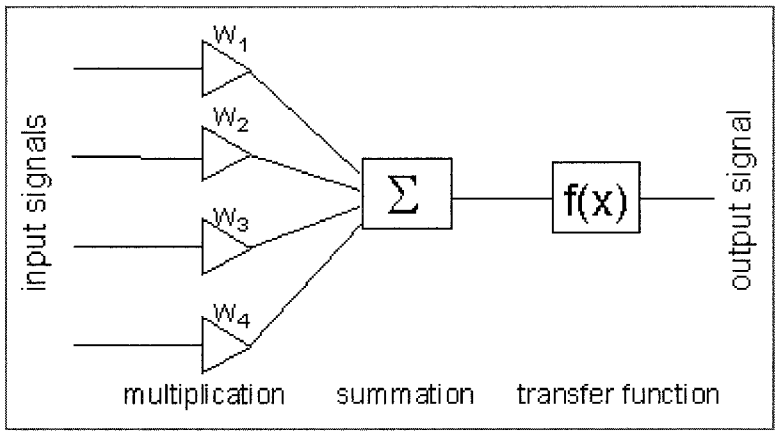


Figure 2.17: Transformation of input signal into output signal via weighing and transfer function (from wwww.appliedsensors.com)

The most commonly used transfer function is the Sigmoid (or logistic function) given by:

$$f(x) \equiv \frac{1}{1+\exp(-x)}$$

It is in effect linear for values close to zero and flattens out for large positive or negative values (Figure 2.18).

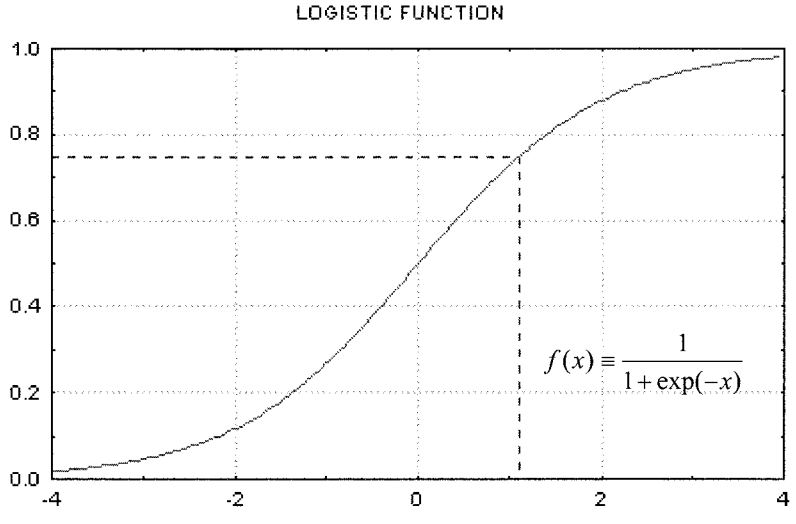


Figure 2.18: Plot of the logistic sigmoid activation function given by (2.1)

The neurones are organised into layers and interconnected thus forming the network. A typical feedforward network is shown in Figure 2.19. In sensor array applications, the input layer generally has one neurone for each variable (sensor response) and their output layer has one neurone for each predicted property. The number of hidden layers and neurones in each one generally depends on the complexity of the problem at hand. This process of finding the optimal architecture generally involves multiple trials and errors.

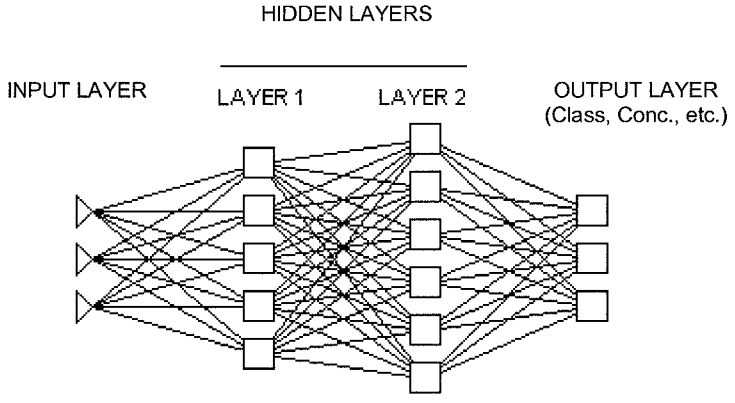


Figure 2.19: Representation of a feed forward neural network with two hidden layers

ANNs can be characterised and described by the following concepts: the type of NN model; their structure (number of neurones, number of layers); the learning algorithms as well as the type of application (classification, regression problems). There are three basic types of neural networks commonly used: the Multilayer Perceptron, the Radial Basis Function (RBF) and the Kohonen Networks (KNN).

1) Multilayer Perceptron

The MLP model is the most common network architecture in use today. As described above, the units are arranged in a layered feed-forward topology and important issues in MLP design include specification of the number of hidden layers and the number of units in these layers (Haykin, 1994; Bishop, 1995). Training is most often carried out using the Back Propagation algorithm as discussed in the next section. The

advantages of MLP over other Network types mainly lies in their compactness and rapidity of execution, although the training procedure can be relatively slow. MLP are discussed at length in many reviews and textbooks (Fausset, 1994; Bishops, 1995; Gardner and Bartlett, 1999). An example of application of a MLP (with BP algorithm) to sensor array data for the identification of household chemicals is shown in Figure 2.20.

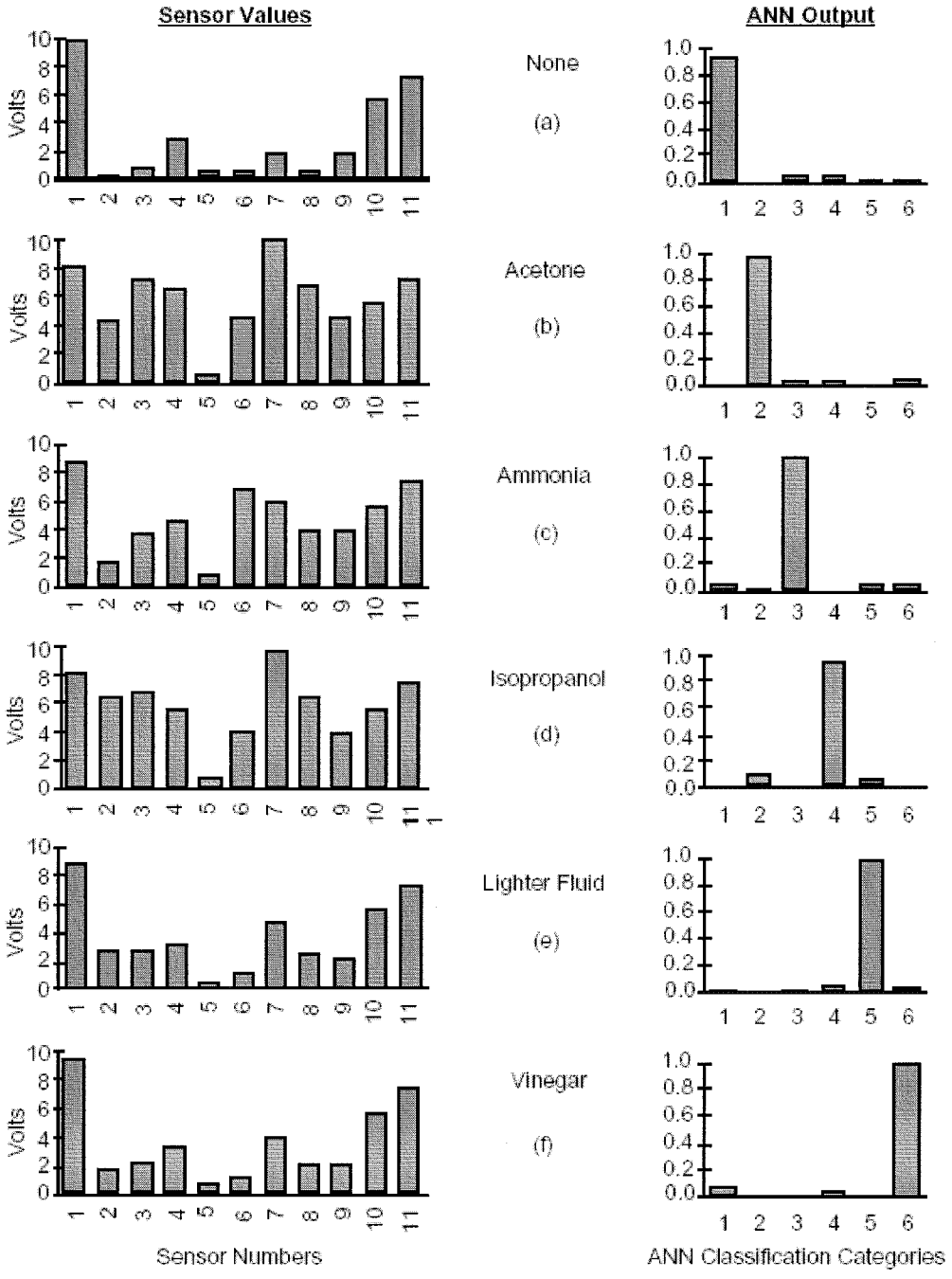


Figure 2.20: Example of sensor response patterns and artificial neural network classification to a range of household chemicals (from Hashem *et al.*, 1996)

2) Radial Basis function

In a Radial Basis Function network the unique hidden layer consists of radial units (rather than sigmoids), each actually modelling a gaussian (bell-shaped) response surface (Figure 2.21). Since these functions are non-linear, it is not necessary to have more than one hidden layer to model any shape of function. In this case the "weights" and "thresholds" are entirely different to those in a linear unit. Radial weights are in effect the centre point and Radial thresholds correspond to the radius (or deviation). The final output layer is a simple linear transformation which can be optimised using traditional linear techniques. RBF can therefore be trained extremely quickly. As Gardner and Bartlett (1999) remarked, RBF can be an attractive method because it always finds a solution to a classification problem. However it has not been widely used by e-nose researchers. Some of the major drawbacks associated with RBF are that they tend to be slower to execute and are more time consuming than MLP. Additionally they are more sensitive to the "curse of dimensionality" and have greater difficulties if the number of input variables is large.

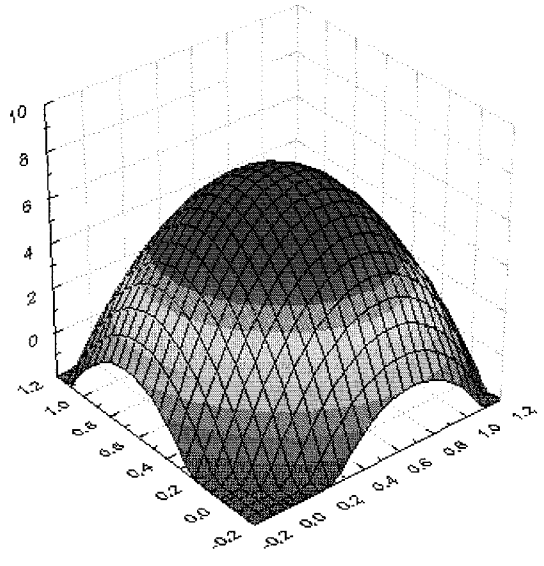


Figure 2.21: Response surface of a single radial unit

3) Kohonen Networks

Kohonen Networks (KNN) are designed primarily for unsupervised learning and are therefore used for different type of applications. Contrary to other techniques based on supervised learning which try to establish a mapping between inputs and outputs, KNN do not use outputs but instead try to learn the structure and patterns in the input data. KNN can learn to recognise clusters of data and highlight similarities between identified classes. Therefore the technique can be used as a visualisation tool to examine the data in data analysis as well as for classification problems. Another possible use includes the detection of unusual new data (e.g. anomalies) which can be useful for early warning applications in the process industry. An example of KNN architecture is shown in Figure 2.22

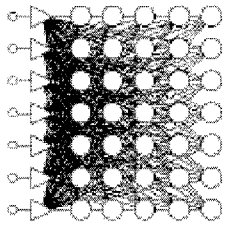


Figure 2.22: Representation of a Kohonen Neural Network architecture.

Training

Once the initial step of designing a specific Network architecture is completed, the ANN can be trained. There are two types of training used in Neural Networks, with different types of networks using different types of training algorithms. These are supervised and unsupervised training, of which supervised is the most common and will be discussed below.

In supervised training, the weights are adjusted through an iterative process (learning algorithm), so that the calculated output values for a set of input values are as close as possible to the known target output values. This iterative training process is

characterised by the number of epoch, which is the number of times the entire training set is fed through the network and used to adjust the weights and thresholds. The best known example of supervised learning algorithm is back propagation (BP), devised by Rumelhart *et al.*, 1986, which uses the training data to adjust the weights and thresholds in proportion to the error it is causing. After the learning phase and an evaluation phase using a verification data set, the new network can then be used to generate predictions where the output is not known.

Other supervised learning algorithms differ in the way the error is minimised. However all techniques use a strategy designed to travel towards a minimum as quickly as possible. Commonly used sophisticated techniques for non linear function optimisation include conjugate gradient descent (Bishop, 1995), Levenberg-Marquardt (Shepherd, 1997), and the often recommended Quasi-Newton algorithm. The major difference between BP and Quasi-Newton is that BP adjusts the network weights after each case, whereas Quasi-Newton is a batch algorithm that works out the average gradient of the error surface across all cases, before updating the weights once at the end of each epoch (Bishop, 1995; Statistica NN user manual). Unsupervised learning techniques alter weights and/or thresholds using input only training sets (output values are not required). A well-known example is the Kohonen algorithm, typically used in Kohonen Networks but also in radial basis function (RBF) and Regression networks (Statistica NN User Manual).

In many cases Neural Networks can produce highly accurate predictions and outperform statistical multivariate techniques. One major advantage is that they can in theory be used to approximate any continuous function. They represent however, a typical a-theoretical research approach and it is virtually impossible to interpret the solution in traditional analytical terms. A detailed view of the relationship between variables and measurements can often be acquired using a PCA or PLS model and it is therefore good practice to perform such analysis before using ANN's (www.appliedsensors.com).

Summary

The analysis of sensor array data involves two principal steps:

The first step is the selection of a pre-processing method. This stage is not always performed but can be necessary depending on the type of application. For example, normalisation is a common pre-processing technique that can be useful for classification purposes. However this can be detrimental for quantitative applications since it tends to remove the concentration information from the sensor response.

The second and major step is the use of an appropriate pattern recognition technique that can extract useful information from a multicomponent set of data. Therefore multivariate statistical analysis techniques are normally used. Regression techniques such as MLR, PCR and PLS are the classic choices most reported in the literature. Classification methods include PCA, CA and DFA. PCA is a very popular technique which is particularly adapted to sensor array data analysis. It allows the researcher to deal with co-linear variables and therefore works particularly well with sensors that are correlated. PCA is a linear recognition technique that is generally used to reduce the dimensionality of the data and can be combined with other pattern recognition techniques such as ANN. It also provides a useful tool for the optimisation of sensor arrays and identification of redundant sensors.

New emerging multivariate analysis techniques such as Artificial Neural Networks have been increasingly popular. ANN such as MLP with BP have particularly proved well suited in a number of successful sensor array applications. Other popular networks include Kohonen's self organising maps (SOM). It appears from the most recent literature that there is a growing tendency to use such techniques. However, the justification of their interest over the more traditional statistical multivariate techniques - which may allow for a better understanding of the relationship between sensor responses and the dependent variable(s) - is not always cleary vindicated.

With the immensely wide range of ANN that can be applied to sensor array systems, the choices of a particular type of NN may sometimes appear quite arbitrary.

Furthermore the number of hidden layers, number of neurons in each layer and other parameters such as the type of activation function used are not always fully reported. As a result objective comparison of such methods and reported results can be difficult. With both sensor array technology and Neural Network modeling still in their early stages of development the questions of which, when and why NN should be used are still unanswered. At present there is no systematic or recommended approach to the analysis of sensor array data and the selection of an optimal protocol remains largely based on trials and errors.

A summary table (Table 2.4) is given that lists a range of gas sensor array applications found in the literature, with particular reference to the type of pattern recognition methods that have been used.

Table 2.4: Some areas of application of sensor arrays and pattern recognition techniques investigated

Application	*Pattern recognition technique*	*Reference*
Microbilology	*DA*	*Namdev* et al., *1998*
	PCA	*McEntergart* et al., *2000*
		Liden et al., *1998*
	BP	*Bachinger* et al., *1998*
		Gibson et al., *1997*
		Gardner et al., *1998*
		Liden et al., *1998*
		Holmberg et al., *1998*
Food and drink	*BP*	*Singh* et al., *1996*
		Gardner et al., *1989*
		Nakamoto et al., *1990*
		Gardner and Hines, 1997
		Brezmes et al., *2001*
	CA	*Gardner, 1991*
		Aishima et al., *1991*
	Fuzzy NN	*Singh* et al., *1996*
		Moriizumi et al., *1992*
	RBF	*Evans* et al., *2000*
	GA	*Fekadu* et al., *1993*
		Gardner and Hines, 1997
	PCA	*Gardner, 1991*
		Schweizer-Berberich et al., *1994*

		Eklöv et al., *1998*
		Capone et al., *2001*
	Independent component analysis	*DiNatale* et al.,*2002*
Fire detection	*Probabilistic NN*	*Rose-Pehrsson* et al., *1999*
Air, Gas Mixtures, Vapours	*CA*	*Byun* et al., *1997*
	Transformed CA	*Nayak* et al., *1992*
	DA	*Delpha* et al., *2001*
		Niebling, 1994
		Pardo et al., *2000*
		Nicolas et al., *2001*
	PCA	*Delpha* et al., *2001*
		Nicolas et al., *2001*
		Pardo et al., *2000*
		Persaud et al., *1996*
		Persaud et al., *1999*
		Llobet et al., *1997*
		Wilson et al., *2000*
	PCR	*Nicolas* et al., *2001*
	MLR	*Nicolas* et al., *2001*
	PLS	*Nicolas* et al., *2001*
	Wavelet Transform	*Llobet* et al., *2001*
	BP	*Niebling, 1994*
		Lu et al., *2000*
		Niebling and Schlachter,1995
		Martin et al., *2001*
		Llobet et al., *1997*
	Fuzzy C-Means Algorithm	*Ping and Jun, 1996*
	RBF	*Ping and Jun, 1996*
		Byun et al., *2000*
	SOM	*Ping and Jun, 1996*
		Ortega et al., *2000*
	GA	*Srivastava* et al., *1998*
		Pardo et al., *2000*
Water & Waste Water	*CC*	*Stuetz* et al., *1999*
	Canonical Discriminant	*Stuetz* et al., *1999*
	DA	*Stuetz* et al., *1998*
	PCA	*Baby* et al., *1999*
		Gardner et al., *2000*
	MLP	*Gardner* et al., *2000*
	LVQ	*Gardner* et al., *2000*
	fuzzy ARTMAP	*Gardner* et al., *2000*

Other comparative studies where different analysis techniques have been applied to particular applications are also reported in Shaffer *et al.*, 1999; Romain *et al.*, 2000; Hierlemann *et al.*, 1995; Getino *et al.*, 1999; Kalman *et al.*, 1997; Brezmes *et al.*, 2001; Di Natale *et al.*, 1997.

2.1.5 Sensor arrays for environmental monitoring

Environmental monitoring has recently become an area of growing interest for electronic nose manufacturers. At present, conventional measurement systems remain largely limited by environmental factors, short lifetime, fouling problems and the need for chemical reagents or frequent calibrations (Bourgeois and Stuetz, 2000). Emerging technologies such as non-specific sensor arrays could provide a suitable solution to a wide range of environmental applications. With the recent research efforts and investments, as well as the latest developments in computing technologies, so called electronic noses have become commercially available and have demonstrated their ability to discriminate between samples of various nature as well as monitor changes in quality with time. The list of potential applications and the number of system designs reported in the literature is continuously growing, encouraged by the development of new sensor materials and the apparition of smaller, more versatile and more sensitive devices.

Original work carried out with laboratory-based systems and more recent field investigations have shown promising results and great potential. However, the number of trials carried out under realistic conditions is relatively limited. It appears that despite the proven versatility and potential for non-invasive and on-line implementation of electronic noses, the technology has yet to be embraced by the end-user and reported examples of real-size commercial applications are still marginal. Consequently, there is a case for environmentally relevant sensor array studies to be reviewed and for a number of practical and fundamental issues to be addressed before such systems become more widely accepted. In this section we discuss the specific challenges associated with a range of environmental applications.

2.1.5.1 Advances in sensor technology

The various types of sensors commonly used for sensor arrays and their respective principles, designs and applications have been described extensively in the literature. Wolff *et al.* (2001) presented the interest of Surface Acoustic Wave Sensors (SAW's) which he described as a first choice for sensing in harsh environments. Other classes of sensors applicable to environmental studies include Bulk Acoustic Wave (BAW's), Conducting Polymers (CP's), Metal Oxide Sensors (MOS), and Quartz Cristal Microbalance (QCM's), and have also been reviewed (Wilson *et al.*, 2001; Albert *et al.*, 2000; Vig, 2001; Lee and Lee, 2001), with many studies focussing on the interest and development of a particular sensor class for the detection of target environmental gases or substances such as NOx, SOx, CO, CO_2, O_3, H_2S, NH_3 and VOC's (Lee and Lee, 2001; Buhlmann *et al.*, 1998; De Wit *et al.*, 1998; Carotta *et al.*, 2000; Becker *et al.*, 1999; 2000; Dejous *et al.*, 1995; Rap *et al.*, 1995). Figure 2.23 shows an example of tin oxide sensor response to CO/O_3 gas mixtures.

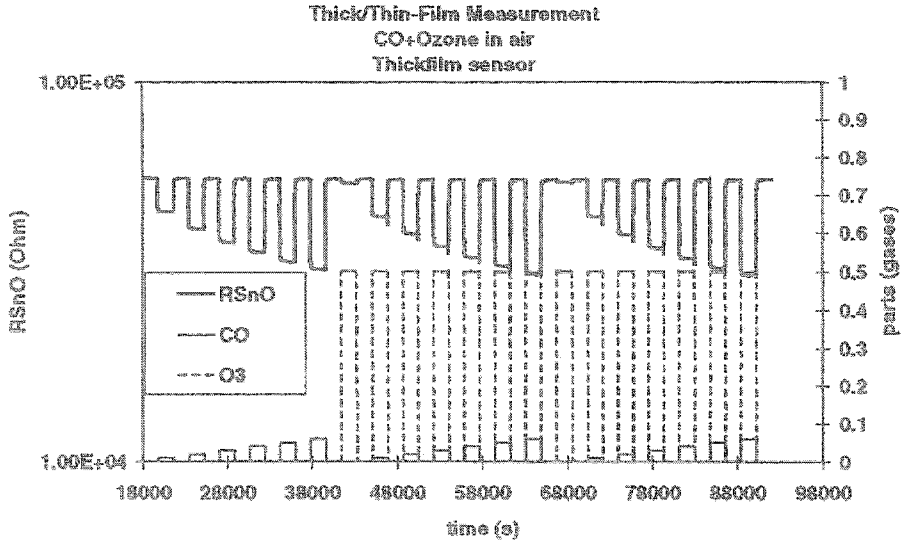


Figure 2.23: Variations of a thick film tin oxide sensor (RsnO) response to CO/O_3 gas mixtures under constant flow conditions at 500^0C. (from Becker *et al.*, 2000)

Not surprisingly these advances in sensor technology have been the key factor in the development of sensor array systems. Since totally selective sensors based on key-lock interactions do not exist for the detection of hazardous organic substances, the cross-selectivity of selected elements can be exploited in a sensor array by applying different algorithms of multicomponent analysis and pattern recognition (Hierlemann *et al.*, 1995). However, this is still a relatively new discipline, and less effort has so far been devoted to the implementation of fully operational devices for environmental applications as seen in other fields such as the foodstuff, beverages and perfumes industries. Only in recent years, have technological progresses and the search for new markets, eventually contributed to a more widespread assessment of commercial or prototype instruments for environment-orientated studies.

2.1.5.2 Single substances

Early investigations related to the environment include the detection of fuel mixtures (Lauf *et al.*, 1991) and oil leaks (Shurmer, 1990). In more recent studies, Sugimoto *et al.* (1999), Ueyama *et al.* (2001) and Ogawa and Sugimoto (2001) reported the recent progress in detecting low levels of hydrocarbons (oils, petrol) in river water samples using QCM-based devices and advanced humidity and temperature control systems. Other field investigations (Traversa *et al.*, 2000; Carotta *et al.*, 2001) have used MOS sensors for air quality monitoring in an urban environment. The good correlations between the sensor responses and in-situ conventional NOx and CO instruments measurements (Figure 2.24 and Figure 2.25), demonstrated the interest of the device in terms of sensor performance and selectivity under real operating conditions. Still, further development would require the long-term stability and calibration requirements (duration, frequency, amount of data needed) to be addressed.

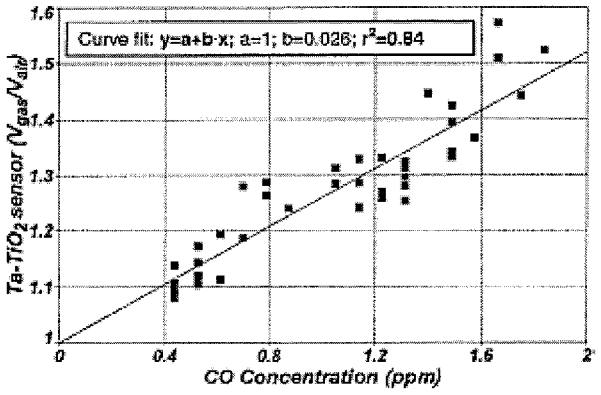


Figure 2.24: Correlation between the response of the Ta-TiO_2 sensor and the CO concentration measured during a field test using a conventional environmental monitoring station (from Carotta *et al.,* 2001)

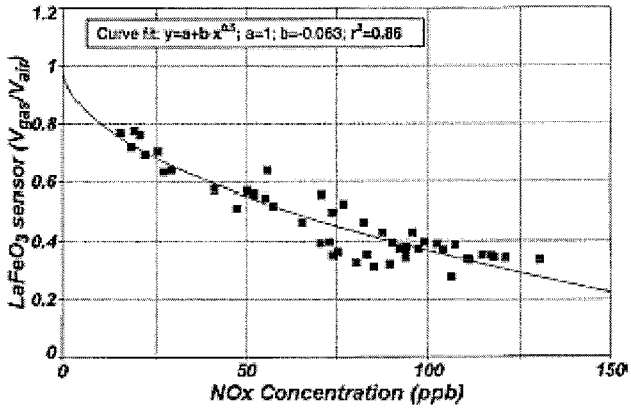


Figure 2.25: Correlation between the response of the $LaFeO_3$ sensor and the NOx concentration measured during a field test using a conventional environmental monitoring station (from Carotta *et al.,* 2001)

Similarly, volatile organic compounds represent an important class of air pollutant and the analysis of organic vapors (hydrocarbons, chlorinated compounds and alcohols) using sensor arrays is now well documented. In particular, compounds such as ethanol, propanol, butanol, acetone, toluene, xylene, oxylene, n-octane, methane, cyclohexane, trichloromethane, tetrachloromethane, tetrachloroethylene, have all been successfully investigated using tin oxide based sensor arrays (Faglia *et al.*,

1997; Llobet *et al.*, 1997; Getino *et al.*, 1999; Orts *et al.*, 1999; Szczurek *et al.*, 1999; Lee *et al.*, 2000) as well as BAW (Hierlemann *et al.*, 1995) CP's (Bartlett & Li-chung, 1989; Hatfield *et al.*, 1993; Furlong and Stewart, 2000) and QCM (Dickert *et al.*, 2000). Other environmentally relevant applications include the detection of insecticides (Baby *et al.*, 2000), nerve and blister agents (McGill *et al.*, 2000), refrigerant gases (Delpha *et al.*, 2001) and pollen (Kalman *et al.*, 1997).

Although new sensor materials and designs are continuously reported, the major limitation of currently available sensors remains their sensitivity to changes in temperature, humidity and flow rate. Consequently much of the original work has been carried out with laboratory based prototype systems under carefully controlled conditions. In practice, there are two possible approaches to deal with these effects. As demonstrated by Ueyama *et al.* (2001) and Ogawa and Sugimoto (2001), careful system design and sample pre-conditioning can help minimise changes in the relative humidity (RH) of the headspace sample (Figure 2.26). But when considering the practical application of a sensor array, this can make the overall instrument, more complex and expensive and can also affect its portability or limit sample throughput. The second approach is to measure these parameters and calibrate the sensors under varying humidity levels in order to compensate for changes in subsequent data analysis. This parametric compensation approach has been favored in a number of applications where RH was used as an input to artificial neural networks (Hierlemann *et al.*, 1995; Faglia *et al.*, 1997; Orts *et al.*, 1999; Dickert *et al.*, 2000).

In the field, humidity, temperature, wind and interfering gases are constantly changing. Therefore, despite the relative success in detecting or identifying individual chemicals (or at best a combination of known substances) in the laboratory, many of these experiments do not yet reflect the reality of most environmental applications where complex mixtures must be analysed rapidly in ever changing background conditions.

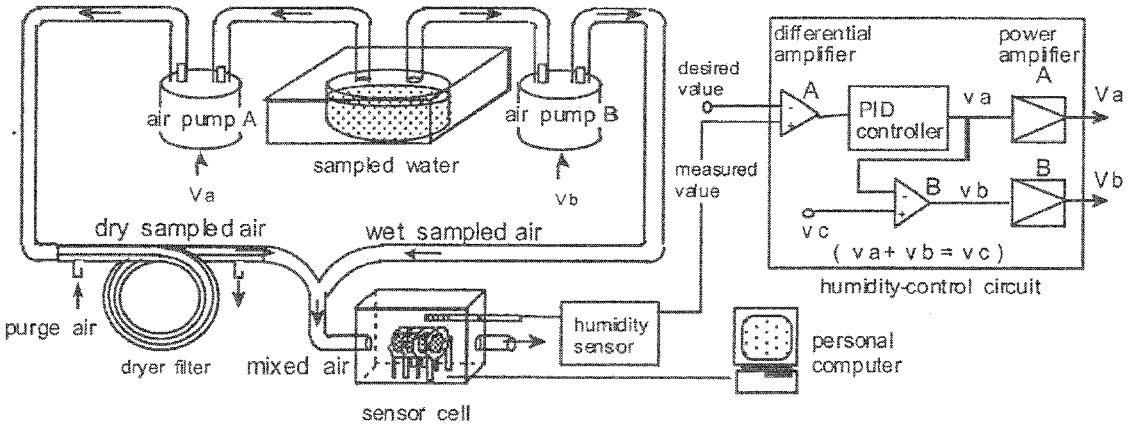


Figure 2.26: Diagram of humidity control system developed by Ogawa and Sugimoto (2001) for the detection of petroleum hydrocarbon in water samples. The sensor cell temperature is also controlled using a heating system and Peltier element (not represented).

2.1.5.3 Mixtures

Reported developments in complex odour analysis have typically been limited to quality assessment in the food, drink and perfume industries and so far fewer attempts have been made to characterise and measure complex odours in the environment. The approaches reported in the previous section, which in effect are surrogates for existing analytical techniques, have the advantage of rapidity, objectivity and selectivity and can help provide an accurate description of chemical composition. This can be useful for a number of application but they also have the following disadvantage when applied to broader environmental odour measurements (Preti *et al.*, 1993; Gostelow *et al.*, 2001):

- Most environmental odours are complex mixtures of components and single substance analysis may not provide a representative picture of the odour as perceived by humans
- The limit of analytical detection may be higher than the threshold of smell
- Interactions between different odorants and interfering background substances may lead to synergistic or antagonistic effects

Cross-reactive sensor arrays present a real interest since they naturally perform an integration to yield a unique response pattern (or fingerprint) for complex but distinctive odours, without requiring the mixture to be broken down into its individual components. This is advantageous when the only required information is the composite composition of the odour of concern (Albert *et al.*, 2000). In a review of odour measurement techniques for sewage treatment works, Gostelow *et al.* (2001) listed some examples of sensor arrays application to environmental odour problems (Table 2.5).

Table 2.5 Some application of sensor arrays to environmental odour problems (Gostelow *et al.*, 2001)

Authors	Application	System type	Comments
Hobbs *et al.* (1995)	Assessment of odours from livestock wastes	20 element conducting polymer array	Intrument found to be insensitive compared to olfactometry. Able to discriminate between odours at high concentrations
Persaud *et al.* (1996)	Measurement of odourous components of pig slurry	20 element conducting polymer array	Instrument capable of discriminating between different odour components and produced output proportional to odorant concentration
Misselbrook *et al.* (1997)	Comparison of electronic noses and olfactometry for cattle-slurry related odours	20 and 32 element conducting polymer arrays	Average sensor output explained ~60% of the variance in odour concentration. Odour concenrations in the range 100-1000 ou m^{-3} considered
Stuetz *et al.* (1998b)	Detection of tainted water samples	12 element conducting polymer array	Tainted samples successfully identified at concentrations as low as 1pg.l^{-1}
Stuetz *et al.* (1998a); Stuetz and Fenner (1998)	Comparison of electronic nose against olfactometry for 10 sewage treatment works	12 element conducting polymer array	Strong correlation between nose output and odour concentration found when odour from single source is considered. Poorer correlations between nose output and odours from several sources

In a more recent study, Romain *et al.* (2000), successfully used an array of 12 commercial tin oxyde sensors to identify real malodour samples collected in the field at various concentration and under a wide range of operating and weather conditions. These developments follow results from preliminary investigations on synthetic

single-components odours and the study of the influence of humidity and temperature (Romain *et al.*, 1997).

These examples of environmental odour measurement illustrate some of the possible future application of electronic noses for air quality and annoyance odour assessment. However, some obstacles remain in their application, not least being the need for calibration against olfactometric measurements (Stuetz and Fenner, 1998). As Gostelow *et al.* (2001) pointed out, development has to a large extent focused on the discriminatory capabilities of sensor arrays. Future development would also need to address the relationship between sensor output and odour intensity (Stuetz and Nicolas, 2001).

Subsequent investigations involving real wastewater sample (Stuetz *et al.*, 1999a, b), demonstrated that a commercial array of 12 CP sensors could potentially be used for monitoring organic pollution as a good correlation between the sensors output and the 5-day Biochemical Oxygen Demand test (BOD_5) was observed for periods of 4 weeks or less. The findings showed the enormous potential of the technique and suggested that it could be used for predicting organic load content and for process control at a wastewater treatment plant. In a similar study using an array of 12 MOS sensors, Dewettinck *et al.* (2001) confirmed the ability of non-specific sensor arrays to monitor routine parameters (COD, TSS and VSS) in a treated domestic effluent.

2.1.5.4 Continuous monitoring in the environment

While measurements and detection of environmental pollutants using sensor arrays can be successful under laboratory controlled conditions, continuous in-situ monitoring remains the most challenging aspect of environmental sensing. The following examples show some of the latest developments in this field.

Air quality and malodours

Despite the relatively limited number of reported studies, air quality monitoring is an area of growing interest where important field-based knowledge is currently being gained. Recently, Persaud *et al.* (1999) used a hybrid sensor array consisting of 20

CP's and 6 QCM's to continuously monitor the environment of a confined system over a 6-month period. The study was carried out in the MIR space station during the MIR-95 and DARA MIR-97 missions, and showed that the system was capable of monitoring the changes in air quality in real-time as the cosmonauts carried out their daily duties. It also proved useful in detecting simulated pollution as well as leakage and fire events. Notably it was reported that after 1.5 year the system showed little drift or degradation. Carrotta *et al.* (2001) also demonstrated that carefully selected sensors can be used for real time assessment of CO and Nox concentration. Menzel and Goschnick (2000) proposed a dynamic signal processing approach for real time assessment of air quality in a rapidely changing environment.

Continuous monitoring of malodours emitted by industrial and waste processing installations cannot be performed with conventional measurement techniques. Because odour emissions are often indicative of a process performance, real-time measurement of odour releases has become highly desirable. Similarly, the control of odour abatment processes as well as the detection and identification of odour annoyances reported by the public would be of interest to plant managers. Nicolas *et al.* (2001) have used an expert system for the detection of a single odour event corresponding to the daily emptying of the settling pond of a sugar factory. The authors also investigated a number of application for olfactive pollution monitoring using MOS sensors arrays (Nicolas *et al.,* 2000; Romain *et al.,* 2000). These included the detection and identification of olfactive annoyances originating from printing houses, paint shop, wastewater treatment plants, landfill sites, urban waste composting facilities and rendering plant.

Water and wastewater

The quality of water is an important issue for water supply to many different industries and household consumption, and sub-standard quality is a major source of complaints for water companies. Trace concentrations of chemicals can affect the organoleptic properties of supply water. The ability to detect these changes in taste and odour is a complex task which mainly relies on human sensory analysis (Rigal, 1995). Hogben and Stuetz (2001) demonstrated that an array of CP sensors could be

successfully used to detect the presence of organic pollutants such as 2-chlorophenol in water down to 1 ppm. Still, the use of electronic noses to monitor water quality in the field remains a virtually unexplored domain with only a handfull of applications reported so far. A promising area may be the detection of hydrophobic and highly volatile organic compounds in clean water systems as recently reported by Ogawa and Sugimoto (2001) and Ueyama *et al.* (2001).

With regard to monitoring wastewater quality, the recent advances in sensor technology discussed in the previous sections, together with progresses in computing technology and pattern recognition, may also help provide a solution to a yet unresolved problem. The potential of sensor arrays for wastewater monitoring is discussed in section 2.2.5.4., where the technique is compared to the more traditional approaches for on-line monitoring of global organic parameters

2.2 ON-LINE MONITORING OF WASTEWATER QUALITY

2.2.1 Introduction

Real-time monitoring of wastewater quality remains an unresolved problem to the wastewater treatment industry. In order to comply with increasingly stringent environmental regulations, plant operators as well as instrument manufacturers have expressed the need for new standards and improved comparability and reliability of existing techniques.

A review of currently available methods for monitoring global organic parameters (BOD, COD, TOC) is given (published in revised form in Bourgeois *et al.*, 2001). The study reviews both existing standard techniques and new innovative technologies with the focus on the sensors' potential for on-line and real-time monitoring and control. Current developments of biosensors, optical sensors and sensor arrays as well as virtual sensors for the monitoring of wastewater organic load are presented and the interests and limitations of these techniques with respect to their application to wastewater monitoring are discussed.

2.2.2 Principles and classification of existing techniques

In addition to traditional laboratory based analytical techniques used in the water industry, recent years have seen the development of a range of innovative monitoring equipment. Although only a small number has yet reached the market or has been accepted, there is already a great diversity of techniques and technologies available, both commercially and in research laboratories, which are reported in the literature. As a consequence, different schemes have been used in an attempt to classify existing sensors and analysers according to their respective properties.

Lynggaard-Jensen (1999) listed 8 different sensor/analyser properties (Table 2.6) which should be taken into consideration before their introduction into wastewater

systems (i.e. for monitoring or process control). Indeed key features such as the cost of ownership, ease of use, placement of the sensors, as well as the response time, will influence the consumer's choice. Other technical aspects such as the principle of measurement, reliability, accuracy and detection limits will also dictate whether or not the technology will be accepted and promoted as a standard (or alternative) method by the end user and relevant authorities. It is, therefore, evident that both the performance characteristics (range, linearity, accuracy, response time, limit of detection, etc.) and the intrinsic properties of the sensors (single or multi-parameter, need for external sampling and filtration, intrusive/non-intrusive) are of major importance when looking at existing and new methodologies for wastewater systems.

Table 2.6 Relevant sensor properties (after Lynggaard-Jensen, 1999)

	Property	Example
1	Placement of sensor	*In-situ*, at-line, in-line; on-line, off-line
2	Principle of sampling	External sampling, no external sampling
3	Principle of filtration	Filtration, no filtration
4	Principle of sample treatment	Continuous, batch
5	Principle of measurement	Photometric, colorimetric, enzymatic, titrimetric...
6	No. measurands	Single parameter, multi parameter
7	Need for supplies	Consumables, no consumables
8	Service intervals	Long, medium or short intervals

2.2.3 Wastewater monitoring

Monitoring of wastewater quality parameters is currently a subject of growing concern both in the United Kingdom and internationally. Indeed, a number of European regulatory measures and recommendations, such as the 91-271 EEC directive, have put pressure on the water and wastewater treatment industries with

respect to discharge requirements. Traditionally, the quality of treated wastewater is defined by the measurement of global parameters such as Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Organic Carbon (TOC), and Total Suspended Solids (TSS) (Thomas *et al.*, 1997; Wacheux, 1998; Cecile, 1998). In addition to providing vital information on the quality of treated wastewater and treatment efficiency, these procedures demonstrate that a wastewater treatment plant meets statutory requirements. In order to comply with these regulations on a permanent basis, and because of the spatial and time dependant variability of wastewater characteristics, on-line monitoring of the above parameters is clearly needed. At present, however, there is often infrequent monitoring of wastewater quality at most treatment plants. The automation of wastewater systems is not as developed as other process industries, mainly because of the hostile environment in which sensors have to be located (Lynggaard-Jensen, 1999). The lack of sensors suitable for on-line real-time monitoring/control is often due to uncertainties regarding the reproducibility/reliability of existing methods, whereas, more sensitive and standardised laboratory-based techniques are time consuming and require sample collection and retrospective analysis. Furthermore, straightforward extrapolation of laboratory measurements is not satisfactory as a grab sample (taken on a daily basis in the best case scenario) and is unlikely to provide a meaningful and high resolution picture of the nature of, and variation in wastewater quality.

In Europe, the owners and operators of treatment plants together with the producers of monitoring equipment have expressed an urgent need for new standards and the improvement of comparability, reliability and quality of existing techniques, as well as, to develop new, fast-responding technologies (Colin and Quevauviller, 1998; Pouet *et al.*, 1999). In this perspective, and with the growing need for on-line, and real-time, monitoring of wastewater quality, it became necessary to present and discuss the state of the art of wastewater monitoring techniques for global parameters.

2.2.4 Monitoring organic pollution in wastewater : Standard methods

Gross discharges of untreated or insufficiently treated wastewater effluent into receiving waters are generally considered to be one of the most important and common form of water pollution (Reynolds, 1999). Because of the complexity and variability of wastewater quality, 'Blanket' determinations, which measure a range of similar substances rather than individual compounds, are usually carried out. When considering organic substances in wastewater, it is customary to measure the amount of oxygen required to oxidise these substances, this then provides a measure of the strength or polluting potential of the sample (An *et al.*, 1998; Khan *et al.*, 1998a; Khan *et al.*, 1998b; Guwy *et al.*, 1999).

2.2.4.1 Biological Oxygen Demand

Traditionally, organic pollution is measured by the standard, off-line, five-day Biological Oxygen Demand test (BOD_5), mainly as proof of compliance with relevant legislation (Guwy *et al.*, 1999). This test is an empirical test indicator of biological activity. It is defined as the potential for removal of oxygen from water by aerobic heterotrophic bacteria which utilise organic matter for their metabolism and reproduction (Brookman, 1997). In practice it is essentially a measure of the amount of dissolved oxygen required for the biochemical oxidation of organic compounds in 5 days. This gives an indication of the waste biodegradability and is, therefore, a desirable measurement in biological treatment processes (Ekama *et al.*, 1986; Germili *et al.*, 1991; Guwy *et al.*, 1999). The standardised test was first published in *Standard Methods* in 1917 (Young and Clark, 1965). There are many problems associated with the application of the test and the meaning of the results, which are well documented, and its limitations have been extensively reviewed (Schroeder, 1977; Grady and Lim, 1980; Metcalf and Eddy, 1991). Although the test has been refined over the years, the basic approach of using a dilution technique has remained essentially unchanged (Logan and Wagenseller, 1993). Unfortunately these dilutions reduce the concentration of substrates and micro-organisms in the samples, thereby decreasing overall kinetic rates (Logan and Wagenseller, 1993). This, in conjunction

with the arbitrary time period of 5 days, may not reflect the conditions in the treatment process (Logan and Wagenseller, 1993) and must be taken into account when interpreting the results.

Another problem with the test lies in the presence of toxic substances in wastewater which can affect (or even kill) the bacteria, even at low concentrations. For instance, heavy metals such as lead, copper, mercury or chromium have been shown to inhibit, or completely prevent the oxidation of organic waste by bacteria (Ademoroti, 1986). The extent to which toxic substances affect BOD values is well documented (Kalabina, 1946; Placak *et al.*, 1950; Ademoroti, 1985). As, in practice, such substances are rarely monitored, it is often difficult to tell if a decrease in BOD values is only due to a decrease in the organic load. Although the BOD test still provides the best estimate of the reactivity of the contaminants in the natural environment, it remains insensitive and imprecise at low concentrations (Khan *et al.*, 1998a). Last but foremost, the BOD test is slow to yield information and is also labour intensive, because of the necessary dilutions and other manipulations. Since the method takes 5 days to complete, it is not suitable for process control and real-time monitoring where rapid feedback is essential. In the case of a pollution event or an operation problem, adverse conditions would be unknown and perhaps persist for 5 days until the outcome of the BOD test is known (Logan and Wagenseller, 1993). Despite all these limitations, the 5 day BOD test is still extensively used and preferred over other tests (Clark, 1992; Guwy *et al.*, 1999). Interestingly, Logan and Patnaik (1997) noted the crude character of the standard method, when compared with modern analytical techniques. Yet a substantial amount of time and money is still devoted to measuring BOD at wastewater treatment plants. The only recent improvement involved measuring the Dissolved Oxygen (DO) using a DO probe which often requires daily calibration (Winkler method) to ensure probe accuracy. Finally, the method requires experience and skills, and it generally has an uncertainty of 15-20% in the results. It is argued that the heterogeneity of microbial populations in wastewater treatment plant systems and their widely differing responses to substrates, in combination with the variability of wastewater compositions could be responsible for these discrepancies (Iranpour *et al.*, 1999).

2.2.4.2 Chemical Oxygen demand

In an attempt to overcome the difficulties in achieving reproducible results, as well as speed and precision, a series of wet chemical techniques have been developed (Reynolds, 1999). The use of chemical oxygen demand (COD) as an analytical parameter to monitor organic pollution has become more common in recent years. The COD test is now widely used as a means of measuring the organic strength of domestic and industrial waste, often replacing BOD as the primary parameter in wastewater. It is based upon the fact that most organic compounds can be oxidised by the action of strong oxidising agents under acid conditions (Sawyer *et al.*, 1994). As for BOD, it essentially consists of measuring the amount of oxygen required. In the case of COD however, organic matter is converted to carbon dioxide and water regardless of the biological assimilability of the substances. For example, glucose and lignin are both oxidised completely (Sawyer *et al.*, 1994).

The major advantage of the COD test is that the results can be obtained within a relatively short time (approx. 2 hours instead of 5 days for the BOD_5). Additionally, it was shown that the presence of toxic substances would not affect COD measurements (Ademoroti, 1985). In many cases a linear relationship between COD and BOD can be established (Ademoroti, 1986), allowing COD data to be interpreted in terms of BOD values. However this relationship may be time dependent and is very likely to be affected by changes in wastewater quality. Such prediction requires the use of a solid experience-based model and reliable correlation factors (Sawyer *et al.*, 1994), with particular care being given to sudden changes to wastewater composition. When used in conjunction with BOD, the COD test can provide an indication of the biodegradability of the wastewater (the BOD/COD ratio is frequently used) and can also be helpful in the detection of toxic conditions.

One of the main limitations of the COD test is its inability to differentiate between biodegradable and biologically inert organic matter on its own. Therefore, its use as a parameter in controlling biological treatment plants remains questionable. In addition, the use of chemicals such as acid, chromium, silver and mercury produce liquid hazardous waste which requires disposal. This has largely been responsible for

the lack of uptake by the water industry. Finally, limited precision and accuracy of the test at values less than 5 mg/l was reported by APHA *et al.* (1989).

2.2.4.3 Total Organic Carbon

Since its appearance in the 1970s, the use of total organic carbon (TOC) as an analytical parameter has become more common. This is particularly true in the case of industrial wastewater where it is often considered as the most relevant or 'true' parameter for the global determination of organic pollution (Thomas *et al.*, 1999). One of the major advantages lies in the rapidity of the test. Determination can be carried out in triplicates within minutes, allowing a much greater number of measurements than is possible with BOD or COD tests.

Two main techniques are usually used for the conversion of organic carbon to carbon dioxide for TOC determination (Thomas *et al.*, 1999). The first, which is sometimes called Wet Chemical Oxidation (WCO), oxidation is performed at low temperature by ultra-violet light and the addition of persulphate reagent, after removal of inorganic carbon by acidification and aeration. The second and more recent method, uses a catalyst at high temperature (650-900°C) and is known as HTCO (High Temperature Catalytic Oxidation). Recent studies have, however, shown significant differences and conflicting results between the two techniques (Hedges and Lee, 1993; Sharp *et al.*, 1993; Thomas *et al.*, 1999). As a result, both methods are still being investigated and their accuracy is still subject to controversy (Statham, 1997). In addition, there is still considerable resistance to its use for municipal wastewater because of the difficulty faced in obtaining good TOC-BOD_5 correlations (Viraghavan, 1976; Wilson, 1997). This is particularly important in studies on the kinetics of wastewater treatment processes (Wilson, 1997). In fact, TOC only measures the content of organic compounds, not other substances that may contribute to BOD (APHA, 1992).

Given these considerations, it is clear that the standard methods of monitoring wastewater BOD, COD and TOC all pose some problems to the end user as well as

the legislator. Amongst the most important limitations are the time required (5 days in the case of BOD_5), the lack of reliability and the difficulty in achieving reproducible results. Although both COD and TOC are more effective in terms of speed and accuracy, they are unable to differentiate between biodegradable and non-biodegradable matter. Furthermore, these methods are mainly based on sample collection and retrospective analysis. In order to comply with the regulations and establish standards for environmental protection, wastewater treatment plants need to implement automated measuring techniques. Sensors and other analytical tests in continuous or sequential mode would be suitable for alarm systems, and would facilitate process control and plant operation strategy.

Following the increasing demand for real-time monitoring, several innovative techniques and prototypes have been developed on a laboratory scale in some universities and a considerable amount of literature has been published. In recent years laboratory based commercial instruments have become more widely available and manufacturers are now offering some devices adapted to on-line applications that can perform rapid monitoring of wastewater organic strength (Figure 2.27).

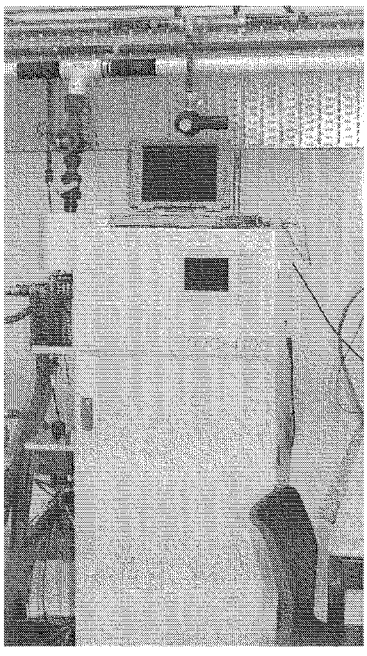


Figure 2.27: Example of At-line commercial TOC measurement system: TOC4100, Shimatzu UK.

However, there is still a gap between the research and development of new techniques, and their effective implementation by the end user. The water industry remains slow in taking up new technologies because of the lack of recognised and standardised methods or instruments that would satisfy all their practical requirements. Consequently, in order to improve the comparability, reliability and quality of measurements obtained from in-situ on-line sensors/analysers, new projects such as the ETACS (European Testing and Assessment of Comparability of on-line Sensors/analysers) project have recently been initiated (project 3256 ETACS). This project aims to achieve standardised validation of methods for in-situ on-line measurement of water quality determinants such as BOD, COD and TOC (Jacobsen and Lynggaard-Jensen, 1998) which should help fill the wide gap that currently exists between laboratory or field research results of sensor technology and their implementation by the end user.

2.2.5 Alternative new techniques

There is a plethora of new techniques and technologies available for monitoring changes in organic load in wastewater. The purpose of these is to obtain specific information about changes in wastewater quality (particularly at the inlet) and to ensure treatment efficiency for compliance assessment and process control. Accordingly, several instrument manufacturers are now offering devices to perform rapid monitoring of wastewater BOD, but little experience with the technology has yet accumulated (Iranpour *et al.*, 1997). Here we present a selection of more or less developed systems that we have come across in the literature (Tables 2.7, 2.8 and 2.9).

In order to meet the requirements for real-time and on-line monitoring in the real world, all methods must be simple, fast and reliable, as well as cost effective (reagents, maintenance, energy, operator's time, etc.). With this in mind, several studies have been carried out to test and compare various types of analysers (Londong and Wachtl, 1996; Osbild and Vasseur, 1998; Gernaey *et al.*, 1999). Brookman (1997) noted the wide range of existing methods for the determination of BOD. These include the use of a microbial sensor (Riedel *et al.*, 1990);

potentiometric stripping analysis (Fayyad *et al.*, 1987); incubation in acidified N/80 permanganate for 4 hours to give the permanganate value (Lowden, 1981); and the use of a scanning optical sensor based on the oxygen quenching of luminescence (Li *et al.*, 1994). Accordingly, it is possible to distinguish between a few major types of techniques: biosensors and chemical, optical, or even "virtual" (software, modelling) sensors.

Table 2.7 Current techniques in monitoring wastewater organic load: Biosensors

Parameter	Description	Applications	Range & Response time	Interests	Limitations	Authors
BOD	Yeast impregnated on membrane	Domestic wastewater Food, pharmaceutical and pulp industry	30 minutes	Suitable for sterilised and artificial samples Uses commercially available instrument	Sample handling Filtration needed (-clogging problem) Temp.-controlled lab. "Soluble BOD" only	Iranpour *et al.* (1997).
(B)OD	Microorganisms isolated from activated sludge (immobilized)	Wastewater	6 minutes (6.6 - 32.8 mg L OD)	Short response time Not affected by variations in WW quality Uses realistic mixture of microbes (not just one species)	Influence of temp. and pH Storage at 4°C Limited lifetime and stability	An *et al.* (1998)
BOD	Luminous bacteria in suspension	Synthetic, municipal and food factory wastewater	15 minutes (5 - 120 mg L (linear))	Temp range (18- 25 °C) Disposable biosensor Easy storage and reactivation of cells	Single use Difficult for on-line (preparation and maintenance of cultures) Only one species Ph sensitive	Hyun *et al.*(1993)
BOD	Biofilm with preozonation of refractory organic compounds	Artificial wastewater, River waters	~ 5 minutes + 20 minutes (DL:0.2mg L)	Rapidity Sensitivity	Sample pretreatment Chemicals needed Ph and temp. sensitive	Chee *et al.* (1999)
N-BOD	Disposable microbial sensor	Ammonium standard solution and municipal wastewater	6 to 12 minutes	Specificity Short response time Measures nitrification and inhibition of nitrification Possible automation	Short lifetime (3 days) and stability Storage Long recovery time (45 min to 2 hrs) and calibration	Konig *et al.*(1999)

Table 2.7 (cont.) Current techniques in monitoring wastewater organic load: Biosensors

Parameter	Description	Applications	Range & Response time	Interests	Limitations	Authors
Headspace BOD ($HBOD_5$)	Similar to BOD_5: Oxygen demand from liquid phase and container headspace	Municipal wastewater	5 days	Representative of real conditions No dilutions Inexpensive Similar to standard method	Time required, not suitable for on-line monitoring Sample transfer for DO measurements	Logan & Wagenseller (1993)
$HBOD_3$	GC-based HBOD test	Municipal wastewater	3 days	Improved HBOD method (no sample transfer) Attractive alternative to BOD_5 test	Time required	Logan & Patnaik (1997)
BDOC (Biodegradable Dissolved Org. Carbon) SBOD (Soluble BOD under certain conditions)	Biological (inoculum from the sample)	Reclaimed and municipal wastewater	28 days (also 5 days) (4 - 15 mg L of DOC)	Specificity Insensitive to nitrifiation and inorganic oxidations Good precision and accuracy at low DOC concentrations	Filtration and sample handling Dilutions (for unknown samples) Complicated and time consuming lab-based technique	Khan *et al.* (1998a, 1998b)
Bacterial biomass-COD	AODC (Acridine Orange stain Direct Counting of bacteria)	Wastewater and wastewater sludges	Few minutes (5 - 10)	Specificity Useful for predictive mathematical models Not affected by particulate matter	Specificity Chemicals and sample handling Equipment Time consuming	Munch & Pollard (1997)
Readily biodegradable COD (RBCOD)	Single-OUR method	Wastewater	Few tens of minutes	Specificity Useful information for mathematical models	Specificity Calibration and dilutions required for high COD samples Temp sensitive	Xu & Hasselblad (1996)

2.2.5.1 Biosensors

In the field of biosensing, respirometric techniques have been a popular method of measuring biodegradable organics and a detailed description of different systems can be found in the *IAWQ Scientific and Technical report 7* (Spanjers *et al.*, 1998). Although some techniques use free cells in solution (Hyun *et al.*, 1993), the basic structure of microbial (BOD) sensors generally consists of an oxygen probe and immobilised micro-organisms. Tan and Qian (1999) reported that the microbial system used in the preparation of (BOD) biosensors usually consists of a single type of micro-organisms with species varying from one system to another. Similarly, the medium and the methods of immobilisation are generally different, and membranes of different types and pore sizes are used (Iranpour *et al.*, 1997; An *et al.*, 1998; Qian and Tan, 1999). Biosensors based on immobilised bacteria are now starting to be commercially available for operational use (Lynggaard-Jensen, 1999). An example of such biosensor is shown in Figure 2.28.

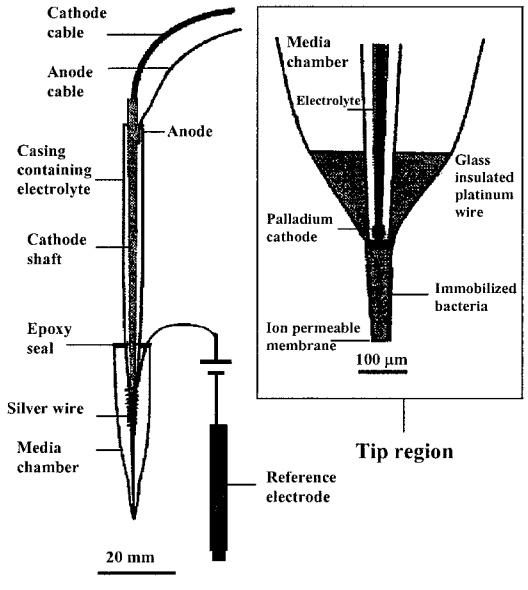


Figure 2.28 Biosensor based on immobilised bacteria (Lynggaard-Jensen, 1999)

Some of the major problems currently encountered when using biosensors for monitoring BOD have been discussed by Iranpour *et al.* (1999), Tan and Qian (1999) and Qian and Tan (1999). In particular, they pointed out the risks associated with the use of infectious (or non-infectious) microbes, both for the environment and for the users. Additionally, biosensors that use single species of micro-organisms, and of which the measurements are essentially based on the correlation between the concentration gradient of Dissolved Oxygen (DO) across the biofilm membrane composite and the BOD_5 equivalence of glucose-glutonic acid (GGA) BOD check solutions (according to Japanese Industrial Standards, 1993), are unlikely to be representative of the conditions that prevail in wastewater treatment plants (Figure 2.29). This is in contrast with the conventional BOD_5 test. However, biosensors offer the possibility to measure sterilised and synthetic samples which usually require seeding and acclimation of the micro-organisms. Thus, Qian and Tan (1999) and Tan and Qian (1999) showed the advantage of thermally killed and multi-species systems (now commercially available) and expressed the need for more complex BOD check solutions. These would preferably contain solutes with different molecular sizes and structures commonly found in wastewater.

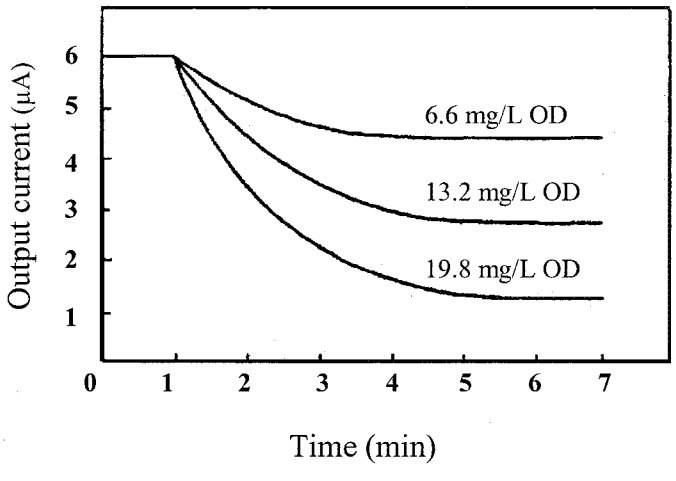


Figure 2.29 The typical response curves of the OD biosensor for GGA standard solution (An *et al.*, 1998).

Table 2.8 Current techniques in monitoring wastewater organic load: optical sensors and non-specific sensor arrays.

Parameter	Description	Applications	Range & Response time	Interests	Limitations	References
BOD	Non- specific sensor array (electronic nose)	Domestic wastewater	Less than 10 minutes.	Rapidity Non-invasive No reagents Versatile technique Good potential for on-line monitoring Uses commercially available instrument	Relationship is Source/site specific and time dependant Further development needed	Stuetz *et al.* (1999a, 1999b)
BOD* COD**	Oxydation by hydrogen peroxide with UV light	Diluted raw seawage, synthetic sewage and food industry effluents	55 minutes (0 - 54 mg L* 25 - 150 mg L**)	Fully automated Can be used for on-line monitoring Uses non toxic reagents	Limited range Range and correlation are source dependant	Guwy *et al.* (1999)
BOD COD* TOC** TSS Nitrates and anionic surfactants	U.V. spectral measurements and multivariate calibration	Wastewater surface waters	Few minutes 0 - 250 mg L 0 - 500 mg L* 0 - 150 mg L**	Rapiditiy Versatility Good potential for automation and on-line field monitoring of TOC, COD and nitrates Commercialy available	Sample handling Acquisition of reference spectra and calibration necessary for samples of different origin Not as good for BOD	Thomas *et al.* (1996, 1997)
BOD, nitrates, (TOC and COD)	Optical scattering (fluorescence)	Wastewater and surface water	Real-time	Real-time measurement Non-invasive Potential for screening and on-line monitoring	Still in infancy Research needed Fluorescence affected by pH and temp. *Correlation with BOD is plant site selective*	Reynolds (1999) Reynolds & Ahmad (1997) Ahmad & Reynolds (1999)
BOD	UV absorption (280nm)	Farm wastes	Few minutes (100 - 10000 mg L)	Indicates BOD range and dilution required for true BOD5 Potential for rapid gross measurement in the field	Poor sensitivity Uses only one wavelength Interferences from particles and toxic metals	Brookman (1997)
COD TOC	UV absorption	municipal wastewater	1 minute	No chemicals Relatively inexpensive On-line and real-time Sensitivity	Immerged sensor (fouling) Influence of suspended particulate material	Matsche & Stumwohrer (1996)

Biosensor measurements usually take a few minutes or hours. While this can be an advantage over conventional methods, less reactive substances of low diffusivity and/or enzymatic oxidation rates may be neglected during this short period (Tan and Qian, 1999). Because of the biochemical nature of the oxidation reaction involved, temperature and pH are limiting factors that must be controlled accurately (Figures 2.30 and 2.31).

Live cells also demand particular storage conditions and feeding when not in use, whereas the re-activation of dead cells, which allow easier and larger storage (and commercialisation), can be time consuming and take up to a few days. Furthermore, Osbild and Vasseur (1998) remarked that high concentrations of nutrients may also have adverse effects and decrease sensitivity. Finally, biosensors have a short lifetime, from a few days to a few months, which limits their application to continuous on-line monitoring.

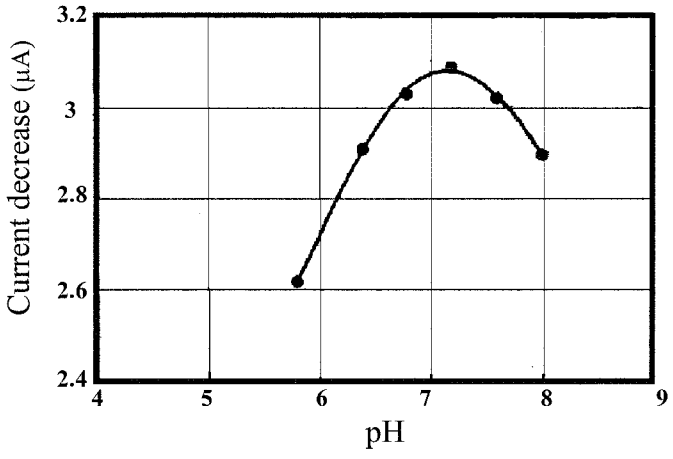


Figure 2.30 Influence of pH. A GGA standard solution of 13.2 mg/L OD was used at 30°C (An *et al.*, 1998).

Table 2.9 Current techniques in monitoring wastewater organic load: Modelling and virtual sensors.

Parameter	Description	Applications	Range & Response time	Interests	Limitations	References
COD, biomass and nitrogen	Modelling from known percentage of organic fractions	Activated sludge		Could help detect anomalies, or predict changes in complicated bioprocesses?	Still in infancy Lack of experience and basic understanding Better characterisation needed	Henze (1992)
BOD	Prediction using artificial neural network	Paper mill effluent	Few hours	Rapid prediction Good accuracy if bias monitored and feedback algorythm applied	Many other parameters needed Rely on operator to enter imput variables Affected by long term changes in WW quality Retraining and updating every 6 months	Masmoudi (1999)
BOD	Mathematical prediction from COD values	Domestic, poultry and brewery wastewaters	~ 2 hours (COD)	Rapid surrogate measurement	Not a real measurement Knowledge and dataset needed for each particular wastewater Assumes a constant and linear relationships exists all the time	Ademoroti (1986)
COD* NH4 NO3** NH4	ANN + multi sensor (pH, temp conductivity, redox potential, DO, turbidity..)	Influent* Aeration basin**		On-line Rapid surrogate measurement For monitoring treatment efficiency only	Approx/estimation Training needed Problem in case of sudden changes in ww composition. Reliable for a short period only	Hack & Kohne (1996)
RQ value (linked to COD/TOC ratio)	off-gas analysis (CO2 and 02)	WW treatment plants		Can be used to measure COD if RQ combined with TOC measurement	Does not distinguish C-oxydation from N-removal Only big changes in nitrification activity can be monitored	Hellinga *et al.* (1996)

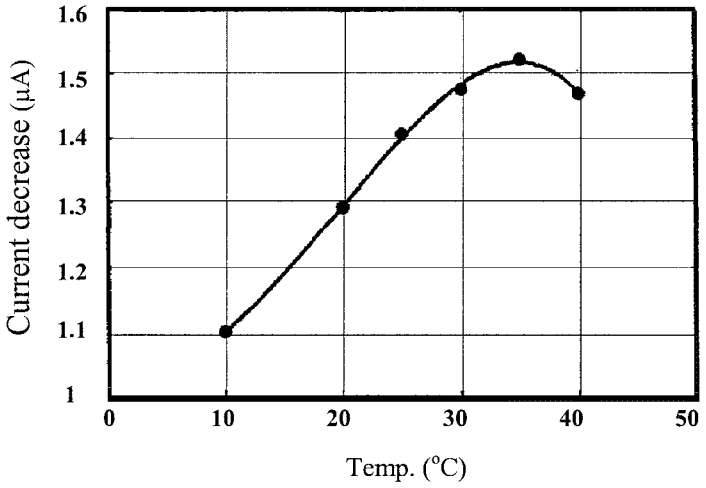


Figure 2.31 Influence of Temperature. A GGA standard solution of 6.6 mg/L OD was used at pH 7.2. (An *et al.*, 1998).

Existing measuring systems can be classified into batch-type and flow through, or flow injection type devices (Yang *et al.*, 1997). Most commercially available BOD sensors are flow-type systems that can be more easily automated, but generally require high maintenance to prevent fouling and clogging. In a recent study, Osbild and Vasseur (1998) presented a detailed review of a few major types of biosensors, along with their characteristics, and also discussed the needs of biosensor technology development. Respirometric type biosensors mainly rely on the assumption that oxygen demand is proportional to the organic content of the wastewater being analysed.

Alternative techniques have recently emerged and have proved useful in characterising wastewater quality. Gernaey *et al.* (1999) and Rozzi *et al.* (2000) recently looked at the development of titration biosensors. This relatively new family of instruments measure the activity and the proton consumption, or production rate, of a microbial population in a reactor vessel. In their reviews, the authors also described new advanced developments of hybrid respirometers where titrimetric techniques are coupled with respirometric techniques. Although indirect characterisations of the biodegradation of readily

biodegradable organic carbon (rbCOD) have been investigated (Rozzi *et al.*, 1997), these new technologies are generally still limited to ammonia and nitrification/denitrification studies (Table 2.10).

Table 2.10 Investigated applications of titration biosensors (after Rozzi *et al.*, 2000)

Conversion	Type	Substrate	Product	Titrant	Application Name
Nitrification	Acidif	NH4+	H+	NaOH	ANITA
Denitrification	Alkal	NO3-	OH-	HCl	DENICON
Denitrification	Alkal	rbCOD	OH-	HCl	DENICON-BND
Dephenol	Acidif	Phenol	H+	NaOH	ANITA
Methanogenesis	Alkal	Acetic acid	HCO3-	Acetic acid	MAID

2.2.5.2 Optical sensors

Optical techniques have a long history of use for chemical analysis and water quality monitoring. These techniques, based on the interaction of light with the sample, can be classified as light absorption measurements (UV-visible spectrophotometry, IR spectrometry) and fluorescence measurements (spectrofluorometry). The characteristic transmission, absorption, fluorescence spectrum or vibrational properties of a chemical species are measured in order to determine its concentration or identity (Scully, 1998).

Over the past few decades, the growing need for reliable on-line monitoring resulted in a considerable amount of research and development that has been devoted to more rapid and new techniques. The appearance of a wide range of near on-line instruments has been encouraged by the "optoelectric revolution" and associated lower costs, as well as the improvement of existing devices, light sources, detectors, and new optical material. The

uses and limitations of these techniques in water quality and bioprocesses monitoring, as well as the potential of some new concepts in sensor technology, have been extensively reviewed and discussed in recent literature (Scully, 1998; Lynggaard-Jensen, 1999; Reynolds, 1999; Vaidyanathan *et al.*, 1999).

The major advantages of optical techniques lie in their rapidity and versatility. Additionally, their relatively low running costs, the absence of chemicals, and the limited or no sample handling, make them highly desirable for real-time on-line monitoring for a wide range of parameters. Therefore, the range of optical sensors and measurement techniques continues to expand as new optoelectronic devices and material become available, with researchers moving towards waveguide-based systems and integrated optics (Scully, 1998). However a number of problems still need to be overcome, including biofouling of the probe tips, calibration stability and selectivity (Scully, 1998). For example, light absorption/ scattering measurements are often affected by the presence of air bubbles in solutions which can cause interference to the optical signal, and may result in errors. Similarly, agitation and aeration may result in noise in the data, and high suspended particles concentration can be a major limiting factor (Matsche and Stumwohrer, 1996; Brookman, 1997; Vaidyanathan *et al.*, 1999). Some early studies, based on the knowledge that many pollutants and dissolved organic compounds with aromatic structures strongly absorb UV radiation, have used UV spectrophotometry to determine dissolved organic matter in stream water (Grieve, 1985) and also to measure total organic carbon in waste water (Dobbs *et al.*, 1972).

In a recent attempt to estimate BOD_5 by using UV spectrophotometry, Brookman (1997) investigated the absorbance at 280nm for slurry and farm effluents. The approach proved useful for rapid estimations of BOD levels and the indication of dilution ranges for standard BOD_5 tests, but it showed poor sensitivity and was affected by interference from particles and toxic metals. Alternatively, Matsche and Stumworher (1996) showed the good correlation of UV absorption at 260 (or 254) nm with COD (and TOC), and the interests of additional measurements at a second wavelength (e.g. 380nm) to reduce the influence of particulate material. The technique is rapid (1 min.), relatively inexpensive

and offered good sensitivity without the use of chemicals and sample preparation. The major drawback, however, lies in the use of an immersed sensor for on-line monitoring which is prone to fouling (Figure 2.32).

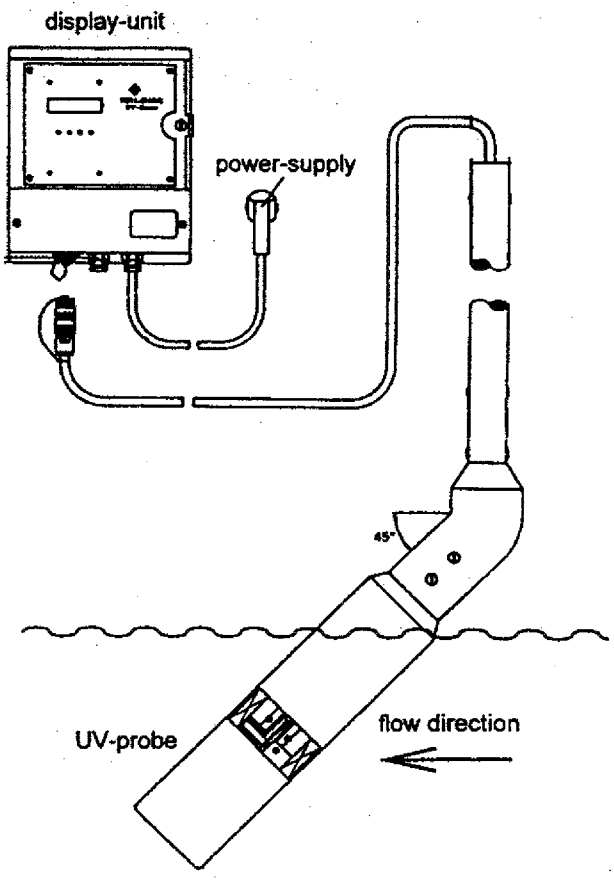


Figure 2.32 UV-probe (Matsche and Stumworher, 1996)

Thomas *et al.* (1996) discussed the significance of UV absorption and showed some of its limitations for the quality control of wastewater when using a small number of wavelengths (one or two). In their studies, the authors demonstrated the potential of UV spectral measurements with a deconvolution method for the estimation and on-line monitoring of specific parameters such as TOC, COD and BOD (as well as nitrogenous

and phosphorous compounds by the addition of chemical reagent). This rapid and versatile technique is now commercially available (Thomas *et al.*, 1996, 1997). Despite the fact that this is a flexible and well developed technology, some inadequacies still remain, associated with sample handling and the need for the acquisition of reference spectra. Finally, not all compounds (e.g. carbohydrates, saturated hydrocarbons) absorb at the specified wavelength and thus cannot be considered with UV spectrophotometry. This may lead to uncertainty in the estimation of factors such as BOD.

Although the technologies discussed above have gone some way to enhance monitoring, they continue to be invasive and are often not robust enough for on-line monitoring applications (Reynolds, 1999). As Reynolds and Ahmad (1997) noted, in most cases research was primarily concerned with establishing a correlation between UV absorbance at 254nm and the dissolved organic matter or carbon in natural waters, with the fouling of optical cell components resulting in a loss of sensitivity, reproducibility, and the need for frequent re-calibration. As a consequence, the use of fluorescence properties of natural waters to determine the presence of organic matter and to measure total organic carbon, or dissolved organic carbon, has been extensively studied in recent years (Fig 2.33) (Green and Blough, 1992; Mopper and Schultz, 1993; Ahmad and Reynolds, 1999).

In particular, Reynolds and Ahmad (1997), Ahmad and Reynolds (1999) and Ahmad *et al.* (1993), have studied the correlation between BOD_5 values and fluorescence intensity. The authors have shown the feasibility of detecting the scattering and fluorescence properties of surface and waste waters using non-invasive remote-sensing technology (Fig 2.34). The effects of some environmental parameters and the advantages over current methods (BOD, COD, TOC and absorption at 254nm) are also discussed. Although the prospects for non-invasive real-time and on-line monitoring of water and wastewater quality may seem promising and are indeed very attractive, the technology is still in its infancy, and further research and development is required. Specifically, research regarding the effects of environmental factors such as temperature, changes in pH and the concentration of metal species needs to be considered (Reynolds, 1999).

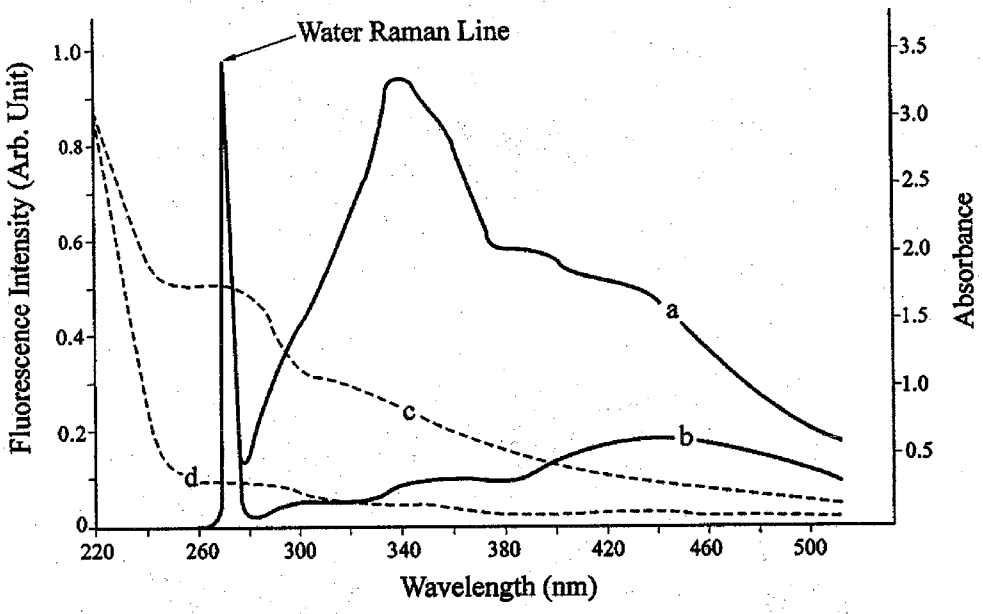


Figure 2.33 Typical fluorescence spectra of: (a) settled sewage; (b) treated effluent, and absorption spectra of: (c) settled sewage, (d) treated effluent (Ahmad and Reynolds, 1999).

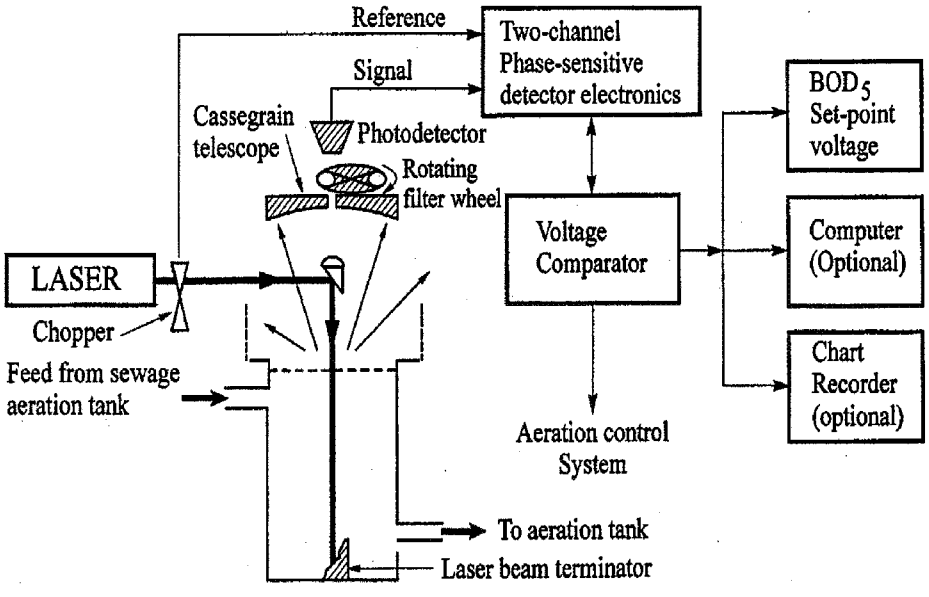


Figure 2.34 Scheme for non-invasive continuous monitoring of water quality for process control in water treatment plants, based on the detection upwelled fluorescence (Ahmad and Reynolds, 1999).

2.2.5.3 Virtual sensors

Software sensors, or 'virtual sensors', have been defined by Jacobsen and Lynggaard-Jensen (1998) as terms used for "signals" obtained from calculations using measured signals from reliable, available sensors in combination with other signals such as on/off indications and time counters. Such "sensors" emerged from the need for rapid evaluation/ prediction of essential parameters for which data is not readily available. In most cases, the lack of real-time/ continuous data is associated with technical limitations or cost restrictions.

Some virtual sensors have been used as filters or as real-time quality controls of datasets and sensor validation procedures. The technique has also proved useful in the field of modelling and the determination of environmental parameters, such as in the case of wastewater treatment plant designs. This allows the rapid prediction of the effect of changes of a number of factors on the variable of interest. One such mathematical tool is the neural network model, which was used for on-line prediction of mill effluent BOD (Masmoudi, 1999). The method uses artificial intelligence to build non-linear multivariate models based on large amounts of data. These techniques are not based on fundamental physical or chemical, or even biochemical principles and must be combined with practical insight to be effective. Similarly, in an early evaluation of the modelling of activated sludge processes, Henze (1992) reported that there are still many problems to be solved with regard to wastewater characterisation for modelling purposes. Indeed there are many difficulties associated with the lack of experience and/or basic understanding.

2.2.5.4 Sensor Arrays

In the last two decades there has been increasing interest in the development of sensor array technology. The ability of so-called 'electronic noses' to discriminate between

samples of various nature, as well as monitor their stability and changes in quality with time, has been discussed in the previous sections. With its potential for real-time, non-invasive and on-line monitoring of water and wastewater quality, the technology acquired growing interest in the water industry. Recent research has demonstrated the ability of electronic noses to detect low levels of organic pollutants and tainting compounds in both waste water and potable water, and that it is possible to distinguish between wastewater samples of different types and origin (Fenner and Stuetz, 1999; Stuetz *et al.*, 1999c).

Other similar research involving the analysis of real wastewater samples has shown that non-specific sensor arrays can be useful in monitoring organic pollution as a good correlation between the sensors' responses and the 5-day BOD test was observed for time periods of 4 weeks or less (Stuetz *et al.*, 1999a; 1999b; 2000). The findings (Figures 2.35 and 2.36) showed the enormous potential of the technique, and suggested that it could be used for predicting BOD values (or other organic load parameters), as well as for monitoring and/or controlling the biochemical activities of a wastewater treatment plant. However, this is still a novel technology and much of the original work was carried out with laboratory based sensor systems, which were developed for other applications such as foodstuffs, beverages and perfumes discrimination. With further development, the technology could serve a number of applications and provide a simple technique for real-time and non-invasive monitoring of wastewater quality.

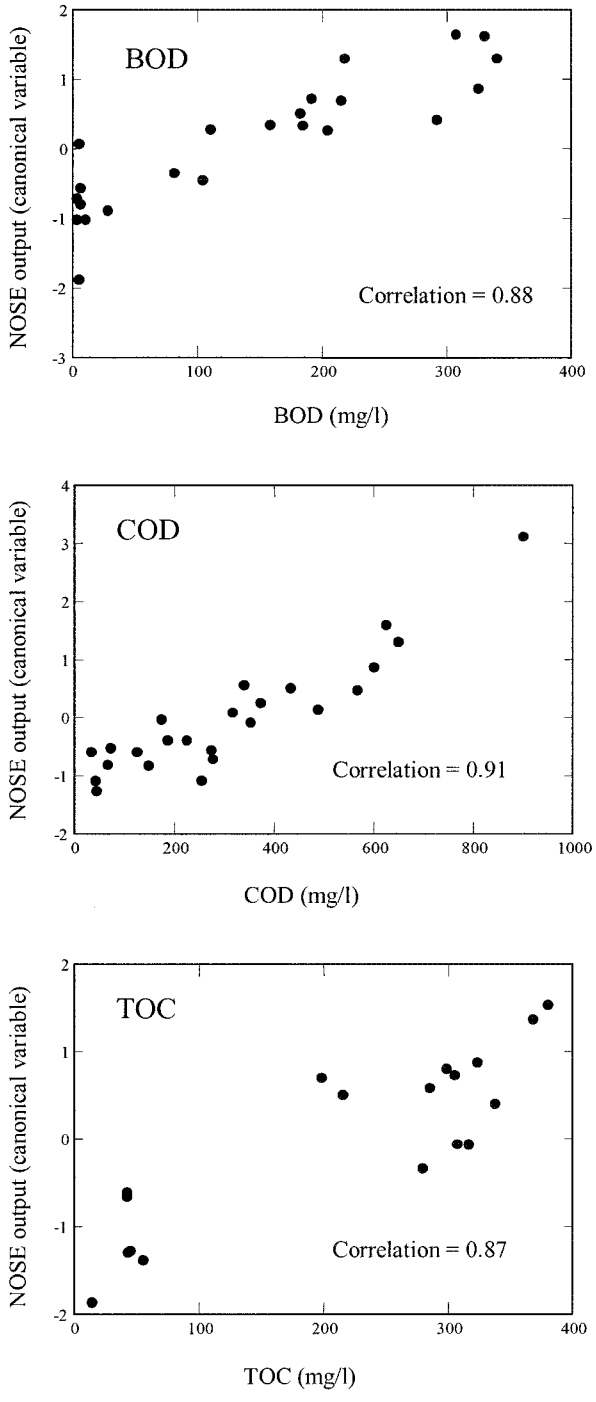


Figure 2.35: Canonical correlation analysis showing relationships between sensor response and BOD, COD and TOC, Using raw, settled and final effluent sewage collected over the same three weeks period (Stuetz *et al.*, 2000).

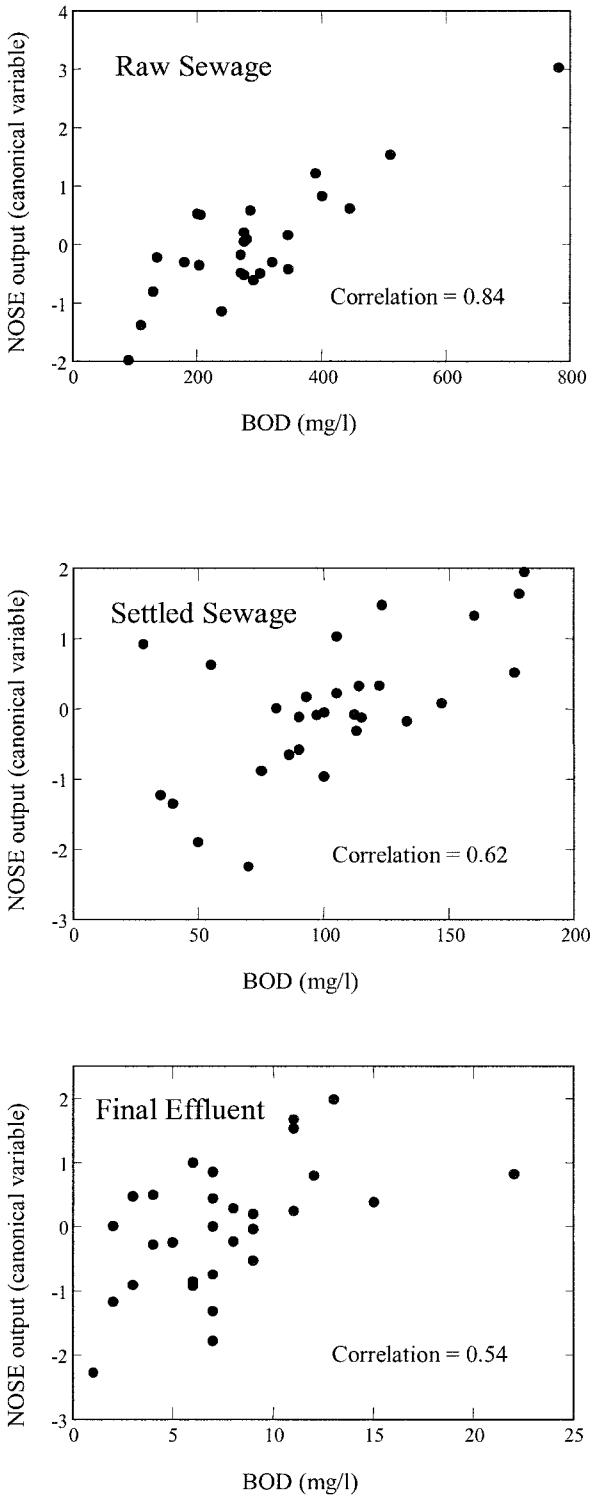


Figure 2.36: Canonical correlation analysis showing relationships between sensor response and BOD for raw, settled and final effluent sewage collected over 5 months (Stuetz *et al.*, 2000).

2.3 CHEMICAL ANALYSIS OF WASTEWATER

The typical composition of wastewater is highly variable and depends on a range of factors such as local diet, cleansing habits, sanitation and collection systems as well as environmental conditions (Parsons and Stephenson, 2002). Consequently, wastewater organic matter is also highly heterogeneous, containing molecules of various molecular weight, ranging from simple compounds like acetic acid, to very complex polymers (Henze, 1992). However, the chemical composition of the organic matter of waste waters and treated waters remains largely unknown (Painter 1973, Nielsen *et al.*, 1992). The principal groups of organic substances found in wastewater are proteins (40 – 60%), carbohydrates (25 – 50 %) and fats and oils (10%). A typical analysis of raw sewage is given in Table 2.11. In a recent study, Dignac *et al.*, (2000) investigated the fate of a range of organic compounds during activated sludge treatment. Molecular analysis of the samples however, only allowed the authors to characterise less than 50% of the organic matter found in raw wastewater and 22% of that present in treated waters.

Table 2.11: Typical Raw Sewage Analysis (from Parsons and Stephenson, 2003)

Parameter	**Units**	**Concentration**
Total dissolved solids	mg/l	250 – 850
Total Suspended solids	mg/l	100 – 350
BOD_5	mg/l	110 – 400
COD	mg/l	250 – 1000
Ammonia	mg/l N	10 – 50
Nitrate	mg/l N	0 – 5
Phosphate	mg/l P	5 – 10
Chloride	mg/l Cl	30 – 100
Sulphate	mg/l SO_4	20 – 50
Alkalinity	mg/l $CaCO_3$	50 – 200
Fat, Oil and Grease	mg/l	50 – 150
VOCs	µg/l	50 – 500
Total coliform	cfu/ml	10^4 - 10^7

An important group of chemicals present in wastewater are the volatile organic compounds (VOC's), particularly chlorinated and aromatic hydrocarbons, originating from various sources including industries, commercial facilities and residential households. The presence of these substances has, in recent years, complicated wastewater treatment because many of them can be persistent and induce toxicities to aquatic and mammalian organisms (Dewettinck *et al.*, 2001; Metcalf and Eddy, 1991).

Quantitative headspace and trace organic analysis (in the range 10^{-12} – 10^{-3} mg/L) are usually carried out using gas chromatographic techniques and mass spectroscopy (Metcalf and Eddy, 1991; Cruwys *et al.*, 2002). However with the great number of substances that can be found in wastewater, few comprehensive studies have so far been reported, with authors typically focussing on a few or family of compounds. In a recent detailed study of wastewater samples collected at four different municipal sewage treatment plants, Nikolaou *et al.*, (2002) investigated the occurrence of a range of VOC's included in the Catalogue I and Catalogue II of the 76/464 EEC directive. Pantsar-Kallio *et al.*, (1999) also reported on the chemical composition of wastewaters, looking at 83 chemical variables in a study of the factors affecting the quality of domestic sewage.

With particular reference to the analysis of the headspace gas of wastewaters, reported work has been mostly concerned with the characterisation and measurement of odour nuisances (Hobbs *et al.*, 1995; Persaud *et al.* 1996). Indeed, odours have been rated as the first concern of the public relative to the implementation of wastewater treatment facilities. (Patterson *et al.*, 1984). Many substances produced in the decomposition of the organic matter present in wastewater have very low human olfactory thresholds and so are perceived as odour nuisances even when their concentration in the air are very low. The major categories of offensive odours are listed in Table 2.12. Other work (Cruwys *et al.*, 2002) focussed on the measurement of volatile fatty acids (VFA's) in the headspace gas of wastewaters. As the author pointed out, the presence of VFA's in a sample matrix is often indicative of bacterial activity and VFA measurements are required to monitor the operation of wastewater treatment plants carrying out digestion, phosphorus removal

or denitrification. GC/MS and GC- Flame ionisation detection (FID) can be used to detect individual organic compounds, and although it can provide valuable and specific information on the quality of a wastewater, this is largely a lab-based and costly technique, which remains unsuited to real-time monitoring of global organic parameters in the field.

Table 2.12: Odorous substances group found in sewage (adapted from Vincent and Hobson, 1998)

Substance	Formula	Odour quality
i) *Volatile Sulphurous Compounds*:		
- Hydrogen sulphide	H_2S	Rotten eggs
- Mercaptan:		
Methyl mercaptan (methanethiol)	CH_3SH	Decayed cabbage
Butyl mercaptan (1-butanethiol)	CH_3CH_2SH	Skunk
Ethyl mercaptan (ethanethiol)	$CH_3(CH_2)_3SH$	Decayed cabbage
- Sulphur Dioxide	SO_2	Pungent, acidic
- Phenyl sulphide	$(C_6H_5)_2S$	Rotten cabbage
- Dimethyl sulphide (methyl sulphide)	$(CH_3)_2S$	Rotten cabbage
- Dimethyl disulphide (methyl disulphide)	CH_3SSCH_3	Rotten cabbage
ii) *Volatile Nitrogenous Compounds*:		
- Ammonia	NH_3	Ammoniacal, pungent
- Amines:		
Methylamine	CH_3NH_2	Fishy
Trimethylamine	$(CH_3)_3N$	Fishy
Putrecine (1,4-diaminobutane)	$NH_2(CH_2)_4NH_2$	Rotten flesh
Cadaverine (1,5-diaminopentane)	$NH_2(CH_2)_5NH_2$	Rotten flesh
- Skatole (3-methylindole)	C_9H_9N	Feacal, repulsive
- Indole	C_8H_7N	Feacal, repulsive
iii) *Volatile fatty acids, alcohols and ketones:*		
- Isopropanol	$CH_3CHOHCH_3$	
- Propionic acid	CH_3CH_2COOH	
- Acetone	CH_3COCH_3	

Chapter 3: AIMS AND OBJECTIVES

CHAPTER 3: AIMS AND OBJECTIVES

The overall aim of this investigation was to determine the extent to which non-specific sensor array technology could be used for wastewater monitoring. Our particular research objectives were as follow:

- To evaluate the application of using a commercial sensor array system to monitor wastewater organic load, in response to the water industry's need for on-line and real time devices.
- To identify and develop an appropriate sampling methodology for on-line monitoring of wastewater quality and to test it in-situ at an operating wastewater treatment plant.
- To study the relationship between the sensor responses pattern and wastewater organic load parameters such as BOD, COD and TOC.
- To identify appropriate data analysis protocols for the real time measurement of changes in the organic content of wastewater.
- To evaluate the potential of the sensor array based system as an early warning device for the detection of pollution incidents and sudden changes in wastewater quality.

Chapter 4: EXPERIMENTAL AND SYSTEM DEVELOPMENT

CHAPTER 4: EXPERIMENTAL AND SYSTEM DEVELOPMENT

4.1 INTRODUCTION

The proposed research aims to investigate the interest and suitability of a commercial sensor array system for on-line monitoring of wastewater quality. However, commercially available instruments are generally designed for laboratory based applications and there are no reports of instruments being developed for continuous use in wastewater treatment plants. Currently available devices generally require manual handling of the sample and therefore, have a limited sampling frequency. Consequently, it was necessary to make a number of hardware and software modifications to the instrument provided for this study. These were implemented with the aim to increase sampling frequency and rapidly produce larger databases for further statistical investigations.

This chapter gives a description of the material used for this research, and reports on the development of an on-line system and its evaluation in the laboratory under controlled conditions. The initial stages of this work have focussed on the design of a suitable headspace sampling apparatus as well as on the selection of an optimal sampling methodology that could be used for real-time monitoring of a continuously flowing wastewater effluent.

4.2 SENSOR ARRAY ANALYSIS AND DATA HANDLING

4.2.1 Instrumentation

All original work and preliminary investigations were carried out using a modified laboratory based eNOSE 5000 sensor array module (Marconi Applied Technologies, UK). The commercial instrument (shown in Figure 2.1) consists of an array of 12 electrochemically grown polypyrrole sensors, maintained in a temperature-controlled sensor chamber. The sensors, originally developed by Neotronics Scientific Ltd (currently Marconi Applied Technologies), were selected by the manufacturer for this particular study. These are sensor type 298; 401; 462; 463; 483; 501; 502; 503; 504; 505; 506 and 601. The main reasons behind the choice of these sensors for wastewater monitoring were:

1) "The conducting polymers are the most stable sensor types and have the longest lifetime. Therefore the choice was justified for such application where there would be prolonged and frequent sample exposure.
2) CP's respond well to volatile organic compounds. VOC's present in wastewater have been shown to provide relevant information on its quality (section 2.3.5.2). The actual sensor coatings were chosen as they gave a range of different responses to a number of chemical standards (ethanol, methanol, hexane, toluene, acetone, cyclohexane, etc.). It was believed that assembling the most versatile array of sensor would provide the greatest amount of information.
3) The "500 series" sensors (501-506), referred to as "CR3" sensors, were developed as hydrophobic sensors and were expected to be less affected by changes in humidity levels". (Marconi applied technologies, personal communication)

Unfortunately we are unable to provide more detailed information regarding sensor coatings and fabrication processes. This is because such industrially sensitive information is proprietary and the manufacturer has requested that it remains confidential. The instrument includes a control module for the measurement of RH, gas flow rate and headspace gas temperature. A third module is also present for the

addition of a MOS sensor array. In this study the corresponding channel was used to continuously measure the temperature inside the instrument. Thus, the temperature of the sensor module, of the headspace gas and of the instrument internal circuitry were accurately and independently monitored.

4.2.2 Acquisition

Filtered zero-grade nitrogen (BOC, UK) was used as a carrier gas for the sample headspace and to purge the sensors between each acquisition. Gas flow rate was controlled using an integrated mass flow controller (Brooks, USA). The sampling protocol for each acquisition cycle consisted of the following sequence, shown in Figure 4.1:

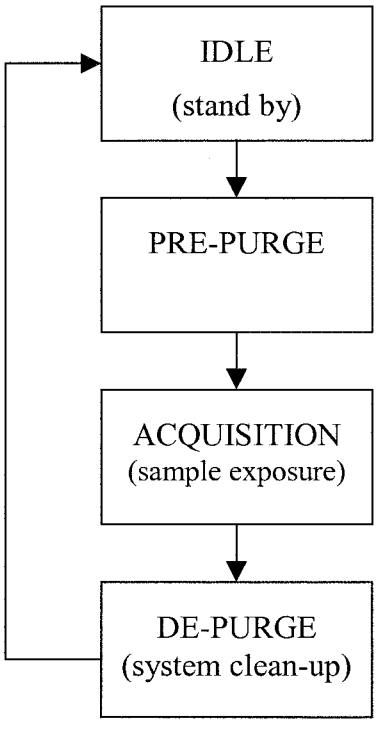


Figure 4.1: Flow diagram of the acquisition cycle using the eNose5000 sensor array

Using the instrument's software, duration and conditions (temperature, gas flow rate) for each step can be adjusted and the user's methods can be saved and recalled at a later stage.

The pre-purge period immediately prior to the acquisition phase was implemented in order to drive out any dead volume of gas in the circuitry and replace it with the true sample headspace sample to be analysed. This reduced any latency phase at the start of the acquisition due to sample transfer delays and avoided potential dilution or cross-contamination problems.

During acquisition, the sensors are exposed to the sample headspace and their change in resistance with time is recorded along with experimental condition measurements (gas flow, temperature, RH, etc).

The de-purge step generally consists of a higher flow of zero-grade nitrogen through the sensor chamber, until the sensors return to their original baseline.

An Idle mode (reduced flow of N_2 across the sensors) was also included in the procedure so as to be able to vary the sampling frequency without altering the selected Pre-purge – Acquisition – De-purge sequence, while keeping gas consumption to a minimum.

4.2.3 Data extraction

A raw data file is created for each acquisition cycle by the eNose 5000 operating software. These Comma Separated Value (".CSV") files contain the averaged sensor resistance values at every second during the acquisition and de-purge periods as well as temperature, gas flow rate, RH measurements and a time and date stamp. From these raw data files, the sensor response at a single time point was extracted and used to determine the change in resistance (% $\Delta R/R$) and characterise the sample headspace (Figure 4.2).

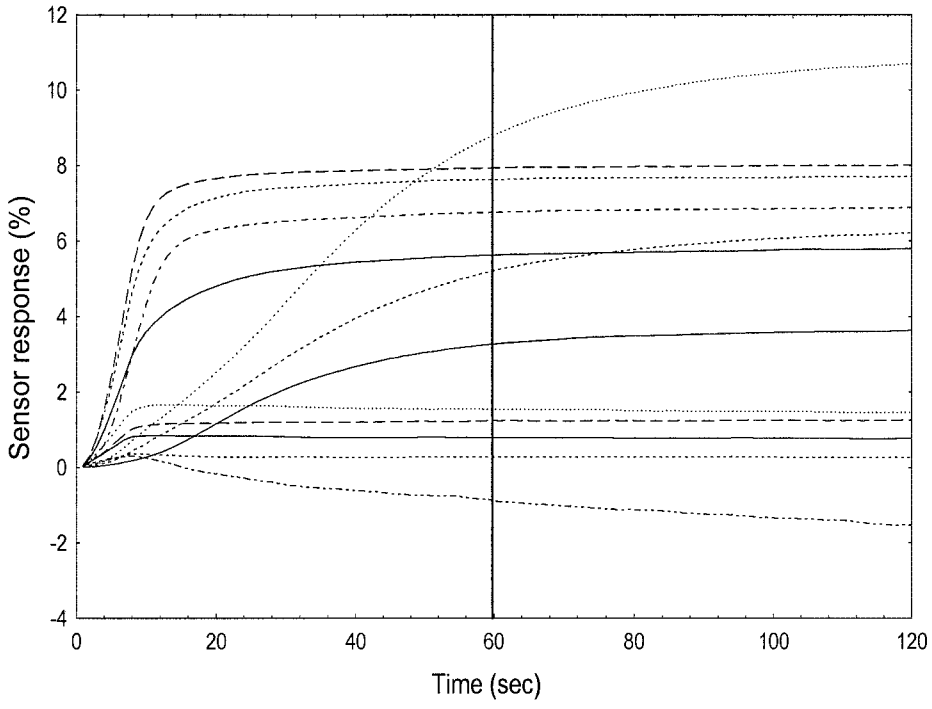


Figure 4.2: Typical response pattern of a chemical sensor array showing the response change (%) of 12 CP sensors and the extraction of a pattern profile at 1 min.

The extracted information is then stored in Excel files for data analysis and feature extraction using statistical packages such as Unistat and Statistica. A more detailed description of the multivariate data analysis and other pattern recognition procedures used in this study will be given in the corresponding sections.

4.3 DEVELOPMENT OF A HEADSPACE GENERATING FLOW-CELL

4.3.1 Design and preliminary assessment

As stated in Chapter 3, one of the first objectives of this study was to identify and test possible sampling techniques which could be adapted to on-line monitoring of a continuously flowing wastewater sample and coupled to a sensor array system. To do so, it was necessary to move away from the traditional and time consuming static headspace sampling of quiescent liquid samples, and propose a dynamic approach that would produce a representative gaseous sample in a rapid and reproducible manner. A precondition seen as essential to the validity of the evaluation trials and other method developments was the ability to have some control over a range of physical parameters such as temperature.

Selected results from the following sections have been published in Bourgeois and Stuetz (2000) and Bourgeois *et al.* (2002).

4.3.1.1 Design

A sampling system was developed in order to generate a headspace gas from liquid samples for continuous sensor array analysis. The sample vessel (flow-cell) shown in Figure 4.3 was designed to allow for a tight control over a range of sampling parameters such as sample temperature, gas flow rate and sparger porosity.

The sample temperature inside the internal sample chamber was regulated to +/- 0.1°C using an external water jacket and a heater cooler system (Haake DC50-K10, Germany). The sample chamber has a volume of 750ml and typically contains 500ml of liquid sample. The sample is sparged with zero grade N_2, producing a headspace gas which is then transferred to the sensor array system via a PTFE transfer line as described in Figure 4.4

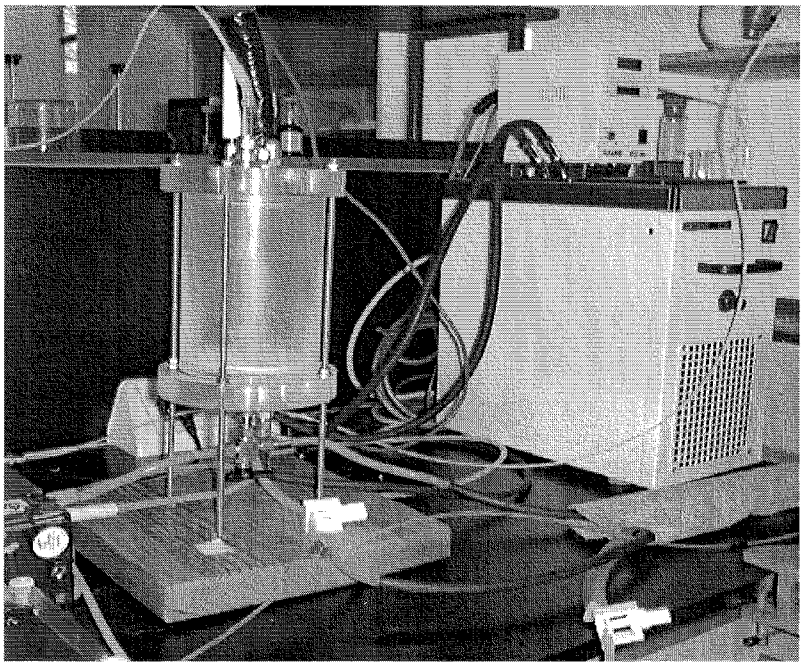


Figure 4.3: Picture of the headspace generating system (left) and temperature control unit (right).

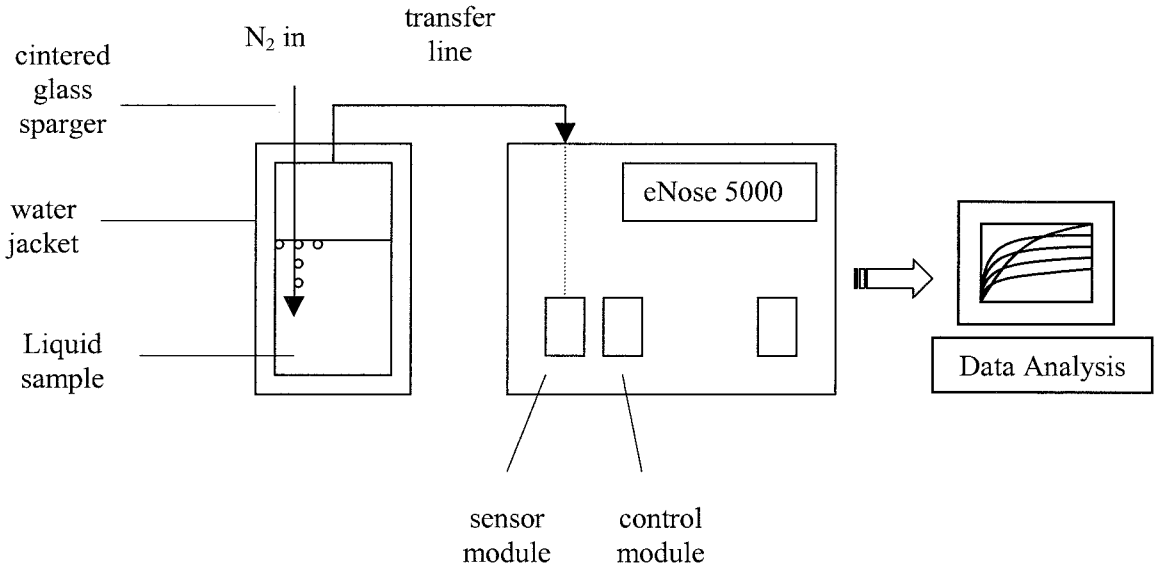


Figure 4.4: Diagrammatic representation of sampling apparatus showing (from left to right): headspace generating flow cell, eNose 5000 and PC for data anlysis.

The fittings of the cell allow for easy changing of temperature probe, sampling ports and sparger types and porosity. Changing the latter allows us to alter the size and amount of bubbles produced during sparging, therefore varying the surface contact between the liquid and gaseous phases.

4.3.1.2 Preliminary assessments

A first series of experiments was carried out in order to observe and account for the system and sensor's behaviour during analysis. These early tests were necessary to further develop a robust sampling methodology and validate the choice of a suitable data acquisition protocol.

The response of the sensors when exposed to a sample and their desorption time during the clean-up period were monitored. At first, long steps were chosen before the length of the acquisition cycle could be reduced so as to increase the measurement frequency. De-ionised (RO) water was initially used to look at the changes in the sensor profiles and ensure that all sensors have returned to their original baseline level at the end of each acquisition cycle. In order to cover both ends of the organic load spectrum, undiluted raw sewage (primary settled effluent, Cranfield university sewage works) was also analysed. Figure 4.5 shows how the time required by the sensors to desorb increase proportionally with the duration of the previous sample exposure. To increase sampling frequency, it is therefore recommended to reduce both the sampling and clean-up step accordingly. Based on experience acquired with a range of different samples and sampling conditions, the total duration of a full acquisition cycle was subsequently reduced.

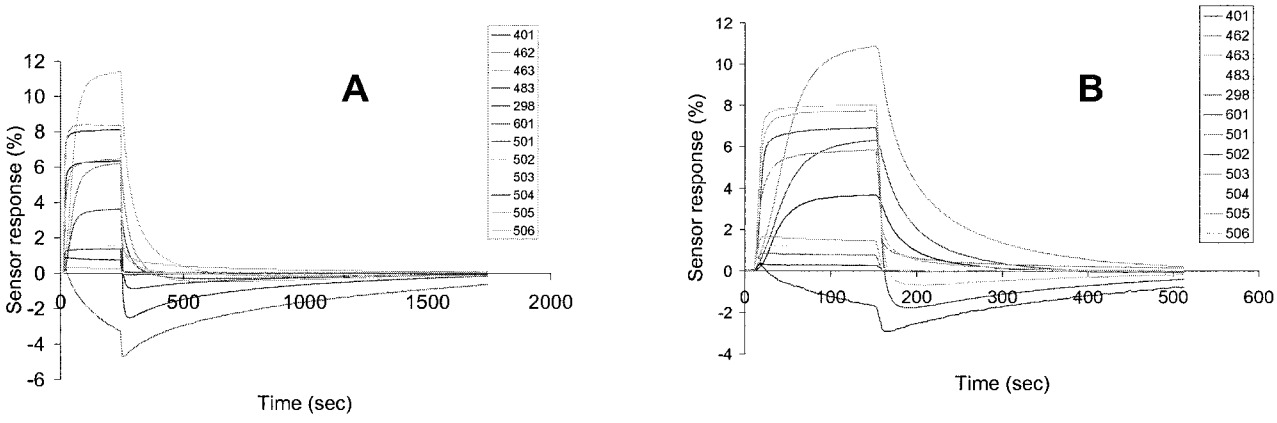


Figure 4.5: Sensor profiles for RO water showing a more rapid return to baseline after shorter acquisition period. Acquisition and clean-up time are: 5min + 25 min (A), 2.5min + 6min (B) respectively.

Additional observations arising from these preliminary evaluations include:

- The need to discard data from the first few measurements due to the sensors' need to "re-acclimatise" and reach running moisture levels if left unused for more than a few hours. Consequently, the instrument was left running with RO water for a few hours before experiments were carried out.
- A second series of experiments using a sparger with different porosity suggested the possible effect of this variable.
- The importance of gas flow rate was also noted when a leak in the circuitry caused a reduced gas flow through the sensor module (circa 60ml/min instead of 100ml/min). This resulted in distinctly different profiles.

4.3.1.3 Temperature control

The initial attempts to improve the sampling system design were concerned with reducing the possibility of condensation during the transfer of the sample headspace.

To avoid condensation, a positive temperature gradient (+5°C steps) was maintained from the flowcell to the sensor array module. This was achieved by using insulating and electrically heating (Astec, UK) the headspace transfer line. A heating and cooling system (Isopad, UK) was also incorporated within the eNOSE 5000 sensor array unit, which was maintained at 35°C +/- 1°C. Figure 4.6 shows an example of the effects of maintaining a controlled temperature gradient during the transfer of the sample headspace to sensor array unit. The sensor responses do however show that some slight sensor variability still remains in the generation of the sample headspace after the introduction of a temperature gradient.

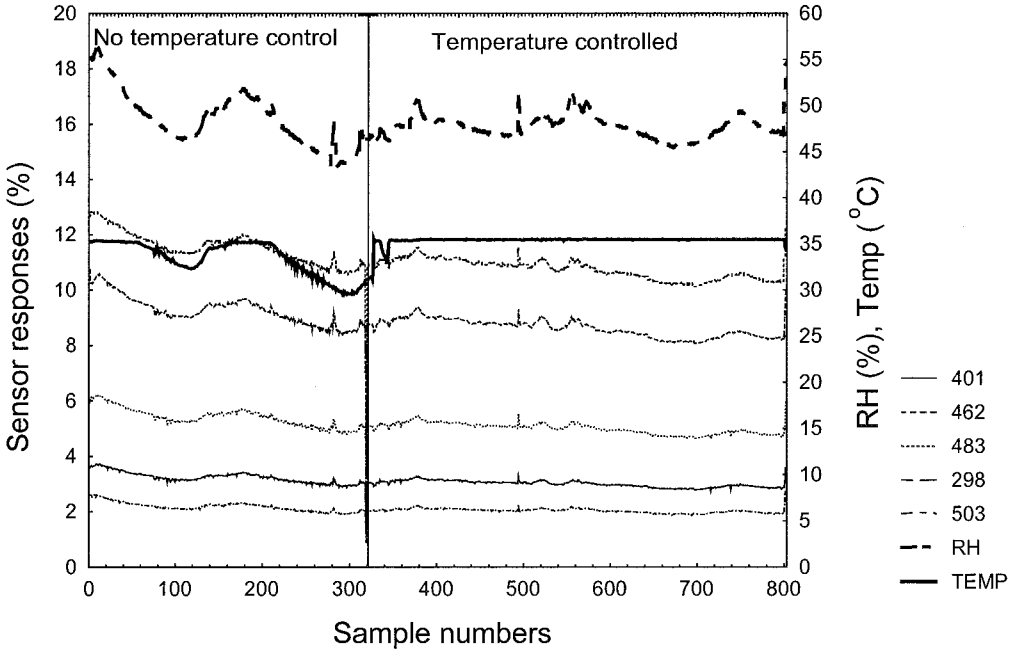


Figure 4.6: Plot of sensor responses showing the effect of system temperature control on relative humidity (RH) and sensor stability.

4.3.2 Prospects for real-time analysis

The work presented in this section aims to assess the application of using the previously described system to monitor the headspace of wastewater samples generated from an external flow cell. Relationships between the sensor array data for RO water and wastewater samples are compared and studied using multivariate statistical analysis.

4.3.2.1 Methods

Sample collection and composition

RO water samples were obtained from a reverse osmosis unit (USF-ELGA). Wastewater samples were collected from the Cranfield University sewage treatment works. Samples were collected from the primary tank (raw sewage) and final effluent outlet (final effluent). Spot samples of approximately 10 litres were collected for the experimental work which were routinely analysed for BOD and COD concentrations according to standard methods (APHA-AWWA-WOCF, 1995).

Sampling apparatus and sensor array analysis

Analysis of the liquid samples was performed using the flow-cell in static condition as described in section 4.3.1. The sample temperature was maintained at 25.5°C, the transfer line at 30°C and the sensor chamber at 35°C. The flow cell was flushed with RO water after each analysis to remove any tainting effect of the previous sample.. The sample protocol for the sensor array consisted of a 4-minute pre-purge, 2.5-minute sensor acquisition and a 6-minunte de-purge of the sensors. The sensor response at 1 minute after the beginning of the acquisition was used to represent the odour profile of the headspace sample.

Data analysis

Multivariate analyses were performed on the data collected using the statistical package UNISTAT. Pattern recognition techniques are used to reduce the dimensionality of the sensor array data, so that relationships between the observation can be explored using one or two dimensions (Persaud *et al.*, 1996; Misselbrook *et al.*, 1997). Unsupervised techniques, such as principal component analysis are used to find hidden relationships between the samples, whereas in supervised techniques such as a discriminant analysis, the training data is used with known properties. These relationships can then be compared and correlated using simple scatter plots.

4.3.2.2 Results and discussion

Flow cell monitoring of wastewater headspace

The sensor responses of headspace gases generated from sparged wastewater sample were used to evaluate the reproducibility of the individual sensor responses for on-line monitoring. Figure 4.7 shows a comparison between the 12 chemical sensors (using sensor responses at 1 min) and the relative humidity of the sparged headspace gas for a raw sewage sample. The 30 replicate responses show that some of the individual sensors change with time and that these changes can be correlated to changes in the relative humidity of the wastewater headspace. These observations support previous results that conducting polymer sensors are sensitive to changes in the sample temperature, humidity and flow rate (Gardner and Bartlett, 1999).

The two distinct clusters seen in Figure 4.8 when performing a PCA on the same raw sewage sensor data (from Figure 4.7) and a similar data set for RO water, shows that the sensor responses of the sample headspace gases (generated from the flow-cell) can be used to separate the two different samples types. However, the observed spread in the separation on either of the principal component axis, shows that the sensor responses are changing, supporting previous observations (Figure 4.7) that the sensor responses are being affected by changes in the relative humidity of the flow-cell generated headspace. In order to minimise the effect that changes in relative

humidity and other environmental variables may have on the sensor response baseline and the magnitude of their responses to a sample, a flow-cell design needs to produce a reproducible headspace gas.

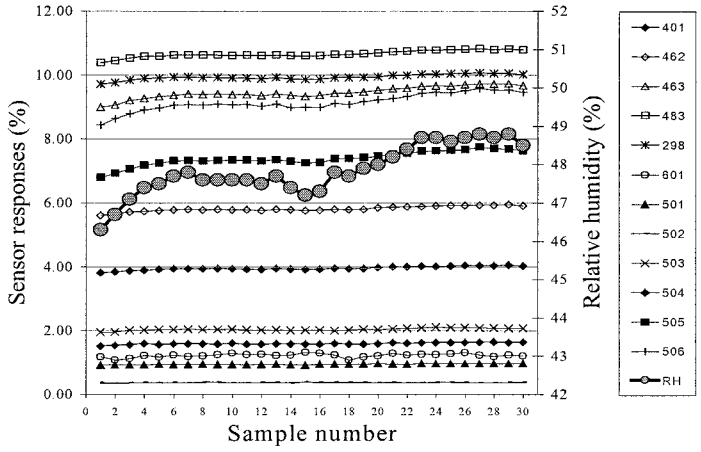


Figure 4.7: Plot of sensor responses (%) and relative humidity (%) for raw sewage from a flow-cell generated headspace

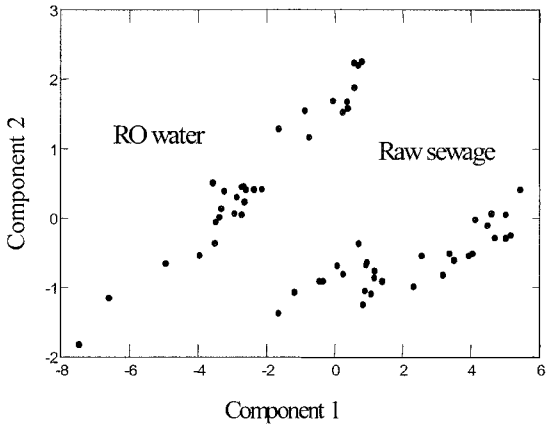


Figure 4.8: Plot of principal components showing separation of raw sewage and RO water from a flow-cell generated headspace.

Development of flow cell sampling system

To minimise environmental changes (i.e. relative humidity) in the formation of a flow-cell generated headspace, a statistical approach was used to evaluate the effect and interactions that different flow-cell parameters (flow-cell temperature, gas flow-rate and sparging porosity) had on producing a stable relative humidity and consequently a more reproducible sensor response profile for a sparged sample. The results are reported in section 4.4.

Figure 4.9 shows an example of a comparison between the 12 chemical sensor responses and relative humidity once a more stable sample headspace generation had been established. The results show that the individual sensor responses remain relatively stable (compared to Figure 4.7) when changes to the relative humidity of the flow-cell generated headspace are minimised. The plot of the first two principal components for raw sewage, final effluent and RO water is shown in Figure 4.10.

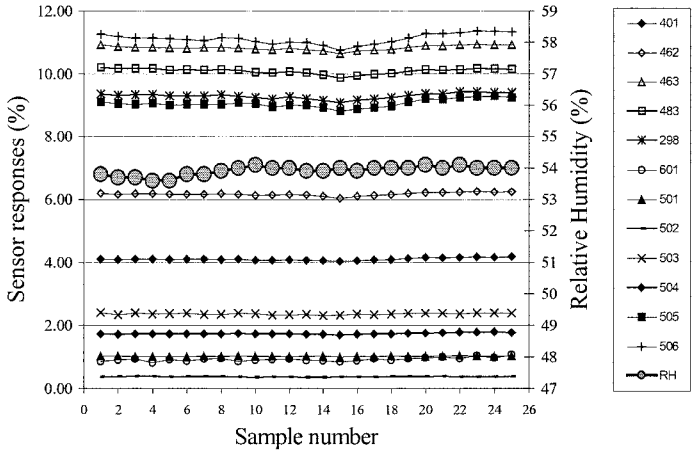


Figure 4.9: Plot of sensor responses (%) and relative humidity (%) for raw sewage from a flow-cell generated headspace

The scatter plot shows that the headspace sensor responses are able to clearly separate the different sample types along the first principal component. This component represents 93 % of the variance in the data set. The spread in the data set along the second principal component (less than 5 % of the variance) suggests that some variation is still present in the headspace generated but that this is only slight. The multiple discriminant analysis (Figure 4.11) of the same data sets (as shown in Figure 4.10) shows how the different sample types can be clearly discriminated into distinct clusters and their group class membership be accurately predicted using a linear statistical approach.

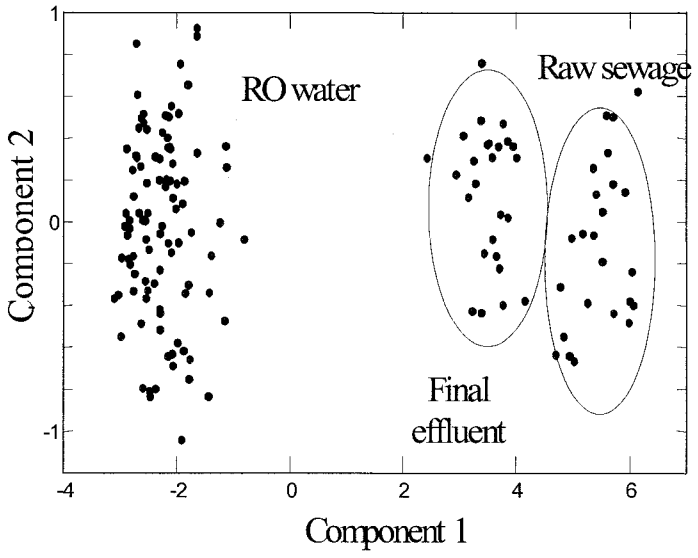


Figure 4.10: Plot of principal components showing separation of raw sewage, final effluent and RO water from a flow-cell generated headspace.

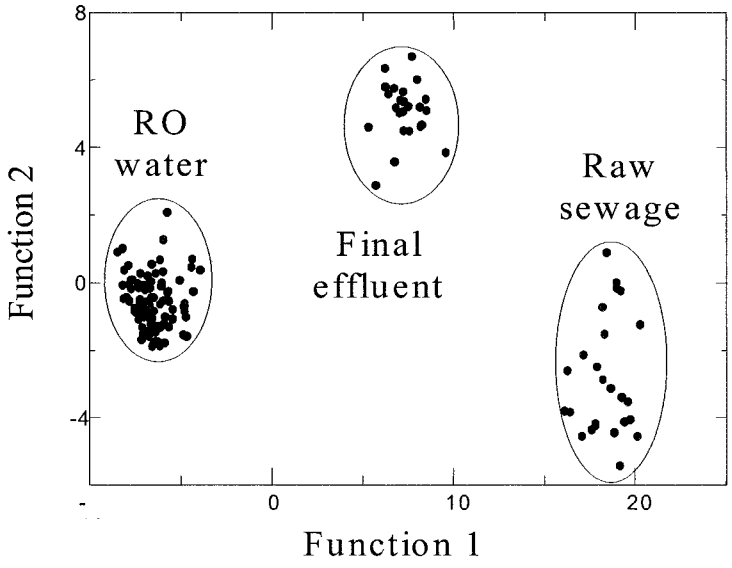


Figure 4.11: Multiple discriminant analysis showing separation of raw sewage, final effluent and RO water from a flow-cell generated headspace

These findings demonstrate that the sampling methodology of using external flow-cell (for generating a sample headspace gas) for subsequent sensor array analysis is reproducible and could provide a sampling technique for real-time monitoring of wastewater quality. To further validate the flow-cell sampling and sensor array analysis technique for assessing wastewater quality in real-time, a dynamic study whereby the wastewater is continuously passed through the flow-cell was evaluated (section 4.3.3)

4.3.3 Semi-static monitoring

Initial analyses were carried out with limited control over the previously described sampling parameters. Variations in the temperature of a quiescent sample (static) had a major impact on the relative humidity (RH) of the headspace produced and consequently affected the sensor responses. More importantly instabilities in RH at a given water temperature significantly reduced reproducibility. The interference effect of temperature and humidity have been previously shown to affect the sensor response patterns of conducting polymer chemoresistors (Gardner and Bartlett, 1995; Gardner and Bartlett, 1999; Ingleby *et al.*, 1999). In principle, these effects can be minimized by careful system design and sample handling. Therefore, semi-static experiments (on-line sampling of a 40 litres batch sample) were carried out in order to study the role of a number of parameters on reproducibility and repeatability over long period of time (few hours to a few days).

Temperature was accurately controlled and an overflow system allowed us to re-circulate large volumes of samples. This "semi-static" approach aimed to limit the eventual effect of continuous sparging of N_2 in a small sample volume over long periods of time. This is because the quality of a given sample is bound to change with time as it is continuously being stripped-off and its more volatile compounds are lost. The liquid sample might eventually become saturated with inert gas.

This experiment provided valuable information for the design and development of an on-line system. Results showed the loss of headspace gases via the overflow tubing, and highlighted the need for a pressure-tight cell for on-line experimentation. Gas flow and the loss of headspace were confirmed to be of paramount importance when inconsistent results (or even no results) were obtained as a consequence of gas leaks during the sampling period. A programmable peristaltic pump was used for dynamic headspace sampling. The system (Figure 4.12) is intended to prevent any loss or uncontrolled dilution of the headspace during the acquisition period and allows us to achieve higher sampling rates.

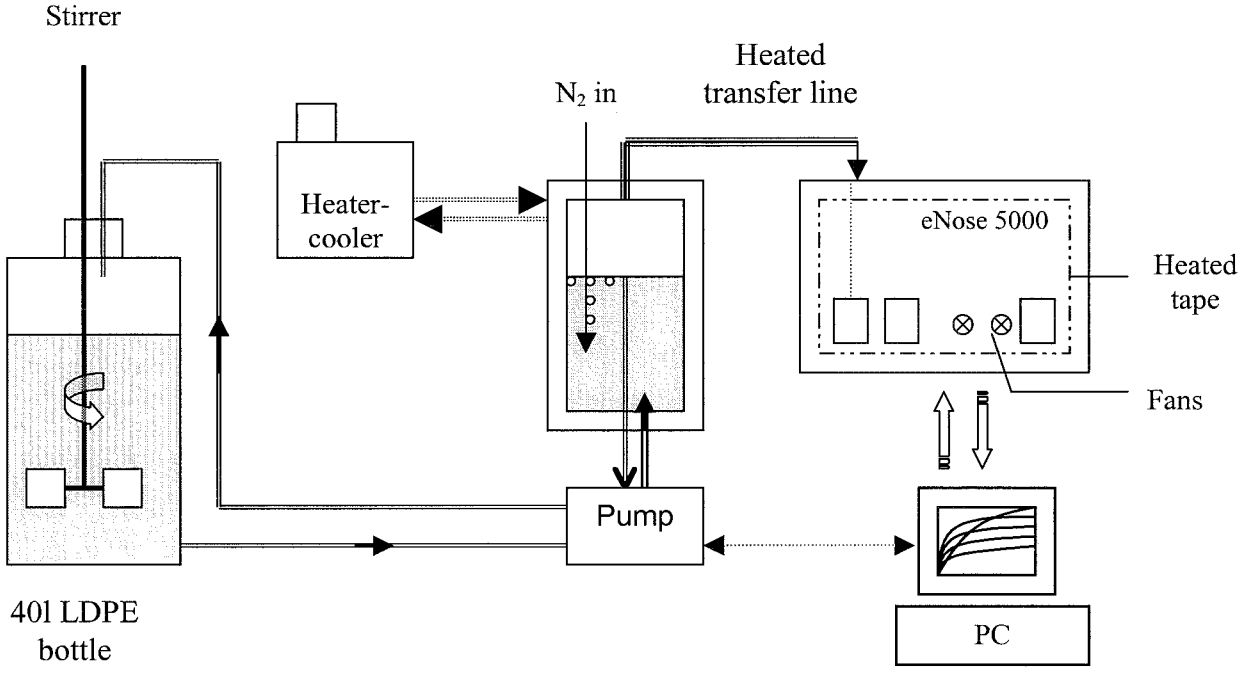


Figure 4.12: Schematic of temperature-controlled monitoring system for continuous analysis of liquid samples.

4.4 MODUS OPERANDI SELECTION FOR ON-LINE APPLICATION

To further optimise our methods and in particular minimise environmental (i.e. RH) changes so as to achieve a more reproducible sensor response, a chemometric approach was adopted. A statistically designed experiment was used to investigate the importance of flow-cell temperature, gas flow rate and sparger porosity on RH stability. Such methods are very appropriate for this type of study because of the number of variables to be considered and the fact that they may interact between each other in a complex manner. Furthermore, a good understanding of the effects of the sampling parameters and their individual and combined effects on RH is essential before any attempt to understand the relationship between the sensor responses and wastewater quality parameters can be made.

4.4.1 Experimental design I

A two level full factorial design was used to determine the optimal sampling parameters and to see what combination of temperature, flow rate and porosity would give us a stable RH, and therefore a reproducible sensor response.

As there were 3 variables that we wished to study at 2 levels (high/low) we needed to carry out $2^3 = 8$ experiments. The design matrix is shown in Table 4.1. Each run was repeated 10 times to allow calculation of the relative standard deviation (RSD%) of RH. Additionally, the experiment was carried out on three different days (1 to 3) to ensure that the results are significant and that further experiments are based on a model which is as robust as possible.

The analysis of the results of the first series of statistically designed experiments on De-ionised water is presented in Table 4.2

Table 4.1: Experimental design matrix

Run	Flow rate of sample across sensors (ml.min)	Porosity of sparger (former B.S. grade)	Temperature of sample (°C)	Relative Humidity
1	75	0	20	?
2	125	0	20	?
3	75	2	20	?
4	125	2	20	?
5	75	2	30	?
6	125	2	30	?
7	75	0	30	?
8	125	0	30	?

4.4.1.1 Main effects

Table 4.2: Experimental conditions and results for 8 R.H experiments on DI-water using the flow cell apparatus.

	Day 1		Day 2		Day 3	
Run	**R.H. (%)**	**RSD (%)**	**R.H. (%)**	**RSD (%)**	**R.H. (%)**	**RSD (%)**
1	42.181	0.27	40.670	0.18	41.679	0.28
2	45.181	0.19	44.347	0.40	45.023	0.59
3	43.431	1.08	42.348	0.60	42.732	1.00
4	46.137	0.36	45.553	0.13	45.480	0.22
5	35.591	1.37	30.033	1.35	29.485	1.56
6	33.405	0.12	32.828	1.09	28.178	0.28
7	31.176	1.21	30.700	0.25	30.503	0.43
8	32.151	0.98	33.055	0.35	32.716	0.16

It appears from Table 4.2 that runs 1 and 4 gave the most reproducible results with an averaged RSD of 0.24% in both cases (run 1 being slightly more consistent). This occured with the temperature parameter at its low level (20°C) and with apparently associated high R.H values. It is not yet clear how the other two parameters (porosity and flow) contribute to this result. To try and understand this, each variable has to be considered separately and their average contribution is found as follows:

The average result is calculated for all of the runs where a given variable is at its low level (simply by summing and dividing by 4). This is then repeated for all the runs with the variable at its high value. The average effect of the variable is then found by calculating the difference of the average contributions at the two different levels. Similarly, this is repeated for all of the variables.The tables of the results for each day are shown in Table 4.3 to 4.5:

Table 4.3: Average contribution and average effect of temperature, gas flow rate and sparger porosity on R.H levels, Day 1.

Variable	Level	Average	Difference
Temperature (°C)	20	44.232	-12.4015
	30	31.831	
Porosity	0	37.672	0.7186
	2	38.391	
Gas flow rate	75	36.845	2.3736
(ml/min)	125	39.218	

Table 4.4: Average contribution and average effect of temperature, gas flow rate and sparger porosity on R.H levels, Day 2.

Variable	Level	Average	Difference
Temperature (°C)	20	43.229	-11.5755
	30	31.654	
Porosity	0	37.192	0.4975
	2	37.690	
Gas flow rate	75	35.938	3.0077
(ml/min)	125	38.946	

Table 4.5: Average contribution and average effect of temperature, gas flow rate and sparger porosity on R.H levels, Day 3.

Variable	Level	Average	Difference
Temperature (°C)	20	43.728	-13.5084
	30	30.220	
Porosity	0	37.480	-1.0117
	2	36.468	
Gas flow rate	75	36.099	1.7501
(ml/min)	125	37.849	

By ranking the average effects (according to their magnitude), the relative contribution of each factor becomes evident (Table 4.6):

Table 4.6: Ranked average effects of temperature, gas flow rate and sparger porosities (relative to RH)

	Day 1	Day 2	Day 3
Temperature	-12.4015	-11.5755	-13.5084
Flow rate	2.3736	3.0077	1.7501
Porosity	0.7186	0.4975	-1.0117

Although the numerical values differ slightly, the results in Table 4.6 clearly show the same relative contribution over the three days, with temperature clearly being the most important factor (by an order of magnitude) controlling the relative humidity and thus affecting the sensor's response. The importance of the flow rate is also suggested with an effect being more than twice that of porosity.

These observations however, remain of secondary importance until the system is stabilised effectively. Therefore, the same method was also used in order to investigate the contribution of the variables on the stability of the system by looking at the RSD (%) (Table 4.7).

Table 4.7: Ranked average effects of temperature, gas flow rate and sparger porosities on the relative standard deviation of RH (%)

Day 1	Day2	Day 3
-0.5685	0.4944	- 0.1565
Flow rate	Porosity	Flow rate
0.446	0.4248	0.1275
Temp.	Temp.	Porosity
0.0689	-0.0948	0.0434
Porosity	Flow rate	Temp.

The variability of these results from one day to another makes it more difficult to reach a conclusion on the contribution of a particular factor to give stable RH values. This is certainly accentuated by the insufficient number of replicates to allow the calculation of a statistically significant RSD. Flow rate and temperature may however, contribute to some extent to a better stability of the system (with their higher and lower levels respectively). Further analysis to see if and how the variables interact with each other is needed.

4.4.1.2 Interactions

Looking at the main effects only showed the unambiguous effect of both temperature and gas flow rate on Relative Humidity levels. However, the contribution of porosity was not so clear and that of each of the three variables to the stability of the results (RSD) remained difficult to establish. Therefore, it is intended in this section to make use of a simple mathematical model to detect and measure any possible interaction between the variables. This model contains a number of coefficients calculated by the method of "contrast patterns" and generated from the design matrix of the experiment (N. Collins, experimental design notes). The results are shown in Table 4.8.

The coefficients are labelled as A with a subscript to define them (e.g. A_1 is the main effect coefficient of variable V1 and A_{12} the interaction coefficient of V1 and V2). They are obtained by summing the results, column by column using the above matrix. For example:

$$A_1(\text{day1}) = (-42.181 + 45.181 - 43.431 + 46.137 - 35.591 + 33.405 - 31.176 + 32.151) \times 1/8 = 0.5619$$

The Final coefficient A0 is the average of all results.

The results of the calculation using the data from Table 4.8 are presented in Table 4.9. It appears that the main effect A1 (temperature) and A3 (flow), and the interaction between them (A13) are the most important. The interaction between all 3 variables is very small. This value will be used as the standard error thereafter.

Table 4.8: Contrast pattern matrix generated for the 8 R.H experiments on DI-water. Where V1 = Gas flow rate, V2 = porosity and V3 = Temperature.

Run	V1	V2	V3	V1V2	V1V3	V2V3	V1V2V3	RH Day 1	RH Day 2	RH Day 3
1	-1	-1	-1	+1	+1	+1	-1	42.181	40.670	41.679
2	+1	-1	-1	-1	-1	+1	+1	45.181	44.347	45.023
3	-1	+1	-1	-1	+1	-1	+1	43.431	42.348	42.732
4	+1	+1	-1	+1	-1	-1	-1	46.137	45.553	45.480
5	-1	+1	+1	-1	-1	+1	-1	35.591	30.033	29.485
6	+1	+1	+1	+1	+1	+1	+1	33.405	32.828	28.178
7	-1	-1	+1	+1	-1	-1	+1	31.176	30.700	30.503
8	+1	-1	+1	-1	+1	-1	-1	32.151	33.055	32.716

Table 4.9: Main effect coefficients and interaction coefficients generated for the 3x8 RH experiments on DI-water.

Coefficient	Day 1	Day 2	Day 3	Average
A1	0.5619	1.504	0.8748	0.9802
A2	0.9844	0.2488	-0.5058	0.2425
A3	-5.5759	-5.7878	-6.754	-6.0392
A12	-0.4319	-0.004	-0.5145	-0.3168
A13	-0.8646	-0.2165	-0.6483	-0.5765
A23	0.4329	-0.4723	-0.8833	-0.3076
A123	0.0166	0.114	-0.3655	-0.0783
A0	38.6566	37.4418	36.9745	37.6910

The RH for any experiment can then be predicted by using the following equation :

$$RH = A_0 + A1.V1 + A_2.V2 + A_3.V3 + A_{12}.V1.V2 + A_{13}.V1.V3 + A_{23}.V2.V3 + A_{123}.\ V1.V2.V3$$

Where V1, V2 and V3 are the coded values for their respective variables which can be determined as follows:

V1 = 0.04 x flow rate ($ml.min^{-1}$) – 4
V2 = porosity (former B.S. grade) – 1
V3 = 0.2 x Temperature (^oC) – 5

For example, assuming a linear relationship applies, the estimated RH value for a gas flow rate of 250 $ml.min^{-1}$ (coded value 6), a porosity of 0 (coded value –1) and a sample temperature of 20°C (coded value –1) would be 53.4 % (using the averaged coefficients).

Experimental results (5 replicates were carried out) gave an average RH value of 46.9 %. Similarly, at 125 ml.min, 20 °C and with a porosity 3 (30 runs) the predicted value of 44.6% did not match with the experimental average of 47.1%. This reflects

the linear approximation and shows that one should not attempt to predict values outside the measured experimental area as the model will not be valid.

4.4.1.3 Significance

By using the average coefficient from days 1,2 and 3 (Table 4.9), we can determine the significance of the effects and interactions. Assuming that a 3 way interaction (V1V2V3) is extremely improbable, we take the average value for the A123 coefficient as the standard error. The t values are then calculated by dividing each coefficient value by the standard error and are give in Table 4.10.

Table 4.10: t values calculated from main effect coefficients.

Coefficient	Coefficient value	t value
A1	0.9802	12.52
A2	0.2425	3.10
A3	-6.0392	77.13
A12	-0.3168	4.05
A13	-0.5765	7.36
A23	-0.3076	3.93
A123	-0.0783	n\a

T values greater than two are statistically highly significant. In this case, the main effects of temperature (A3) and flow rate (A1) are the most important, followed by the interactions between them (A13).

Because minimising the RH variations to ensure steady and consistent sensor response remains the prime purpose of this experiment, supplementary analysis of the data was carried out by Marconi Applied Technologies. The results, desirability plots and comments are attached in Appendix A. These results showed that the system is

non linear and that strong interactions between flow rate and porosity occur. Consequently, further experiments are required before any conclusion can be reached.

4.4.2 Experimental design II

When analyzing the results of the designs discussed above, a prediction equation for the dependent variable can be fitted to the observed responses, and values can then be computed at any combination of levels of the predictor variables. Using Statistica's "design of experiment" functions it is possible to inspect the predicted values for the dependent variable at different combinations of the predictor variables and search for the levels of the dependent variable that produce the most desirable responses (most stable RH) on the dependent variable.

The relationship between predicted responses and the desirability of responses is called the desirability function. Profiling the desirability of responses involves specifying the desirability function for the dependent variable, by assigning predicted values a score ranging from 0 (very undesirable) to 1 (very desirable). Surface and contour plots of desirability show the effect of pairs of IV's on RH.

The results from the first series of experiments suggested that there may be a "saddle" developing as seen in Appendix A. This reflected the limits to the linear approximation of a complex system. Therefore a second experiment was performed in a different area based on these observations.

The plots of desirability obtained for this second series of experiments (Figure 4.13) show where a developing region of stability can be observed. The new data implies that the model is improving as the system is entering a region where the linear approximation holds better. In accordance with the previous batch of experiments, gas flow rate was confirmed to be the most significant factor, followed by porosity and the flow/porosity interaction to affect RH stability. For this particular experimental setup and laboratory condition, a good reproducibility could be achieved under the following conditions: Flow: 170ml/min; Porosity: 0 and

Temperature: 25.5°C. Based on these results, the selected methodologies were applied to the system for further continuous monitoring studies.

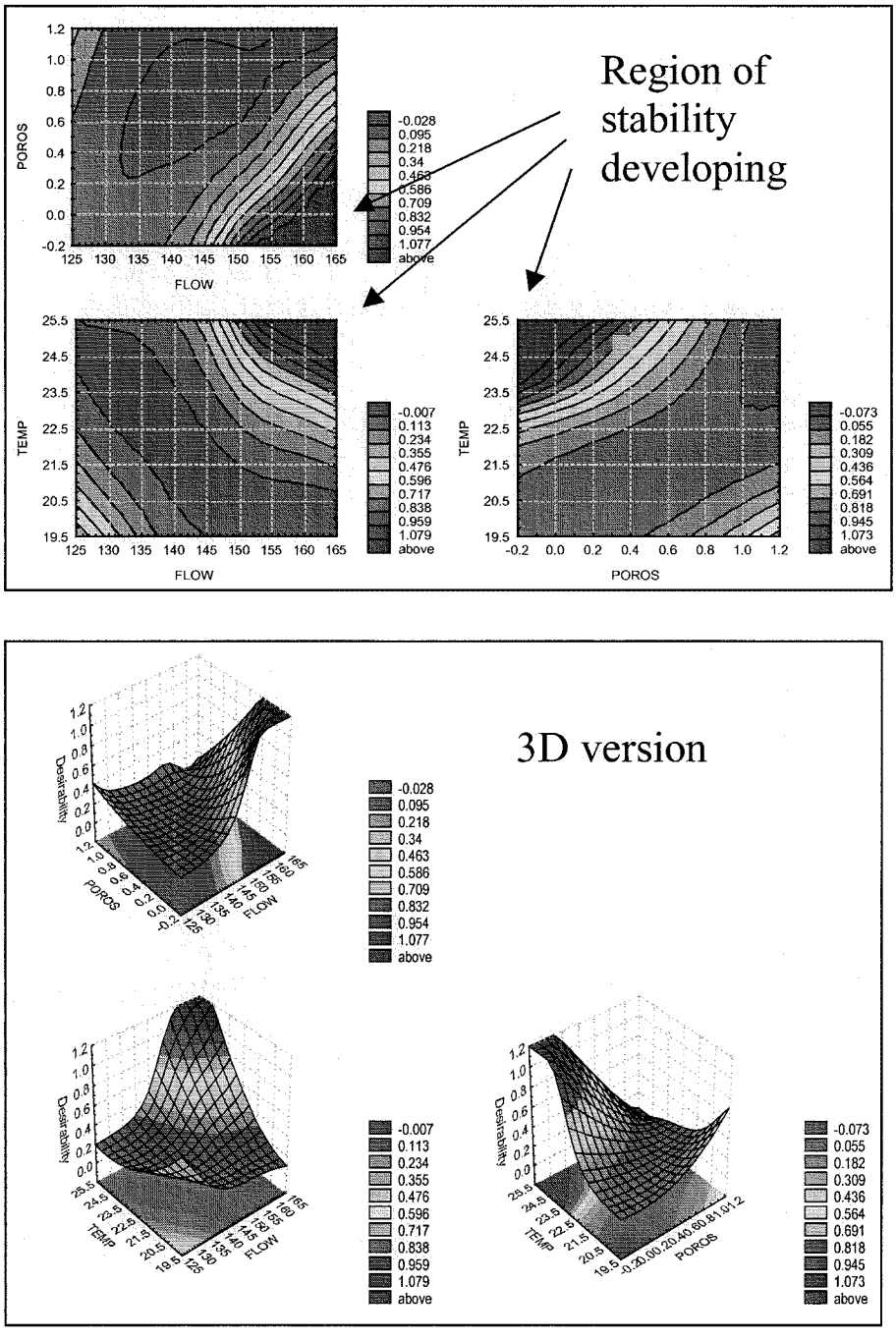


Figure 4.13: Experimental design II: desirability plots showing area of improved RH stability.

4.5 ON-LINE MEASUREMENT OF WASTEWATER ORGANIC LOAD IN A CONTROLLED ENVIRONMENT

4.5.1 Introduction

Results from preliminary studies have shown that our CP-based sensor array system can be used to differentiate between different types of wastewater by measuring the headspace gas generated from liquid samples. Using the methodologies developed in the previous sections, the flow cell sampling system and selected analysis protocol were further validated in a dynamic study. The experiment uses a fully automated system (as described in section 4.3.3), whereby the wastewater is continuously pumped through the flow cell.

The aim of this study is to assess the ability of the technique to provide both qualitative and quantitative information on the quality of wastewater before field experiments could be carried out. Results showing the relationship between the data generated and organic load measurements were presented at the IWA-Instrumentation, Control and Automation (ICA) conference in Malmo, Sweden, in June 2001.

4.5.2 Methods

4.5.2.1 Sample collection

Wastewater samples were collected from the primary tank (settled raw sewage) at the Cranfield University sewage treatment works. For laboratory experiments approximately 80l. was collected at 9.00 am at 1 week intervals. Dilutions were carried out with reverse osmosis (RO) water in 40l. PTFE containers. Solutions were kept at 25°C and constantly homogenised during the course of the experiment. All

samples were analysed for BOD and COD concentration (Standards Methods, APHA/AWWA/WPCF, 1995).

4.5.2.2 Headspace sampling and sensor array analysis

The headspace gas was generated using the temperature-controlled flow-cell by sparging the liquid samples with filtered zero grade nitrogen as described in section 4.3.3. A blank was carried out with RO water over 48 hours. On-line analysis was performed by continuously re-circulating the samples from the 40l containers through the flow-cell using a peristaltic pump. Each 40l sample was repeatedly analysed for periods of over 24 hours.

Sensor array analysis was done using the temperature controlled eNose 5000. The carrier gas temperature could not be controlled and was subject to diurnal variations. Analysis of the headspace gas was carried out every 5 minutes as follows: 40 sec pre-purge of sampling line; 1 min acquisition; 3 min 20 sec sensor cleanup.

4.5.2.3 Data analysis

Principal Component Analysis, Multiple Discriminant Analysis and Multiple Linear Regression are used to reduce the dimensionality of the data. The relationship between the sensor output and the samples is investigated and correlated using the statistical package Statistica.

4.5.3 Discrimination

Figure 4.14 illustrates how multivariate analysis techniques can be used to reduce the dimensionality of the sensor's response and discriminate between different samples. This two-dimensional scatter plot of on-line data, shows the good separation between on-line raw sewage, diluted raw sewage and RO water using only eight of the twelve sensors (type 298; 401; 462; 501; 502; 503; 504 and 506).

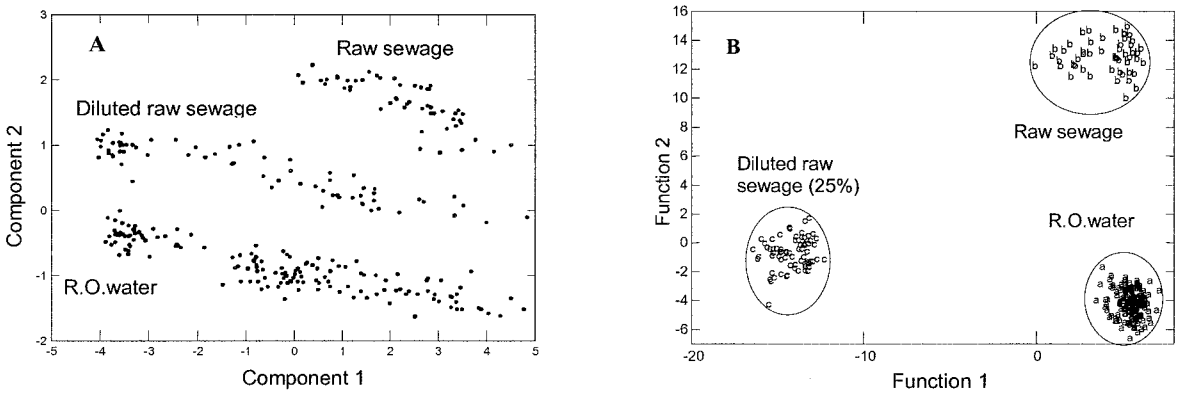


Figure 4.14: Plots of principal components (A) and multiple discriminant (B) showing the separation and classification of reverse osmosis (R.O) water, Raw sewage and diluted raw sewage (25%).

In Figure 4.14 (A), the samples are separated according to their concentration mainly on the y-axis which accounts for 26% of the variance. The drift observed along the component 1 axis (70% of the variance), indicates that variations in the relative humidity (RH) of the sample still remain the major factor affecting the separation with increasing RH values from left to right on the x-axis. A comparison with observations from previous studies (Section 4.3.2), using the same experimental apparatus in a static way, pinpoints the difficulty in trying to control experimental parameters when continuously sampling with a dynamic system. However, the strong linear relationship observed in the data indicates that these variations may be accounted for when processing the sensor responses. For instance, temperature (and/or RH) could be used as an input to artificial neural networks which afford

some parametric compensation. The results from the MDA (Figure 4.14(b) strongly suggest that a linear discrimination model can be built that appears relatively unaffected by the drift in the data resulting from unstable RH conditions (component 1 axis, Figure 4.14(a)).

4.5.4 Multiple Linear Regression

Multiple linear regression (MLR), showed that the relationship between the 12 conducting polymer sensors' responses and the BOD, COD and TOC for raw sewage, diluted raw sewage and RO water is linear (R^2=0.98). Applying the coefficients calculated for each sensor from a training dataset to unknown data (Figure 4.15), demonstrates that it is possible to predict wastewater parameters using simple statistical regression tools.

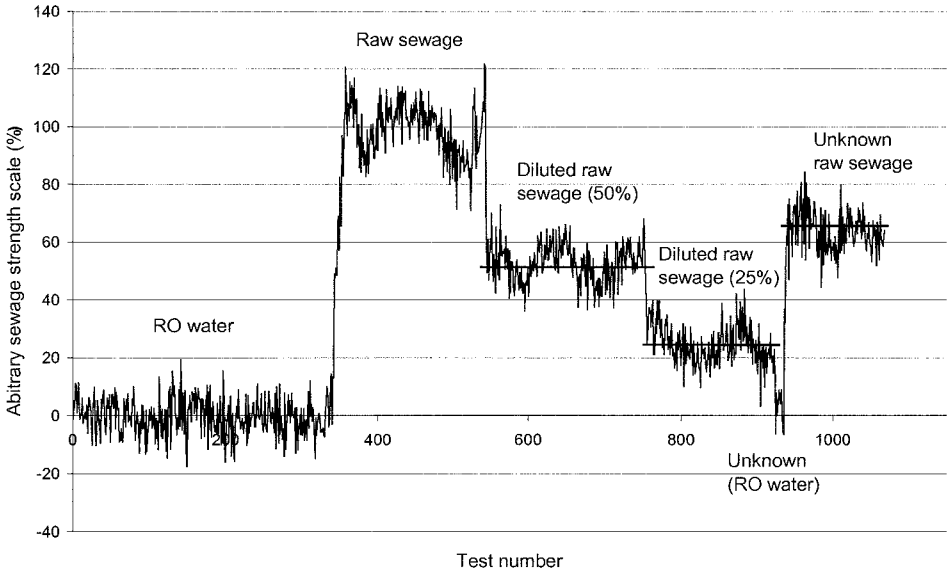


Figure 4.15: On-line prediction of wastewater concentration using MLR. R.O water (0%), raw sewage (100%) and diluted raw sewage (50% & 25 %) were continuously analysed for over 24 hours each (approximately 200 points) and used to calibrate the model.

The plot shows the predicted wastewater concentration expressed as a percentage or dilution factor of the original undiluted wastewater sample. Alternatively, predicted BOD or COD values can be used in the same way to compare the samples. The periodic variations seen in Figure 4.15 have been shown to match the recorded changes in ambient temperature during the course of the experiment. These diurnal changes directly affect the temperature of the carrier gas used and as a result cause variations in the relative humidity of the generated headspace. Despite these fluctuations the relative concentration (BOD: 128 mg/l, COD: 349mg/L) of unknown wastewater samples could still be predicted as 75% and 77% (respectively) of the undiluted training sample concentration (BOD: 172 mg/l, COD: 451mg/l) two days after the training period. Correlation between predicted COD and measured COD was good (R=0.96) with an average prediction error of 19%. These results seem encouraging when compared to the accepted 20-30% error with traditional BOD_5 measurements. However we must keep in mind that COD levels for this experiment were constant and the differences in concentration quite pronounced (0; 112.5; 225 and 450 mg/l), which is generally not the case in real applications.

4.5.5 Drift and limitations

When trying to apply the same coefficients to unknown data between three days and a week after the training period, a gradual drift in the predicted values can be observed (Figure 4.16). Sensor drift is a well-known and documented occurrence in sensor array analysis (Albert *et al.*, 2000) which can be monitored by looking at the sensor baselines prior to analysis. In this study however, it may not be the only factor responsible for the observed decrease in the predicted values, as the drift appears to be more pronounced for raw sewage samples than for clean water. Indeed the relatively short duration of the experiment does not allow for a significant sensor drift to occur and the observed decline could therefore suggest a gradual change in the olfactory quality of the re-circulated wastewater samples with time.

In an attempt to assess the robustness of this approach and understand the observed relationship, training was carried out over the same time period using only a fraction

(50% then 25%) of the original dataset. The strong multiple correlation that prevailed (R^2~ 0.96 in both cases) demonstrates that a strong relationship exists between the sensor array output and the organic strength of the sample. Calculation of the organic content of the training samples could be carried out with some degree of accuracy (Figure 4.17).

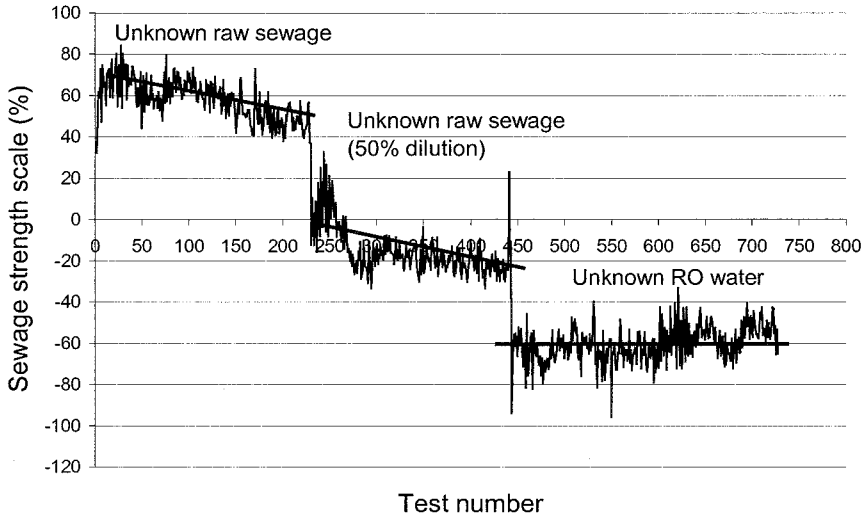


Figure 4.16: Drift observed for on-line prediction of wastewater concentration using MLR between 3 days and a week after the training period

However, it was clear that the model rapidly deteriorated and these coefficients could not be applied to predict values of unknown samples even soon after the calibration. These findings support previous observations by Stuetz *et al.* (1999b, c) that a better correlation exists between the BOD and the sensor response over shorter periods of time. Unknown samples were collected one week after the raw sewage samples used for the calibration of the system. During this time interval the characteristic odour of the influent from the wastewater treatment plant is very likely to have changed independently of its organic content. For instance, a higher proportion of grey water (i.e. from baths, showers..) would make it more difficult to establish a linear correlation with the sensor array's output. As illustrated in Figure 4.16 and 4.17 this

becomes particularly true when limited amount of data is available and when trying to apply the coefficient over longer periods of time.

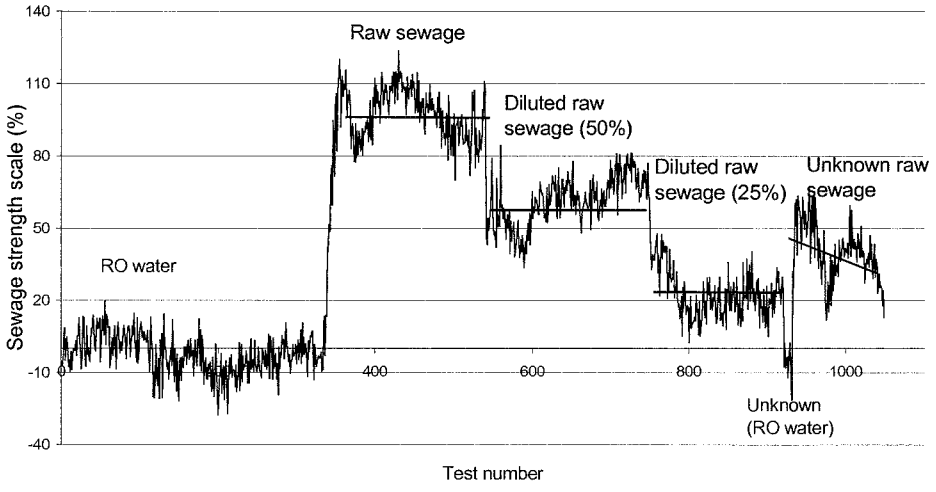


Figure 4.17: On-line prediction of wastewater concentration using MLR and coefficients obtained from 50 % of the original calibration dataset.

4.6 SUMMARY

- A flow cell was built for continuous sampling of wastewater and coupled to an array of 12 CP sensors for analysis.
- The liquid sample is sparged with N_2 to produce the headspace gas sample.
- The effects of temperature and RH on the sensors were important. Variations in RH can be minimised through precise control of temperature, gas flow rate and porosity.
- The system was successfully used in a controlled environment to differentiate between different types and concentrations of wastewater using PCA and MLR.
- The effect of time on the prediction of organic load was important. The linear model could not values for samples collected more than a week after training.

Chapter 5: IN-SITU APPLICATION OF A SENSOR ARRAY AT A WASTEWATER TREATMENT PLANT

CHAPTER 5: IN-SITU APPLICATION OF A SENSOR ARRAY AT A WASTEWATER TREATMENT PLANT

5.1 INTRODUCTION

This chapter considers the progression from early laboratory investigations to full scale trials at an operating wastewater treatment plant. In order to evaluate the application of a non-specific sensor array to monitoring changes in wastewater quality under real conditions, a new system (PROSAT, Marconi Applied Technologies) was installed at the university's own sewage treatment plant. Hardware and software modifications to the system were done to allow real time measurement of headspace gases generated from the flow cell. In addition to testing the instrument in a harsh environment, continuous sampling and the generation of large databases allows appropriate data analysis protocols to be developed. In Chapter 6 and Chapter 7 respectively, traditional multivariate statistical techniques and neural network will be assessed for real-time monitoring application using these databases. In this chapter, typical results obtained from a 12 months field study are presented, showing the effect of a range of natural and accidental changes in wastewater quality on the sensor responses. The system was first implemented on the 4th of December 2000 for evaluation and was eventually left to run continuously from the 23rd of January 2001 until the 5th of December 2001. Selected results have been published in Bourgeois and Stuetz, (2002) and Bourgeois *et al.*, (2002)

5.2 MATERIAL AND METHODS

5.2.1 Sensor array analysis

The system used in this field study is in principle very similar to the one used in laboratory experiments. The PROSAT sensor array module (Marconi Applied Technologies) shown in Figure 5.1 is designed for on-line application in a wide range of industrial processes. The instrument consists of an array of 8 conducting polymers in a temperature controlled sensor chamber (35°C) and has a built in PC for controls and data acquisitions as well as a network connection for data transfer. Unit operation and configuration is set up by menu-driven software. A screen allows continuous display of the response profiles.

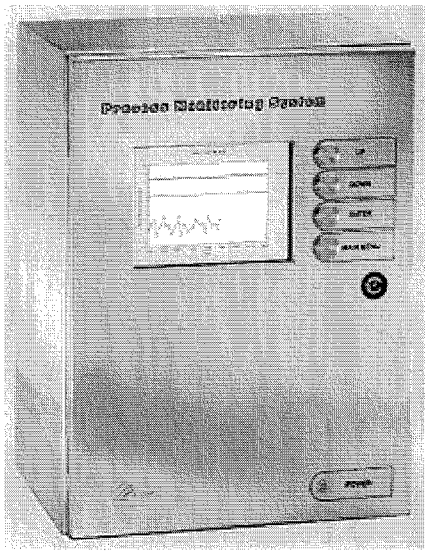


Figure 5.1: PROSAT sensor array system used for continuous monitoring at the wastewater treatment plant.

The choice of sensors for the new system was based on experience gained from the laboratory studies with the eNose 5000 system. Principal component analysis was used to eliminate sensors which did not significantly improve the discrimination between wastewater samples. The following 8 sensors were therefore used in the Prosat module due to their wide selectivity and and good stability over time: Type 298; 401; 462; 501; 502; 503; 504 and 506. Recommendations on the stability and longevity of the sensors were given by the manufacturer. The discriminatory properties of these sensors to different types of wastewater is shown in Section 4.5.3 (Figure 4.14).

A schematic and photograph of the new experimental set-up (pre-sample vessel, flow-cell, Prosat) are shown in Figure 5.2 and Figure 5.3 respectively. Wastewater from the primary settlement tank (BOD: 50 – 200 mg/l and COD: 150 – 400 mg/l) is re-circulated through a ring main in the pilot hall from which the wastewater influent was drawn into the pre-sample vessel at a rate of 20 l/min and mixed continuously. The wastewater sample was then pumped (Watson-Marlow, UK) through the flow-cell at a rate of 200 ml/min, except during sensor acquisition when the flow-cell is sparged with zero-grade N_2 gas to generate a sample headspace for subsequent sensor array analysis. Analysis of the headspace gas was carried out every 5 minutes and consisted of the following sampling protocol: 40 sec pre-purge of headspace sampling line, 1 min sensor acquisition and a 3 min 20 sec sensor de-purge. Sensor responses at the end of the acquisition stage (1min) were used to represent the fingerprint of that headspace sample.

Plots of sensor responses from the 8 conducting polymers are used to monitor for changes in sensor resistance, as different sensors are sensitive towards different compound types. The plots were then used to give a visual indication of any rapid change in the quality of the wastewater. To reduce the dimensions of the multi-dimensional sensor array data, pattern recognition techniques have been previously employed (Hobbs *et al.*, 1995; Persaud *et al.*, 1996; Stuetz *et al.*, 1998) so that relationships between the observations can be explored using one or two dimensions. In this study, principal component analysis (PCA) was used to find hidden relationships between the samples by maximising the variance between two or more

groups of objects with respect to several variables simultaneously (Gardner and Hines, 1997). PCA was performed using the statistical package STATISTICA.

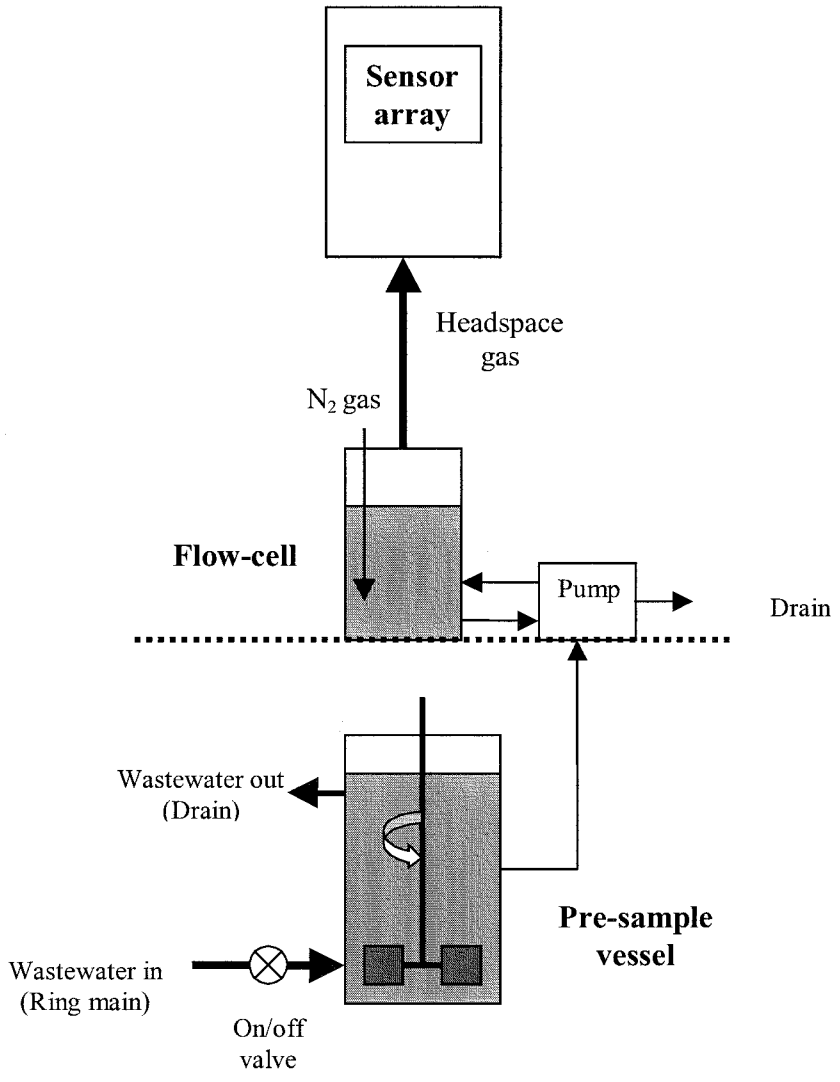


Figure 5.2: Schematic of the on-line monitoring system (Cranfield University Pilot Hall), showing pre-sample vessel, flow-cell and sensor array module.

5.2.2 Total organic carbon and biological oxygen demand

The wastewater from the primary settlement tank was simultaneously analysed for TOC (every 10-30 min) using an on-line TOC-4100 analyser (Shimadzu, UK) as well as for BOD (every 10-30 min) using a RACOD (USF,UK). Both instruments were connected to the ring main as seen in Figure 5.3.

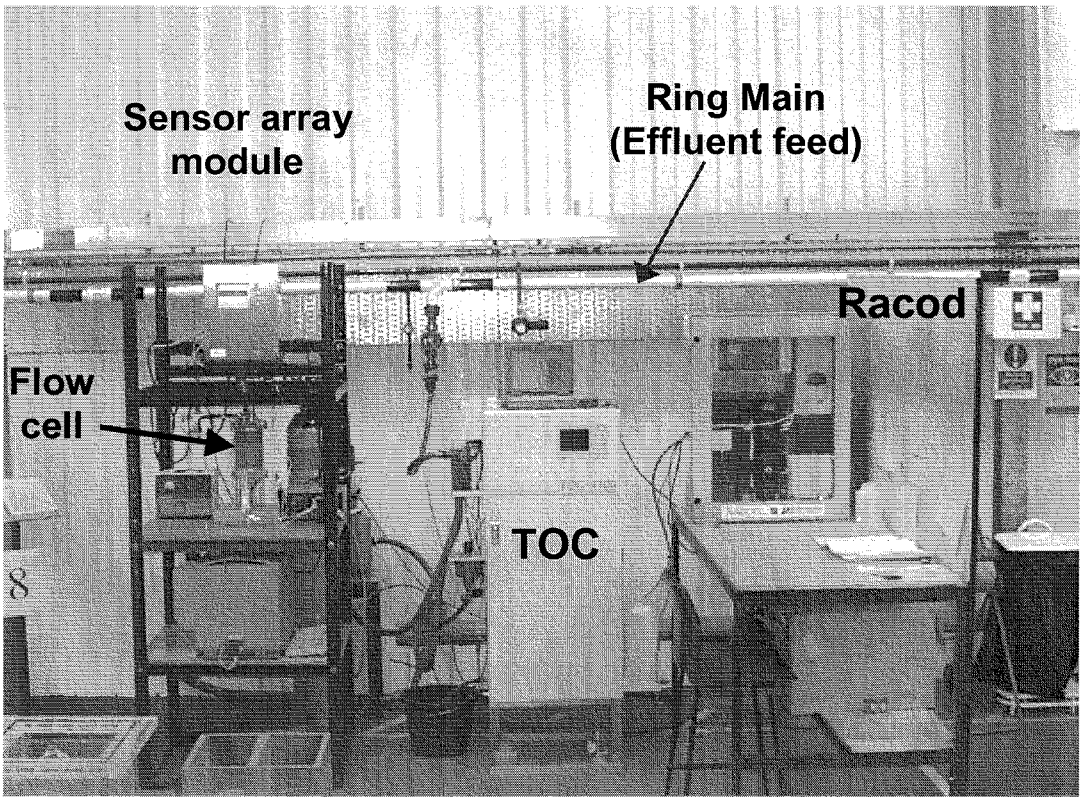


Figure 5.3: Photograph of the on-line monitoring instrumentation at Cranfield University sewage works. From left to right: sensor array module, Shimadzu TOC-4100 and RACOD analyser.

Additional spot samples were taken on a regular basis for standard BOD_5 and COD analysis in the laboratory. In a particular study, BOD_5 and COD measurements were carried out 3-4 times a day, everyday for 3 weeks by a visiting student (Maxime

Porterie, May 2001). It became clear from this work that not enough data could be generated for an adequate data analysis using the traditional manual sampling and standard laboratory analysis approach. Furthermore, concerns arose from the overall quality of the data that could be produced (over 30% RSD for BOD_5), despite our effort to carry out relatively large number of replicates. Consequently, correlation of BOD_5 and COD values was judged unrealistic. Instead, data obtained manually was mainly used to ensure that the TOC and Racod instruments were properly calibrated. Thus we relied on the quality of the data provided by these instruments to give us a better understanding of the underlying relationship between the sensor array data and the wastewater organic content.

The 12-month continuous wastewater monitoring study using the Prosat resulted in the generation of over 100,000 acquisition points. In addition to the sensor responses (60 points per acquisition per sensor), the data files contain measurements of gas flow rate, wastewater temperature, sensor module temperature and humidity and temperature of the gas phase. A summary of on-line measurements for this study is given in Table 5.1 below.

Table 5.1: Summary of on-line quality measurements of primary settled effluent (ring main)

Variable	Frequency	Data files location
8 CP sensor responses	Every 5 min (for 1min)	Prosat hard drive (csv file)
Headspace RH and temperature	Every 5 min (for 1min)	Prosat hard drive (csv file)
Sensors temperature	Every 5 min (for 1min)	Prosat hard drive (csv file)
Gas flow rate	Every 5 min (for 1min)	Prosat hard drive (csv file)
Wastewater temperature	Every 5 min (for 1min)	Prosat hard drive (csv file)
TOC	Every 10-30min	Laptop (xls file)
BOD (Racod)	Every 10-30 min	Gemini Data logger

5.3 RESULTS

5.3.1 Diurnal variations

The analysis of this sensor array data has shown that the individual sensor responses do differ in magnitude and that the response patterns were found to follow a similar diurnal variation. An example of a typical sensor response profile for 8 sensors over a 5-day period (Figure 5.4) shows that the same changes are observed at the same time every day. The increase in the sensor response occurs at approximately 9:00 am, which corresponds with the increase in organic load at the treatment works due to the increased activity at the University.

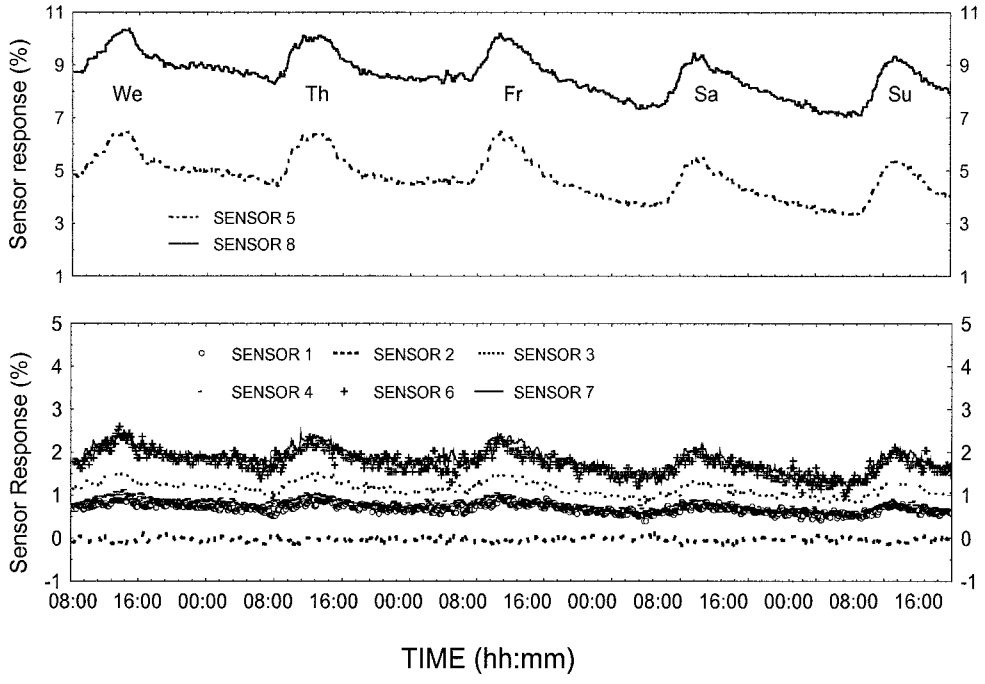


Figure 5.4: Plot of sensor responses over a 5-day period (24/01/01 to 28/01/01), showing diurnal variations in the headspace of the wastewater influent from the ring main.

The sensor profiles normally peak between 13:00 and 14:00, then gradually decline in the afternoon to reach a minimum early in the morning. A close observation of the sensor array profiles also shows a lower response on Saturday and Sunday. This corresponds to the reduced activity on campus during the weekend and the resulting drop in wastewater organic load. As shown in Figure 5.5 the trends in the sensor responses show some similarity to those of the recorded TOC.

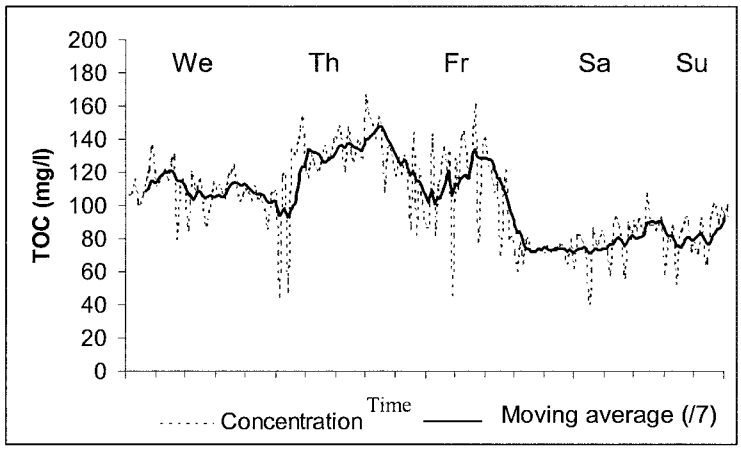


Figure 5.5: Plot of measured wastewater TOC concentration from the ring main for the period of 24/01/01 to 28/01/01 and 7-point moving average.

The same sensor array data from the 5-days continuous measurement period was plotted over a 24-hour axis (Figure 5.6). This plot provides a graphical representation of the dynamic ranges for each sensor and gives an indication as to the normal wastewater quality at a particular time of the day. These observations give a good indication as to the repeatability of the observed diurnal pattern. The approach of superimposing data over specific time periods could be applied to predicting data points that are outside predetermined limits. These outliers could be the result of either potential intermittent discharges in wastewater or other abnormalities in the wastewater processes.

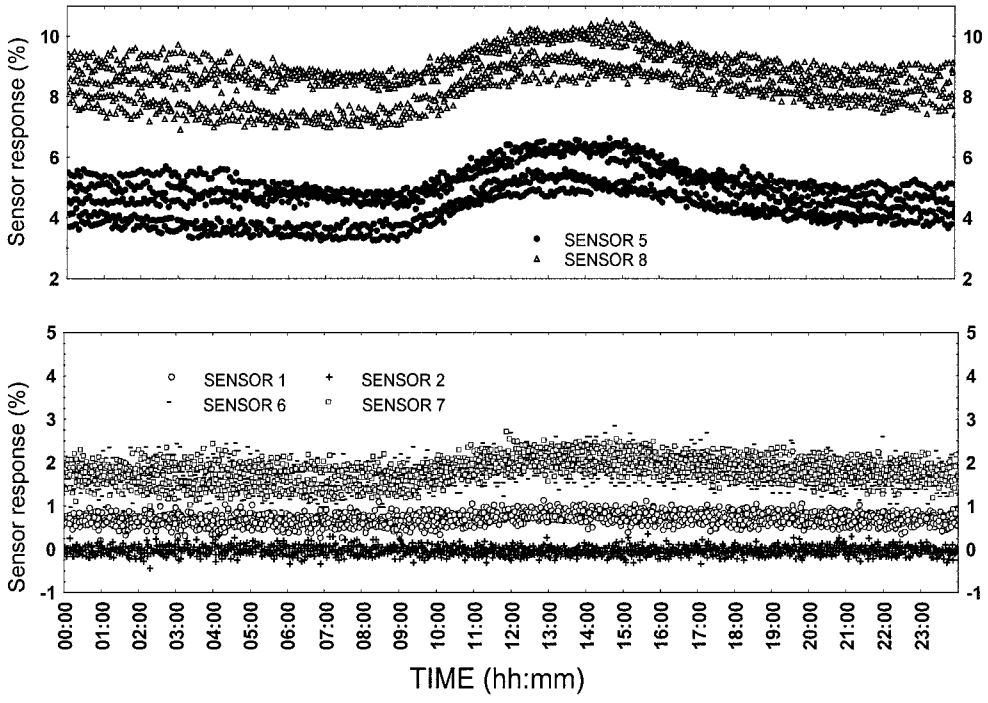


Figure 5.6: Plot of sensor responses for a 5-day period on a 24-hour scale, showing the repeatability of the diurnal patterns at any time of the day.

5.3.2 Multiple Linear Regression

Using MLR to study the relation between TOC and the sensor array data collected in the field supported the results presented in the previous section (4.5.4). Although the correlation (R^2=0.36 over 2 weeks), between the constantly changing odour profile of wastwater and the organic content (TOC), was poor compared to the one observed in the laboratory controlled environment, the model allowed a resonably good prediction (R^2=0.65) over a one week period. Also, as reported earlier, some of the selected sensors respond strongly to changes in RH.

Although temperature and RH could not be realistically controlled with the field-based system, their variations were accurately recorded. The alternative approach to a physical control is to use this information for parametric compensation as suggested in Gardner and Bartlett (1999). To see if the effect of humidity could be artificially reduced, the change in resistance as a function of RH levels was quantified for each sensor, using data obtained with RO water only (800 acquisitions

on 2 different days). It was observed that for most sensors the response to changes in RH was quasi-linear. This relationship could be expressed by an equation of the form:

$$\Delta R/R_{sensor(i)} = a_0 + a_1 * RH$$

As illustrated in Figure 5.7, the regressions (slope equations) for the most sensitive sensors (R^2 >0.70), were then used to subtract the fraction of resistance change thought to be only due to RH from the wastewater data. The idea is that the remaining variations in the sensor response are, in principle, more readily associated with changes in the wastewater quality. However it must be noted that there are many uncertainties with this rudimentary approach (sensor drift, effect of temperature and gas flow rate..) which was therefore only intended as an exploratory way to give some indication as to the feasibility of parametric compensation when processing the data. A more appropriate and detailed statistical study that includes temperature and RH in the data analysis process is presented in Chapters 6 and 7. The results summerised in Table 5.2 illustrate how controlling the relative humidity (both experimentally and numerically) as well as reducing the time intervals improve the linear correlations.

The correlation values are mainly given for TOC as the global organic load parameter of choice. In the case of the laboratory experiments only COD and BOD_5 measurements were made. However, this does not affect the comparison of the R^2 values presented here, since a constant TOC/COD and TOC/BOD_5 ratio was observed during the course of the experiment. Typically a COD to BOD ratio close to 3 is observed for the settled primary effluent at the university's wastewater treatment plant (P. Leclech, R Ormesher, personal communication).

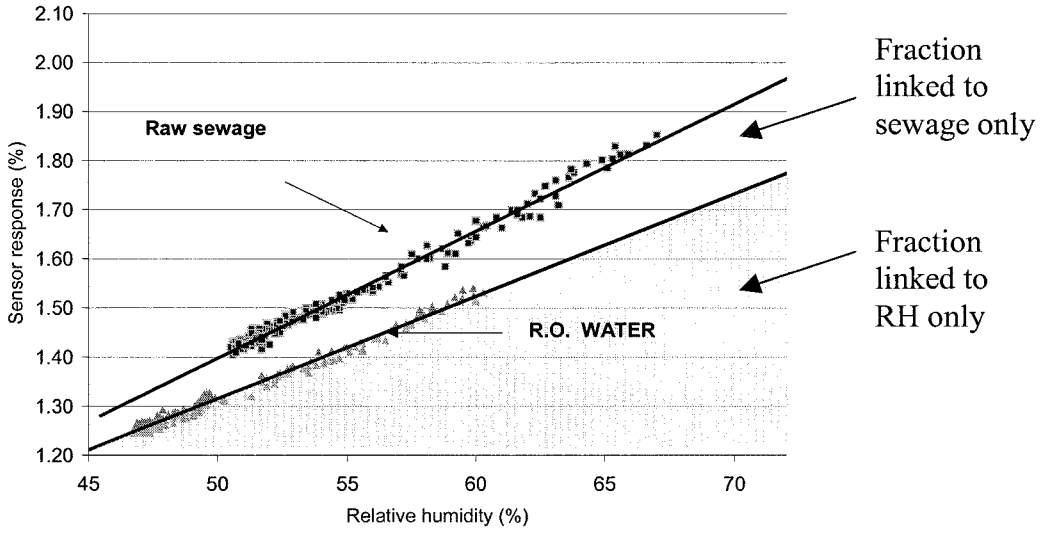


Figure 5.7: Response of sensor 501 Vs RH when exposed to R.O. water and raw sewage (COD= 450 mg/l). Over 200 replicates were carried out in both cases

Table 5.2: Effect of reduced time intervals and RH parametric compensation on multiple correlation (R^2). Note: BOD (COD) vs. TOC ratios remained constant during these experiments

Experimental conditions and time interval	Multiple correlation (R^2)
Lab. temp controlled, (1 week)	0.98 (BOD and COD)
On-line pilot hall (2 weeks)	0.36 (TOC)
On-line pilot hall (2 weeks, RH-compensated)	0.62 (TOC)
On-line, pilot hall (1 week)	0.65 (TOC)
On-line pilot hall (1 week, RH-compensated)	0.69 (TOC)

5.3.3 The effect of rainfall and operating anomalies

In Figure 5.8 the diurnal variations in the headspace of the wastewater can be observed as previously. On the first three days (03/01/01 to 03/03/01) however, continuous heavy rain is affecting the profiles. The lower sensor response matches lower TOC readings for that period and is representative of the typical dilution of the influent caused by important rainfall observed on many occasions. After the rain, a return to the usual more pronounced night/day changes can be seen (days 4, 5 and 6). This effect of rain on the quality of the wastewater arriving at the treatment plant is representative of the nature of the university's wastewater collection network. The combined sewer system has a relatively large catchment area (airfield, car parks, roads, roofs..) for a plant of this size. As a result storm events and prolonged rainfall generally cause a strong and rapid dilution of the domestic influent. In Figure 5.9, the same effect of rain on the sensor profiles can be seen from the 4th day onwards.

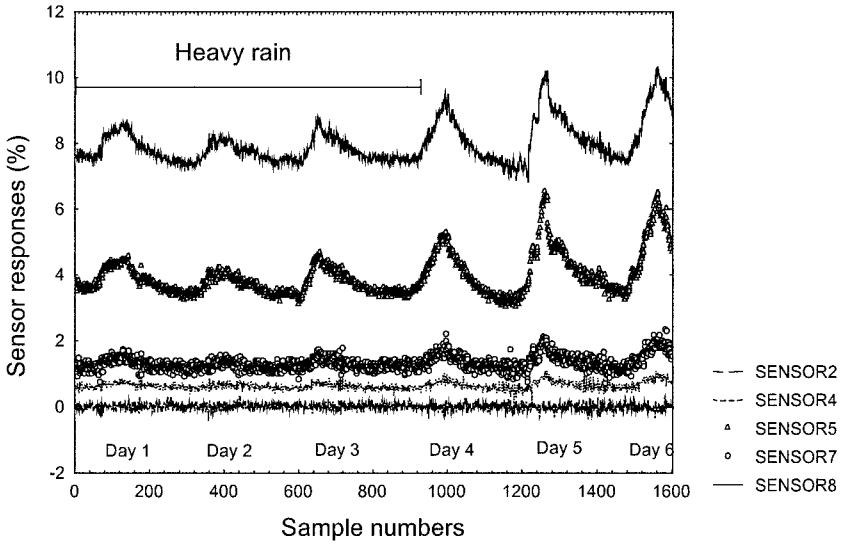


Figure 5.8: Plot of 8 CP sensor responses for 6 days (01/03/01 to 06/03/01) showing diurnal variations in wastewater quality with dilution effect of heavy rain.

Additionally, the severe environmental conditions caused one of the feeding pumps from the primary settlement tank at the sewage works to fail (tripped mains) on the 29th of January. The pump was then manually stopped for maintenance the following day. On both occasions a distinctive change in the profiles can be observed. Although it is not yet clear how such incidents can affect the sensor response, a possible cause could be the change in the suspended solids content of the wastewater as a result of the reduced velocity. Furthermore, the absence of feed from the primary settling tank will rapidly cause the water to stagnate in the pipes. Unusually high COD levels in the ring main water have previously been reported as a result of such pump failures (B. Lodge and R. Ormesher, personal communication). This type of incident are quite unusual and no study of the physical or biological processes that may be responsible for the change in wastewater quality was available.

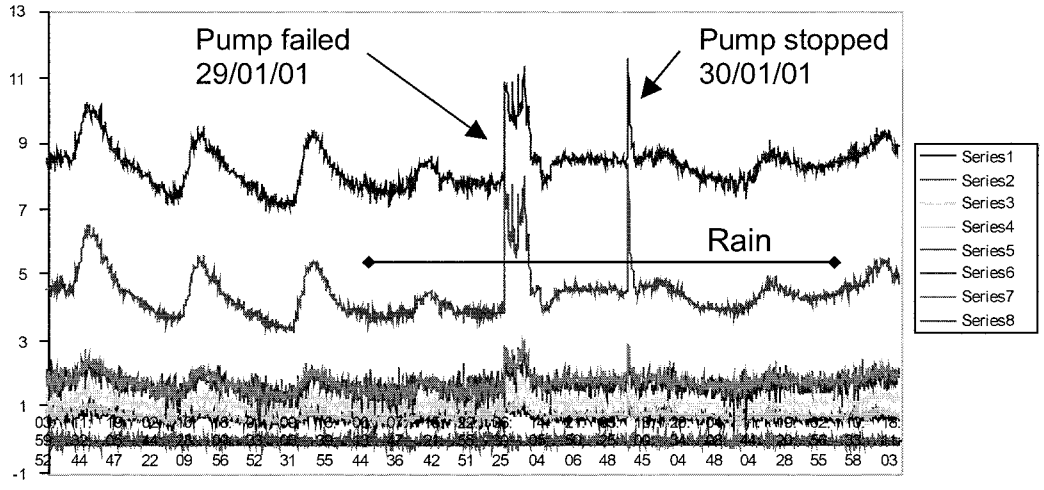


Figure 5.9: Plot of sensor responses (25/01/01 to 01/02/01) showing the effect of heavy rain and the detection of operating anomalies.

5.4 DATA SCREENING AND PREPARATION FOR MULTIVARIATE ANALYSIS

The nature of the data collected with on-line instruments is often quite varied, but unusual or corrupt data is detrimental to the quality and validity of the analysis. Although multivariate statistics may appear as a magic means to extract significant results from a myriad of numbers, it cannot perform miracles (the phrase "Garbage in – Garbage out" is commonly used). The trick with multivariate analysis is to select reliable and valid measurements, choose the appropriate tool, use it correctly, and know how to interpret the output (Tabacknick & Fidell, 1996). In this section the emphasis is put on the selection of continuous data files that can be used to investigate the relationship which may exist between the sensor array profiles and the wastewater organic contents. This consisted of 2 major steps: Data screening, and the association (alignment) of sensor array data with corresponding TOC and Racod values.

5.4.1 Data Screening

On-line measurements can be affected by an almost infinite source of disturbances. These may be related for instance to the hostile environment in which instruments have to be located, power cuts, poor maintenance or calibration, environmental disturbances and other system failures. Indeed quality data is essential to achieving good results, however despite efforts to keep disturbances to a minimum, a majority of data points could not be used.

Figure 5.11 shows the 4 major types of corrupt data that affect almost every measurements. These are: noise, missing values, outliers and drift. This is illustrated in Figure 5.12, which shows the TOC data obtained over a 6-month period.

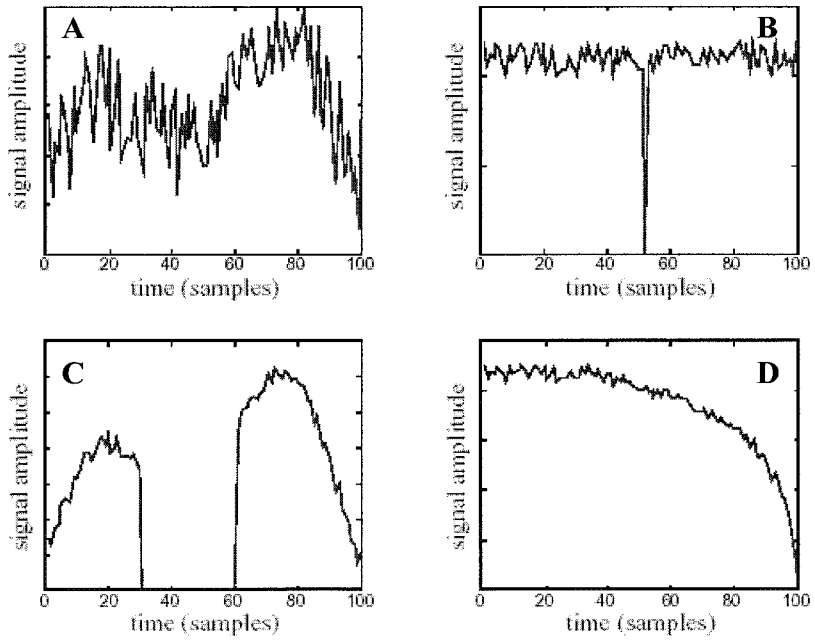


Figure 5.10: Different types of measuring faults: Noisy data (A); Outliers (B); Missing data (C) and Drift (D) (from Rosen, 1998)

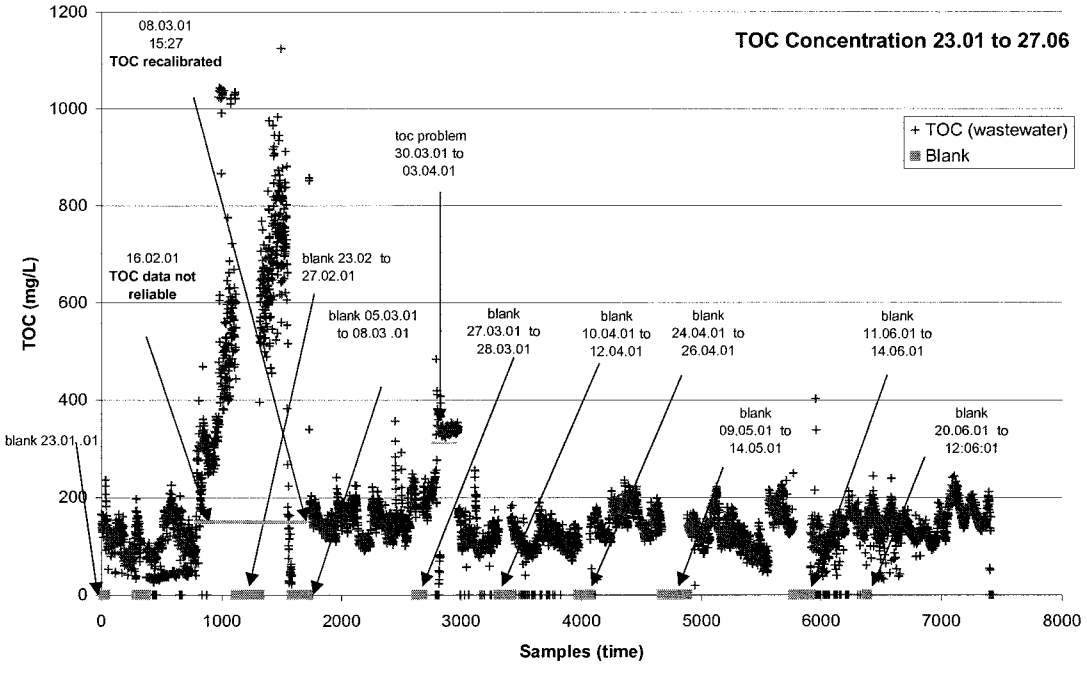


Figure 5.11: On-line TOC data (23.01.01 to 06.07.01) showing the presence of noise, outliers, severe drift and missing data.

5.4.1.1 Noise

Noise is a common problem in almost every measurement system that is difficult to avoid. This disturbance in the measured signal can originate from electromagnetic interferences, inappropriate system setup, poor maintenance or the proximity of heavy electrical equipment. Irregularities within the process can also be considered as noise. In wastewater monitoring, process noise may be caused by inhomogeneous mixing, random variation of, for instance air bubbles and other non-measurable causes. Noise reduction can be achieved by applying analogues or digital filters (Rosen, 1998). Filtering is a relatively large discipline that cannot be covered in detail in this study. Further information on digital filtering can be found in Proakis & Memodakis (1992). Noise was observed in both the TOC and sensor array data and attempts to address this issue have been made that will be further discussed in subsequent chapters.

5.4.1.2 Outliers

In order to ensure consistency of the data, outliers can be removed or replaced. In this study, cases corresponding to outliers such as pollution incidents, pump failures, gas failure etc., were removed from the dataset since they are not representative of the normal conditions that prevail at the wastewater treatment works, and may distort the result of the analysis. Although the amount of data available is not a major issue here, the effect of removing data points on the dynamic properties of the dataset must be carefully considered before carrying out the analysis. With regard to the TOC data, any values outside the range 0-300 mg/l (i.e. normal operating conditions) were discarded.

5.4.1.3 Missing values

Similarly to outliers, missing values can be a serious problem as they distort the dynamic properties of the signal. Especially when multivariate analysis is considered, this can make the whole dataset difficult to use. In particular long periods of missing values can be problematic when continuous and reliable data over long periods of time are required. Missing data from our monitoring systems were mostly due to power cuts, instruments being switched off for maintenance, fouling, blockages or leaks as well as some of the instruments being used for other experiments (particularly in the second half of the 12-month monitoring period).

5.4.1.4 Drift

Because of the discontinuities in the dataset that resulted from missing values and the removal of outliers, the original data was split into smaller subsets of continuous data (see Section 5.4.2). Since the final data files did not exceed 2 weeks in length, drift in the sensor array data was not considered a major issue. However, a fault in the TOC 4100 caused an important drift after a relatively short period of time (Feb. 2001, Figure 5.12). The instrument was repaired and recalibrated on the 8th March and all data prior to this date was discarded as a measure of precaution.

5.4.2 Preparation of data files

Prosat data was acquired over a 12-month period from Jan. 01 to Dec. 01. Although the system was continuously monitoring wastewater throughout most of this period, only data from Jan. 01 to July 01 was used for multivariate analysis. In the second half of the 12-month period, additional experiments were regularly carried out, which in retrospect, split the periods of continuous monitoring into smaller sets that are not ideal for multivariate analysis. Similarly there have been a few times where either the

TOC 4100 or the Racod instruments were not available, as they were temporarily needed for other research.

With the difficulty to generate continuous datasets for each individual instrument also comes the problem of matching (or aligning) corresponding data from these systems into common files that can be directly used for analysis. Differences in sampling frequency further reduced the size of the datasets. After converting the respective files into a compatible format and adjusting the TOC and Racod times to the exact times recorded by the Prosat (used as reference time), the relevant cases were selected (via a Visual Basic macro) using a rule of the form:

"if Datediff (S, TOCtime, PROSATtime) $\leq$ 150 then Select-case=True"

In other words, data was matched to +/– 2 min 30 sec for a Prosat sampling frequency of 1 acquisition every 5 minutes (TOC and Racod measurement frequencies were in the range 10-30 minutes). The number of remaining cases and dates for the final TOC-Prosat and Racod-Prosat files are shown in Table 5.3 and 5.4 respectively. Data set number 12 in the TOC-Prosat file includes all the TOC-Prosat data and will be used for non-dynamic studies.

As Tabachnick and Fidell (1996) rightly pointed out, careful consideration of these issues is time consuming and sometimes tedious. It is common to spend many days in careful examination of the data prior to running the main analysis, but consideration and resolution of these issues before multivariate analysis, is fundamental to an honest analysis of the data. Statistics can often be seen as misleading, Benjamin Disraeli (in Haynes, 2000) once said: "There are three kinds of lies: lies, damned lies and statistics". However, in the majority of cases, this is the consequence of poor pre-examination of the data which can wrongly affect the perception of the results. Therefore particular attention was given to the preparation of the data for this study.

Table 5.3: Continuous TOC-Prosat data subsets

Subset reference number	Date	Number of cases
1	08.03 to 15.03	306
2	28.03 to 31.03	126
3	03.04 to 07.04	174
4	12.04 to 17.04	238
5	20.04 to 22.04	135
6	26.04 to 09.05	571
7	14.05 to 20.05	202
8	22.05 to 24.05	67
9	27.05 to 02.06	244
10	14.06 to 29.06	760
11	04.07 to 06.07	146
12	08.03 to 06.07	2969

Table 5.4: Continuous Racod-Prosat data subsets

Subset reference number	Date	Number of cases
1	30.01 to 23.02	3248
2	27.02 to 13.03	1609
3	28.03 to 07.04	1814
4	12.04 to 17.04	772
5	20.04 to 23.04	507

5.5 DESCRIPTIVE STATISTICS

Multivariate techniques, by their nature, identify complex relationships that are difficult to represent simply. As a result, the tendency is to accept the results without the typical examination one normally undertakes in univariate and bivariate analysis. Such shortcuts can lead to disaster and multivariate analysis requires an even more rigorous examination of the data because of the influence of outliers, violation of assumptions, and missing data can be compounded across several variables to have substantial effects (Hair *et al.*, 1998). Most of the outliers and missing values have been dealt with as described in Section 5.4. However errors and anomalies in the data that are more difficult to detect may remain. Consequently some preliminary descriptive analyses of the data are absolutely necessary if the results of multivariate analyses are to be believed. Such diagnostic measures allow for a better understanding of the data and the basic relationships that may exist. In this section we particularly examine the repartition of the data since a normal distribution is a common assumption for many analysis techniques. Summary tables of the sensor response, RH and concentration's averages, minima, maxima and standard deviation for both the TOC-Prosat and Racod-Prosat files (all subsets) are given in Appendix B.

5.5.1 Normality

As already pointed out, one of the most fundamental assumptions in multivariate analysis is normality, and screening the data obtained in the filed for normality is an important step. Although normality of the variables is not always required and some deviations from this assumption may be acceptable, results are generally better if all variables are normally distributed. In this section normality of the variables is assessed by both graphical and statistical methods.

The graphical tests include the frequency histograms and the more reliable normal probability plots as shown in Figure 5.12 to Figure 5.22, where the distribution of the 8 Sensor responses, RH, TOC and BOD are compared with the line of expected normal distribution. From these graphs it appears that only the TOC data and the responses of Sensor 2 are normally distributed. All the other variables have a slightly positive skewness (i.e. general pile-up of cases to the left). This deviation in the data might reflect environmental changes during the course of the experiment. Indeed it is highly probable that seasonal variations in temperature will affect RH and the sensor responses (as discussed in section 2.1.3) but also the BOD measurement which rely on biological process (section 2.3.4). Alternatively, instrumental drift could also play a significant role on the distribution. These hypotheses are further investigated in the next section with the use of bivariate scatterplots.

In addition to examining the normal probability plots, statistical tests have also been used to assess the normality. Three tests available in Statistica were used on the individual subsets of data (subsets 1 to 11). These are the Shapiro-Wilks' W test, the Kolmogorov-Smirnov test and the Lilliefors test. The results support the observations from the visual examination of the normal probability plots discussed above and are presented in Appendix B.

Usually, transformation of the data to accommodate the non-normality is recommended before carrying out multivariate analysis. Transformations may improve the analysis and may have the further advantage of reducing the impact of outliers (Hair *et al.,* 1998; Tabachnick & Fidell, 1996). However these transformations often make the transformed variable harder to interpret. In our case, because each individual data subset has its own specifications, this would also make it more difficult to find a global model. The relatively large sample size allows us not to transform the data to normal. Instead, the severity of the violation of the assumption of normality (if any) can be tested after or during the analysis with a range of multivariate normality tests.

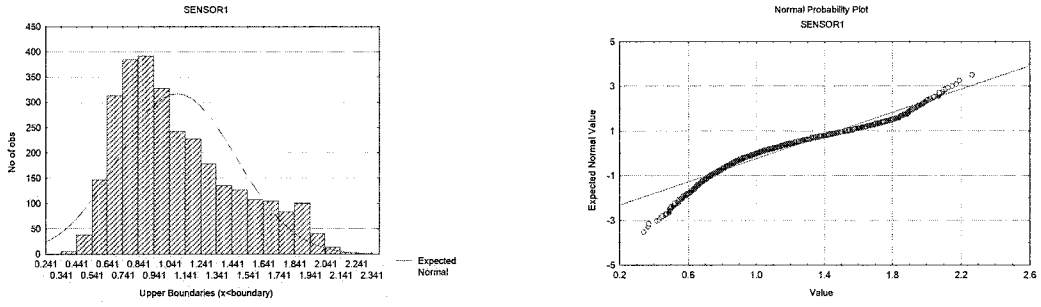


Figure 5.12: Frequency histogram and normal probability plot of sensor 1 data

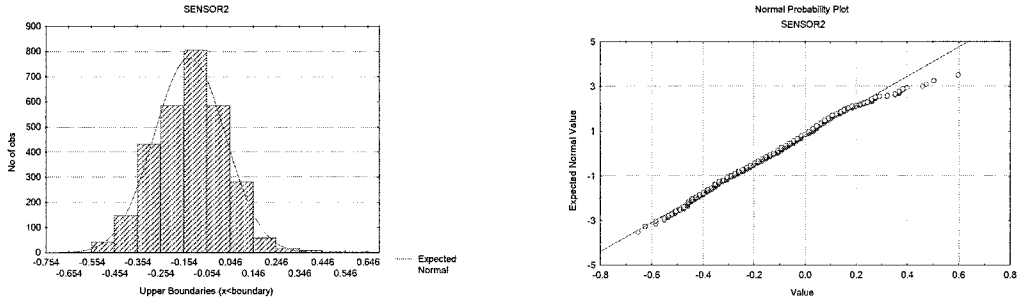


Figure 5.13: Frequency histogram and normal probability plot of sensor 2 data

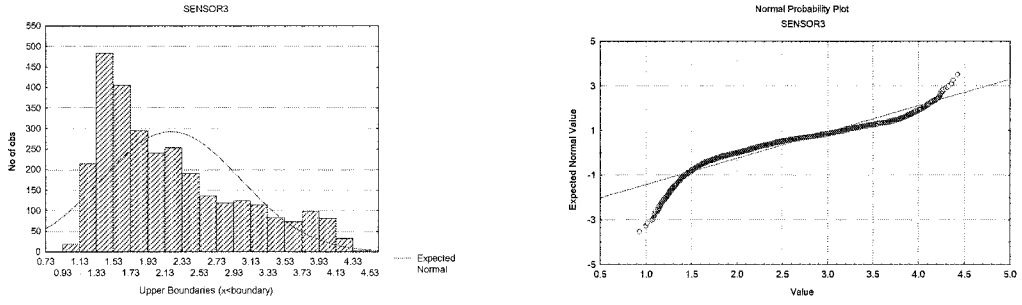


Figure 5.14: Frequency histogram and normal probability plot of sensor 3 data

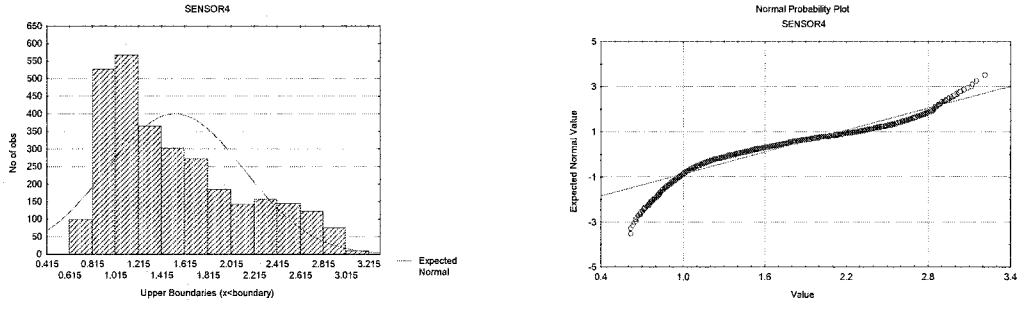


Figure 5.15: Frequency histogram and normal probability plot of sensor 4 data

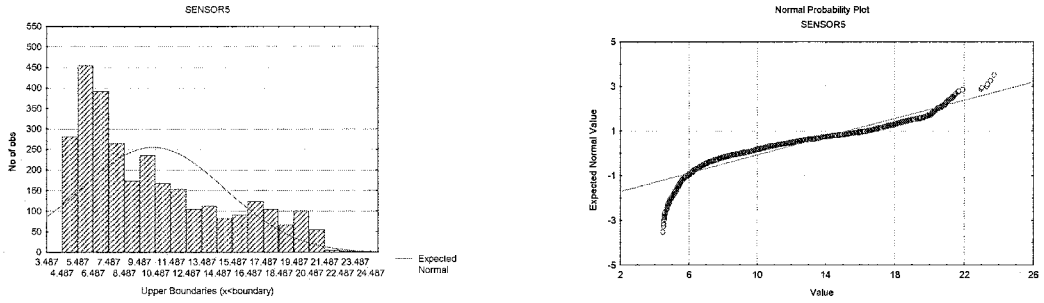


Figure 5.16: Frequency histogram and normal probability plot of sensor 5 data

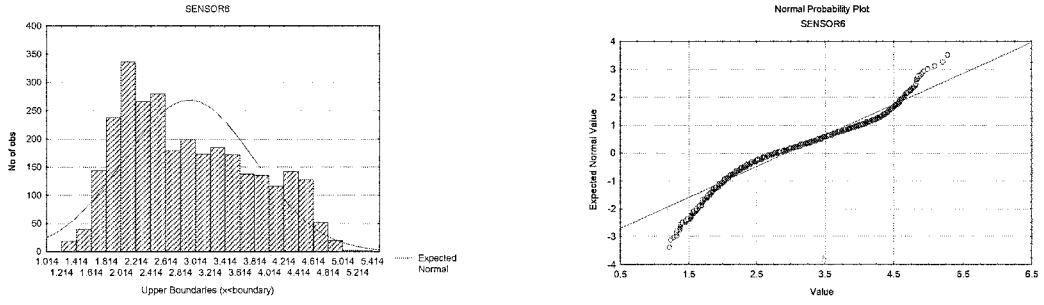


Figure 5.17: Frequency histogram and normal probability plot of sensor 6 data

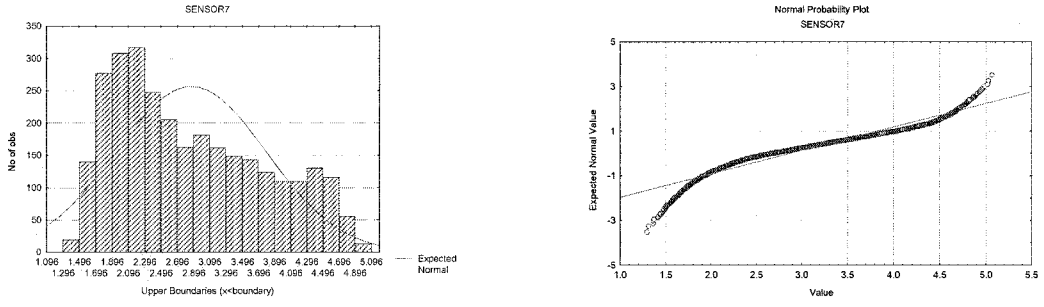


Figure 5.18: Frequency histogram and normal probability plot of sensor 7 data

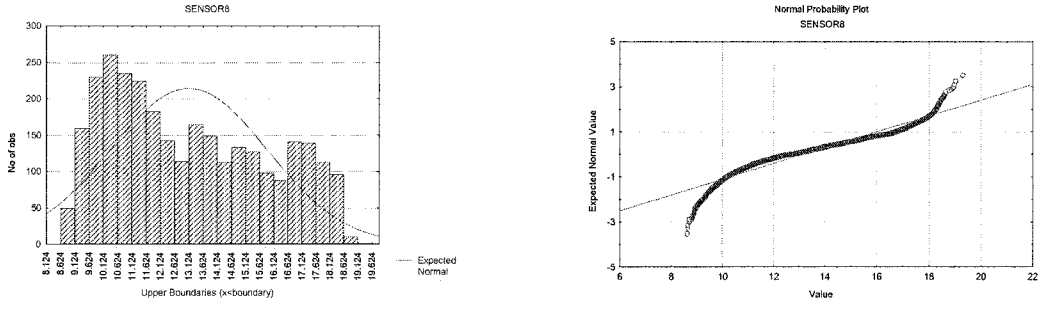


Figure 5.19: Frequency histogram and normal probability plot of sensor 8 data

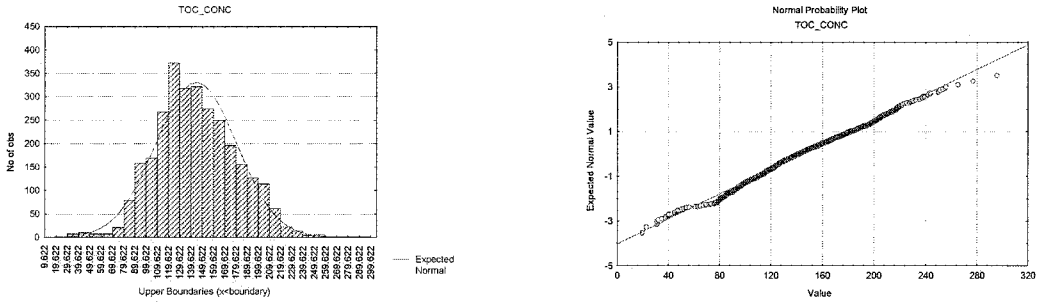


Figure 5.20: Frequency histogram and normal probability plot of TOC data

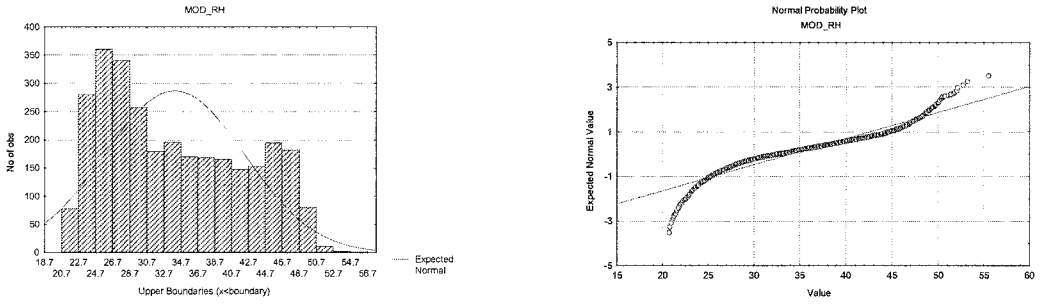


Figure 5.21: Frequency histogram and normal probability plot of RH data

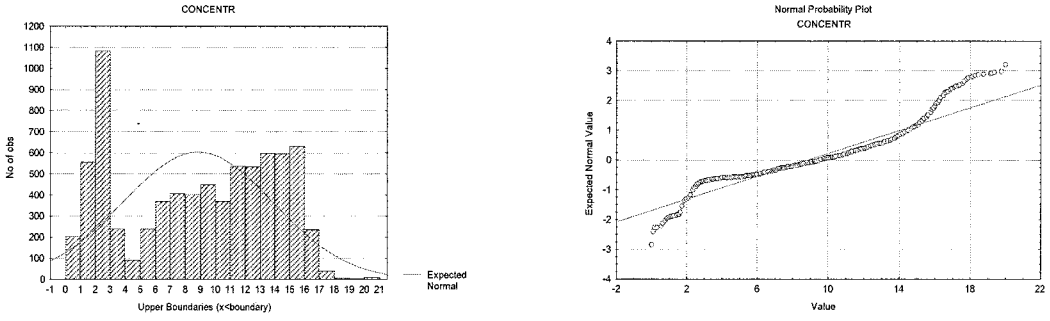


Figure 5.22: Frequency histogram and normal probability plot of Racod's data

5.5.2 Correlations

Because of the a-priori hypotheses about the relationships between the sensor responses on RH, it was necessary to study their correlations. Scatterplots of these variables are generally instructive and may provide valuable clues for the next stages of the analysis. The results presented in Table 5.5 and Figure 5.23 to Figure 5.30 show the strong correlation between the variables. Most sensors display a curvilinear relationship with RH and as a result are strongly correlated with one another. However, both TOC and Sensor 2 show a different pattern. This strongly supports our previous hypothesis that changes in environmental conditions might play a significant role in the distribution of the data.

On the other hand, the correlation between the sensor responses and TOC is comparatively low ($R<0.30$), which suggests that there is no obvious linear relationship between the variables of interest. Finally in Figure 5.31, an example is presented that shows the variation in both TOC and RH values over a few days. Although the two variables are not clearly correlated (Table 5.5), they both follow a similar night and day pattern. These concurrent diurnal cycles must be carefully taken into account during analysis in order to avoid possible misinterpretation of the results.

Table 5.5: Linear correlations between variables (TOC-Prosat all data). Highlighted values significant to $p<0.050$

	RH	Sensor1	Sensor2	Sensor3	Sensor4	Sensor5	Sensor6	Sensor7	Sensor8	Toc
RH	1.00	**0.92**	-0.72	**0.96**	**0.96**	**0.98**	**0.96**	**0.97**	**0.98**	0.29
Sensor1	**0.92**	1.00	-0.68	0.97	0.93	0.94	0.93	0.93	0.93	0.30
Sensor2	-0.72	-0.68	1.00	-0.71	-0.65	-0.72	-0.65	-0.68	-0.73	-0.14
Sensor3	**0.96**	0.97	-0.71	1.00	0.96	0.98	0.94	0.98	0.97	0.29
Sensor4	**0.96**	0.93	-0.65	0.96	1.00	0.98	0.97	0.96	0.97	0.29
Sensor5	**0.98**	0.94	-0.72	0.98	0.98	1.00	0.95	0.98	0.98	0.28
Sensor6	**0.96**	0.93	-0.65	0.94	0.97	0.95	1.00	0.95	0.97	0.27
Sensor7	**0.97**	0.93	-0.68	0.98	0.96	0.98	0.95	1.00	0.98	0.27
Sensor8	**0.98**	0.93	-0.73	0.97	0.97	0.98	0.97	0.98	1.00	0.26
Toc	0.29	0.30	-0.14	0.29	0.29	0.28	0.27	0.27	0.26	1.00

5.6 SUMMARY

- A modified sensor array system was implemented in the field at a wastewater treatment plant and coupled to on-line RACOD and TOC analysers.
- Continuous datasets were created after screening for outliers and missing data.
- Diurnal patterns of activity and the dilution effect of rain were observed.
- MLR on field data supported previous results from the lab. A better correlation is observed over shorter periods of time.
- The use of clean water data to compensate for RH variations improved the correlations. This suggested that RH can be used for parametric compensation or as an imput to multivariate analysis.
- Statistical examination of the datastets have revealed a number of factors which may affect multivariate-based models These are mainly:
  -Effect of RH on the sensors
  -Deviations form the normality assumption
  -Possible non-linearity
  -Concurrent RH and TOC diurnal variations

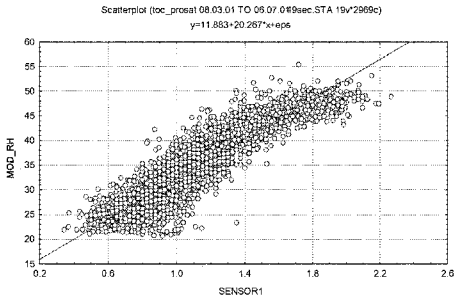


Figure 5.23: Sensor 1 vs RH

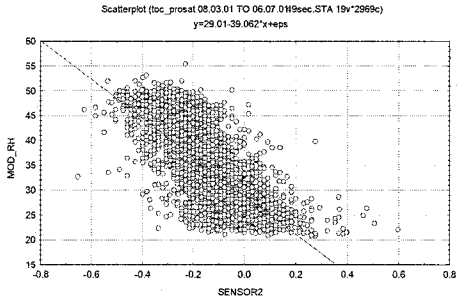


Figure 5.24: Sensor 2 vs RH

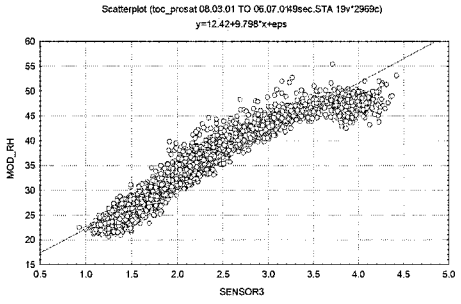


Figure 5.25: Sensor 3 vs RH

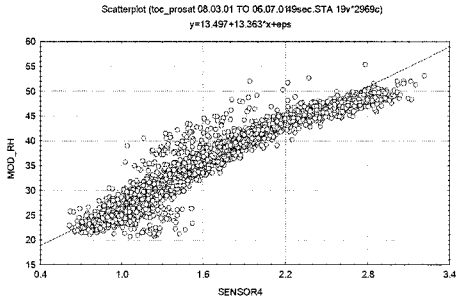


Figure 5.26: Sensor 4 vs RH

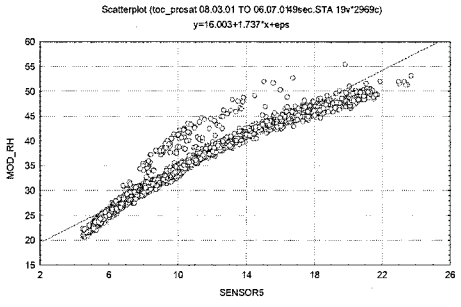


Figure 5.27: Sensor 5 vs RH

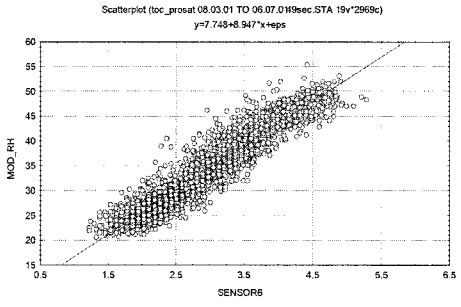


Figure 5.28: Sensor 6 vs RH

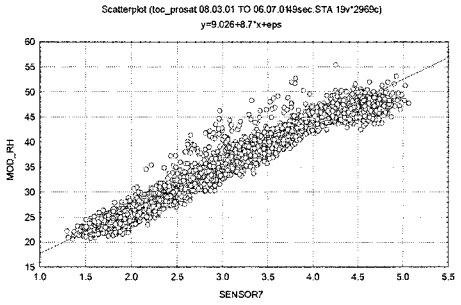


Figure 5.29: Sensor 7 vs RH

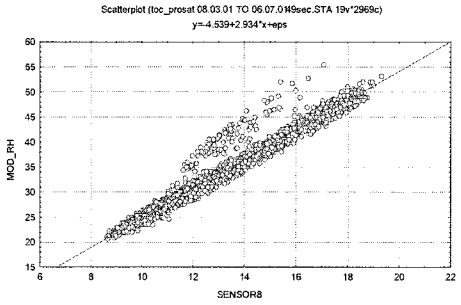


Figure 5.30: Sensor 8 vs RH

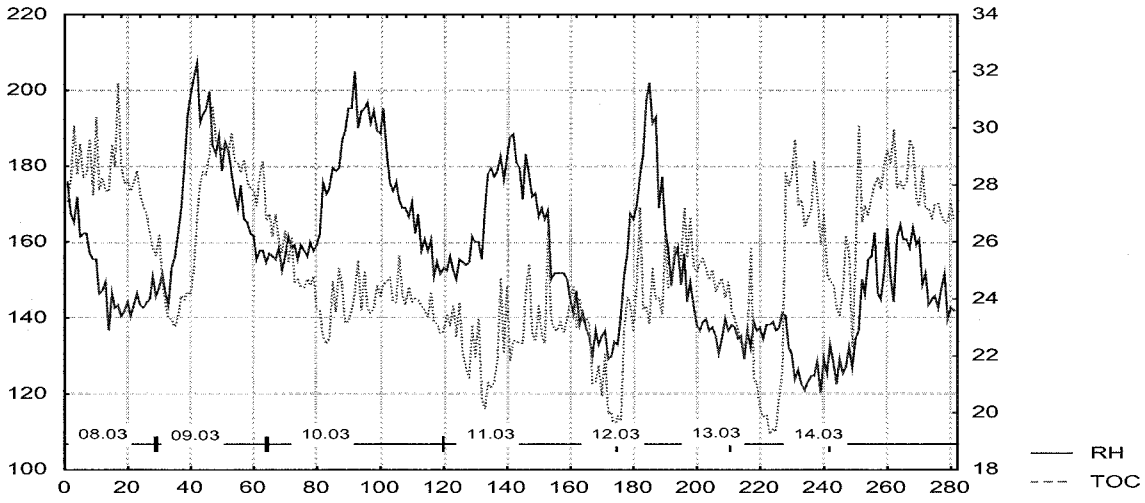


Figure 5.31: Example of TOC and RH diurnal variations (08.03.01 to 14.03.01)

Chapter 6: INVESTIGATION OF ON-LINE DATA CORRELATIONS WITH GLOBAL ORGANIC LOAD PARAMETERS: A STATISTICAL STUDY

CHAPTER 6: INVESTIGATION OF ON-LINE DATA CORRELATION WITH GLOBAL ORGANIC LOAD PARAMETERS: A STATISTICAL STUDY

6.1 INTRODUCTION

In this chapter, multivariate statistical analysis techniques are used to investigate the relationship between the sensor response and the wastewater organic load as given by TOC and BOD. As discussed in section 2.1.4, MVS is particularly suited to the analysis of sensor array data. Despite the increasing popularity of novel pattern recognition techniques based on artificial intelligence, traditional statistical techniques are still the only tools available to the analyst who wishes to gain an understanding of the relationship between the sensor responses and the dependent variables. The choice of techniques for this investigation was cross-validated using the decision tree given in Figure 2.10. All methods used here have been discussed in section 2.1.4.

The problem at hand was to find (and if possible, quantify) a relationship between the dependent variables (TOC, BOD) and the independent variables (8 sensor responses), all variables being metric. "Dependence" techniques are generally used for prediction, whereas "interdependence" examination methods are generally used to study the structure within the data. As seen in section 4.5.4 and 5.3, multiple linear regression was the first obvious choice and gave good results in a controlled environment over short periods of time. This technique is further investigated and compared to other linear (PLS, PCR) and non-linear regression methods (Polynomial

and Factorial Regression). A summary of the Racod-Prosat data investigations is given with all results presented in Appendix C. The plots clearly show the poor quality of the Racod data, which cannot be realistically used in this study. Therefore, only results for the studies carried out on the TOC-Prosat datasets are fully discussed in this chapter.

Correlations (R) between the predicted and measured (true) concentrations were computed as well as the minimum, maximum and mean absolute relative error (RAE) and corresponding standard deviations so as to give an indication of the performance of the models. Hierlemann *et al.* (1995) also used the RAE and Max RAE as perfomance indicators. These measures are considered to be particularly relevant to the characterisation of the reproducibility and predictive ability of the algorithms. Correlations and mean relative error give an overall representation of the performance, whereas minimum, maximum and standard deviation values indicate the uncertainty variability of the estimation, thereby giving an indication of confidence. These were calculated as follows:

$$R(pearson) = \frac{\sum (Cpred. - \overline{Cpred.}) * (Ctrue - \overline{Ctrue})}{\sqrt{\sum (Cpred. - \overline{Cpred.})^2 * \sum (Ctrue - \overline{Ctrue})^2}}$$

$$RAE = \frac{|Cpred. - Ctrue|}{Ctrue}$$

$$\text{Mean } RAE = \frac{1}{n} * \sum \left(\frac{|Cpred. - Ctrue|}{Ctrue} \right)$$

Where n is the number of cases, and *Cpred.* and *Ctrue* are the predicted and measured concentrations respectively

6.2 MULTIPLE LINEAR REGRESSION

We discussed in Section 2.1.4 how MLR can be used to investigate the relationship between a single DV and several IV's. In this study MLR is used in an attempt to predict TOC values from the sensor array data. In MLR it is assumed that the relationship between the 2 sets of variables is linear. In practice this assumption can virtually never be confirmed. Fortunately multiple regression procedures can tolerate minor deviations from this assumption. On the other hand the size of the data set has a great influence on the validity of the analysis, particularly if the data is not normally distributed as we have seen previously. The number of cases (observations) to the number of IV's ratio must be sufficient or the solution will turn out to be ideal - and meaningless (Tabachnick and Fidell, 1996). Although the strict minimum ratio is 5 to 1 to avoid overfitting, Hair *et al.* (1998) recommends that the level should be between 15-20 cases per IV for a model to be generalizable. Consequently, Dataset 8 will not be included in our discussion. With the wide range of parameters likely to influence the quality of our results, a series of different approaches to the use of MLR on our data have been tested:

- Training on data subsets followed by prediction on the same training dataset.
- Generalisation by training on data subsets and prediction over the whole dataset.
- Study of the effect of the size of the training dataset on multiple regression
  1) using continuous subsets
  2) using reduced subsets over a constant period of time
- Training using Blank data to correct the effect of RH (as in section 5.32)
- Training using a reduced number of sensors (3^{rd} generation only: i.e. sensors1-4).

In every case the aim is to find a linear relationship between TOC and the sensor responses of the form:

$$TOCconc = a0+a1*sensor1+a2*sensor2+\ldots+a_n*sensor_n$$

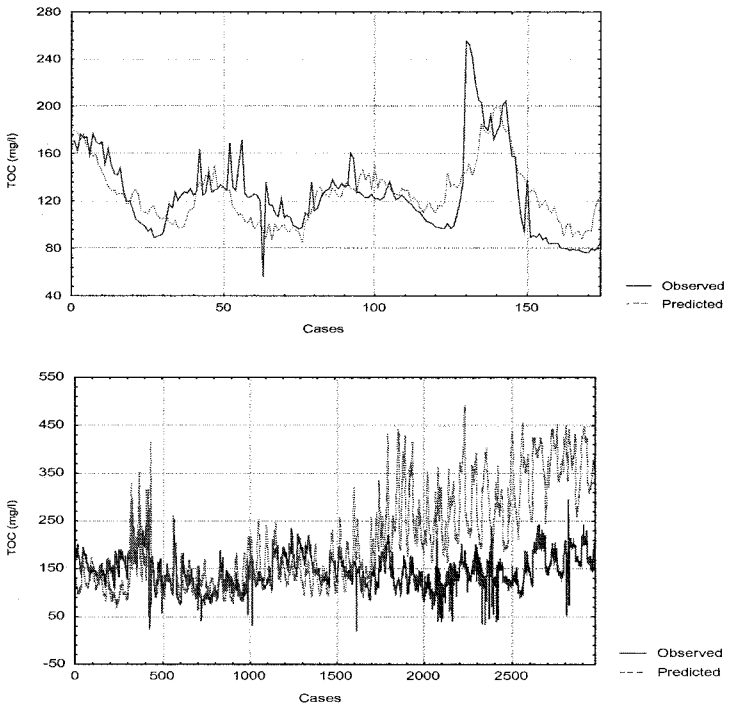
Where n is the number of sensors used in the model.

6.2.1 Sensors 1-8, whole datasets

Figures 6.1 to 6.3 display comparison examples of predicted versus measured TOC concentration, with the top plots showing the results of the MLR-based prediction over the training sets only (i.e. subsets 3, 6 and 9 respectively). The bottom graphs illustrate the application of the respective models to the entire dataset. In Figure 6.4, the whole dataset (subset 12) was used both for training and prediction. The graphs for the remaining subsets (1, 2, 4, 5, 7, 10, 11) are given in Appendix D.

A summary of the performances of the individual models on the training set and when applied to the rest of the data is presented in Tables 6.1; 6.2 and Tables 6.3; 6.4 respectively. As expected it is clear that the models do not perform as well on unknown data as they do on the training set. With the exception of subset 2 (training) and subset 3 (whole set), the average error is less than 19% for training and less than 40% for unknowns. This may be interpreted as acceptable if we consider the accepted 20-30% error on a traditional BOD_5 test and the time-scale over which the models have been applied. However, and despite the great majority of the data being predicted with less than 50% error, the correlations between predicted and measured values are relatively poor. In the few cases where R was slightly higher, the predicted values were also strongly correlated with RH.

It appears that, although sensitive to the diurnal changes, MLR-based models have not been able to adapt to sudden changes in TOC levels nor to the long term (seasonal) changes in wastewater quality. Instead, we can see from Figure 6.4 and Figure 6.5 that the predicted TOC values are mostly distributed around the average measured TOC (~145mg/L) and deviate relatively little from this value. Figure 6.6 shows how this affects the distribution of the absolute prediction error, which is minimal for TOC concentrations close to the average value, but rapidly increases if the concentration increases or decreases. This demonstrates the inability of the model to predict TOC values at both ends of the spectrum.



Error Min	0.00%
Error Max	68%
Error Mean	15%
Error StDev	0.12

<20%	75.28%
<30%	86.20%
<50%	99.42%

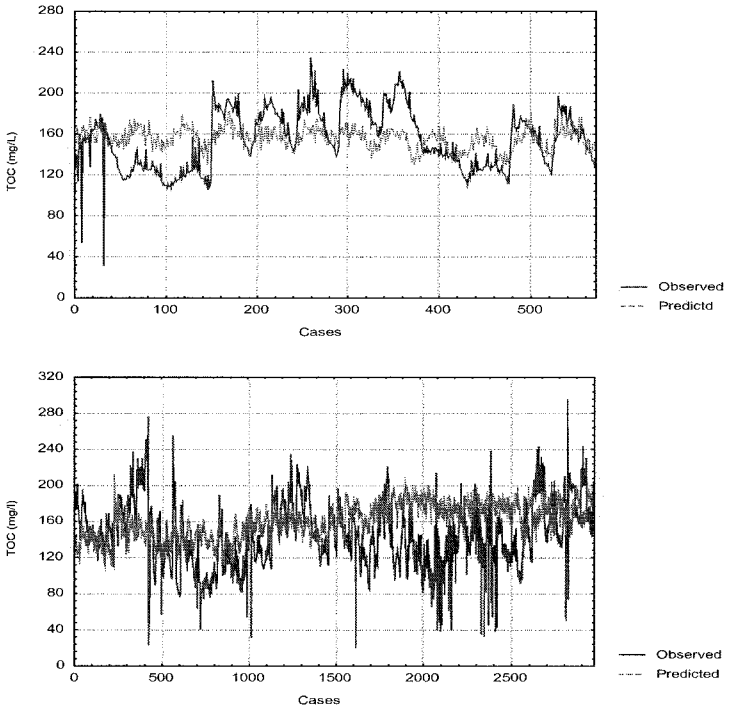
Correlation between:
Toc_pred and toc_conc: 0.71
Toc_pred and Rh: 0.97

Error Min	0.00%
Error Max	1123%
Error Mean	65%
Error StDev	0.78

<20%	29.06%
<30%	42.44%
<50%	57.29%

Correlation between:
Toc_pred and toc_conc: 0.28
Toc_pred and Rh: 0.98

Figure 6.1: Observed and MLR- predicted TOC values. Training set (subset 3, top) and whole dataset (bottom)



Error Min	0.00%
Error Max	397%
Error Mean	16%
Error StDev	0.21

<20%	70.92%
<30%	90.01%
<50%	99.47%

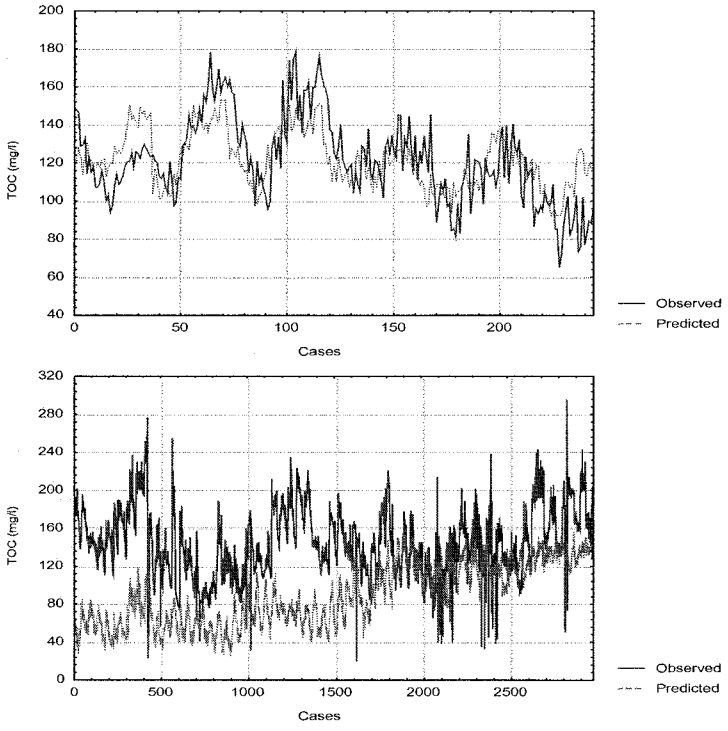
Correlation between:
Toc_pred and toc_conc: 0.34
Toc_pred and Rh: 0.66

Error Min	0.00%
Error Max	804%
Error Mean	28%
Error StDev	0.37

<20%	49.98%
<30%	68.00%
<50%	85.95%

Correlation between:
Toc_pred and toc_conc: 0.12
Toc_pred and Rh: 0.61

Figure 6.2: Observed and MLR- predicted TOC values. Training set (subset 6, top) and whole dataset (bottom)



Error Min	0.00%
Error Max	72%
Error Mean	12%
Error StDev	0.10

<20%	84.83%
<30%	96.72%
<50%	99.18%

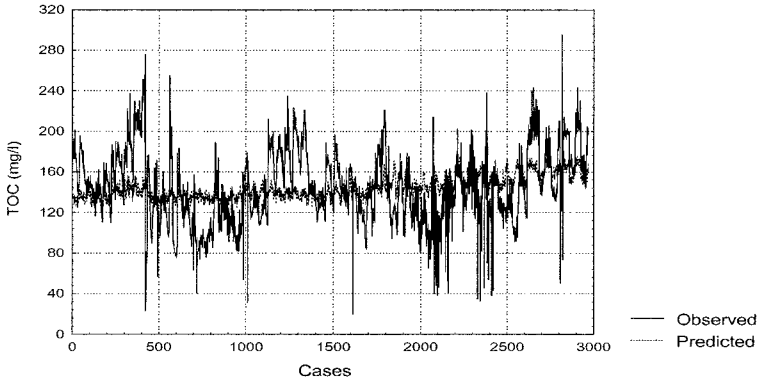
Correlation between:
Toc_pred and toc_conc: 0.65
Toc_pred and Rh: 0.94

Error Min	0.00%
Error Max	456%
Error Mean	40%
Error StDev	0.25

<20%	22.86%
<30%	36.27%
<50%	62.14%

Correlation between:
Toc_pred and toc_conc: 0.24
Toc_pred and Rh: 0.95

Figure 6.3: Observed and MLR- predicted TOC values. Training set (subset 9, top) and whole dataset (bottom)



Error Min	0.00%
Error Max	611%
Error Mean	21%
Error StDev	0.31

<20%	62.98%
<30%	80.59%
<50%	93.97%

Correlation between:
Toc_pred and toc_conc: 0.34
Toc_pred and Rh: 0.75

Figure 6.4: Observed and MLR- predicted TOC values. (whole dataset)

Table 6.1: MLR predictions RAE and correlations on training sets

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	0.33	6.95	0.68	1.91	0.41	3.97	5.84	0.19	0.72	4.12	0.19
Mean RAE	0.09	0.32	0.15	0.18	0.10	0.16	0.13	0.06	0.12	0.19	0.05
RAE StDev	0.07	0.88	0.12	0.16	0.09	0.21	0.41	0.04	0.10	0.35	0.04
R (pred vs obs)	0.48	0.34	0.71	0.38	0.33	0.34	0.51	0.83	0.65	0.46	0.86
R (pred vs RH)	0.09	-0.41	0.97	0.46	-0.32	0.66	0.84	0.91	0.94	0.85	0.86

Table 6.2: Fraction of cases predicted with an RAE < x%, using MLR on training sets

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
<10%	0.63							0.79			0.90
<15%								0.97			0.95
<20%	0.92	0.75	0.75	0.65	0.87	0.71	0.87	1.00	0.85	0.75	1.00
<30%		0.87	0.86	0.85	0.96	0.90	0.94		0.97	0.88	
<40%	1.00										
<50%		0.92	0.99	0.99	1.00	0.99	0.99		0.99	0.95	

Table 6.3: MLR predictions RAE and correlations on all data

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	7.93	7.93	11.23	1.91	4.41	8.04	5.84	8.14	4.56	5.81	7.56	6.11
Mean RAE	0.31	0.31	0.65	0.18	0.23	0.28	0.21	0.32	0.40	0.20	0.40	0.21
RAE StDev	0.40	0.40	0.12	0.16	0.22	0.37	0.29	0.39	0.25	0.27	0.47	0.31
R (pred vs obs)	0.22	-0.15	0.28	0.29	0.28	0.12	0.31	0.20	0.24	0.31	0.05	0.34
R (pred vs RH)	0.81	-0.08	0.98	0.72	0.53	0.61	0.91	0.94	0.95	0.76	-0.41	0.75

Table 6.4: Fraction of cases predicted with an RAE < x%, using MLR on all data

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
<10%	0.27	0.25										
<20%	0.49	0.38	0.29	0.65	0.48	0.50	0.59	0.41	0.23	0.62	0.38	0.63
<30%			0.42	0.85	0.72	0.68	0.79	0.61	0.36	0.80	0.52	0.81
<50%	0.74	0.60	0.57	0.99	0.97	0.86	0.96	0.86	0.62	0.96	0.72	0.94

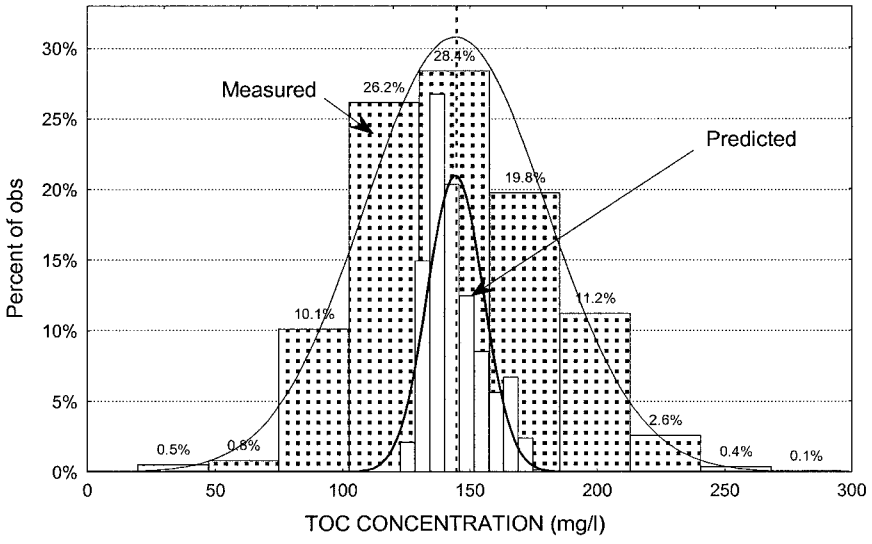


Figure 6.5: Comparative frequency histograms showing the distribution of observed and predicted TOC values. (dashed line: 145mg/l)

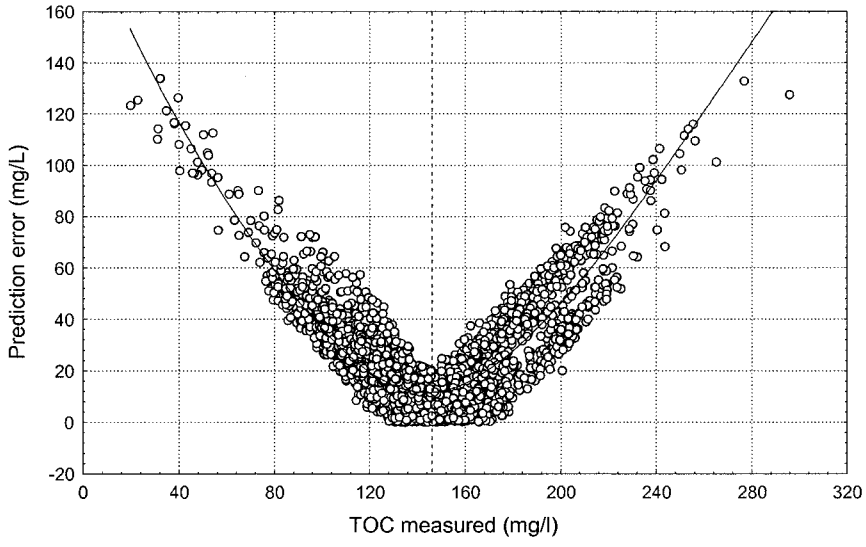


Figure 6.6: Absolute prediction error vs. measured TOC concentration (dashed line: 145mg/l)

MLR assumes that the residuals (predicted minus measured values) are distributed normally. Therefore, it is good practice to review their distribution before drawing any conclusion. We can see from Figure 6.7(a,b), that there is no apparent deviation from this assumption. However, further graphical analysis of the residuals allows us to detect patterns of deviation more easily. Although the residuals in Figure 6.7(b) appear normally distributed, the overall curved shape observed in Figure 6.8(b) suggests that some slight deviation from the normality assumption may be present. In addition, the pseudo triangular shape in Figure 6.8 (a) may indicate the presence of unequal variances (heteroscedasticity). Heteroscedasticity can occur when some of the variables are skewed and others are not. This was shown to be the case with our data (Section 5.5.1) where Sensor 2 and TOC are the only variables displaying a clear normal distribution.

Heteroscedasticity is one of the most common assumption violations (Hair *et al.*, 1998) and failure of homoscedasticity in regression does not invalidate our analysis so much as weaken it (Tabachnick and Fidell, 1996). In this example, the errors of prediction slightly increase as the predicted values decrease. Still, this deviation is not obvious enough to cause concern. From Figure 6.8(a), no deviation from the linearity assumption (i.e. curvilinear shape) can be observed. Thus, we can reasonably assume that any effect of deviations from these assumptions on the linear regression technique investigated, will be within acceptable limits.

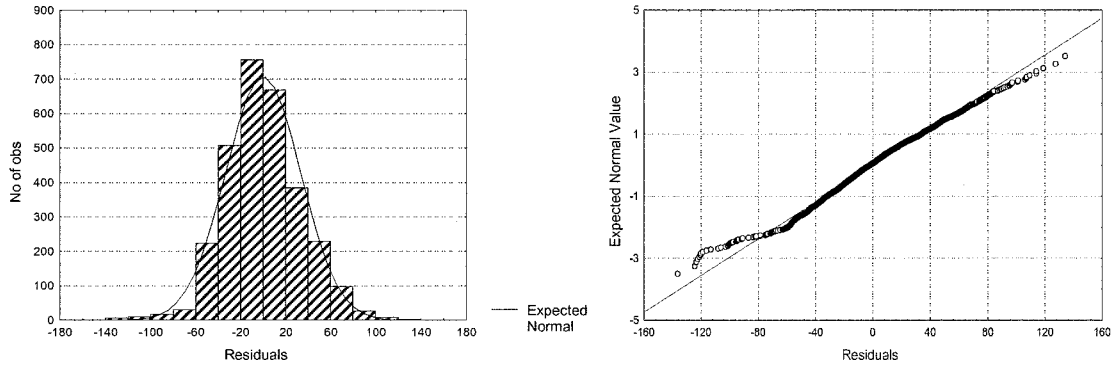


Figure 6.7: Distribution of residuals: Frequency histogram (a) and normal probability plot (b)

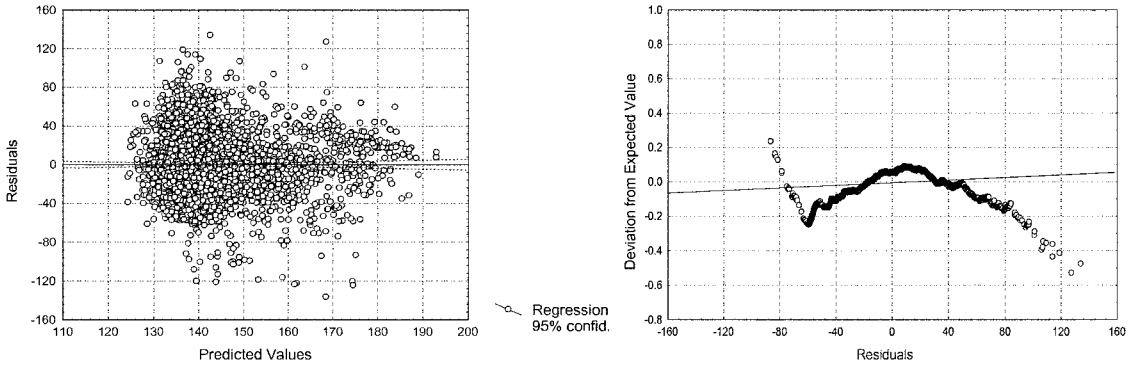


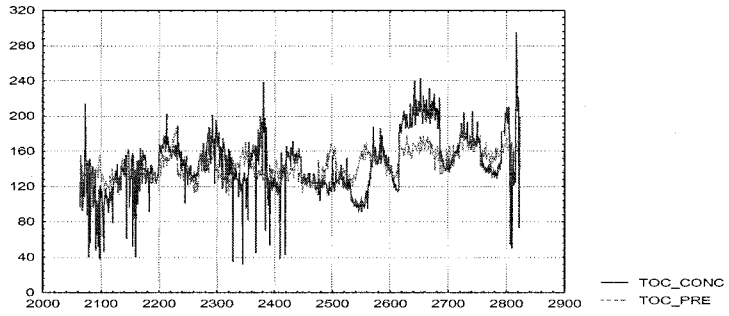
Figure 6.8: Graphical examination of residuals: plot of residuals vs predicted values (a) and detrended normal plot of residuals (b)

6.2.2 Sensors 1-8, reduced datasets.

Following the observed difference in the predictions when using MLR trained on different datasets, a simple study was carried out to establish whether the amount of training data points and/or the time-scale over which training was carried out has a significant effect on the outcome of the analysis. This was carried out on Dataset 10 which is the largest continuous set of data. In Figure 6.9 to Figure 6.11, the length of time over which the MLR is performed was reduced from 14 days (original data) to 7 days and 4 days. Models were then applied to the whole dataset (2 weeks). Comparison of correlations between predicted and measured TOC concentrations show a tendency for R to rapidly decrease. On the other hand, reducing the number of training points from 780 to 360 and to 78 only, did not seem to have a major influence on the correlations or the prediction errors if training was carried out over the two-weeks period (Table 6.5). These results tend to support the findings discussed in 4.5.5 where time-dependent relationships were suggested. Here it appears that reducing the duration of training has more influence than just reducing the amount of data used. Although there are obvious limitations to how little data can be used for training (see Section 4.5.5), the long term validity of MLR based models strongly depends on how long the training lasted. Thus, it is important not to overstretch the prediction for too long after the end of training, particularly if training was carried out over a comparatively short period of time. We have seen from Figure 6.10 that even a one-week trained MLR model performs poorly if projected one week ahead. This was also the case in Section 4.5.5 where the model collapsed when trying to predict the organic content of samples collected a week later, despite the laboratory controlled experimental conditions.

Table 6.5: Effect of duration and number of cases used for training

	Two weeks (760 cases)	Two weeks (360 cases)	Two weeks (78 cases)	One week (380 cases)	4 days, (220 cases)
Min RAE	0.00	0.00	0.00	0.00	0.00
Max RAE	4.12	2.88	3.68	4.19	4.19
Mean RAE	0.19	0.20	0.21	0.23	0.19
Error StDev	0.35	0.35	0.37	0.45	0.39
R (pred vs obs)	0.46	0.47	0.40	0.18	0.15

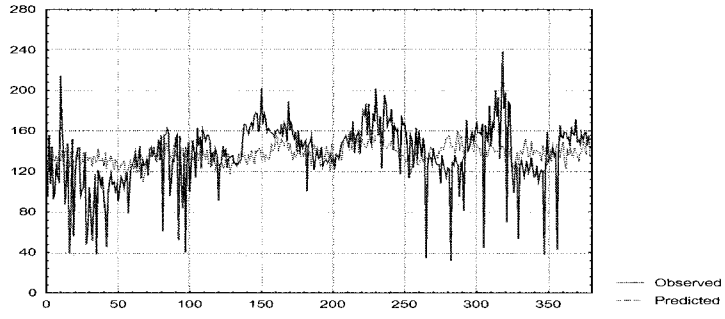


Error Min	0.00%
Error Max	412%
Error Mean	19%
Error StDev	0.35

<20%	74.74%
<30%	88.16%
<50%	94.74%

R toc_conc and toc_pred : 0.46
R toc_pred and RH : 0.85

Figure 6.9: Observed and predicted TOC. Training over two weeks (760 cases)

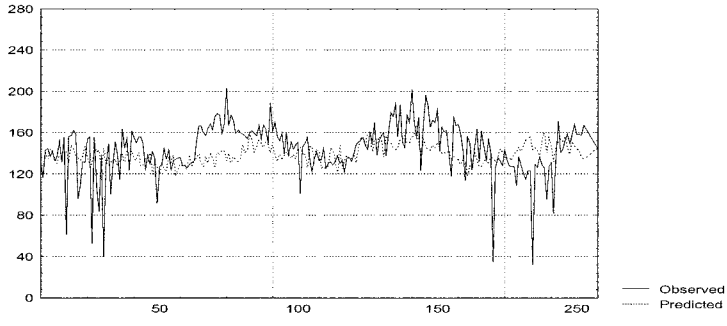


Error Min	0.00%
Error Max	419%
Error Mean	23%
Error StDev	0.45

<20%	75.73%
<30%	86.54%
<50%	93.14%

R toc_conc and toc_pred : 0.18
R toc_pred and RH : 0.77

Figure 6.10: Observed and predicted TOC. Training over one week (380 cases)



Error Min	0.00%
Error Max	419%
Error Mean	19%
Error StDev	0..39

<20%	77.50%
<30%	90.63%
<50%	95.94%

R toc_conc and toc_pred : 0.15
R toc_pred and RH : 0.76

Figure 6.11: Observed and predicted TOC. Training over 4 days (220 cases)

6.2.3 Humidity and reduced number of sensors

Results discussed in section 5.3.2 suggested that the effect of RH could be reduced by using information acquired on clean water samples. Therefore RH values were incorporated into a MLR-based prediction. For each subset, blank data preceding the period of continuous wastewater monitoring was used to adjust the response of individual sensors as a function of RH using the same procedure described in section 5.3.2. The correlations between predicted and measured TOC are given in Table 6.6. As opposed to our previous laboratory investigations, the improvements are negligible. A second approach was also used to try and deal with the effect of RH and find a better solution. In Section 5.5.1, the distribution of relative humidity was discussed. From Figure 5.25 we can see that a higher proportion of data are within the range 23% to 31%, with a maximum number of cases with RH values comprised between 25% and 26% (186 cases). This group of data was used to investigate the effect of selecting data with constant RH values on MLR. Applying the model to the same set gave an average RAE of 18% with a maximum error of 75%. 62% of the data was predicted with less than 20% error and 80% with less than 30%. Clearly, this approach was unsuccessful and gave worse results than MLR on the continuous sets. In Table 6.6 we also give the results for MLR trained with the 3rd generation sensors only (Sensors 1-4) which are theoretically less sensitive to RH. This made the correlations worse.

Table 6.6: Effect of RH correction and reduced number of sensors on R

	Sensors 1-8	Sensors 1-8 (RH-corrected)	Sensors 1-4
Data set 1	0.48	0.48	0.42
Data set 2	0.34	0.35	0.22
Data set 3	0.71	0.71	0.68
Data set 4	0.38	0.41	0.27
Data set 5	0.33	0.33	0.13
Data set 6	0.34	0.34	0.26
Data set 7	0.51	0.54	0.47
Data set 8	0.83	0.85	0.71
Data set 9	0.65	0.65	0.63
Data set 10	0.46	0.50	0.43
Data set 11	0.86	0.86	0.83
Data set 12	0.34	0.39	0.31

6.3 PARTIAL LEAST SQUARE

Disregarding MLR, which did not provide good predictions of the TOC concentration, the use of PLS was investigated. PLS is a tool extensively used for quantitative analysis in chemometrics and sensor array applications. Di Natale *et al.* (1997) noted that although also based on a linear approach, it usually achieves results which are substantially better than those obtained with MLR. Therefore, we applied PLS to all TOC-Prosat subsets using sensors 1-8 as the predictor variables. The results are summarised in Table 6.7 to Table 6.10. Indeed no significant improvements were achieved and generally the performances of the prediction models on the whole dataset were very similar to those of MLR or often worse when applied to the training set only. The relatively poor predictions are illustrated in Figure 6.12 and Figure 6.13 which show a great similarity with MLR results presented in Figure 6.1 and Figure 6.2.

A possible explanation for these large errors in the prediction could be non-linearities in the relationship between the two sets of variables which may be too important for MLR or PLS to handle. This is particularly probable in solutions such as wastewater which are characterised by the presence of many different (and sometimes unstable) compounds and for which a deviation from linearity can be expected. Since MLR and PLS cannot map the full extent of the relationship among the IV's and the DV's, a different approach involving non-linear regression techniques may be required.

Table 6.7: PLS predictions RAE and correlations on training sets

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	0.38	7.32	0.72	1.87	0.41	4.11	8.97	0.23	0.63	4.12	0.24
Mean RAE	0.09	0.34	0.16	0.19	0.10	0.16	0.13	0.06	0.13	0.19	0.05
RAE StDev	0.07	0.93	0.13	0.17	0.09	0.21	0.42	0.05	0.10	0.35	0.05
R (pred vs obs)	0.44	0.18	0.68	0.26	0.33	0.24	0.49	0.76	0.62	0.46	0.83
R (pred vs RH)	0.04	-0.94	0.96	0.74	-0.32	0.97	0.86	0.97	0.98	0.85	0.87

Table 6.8: Fraction of cases predicted with an RAE < x%, using PLS on training sets

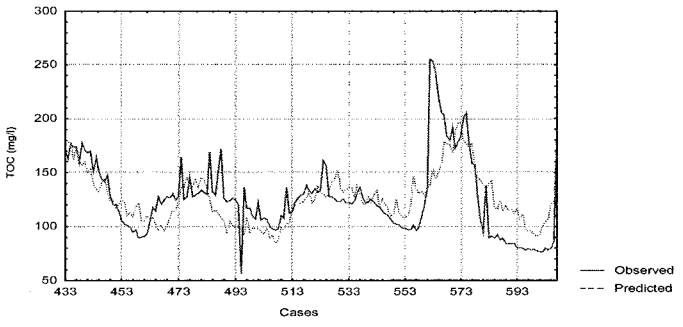
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
<10%	0.62	0.48	0.41	0.28	0.57	0.37	0.60	0.73	0.48	0.46	0.88
<20%	0.92	0.77	0.71	0.61	0.87	0.68	0.86	0.99	0.82	0.75	0.99
<40%	1.00	0.91	0.93	0.96	0.99	0.96	0.98	1.00	0.98	0.93	1.00

Table 6.9: PLS predictions RAE and correlations on all data

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	7.45	8.50	10.27	5.90	4.41	7.71	5.97	7.96	4.81	5.81	6.52	6.11
Mean RAE	0.28	0.47	0.59	0.22	0.23	0.29	0.21	0.31	0.29	0.20	0.22	0.21
RAE StDev	0.39	0.48	0.70	0.29	0.22	0.40	0.30	0.41	0.25	0.27	0.33	0.31
R (pred vs obs)	0.26	-0.28	0.28	0.30	0.28	0.28	0.32	0.24	0.28	0.31	0.31	0.34
R (pred vs RH)	0.29	-0.98	0.98	0.95	0.53	0.98	0.89	0.97	0.98	0.76	0.94	0.75

Table 6.10: Fraction of cases predicted with an RAE < x%, using PLS on all data

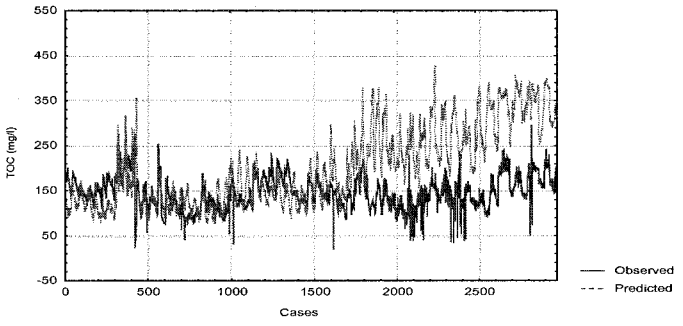
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
<10%	0.28	0.16	0.14	0.29	0.21	0.26	0.34	0.26	0.18	0.34	0.32	0.36
<20%	0.52	0.31	0.30	0.57	0.48	0.50	0.61	0.45	0.37	0.62	0.59	0.63
<40%	0.79	0.54	0.54	0.91	0.88	0.77	0.94	0.76	0.75	0.91	0.90	0.90



Error Min	0.00%
Error Max	72%
Error Mean	16%
Error StDev	0.13

<10%	40.80%
<20%	70.69%
<40%	93.10%

Correlation between:
Toc_pred and toc_conc: 0.68
Toc_pred and Rh: 0.96

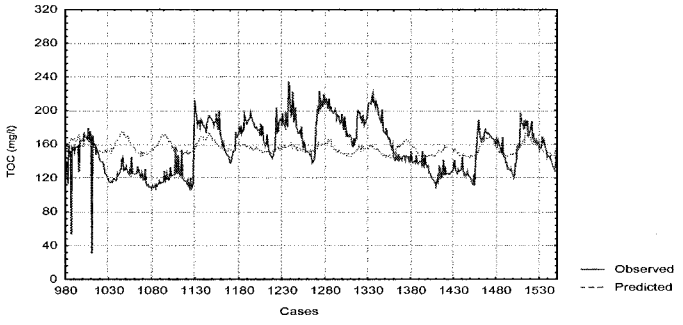


Error Min	0.00%
Error Max	1027%
Error Mean	59%
Error StDev	0.70

<10%	14.42%
<20%	30.11%
<40%	53.65%

Correlation between:
Toc_pred and toc_conc: 0.28
Toc_pred and Rh: 0.98

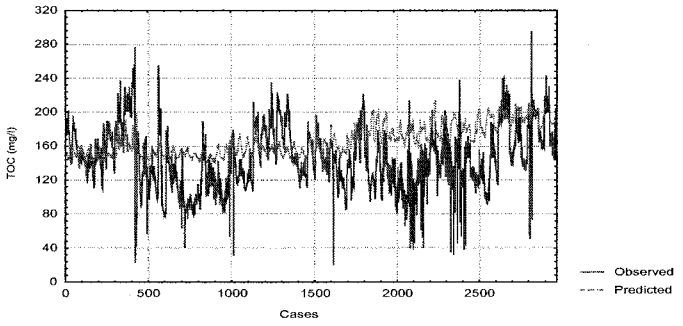
Figure 6.12: Observed and PLS-predicted TOC values. Training set (subset 3) and whole data set (top and bottom respectively)



Error Min	0.00%
Error Max	771%
Error Mean	29%
Error StDev	0.40

<10%	26.17%
<20%	50.12%
<40%	77.30%

Correlation between:
Toc_pred and toc_conc: 0.28
Toc_pred and Rh: 0.98



Error Min	0.00%
Error Max	411%
Error Mean	16%
Error StDev	0.21

<10%	37.30%
<20%	68.13%
<40%	96.32%

Correlation between:
Toc_pred and toc_conc: 0.24
Toc_pred and Rh: 0.97

Figure 6.13: Observed and PLS predicted TOC values. Training set (subset 6) and whole data set (top and bottom respectively)

6.4 POLYNOMIAL REGRESSION

In the previous section we showed how the use of linear regression techniques such as MLR and PLS was unsatisfactory. The next logical step in our search of a suitable model for the prediction of wastewater organic load was to investigate the potential of non-linear techniques. Here, polynomial regression is first evaluated on both training and total datasets using all sensors as well as sensors 1-4 only as input variables, so as to allow comparison with previous studies. The results obtained using a 2^{nd} degree polynomial regression with sensors 1-8 as IV's are presented in Tables 6.11 to Table 6.14. Graphical representations and results obtained using sensors 1-4 as IV's are given in Appendix E.

For all subsets, most training points can be predicted with less than 50% error, and with an average error of less than 29% (33% using sensors 1-4 only). As previously observed with MLR reducing the number of sensors from 8 to 4 did not improve the results. Indeed these results are very similar to those obtained with MLR and PLS in the same conditions as shown in Figure 6.14 and Figure 6.15. Again polynomial regression tends to follow the diurnal pattern but cannot be applied to unknown data acquired a few weeks or a few months later. The same influence of RH can also be observed.

In an attempt to find a better model based on this technique, we gradually increased the degree of the polynomial from 2 to 8 using all 8 sensor responses as IV's. A summary table (Table 6.15) shows the correlation between predicted and measured TOC for each individual training set. Table 6.16 shows a comparison of the RAE's for the different polynomials. As illustrated in Figure 6.16 and Figure 6.17, increasing the degree of the polynomial did not improve the prediction over the whole dataset, despite prediction results showing that higher polynomial degrees give significantly better correlations when the models are applied to the training sets only. The increased dependence of higher degree polynomial regression models on the training set and the resulting poor generalisation strongly suggest that overfitting may be present.

Table 6.11: 2nd degree polynomial regression predictions RAE and correlations on training sets (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
Min RAE (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE (%)	0.33	6.70	0.91	1.90	0.45	4.25	5.49	0.18	0.71	4.05	0.18
Mean RAE (%)	0.09	0.29	0.14	0.17	0.10	0.15	0.13	0.06	0.12	0.19	0.04
RAE StDev	0.07	0.81	0.13	0.16	0.09	0.22	0.39	0.04	0.10	0.34	0.04
R (pred vs obs)	0.54	0.49	0.75	0.44	0.38	0.42	0.55	0.84	0.67	0.50	0.89
R (pred vs RH)	0.11	-0.32	0.91	0.39	0.26	0.56	0.77	0.89	0.91	0.76	0.81

Table 6.12: Fraction of cases predicted with an RAE < x%, using 2nd degree polynomial regression on training sets (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
<20%	0.93	0.83	0.78	0.65	0.89	0.74	0.89	1.00	0.85	0.74	1.00
<30%	1.00	0.87	0.87	0.88	0.96	0.91	0.95		0.96	0.89	
<50%		0.93	0.99	0.98	1.00	0.99	0.99		0.99	0.95	

Table 6.13: 2nd degree polynomial regression predictions RAE and correlations on all data (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
Min RAE (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE (%)	5.13	10.94	17.39	35.71	3.79	7.37	5.49	8.06	4.62	5.95	8.47	6.32
Mean RAE (%)	0.62	0.76	1.08	1.62	0.34	0.31	0.24	0.34	0.28	0.29	0.99	0.21
RAE StDev	0.59	0.65	1.33	2.85	0.24	0.35	0.31	0.40	0.23	0.33	0.93	0.32
R (pred vs obs)	-0.22	-0.07	-0.25	0.29	0.04	-0.19	0.31	0.22	0.20	-0.03	-0.19	0.40
R (pred vs RH)	-0.92	0.32	-0.25	0.80	-0.44	-0.65	0.84	0.94	0.79	-0.36	-0.89	0.68

Table 6.14: Fraction of cases predicted with an RAE < x%, using 2nd degree polynomial regression on all data (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
<20%	0.31	0.16	0.26	0.34	0.31	0.46	0.53	0.36	0.39	0.51	0.22	0.65
<30%	0.43	0.23	0.36	0.48	0.47	0.62	0.73	0.55	0.60	0.67	0.29	0.82
<50%	0.57	0.36	0.49	0.63	0.80	0.79	0.88	0.82	0.90	0.85	0.39	0.94

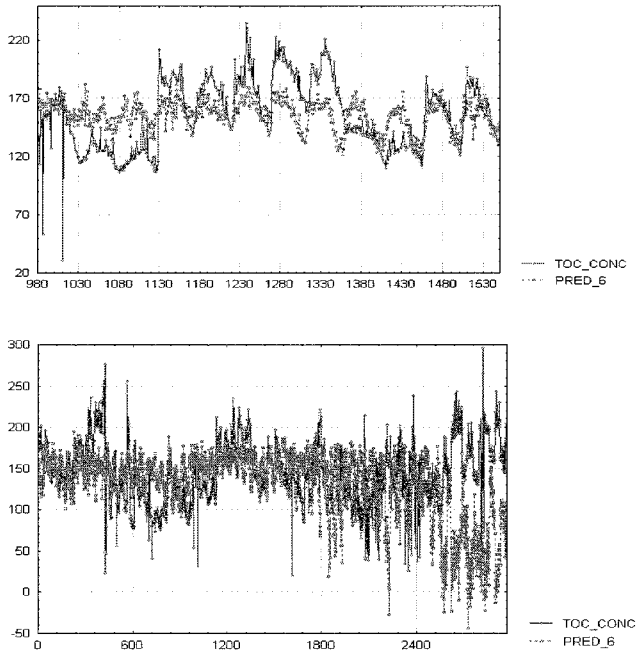


Figure 6.14: Observed and predicted TOC values using a 2nd degree polynomial regression model. Training set (subset 6) and whole data set (top and bottom respectively)

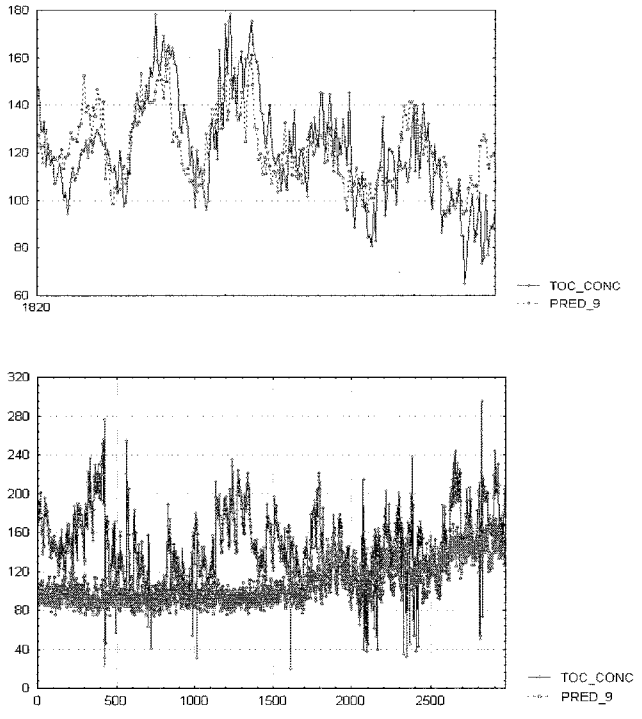


Figure 6.15: Observed and predicted TOC values using a 2nd degree polynomial regression model. Training set (subset 9) and whole data set (top and bottom respectively)

Table 6.15: Correlations (R) between predicted and observed TOC with polynomials of varying degree for each dataset

	2nd degree		3rd degree		4th degree	5th degree	6th degree	7th degree	8th degree
	training set	whole dataset	training set	whole dataset					
Subset 1	0.54	-0.22	0.59	0.30	0.59	0.60	0.63	0.64	0.64
Subset 2	0.49	-0.07	0.60	0.13	0.67	0.74	0.78	0.80	0.81
Subset 3	0.75	-0.25	0.75	-0.29	0.76	0.78	0.79	0.80	0.81
Subset 4	0.44	0.29	0.47	0.29	0.51	0.55	0.57	0.58	0.58
Subset 5	0.38	0.04	0.44	0.27	0.51	0.62	0.65	0.74	0.74
Subset 6	0.42	-0.19	0.45	0.23	0.48	0.49	0.50	0.51	0.51
Subset 7	0.55	0.31	0.59	0.29	0.66	0.71	0.73	0.74	0.75
Subset 8	0.84	0.22	0.88	0.26	0.89	0.95	0.96	0.96	0.97
Subset 9	0.67	0.20	0.69	0.22	0.72	0.73	0.74	0.76	0.76
Subset 10	0.50	-0.03	0.52	0.24	0.53	0.55	0.56	0.56	0.60
Subset 11	0.89	-0.19	0.90	0.21	0.90	0.91	0.91	0.91	0.91
Set 12	0.40	0.40	0.46	0.46	0.48	0.49	0.49	0.50	0.51

Table 6.16: Predictions errors using dataset 12 (sensors 1-8) and polynomial regression model of varying degree

	2nd degree	3rd degree	4th degree	5th degree	6th degree	7th degree	8th degree
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	6.32	6.21	6.49	6.54	6.59	6.64	6.62
Mean RAE	0.21	0.20	0.20	0.20	0.20	0.19	0.19
StDev	0.32	0.31	0.32	0.32	0.32	0.32	0.32

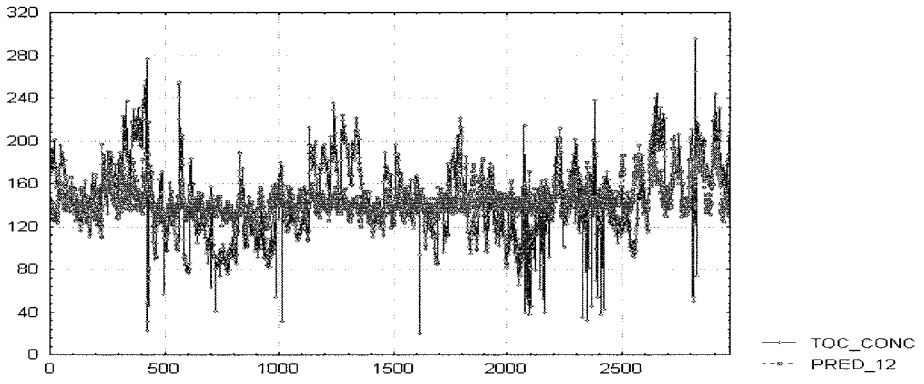


Figure 6.16: Observed and predicted TOC using a 2nd degree polynomial and sensors 1-8 as IV's (all data)

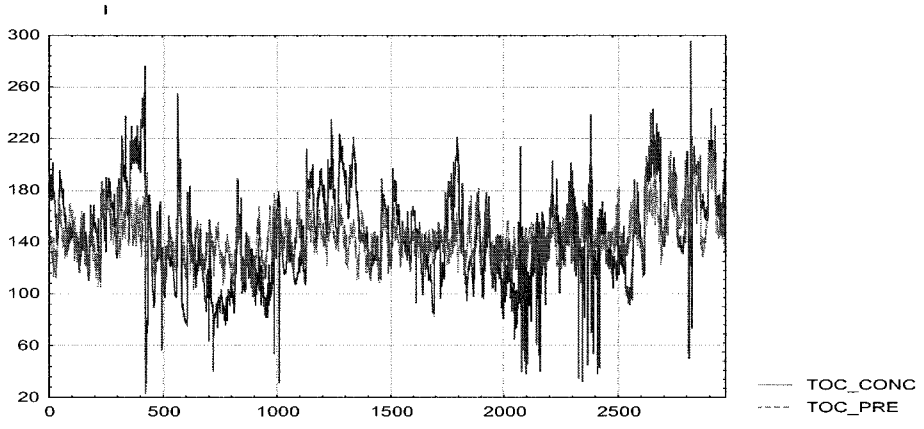


Figure 6.17: Observed and predicted TOC using a 8th degree polynomial and sensors 1-8 as IV's (all data)

6.5 FACTORIAL REGRESSION

In section 2.1.4.1 we discussed how polynomial regression-based models include the main effects and higher order effects for the sensor responses. However, this technique does not take into account the interaction effects between IV's. On the other hand factorial designs do include interactions between the predictor variables. Therefore the use of factorial regression was also investigated. For this study a fractional factorial design to degree 2 would include all main effects and 2-way interactions between each sensor. A full factorial design (i.e. degree 8) includes all main effect, 2-way, 3-way, 4-way, 5-way, 6-way, 7-way and 8-way interactions since there are 8 independent variables!

As a result of the large number of coefficients, factorial regression is a very adaptable technique. Results for the full factorial design are presented in Tables 6.17 and 6.18. Graphical representations are given in Appendix F. In comparison to the other techniques the models performed very well on the training sets. In a number of cases (subsets 2, 3, 5, 7, 8 and 11) all points were accurately predicted, while on the remaining subsets, the mean error was less than 18% (except subset 9 with 51%). But more importantly factorial analysis gave extremely poor results when applied to the rest of the data. Correlations between predicted values and RH were also very low.

The poor generalisation of the full factorial regression model is typical of an overfitting problem. As a result of the great adaptability of the technique, the models provide a very good fit on the training set but are also very dependent on these very datasets. This problem is similar to that of polynomial regression, only more pronounced. Although this is a non-linear approach, factorial regression does not show any improvement on other techniques. On the contrary it is a much more computationally demanding approach which is not well suited to real-time monitoring applications.

Table 6.17: 8^{th} degree Factorial regression predictions RAE and correlations on training sets and all data (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
Min RAE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE	9.95	0.00	0.00	0.94	0.00	4.19	0.03	0.00	3.02	3.32	0.00	6.81
Mean RAE	0.11	0.00	0.00	0.14	0.00	0.12	0.00	0.00	0.51	0.17	0.00	0.18
RAE StDev	0.57	0.00	0.00	0.17	0.00	0.20	0.01	0.00	0.58	0.27	0.00	0.30
R pred vs obs (training set)	0.66	1.00	1.00	0.72	1.00	0.68	1.00	1.00	0.20	0.65	1.00	0.56
R pred vs obs (Set12)	0.28	0.23	0.26	0.27	0.27	0.24	0.23	0.04	-0.20	-0.19	0.24	0.56
R pred vs RH (Set12)	0.57	0.38	0.29	0.54	0.53	0.44	0.43	0.00	0.18	-0.26	0.79	0.49

Table 6.18: Fraction of cases predicted with an RAE < x%, using 8^{th} degree factorial regression on training sets (all sensors)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set12
<10%	0.70	1.00	1.00	0.60	1.00	0.57	1.00	1.00	0.29	0.44	1.00	0.40
<20%	0.90			0.76		0.84			0.41	0.76		0.71
<40%	0.99			0.91		0.98			0.57	0.94		0.93

6.6 PRINCIPAL COMPONENTS AS INDEPENDENT VARIABLES

Despite our efforts to try and find a regression model based on the response of the eight CP sensors as input variables, no satisfactory solution has been found. In MLR the aim was to capture the correlation between the dependent and independent variables without considering the internal structure of the data. PLS on the other hand considered both the correlations between TOC and the sensor responses as well as the internal structure of the sensor responses in a one-step process.

Here, a 2-step approach is investigated, where only the internal structure of the IV is considered. A PCA was first performed on the IV's and the outputs (scores) were then used as an input to linear and non-linear regression models instead of the original sensor responses. In PCR, TOC is regressed onto the PC scores using MLR, whereas in the non-linear approach, polynomial regression (degree 2 and 3) will be used. The scores may be better suited for regression analysis since they are orthogonal, which solves problems associated with colinearity. In addition, reducing the number of PC's to those accounting for most of the variance may help improve generalisation of the models (see section 2.1.4.2).

In this study, we used subset 12 for the analysis. Figure 6.18 shows the plot of Eigenvalues which illustrates how much variance is extracted by the successive factors. This plot gives us a good indication of how many PCs can be retained. The point where the continuous drop in eigenvalues levels off suggests the cut-off where only random noise is being extracted by additional PCs. Here, that point could be at factor 2 or 3. According to Table 6.19, the first three PCs account for 98.4% of the variance in the data, with component 1 alone representing 91.17%. Additional information can be obtained by plotting the first 2 components loadings for each variable.

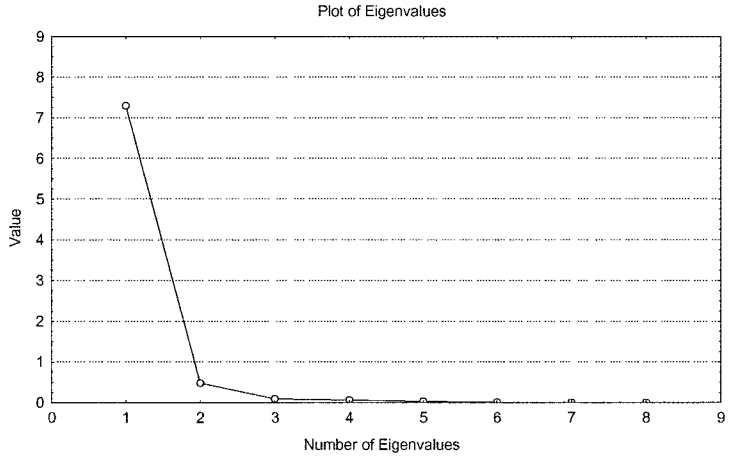


Figure 6.18: Plot of Eigenvalues for the eight extracted factors using PCA

Table 6.19: Eigenvalues, variance and cumulated values

	Eingenvalues	% total variance	Cumulated Eigenvalues	Cumulated varainace (%)
1	7.2934	91.1679	7.2934	91.1679
2	0.4823	6.0288	7.7757	97.1967
3	0.0960	1.2004	7.8718	98.3971
4	0.0671	0.8392	7.9389	99.2363
5	0.0359	0.4492	7.9748	99.6855
6	0.0133	0.1663	7.9882	99.8518
7	0.0064	0.0800	7.9946	99.9318
8	0.0055	0.0682	8	100

Figure 6.19 shows how the smallest loadings (components 1 and 2) are given for sensor 2. This illustrates the difference in the quality of the data provided by sensor 2 compared to the other sensors which tend to behave in a similar way and account for most of the variance in the data. This supports the observations in Sections 5.5.1 and 5.5.2 where sensor 2 was shown to be less humidity correlated and normally distributed. Although we could probably only retain the first two components for our analysis, the variance accounted for by component 1 is strongly associated with variations in humidity levels. Therefore the first 3 PC were used as inputs. The results of the regression analysis are presented in Table 6.20 and 6.21 and Figure 6.20 to Figure 6.22. We can see how the outcome is very similar for the 3 models tested and comparable to those of MLR, PLS or Polynomial regression alone. No improvement was gained as a result of using PCA as a first step prior to regression studies.

Again one of the major limitations appears to be associated with the effect of changing RH conditions during the data acquisition period. Figure 6.23 shows the relative prediction error for PCR together with recorded changes in RH for each case. The RH value itself does not seem to have any particular influence on the relative error and as seen in Figure 6.24 there is no particular value of RH where the models perform better. However it was demonstrated from laboratory studies (Section 4.4), that the stability of RH is a significantly important parameter. On the other hand, we saw in Section 6.2.1 that the value of the measured TOC has a major impact on the performance of the regression model. The same curvilinear relationship (as seen in Figure 6.6) between the prediction error and measured TOC can be observed for all models previously discussed on both TOC and Racod datasets.

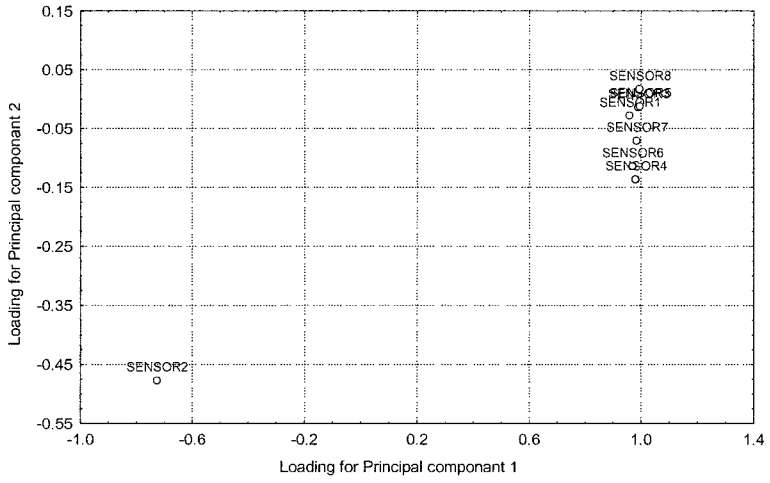


Figure 6.19: Plot of factor loadings. Factor 1 vs. Factor 2 extracted with PCA

Table 6.20: Predictions RAE and correlations using the first 3 PC's as IV's for different regression techniques (all data)

	MLR	Polynomial regression (2^{nd} degree)	Polynomial regression (3^{rd} degree)
Min RAE	0.00	0.00	0.00
Max RAE	6.39	6.49	6.32
Mean RAE	0.21	0.21	0.21
StDev	0.31	0.32	0.31

Table 6.21: Fraction of cases predicted with an RAE < x%, using the first 3 PC's as IV's for different regression techniques (all data)

	MLR	Polynomial regression (2^{nd} degree)	Polynomial regression (3^{rd} degree)
<20%	0.62	0.62	0.63
<30%	0.80	0.80	0.82
<50%	0.94	0.94	0.94

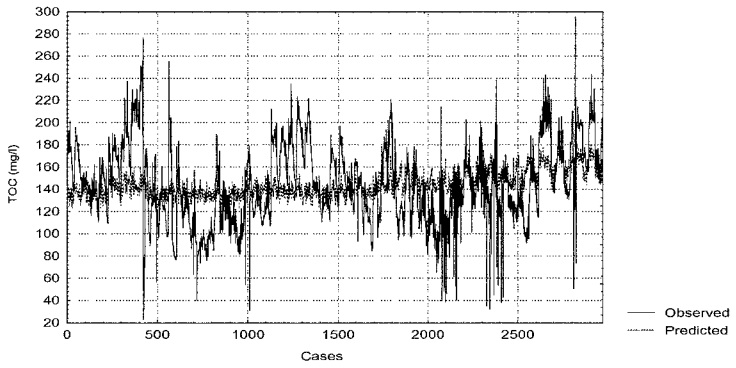


Figure 6.20: Observed and Predicted TOC values using PCR (all data)

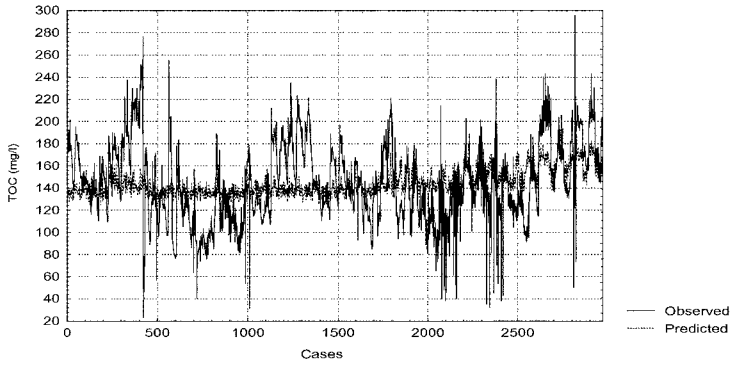


Figure 6.21: Observed and Predicted TOC values using 3 PC's + 2nd degree polynomial regression (all data)

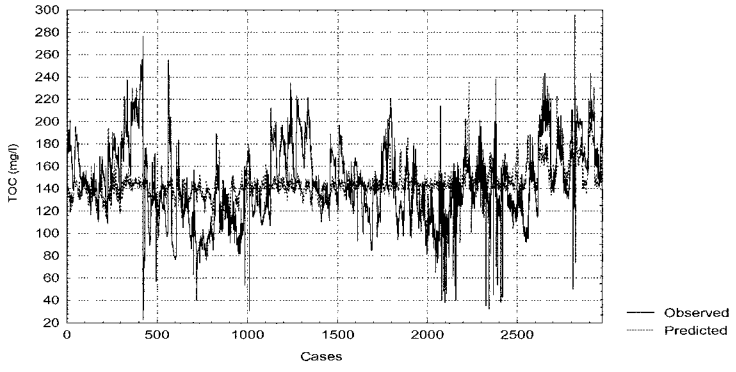


Figure 6.22: Observed and Predicted TOC values using 3 PC's + 3rd degree polynomial regression (all data)

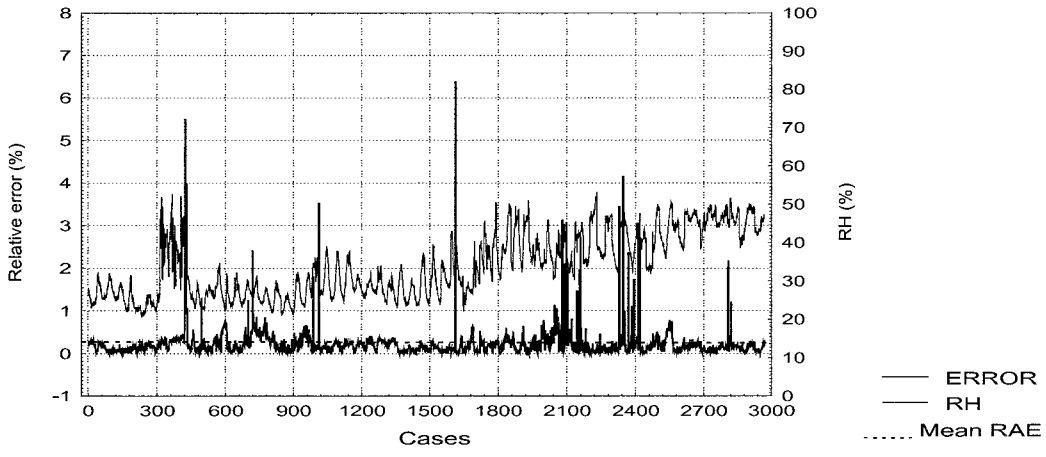


Figure 6.23: Variations in RH and prediction error (RAE) vs time using PCR (all data)

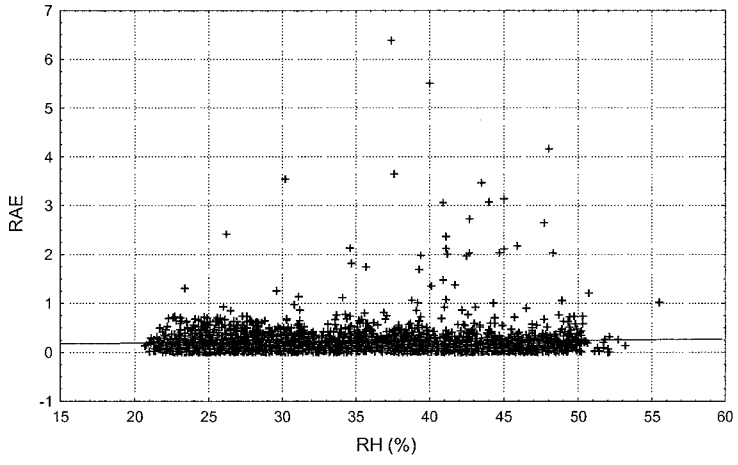


Figure 6.24: Plot of Absolute Relative Error (RAE) vs Relative Humidity (PCR, all data)

6.7 SUMMARY

- Linear (MLR, PLS, PCR) and non-linear (polynomial and factorial regression) multivariate sattistical analysis techniques have been used to investigate the relationship between wastewater organic load and sensor array data.
- MLR and PLS models performed relatively poorly and could not adapt to sudden changes in wastewater quality nor could they be generalised to periods of time of a few weeks or a few months.
- Predictions were limited to a narrow range centred around the average observed TOC concentration.
- Predicted values generally showed a better correlation with RH than TOC. These results reemphasised the importance of RH and duration of training.
- Graphical analysis of the residuals suggests a possible heteroscedasticity, although this was not obvious and does not invalidate the analysis.
- Polynomial and factorial regression were prone to overfitting and showed no improvement over the linear methods.
- No improvement was gained from using PCA prior to MLR or polynomial regression.

Chapter 7: ARTIFICIAL NEURAL NETWORKS FOR THE PREDICTION OF SEWAGE STRENGTH

CHAPTER 7: ARTIFICIAL NEURAL NETWORKS FOR THE PREDICTION OF SEWAGE STRENGTH

7.1 INTRODUCTION

In Chapter 6 we initially used a traditional approach and investigated the performances of multivariate statistical analysis techniques as quantitative tools for the prediction of TOC. Despite providing valuable insight on the quality of the data and helping in the understanding of the relationships studied, some major limitations became apparent. Among these, the difficulty in coping with the effect of RH (and temperature) as well as time related changes in wastewater odour profiles (seasonal changes, sensor drift) have been exposed. Therefore new and more flexible analysis techniques were needed. Neural Networks are free of the traditional assumptions and are well-suited for both linear and complex non-linear relationships. They are a perfect tool for exploratory analyses where the goal is to establish if a relation exists between a set of variables.

In this chapter the use of ANN for the prediction of TOC is evaluated. The principles of ANN have been discussed in Section 2.1.4.3 where it was pointed out that in many cases ANN can produce highly accurate predictions and outperforms statistical multivariate techniques. It is hoped that such methods will lead to a solution and will be able to cope with redundant or intercorrelated IV's, non-linearities and noise in the data, as well as those limiting factors mentioned above. With the variety of types of ANN available and the lack of general guidelines, the choice of a particular model for specific applications is still as much a trial and error decision as any.

A review of the recent literature (Table 2.4) showed that MLP with BP algorithm and Kohonen networks are the most commonly used and have proved to be well suited in a number of applications involving sensor array data analysis. Based on these observations, these two techniques and some variations thereof have been applied to our on-line data in a non-time series analysis. Data preparation and preliminary statistical examination have been performed and were discussed in Sections 5.4 and 5.5.

Among the few pre-selected datasets, Subset 6 was chosen for this study. The two main reasons for this choice are its relatively large and uninterrupted set of cases as well as the acquisition of data during two weekends. We have seen previously (Section 5.3.1) how the responses for both TOC and Prosat are significantly lower on Saturdays and Sundays. This characteristic may present an interest when training ANNs. The cases in the data were randomly divided into 3 subsets for the purpose of the analysis: training set, verification set and test set. The training set is used in training the network, the verification set is used to keep an independent check on the progress of training, and the test set is used to perform a final check on unknown data at the end of a sequence of experiments (i.e. prediction performance evaluation).

Similarly, the variables are divided into inputs and outputs. The observed TOC concentration was used as the target output (referred to as DV in MVS) and the 8 CP sensor responses as inputs to the network. Other potential inputs which have been considered in this study include RH, water temperature and gas flow rate. As noted in Bishop (1995), because of the limited response range of the commonly used transfer functions, neural solutions generally require pre-processing and post-processing stages to be used in real applications. In other words, numeric values have to be scaled into a range which is appropriate for the network. Thus, the data was scaled linearly so as to have the same maximum and minimum, using the built in Statistica "Minimax" transformation algorithm

7.2 MULTILAYER PERCEPTRON

We discussed in Section 2.1.4.3 how MLP is the most commonly used type of ANN. It is also one of the easiest and most rapid ones to implement. However, one of the most important issues, once the neural network architecture has been selected, is the choice of free parameters of that architecture. These issues that are particularly important when designing MLPs include the number of hidden layers, the number of units in these layers as well as the type of activation functions and error functions. The choice of these parameters depends on the complexity of the problem being modeled, which is generally not known in advance. This requires proceeding by experimentation, which is a time-consuming process.

Statistica's neural networks "Automatic Network Designer" was used to assist us in the optimisation process. By letting the built-in algorithm run overnight, it can conduct thousands of tests and consequently stands a better chance of designing a better network than even a skilled experimenter. More details on this procedure can be found in the software's user manual. On the other hand, the input variables were selected manually instead of using the Genetic Algorithm-based input selection (G.I.S') facility available with Statistica NN. ANN such as MLP can learn to ignore useless variables and are generally not affected by redundant variables.

With these considerations in mind, the following conditions were specified to the "Automatic Network designer":

- Network architecture: MLP
- Inputs: 8 CP sensor responses
- Output: TOC concentration.

A 3-layers MLP (i.e. one hidden layer) with 4 units in the hidden layer was selected as the best design for this problem (noted 8-4-1 MLP) and is represented in Figure 7.1. This is consistent with the accepted view that a one hidden layer MLP can approximate any continuous function. Increasing the network size may improve its ability to find a relation but also makes it more prone to overfitting.

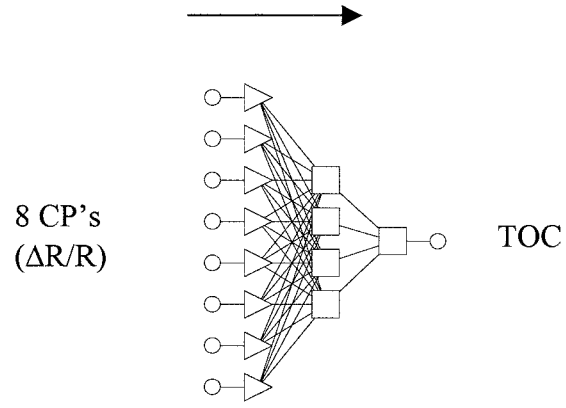


Figure 7.1: Illustration of an 8-4-1 MLP network showing the 8 inputs units, 4 hidden units and 1 output unit.

The next step was then to train the selected network using the training dataset. The Logistic function was chosen as the activation function since it is the most commonly used and performs adequately in most applications. Training was carried out using the Quasi Newton Learning algorithm to adjust the weights and thresholds. Quasi Newton is the recommended technique for most networks with a small number of weights and the most popular algorithm in non-linear optimisations. Preliminary tests using BP confirmed the assumption that Quasi Newton performs significantly better. In particular BP tended to be slower to converge on an error minimum and was more sensitive to noise in the training data.

The error function (used in training and reporting the error) can also have a significant effect on the performance of training algorithms (Bishop, 1995). Here the sum of the squared differences between the measured and predicted TOC values (Sum-Squared function) was used as the error function. This is the standard error function, and the most appropriate for regression problems (Statistica NN user manual). In order to avoid overfitting and determine when training should stop, cross-verification of the error on the verification subset was performed. The algorithm was instructed to stop training when a significant deterioration in the error on the verification set is detected. In other words, if the network over-learns the

training data, the training error may continue to decrease, but the verification error reaches a minimum and begins to rise. The algorithm will then cease training.

7.2.1 MLP with ΔR/R as input

Figure 7.2 and Table 7.1 show the results when using the 8 relative sensor responses as inputs to the 8-4-1 MLP described above. The performance of the network appears to be marginally better than those of the previously discussed multivariate regression models with an average error less than 14%, and more than 91% of the data predicted with an error inferior to 30%. However, it is still apparent that the network follows a diurnal pattern but generally fails to accurately predict lower weekend TOC levels or higher weekdays values. Attempts to increase the complexity of the network by adding a second hidden layer did not show any improvement, and a third attempt to include RH and/or water temperature as inputs to the model also proved of limited interest. As in MVS studies, the results suggest that the strong influence of RH variations on the sensors, tend to have a masking effect on any TOC-relevant information that may exist within the response profiles.

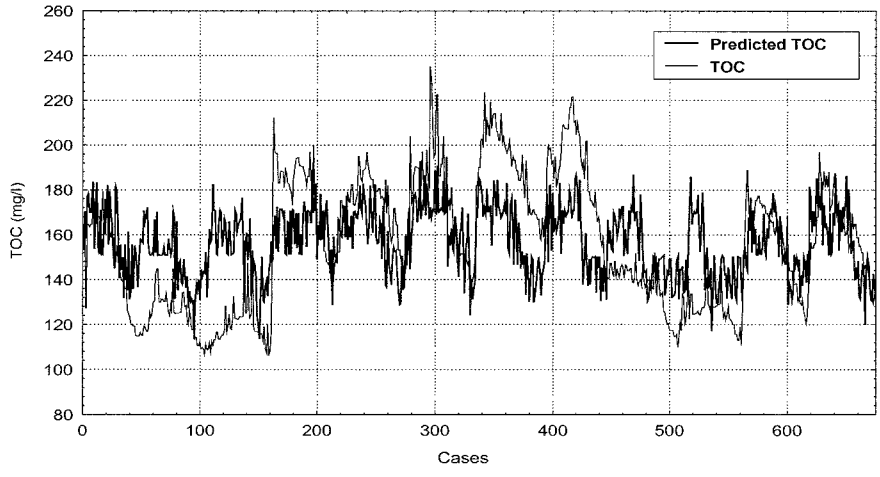


Figure 7.2: TOC prediction with an 8-4-1 MLP and 8CP's ΔR/R as inputs. Training: 475 cases, Verification: 100 cases and Test: 100 cases.

Table 7.1: TOC Prediction statistics for 8-4-1 MLP using the 8 sensors ΔR/R as input

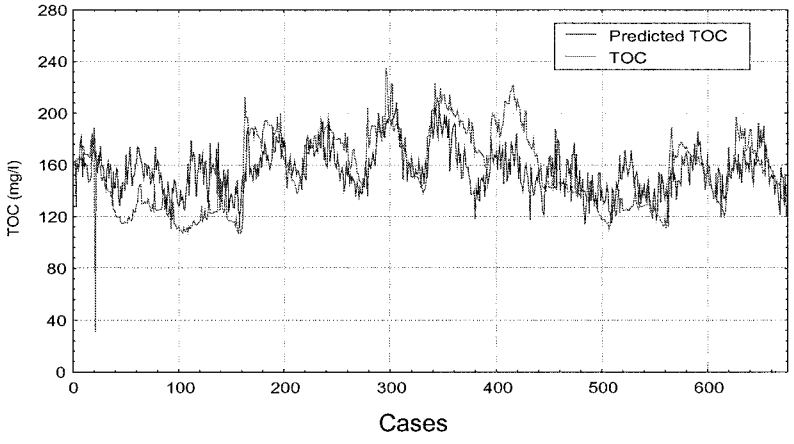
Test set		All Cases (R= 0.49)	
RAE	% Cases	RAE	% Cases
<10%	48	<10%	42.5
<20%	79	<20%	78.1
<30%	91	<30%	91.2
Max	50.5	Max	64.8
Min	0.2	Min	0.01
Average	13.1	Average	13.9

7.2.2 Noise Reduction

An important and common issue with most on-line data is the effect of noise on the quality of the analysis. Noise can be of different origins and there is a variety of techniques available to cope with noisy data. Commonly used noise reduction techniques include the application of smoothing functions (e.g. exponential smoothing), mathematical filters (e.g. Fourrier analysis) and the powerful but more complex wavelet transform analysis. However these approaches generally require a prior knowledge of the location of the relevant information in the frequency spectrum. The combined effect of RH on the sensor responses together with the synchronous diurnal pattern displayed by all variables make this task more difficult. Pertinent TOC concentration information may be hidden in lower frequencies as well as in high frequencies. There is therefore an important risk to remove this information when applying such filters without any knowledge of its whereabouts. Consequently, we used a simpler approach to reduce the noise in the sensor array data. With one Prosat acquisition every 5 minutes and a TOC measure every 30 minutes, we were able to calculate an average sensor response value from the two acquisitions closest in time to that of TOC (maximum 5 min before or after a TOC measurement).

Unfortunately, a major disadvantage with this approach is its inability to deal with a sudden change in TOC concentration if it occurred within this five-minute interval. Thus, a second approach was also tested. A median filter was applied to the sensor profile itself during the sensor response extraction phase (discussed in Section 4.2.3): An average value was computed from the sensor responses at 59 sec., 1 min. and 1 min. 01 sec. for each acquisition. In retrospect, this second approach showed no real interest since the individual response profiles are themselves quite smooth and virtually noise free as seen in Figure 4.2. This suggests that the origin of the noise is more experimental rather than strictly instrumental.

Figures 7.3 and 7.4 show a comparison of the predictions (8-4-1 MLP) using our first averaging approach reported above, and Statistica's simple exponential smoothing function respectively. Comparison with the results obtained from the original noisy data (Figure 7.2) show some improvement. In Figure 7.3, a better approximation of the lower weekend TOC levels is achieved although important variations still remain. In Figure 7.4, however, exponential smoothing gives a significantly better fit. The network is able to predict diurnal variations in TOC concentration as well as give a relatively good approximation on both weekends and weekdays. These observations demonstrate the negative effect of noise on the prediction and also underline the difficulties associated with the use of real on-line field data as a result of experimental or process noise (see Section 5.4.1.1).

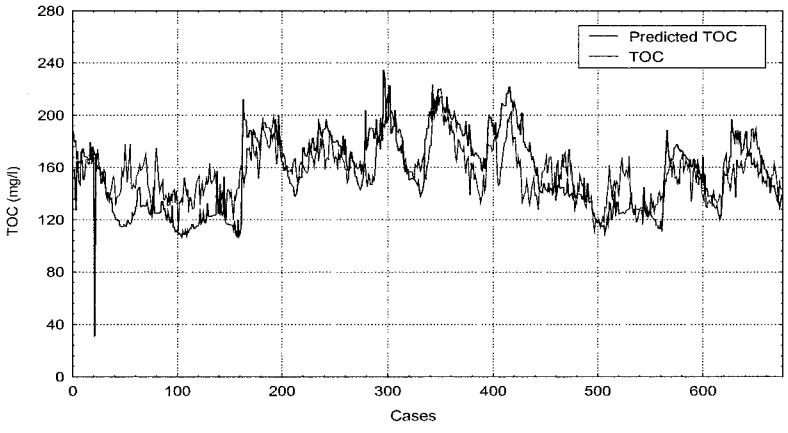


Average RAE	0.13
StDev	0.10
Min RAE	0.00
Max RAE	0.51

<10%	45.2%
<20%	77.7%
<30%	93.6%

R = 0.55

Figure 7.3: TOC prediction with an 8-4-1 MLP and averaged (n = 2) 8CP's ΔR/R as inputs. Training: 475 cases, Verification: 100 cases and Test: 100 cases.



Average RAE	0.11
StDev	0.09
Min RAE	0.00
Max RAE	0.54

<10%	54.4%
<20%	83.4%
<30%	96.1%

R = 0.71

Figure 7.4: TOC prediction with an 8-4-1 MLP and exponentially smoothed 8CP's ΔR/R as inputs. Training: 475 cases, Verification: 100 cases and Test: 100 cases.

7.2.3 MLP with Reduced number of sensors.

The strong correlation between RH and most sensor responses has been established in previous sections. In a similar approach to that of Section 6.2.3 a model was trained with a reduced number of sensors. We used sensors 1, 2 and 6 only as inputs to a 3-5-5-1 two hidden layers MLP network. These sensors were chosen because they are the least correlated with RH on this dataset. We can see from Figure 7.5 that removing the most humidity sensitive sensors did not improve the prediction. On the contrary, this had the opposite effect with the predicted TOC having a very limited range and showing a homogenous pattern independently of the day of the week. These results actually reflect the nature of the selected sensor responses which also show no significant differences between the weekdays and weekend profiles.

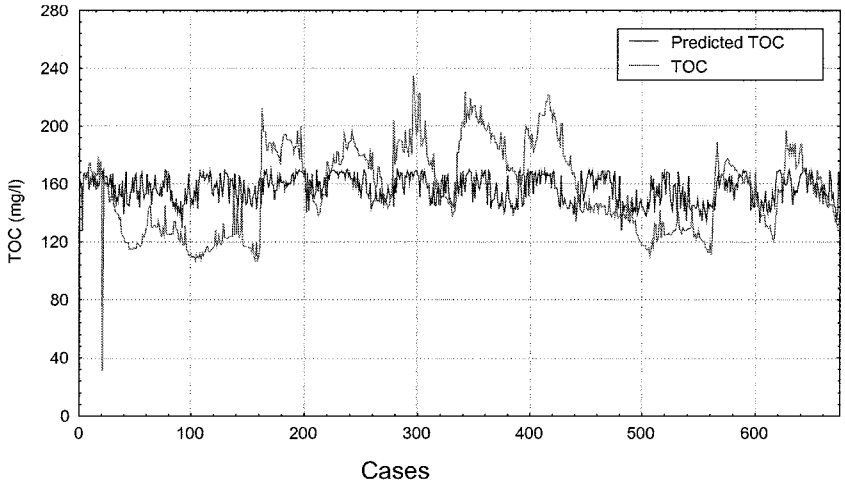


Figure 7.5: TOC prediction with a 3-5-5-1 MLP and sensors 1,2 and 6 (ΔR/R) as inputs. Training: 475 cases, Verification: 100 cases and Test: 100 cases.

7.2.4 Calibration with blanks

In order to improve training and provide the network with additional information, data obtained from continuously flowing tap water was used as a blank to calibrate the model. Figure 7.6(a) shows that the network is able to give a good binary approximation of the TOC levels. However, examination of the changes in the water temperature during the same period (Figure 7.7b) also displays a marked difference between blank data and wastewater data. This compromises any conclusions on the ability of the network to learn differences in the organic load of a water sample under changing temperature conditions.

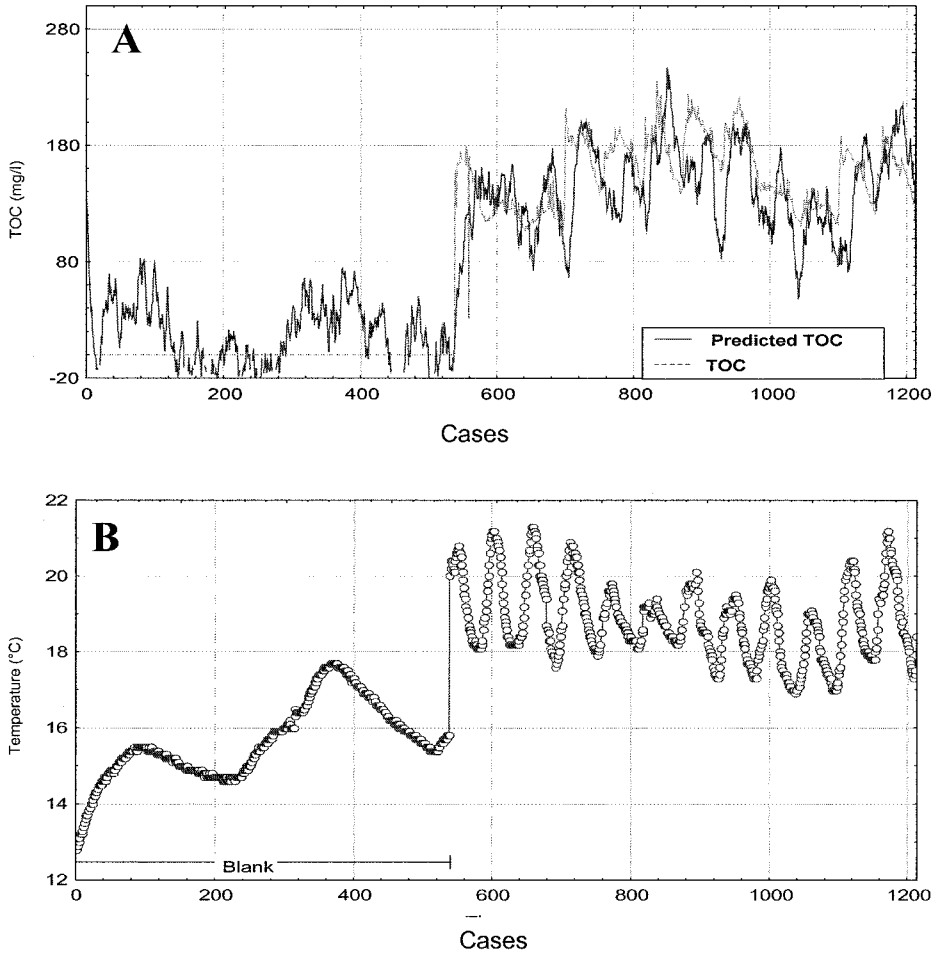


Figure 7.6: TOC prediction with a 8-4-1 MLP and 8 CP's ΔR/R as inputs (A) after training with blank data. Corresponding water temperature shown in (B)

7.2.5 Principal Components extraction

Selection of input variables is a critical part of neural network design. The performance of a network can be improved by eliminating unnecessary variables. We saw in Section 7.2.3 that an arbitrary reduction of the number of variables failed to improve prediction. The alternative approach to reducing the dimensionality of the data is to perform a PCA and use the extracted principal components as inputs to the network. The principles of PCA and how it can reduce the number of variables while retaining as much information as possible have been discussed in previous chapters.

A PCA was carried out on the 8 sensor responses as a pre-processing step prior to MLP training. Figure 7.7(A) shows the plot of Eigenvalues expressed as a percentage of the total extracted variance, with the first five PCs accounting for 99.5% of the total variance. In Figure 7.7(B) the strong correlation (R = 0.97) of the first PC (77.8% of total variance) with humidity can be observed. Similarly, a PCA was also performed on the 8 sensor responses after exponential smoothing. The effect of noise reduction on the extraction of variance can be seen in Figure 7.8(A) Here three components only can explain 99.4% of the total variance, with the first PC alone accounting for 95.4%. Figure 7.8(B) shows how the first PC is still strongly associated with RH variations (R = 0.90), although this relation appears more scattered.

The 8 extracted components were first used as inputs to an 8-4-1 MLP Network. Then, following the observation of a strong relationship between the first PC and RH (Figure 7.7 and 7.8), only components 2 to 8 were retained as inputs to a 7-4-1 MLP. Finally the effect of removing the three least important PCs (generally associated with noise) in a 5-4-1 MLP was also investigated. The results of the predictions are presented in Figures 7.9, 7.10 and 7.11 respectively. Comparison of Figures 7.9 and 7.10 shows the importance of the first component on the accuracy of the prediction despite its association with RH variations. On the other hand keeping only the five most important PCs did not significantly reduce the performance of the network. This is of real interest since it allows us to significantly reduce the complexity of the model (and computing time) while providing comparable results.

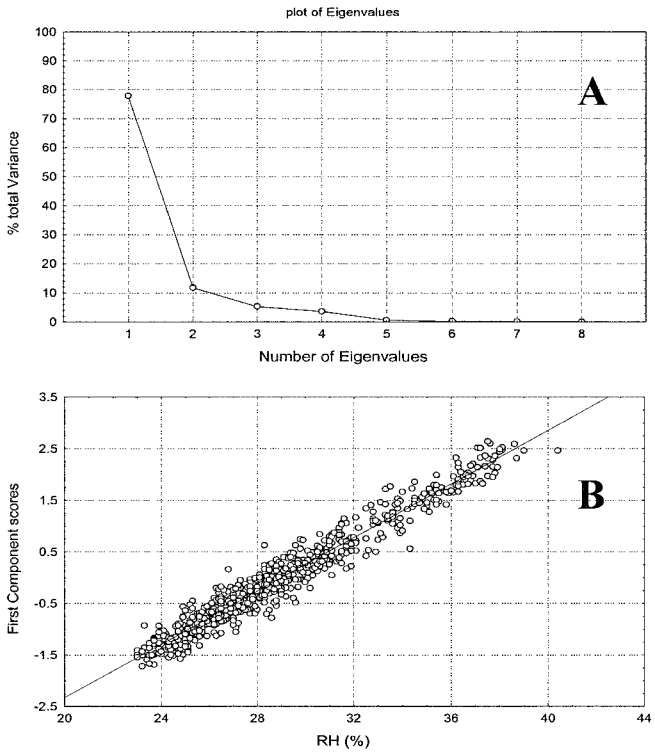


Figure 7.7: Plot of extracted Eingenvalues for relative sensor responses (A) and plot of first component vs RH (B)

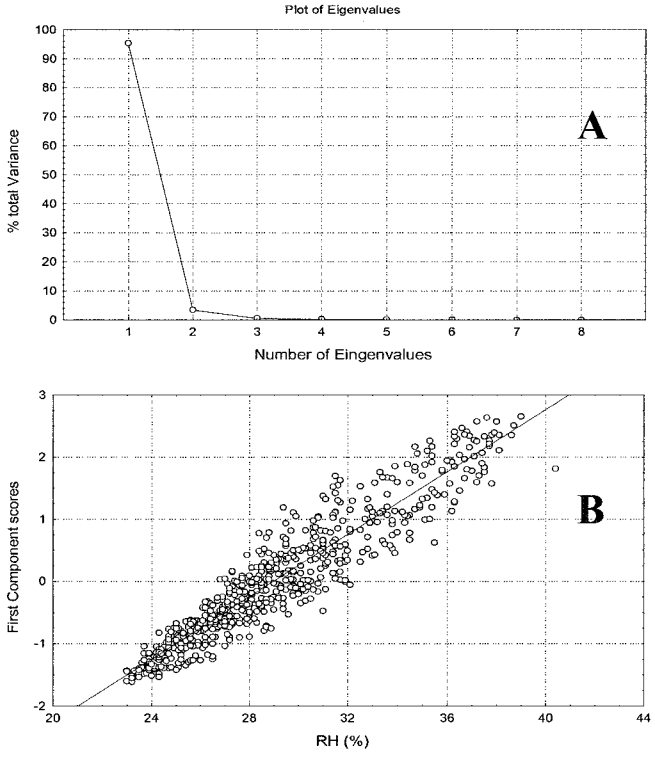
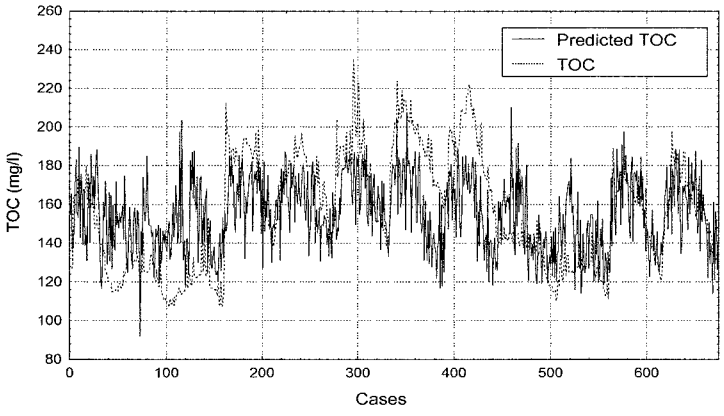


Figure 7.8: Plot of extracted Eingenvalues for smoothed sensor responses (A) and plot of first component vs RH (B)

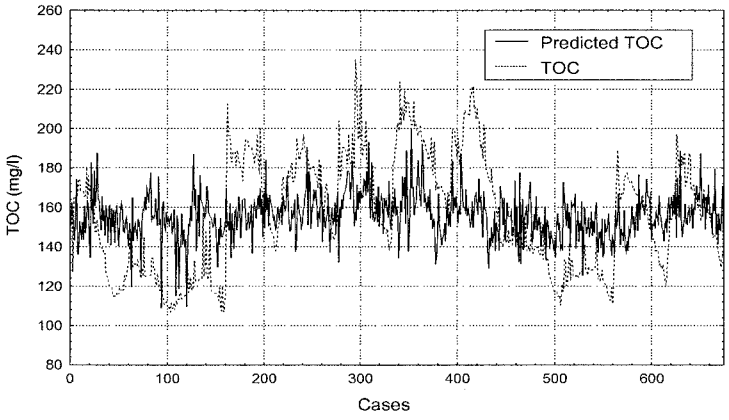


Average RAE	0.14
StDev	0.11
Min RAE	0.002
Max RAE	0.79

<10%	44.6%
<20%	77.8%
<30%	92.2%

R = 0.49

Figure 7.9: TOC prediction with a 8-4-1 MLP, using the 8 extracted PC's as inputs (unsmoothed data)

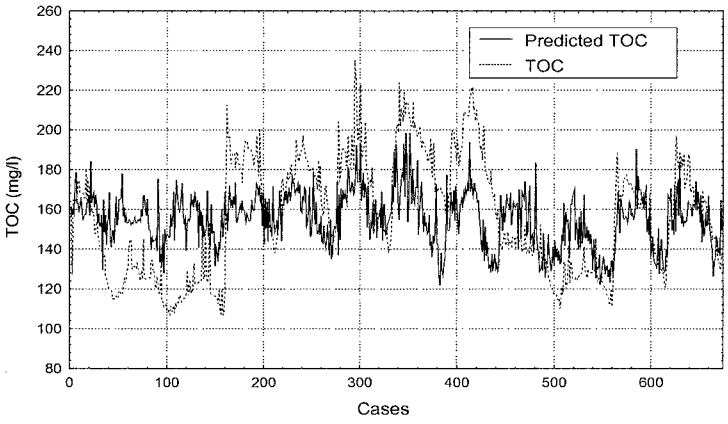


Average RAE	0.16
StDev	0.10
Min RAE	0.008
Max RAE	0.53

<10%	33.9%
<20%	71%
<30%	89.6%

R = 0.30

Figure 7.10: TOC prediction with a 7-4-1 MLP, using components 2 to 8 as inputs (unsmoothed data)



Average RAE	0.14
StDev	0.10
Min RAE	0.000
Max RAE	0.58

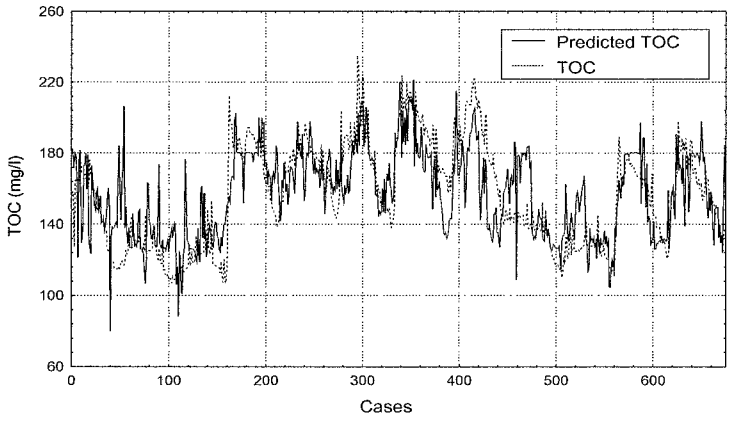
<10%	40.7%
<20%	77.2%
<30%	91.3%

R = 0.47

Figure 7.11: TOC prediction with a 5-4-1 MLP, using components 1 to 5 as inputs (unsmoothed data)

The same approach was then repeated with the data exponentially smoothed prior to PCA extraction. Figures 7.12, 7.13 and 7.14 show the results of the 8-4-1 MLP (all components), 7-4-1 MLP (components 2 to 8) and 5-4-1 MLP (components 1 to 5) respectively. As discussed in Section 7.1.2, the results achieved after noise reduction are much better than those obtained from noisy data. Again, comparison of Figures 7.12 and 7.13 show the importance of the first component on the accuracy of the prediction, although the decrease in performance seems relatively less pronounced. Removing components 6 to 8 on the other hand, had a more important effect than doing so on the noisy data. This difference can be explained by the fact that noise has already been removed, thus leaving more pertinent information in those components generally associated with noise.

These results demonstrate that more benefits can be gained from noise reduction via exponential smoothing than by reducing the dimensionality of the data with a PCA. The dimensionality of the data is comparatively not so much of an issue and there is no interest in performing a PCA after a noise reduction step. However PCA can still be useful if performed on noisy data and a 5-4-1 MLP using only the first five PC's gives comparable results to a full 8-4-1 MLP. This approach has the double advantage of reducing the dimensionality of the problem while at the same time removing some of the noise which is mostly present in the last PC's. However, since computing power and the complexity of the model is not an issue here, the first approach of using exponential smoothing without dimension reduction appears to be a better choice. Parallel studies on different datasets invariably supported these observations. This is illustrated in Figure 7.15 (09.04 to 16.04.01) which confirms the interest of exponential smoothing as a pre-processing step. The 8-4-1 MLP network gave a good prediction of wastewater TOC despite a number of missing values in the dataset. The results can be compared to those from Figure 7.4

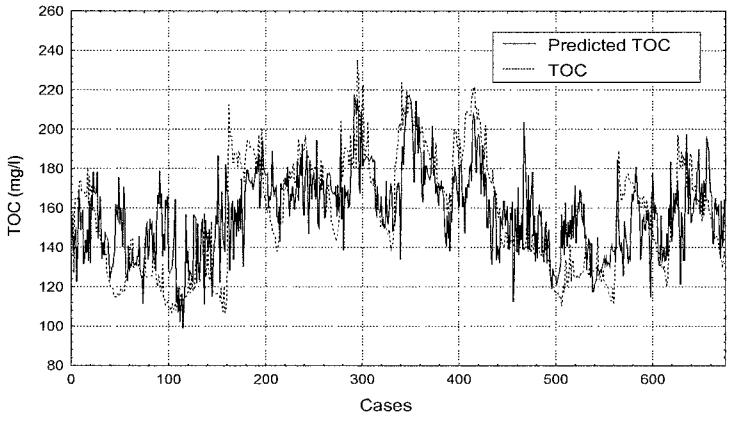


Average RAE	0.10
StDev	0.09
Min RAE	0.009
Max RAE	0.77

<10%	61%
<20%	85.6%
<30%	96.4%

R = 0.74

Figure 7.12: TOC prediction with a 8-4-1 MLP, using the 8 extracted PC's as inputs (smoothed data)

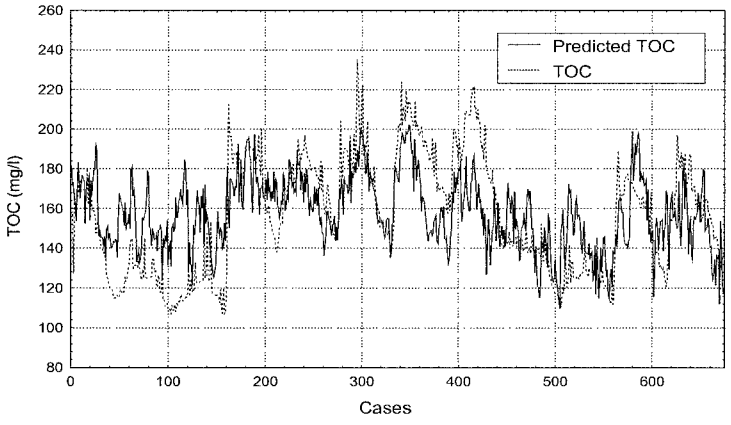


Average RAE	0.12
StDev	0.10
Min RAE	0.003
Max RAE	0.68

<10%	54.5%
<20%	87%
<30%	94.7%

R = 0.66

Figure 7.13: TOC prediction with a 7-4-1 MLP, using components 2 to 8 as inputs (smoothed data)

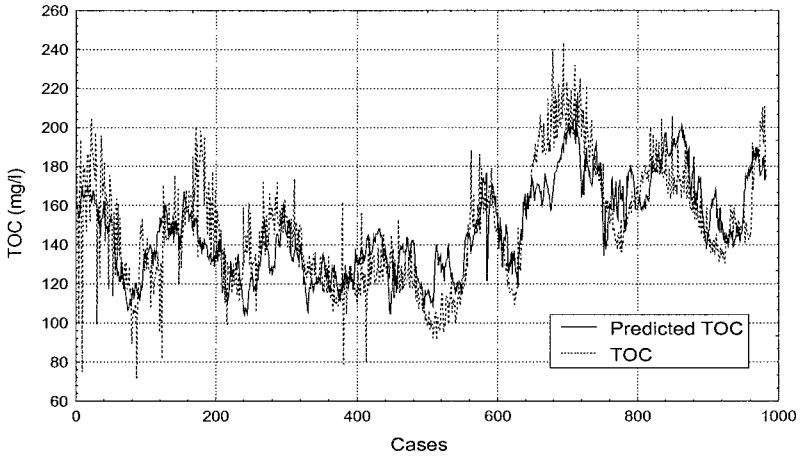


Average RAE	0.13
StDev	0.11
Min RAE	0.00
Max RAE	0.62

<10%	51.9%
<20%	79.2%
<30%	91.7%

R = 0.60

Figure 7.14: TOC prediction with a 5-4-1 MLP, using components 1 to 5 as inputs (smoothed data)



Average RAE	0.11
StDev	0.12
Min RAE	0.00
Max RAE	1.57

<10%	58.2%
<20%	88.8%
<30%	97.2%

R = 0.76

Figure 7.15: TOC prediction with an 8-4-1 MLP and exponentially smoothed 8CP's ΔR/R as imputs. (09.04 to 16.04.01). Training: 575 cases, Verification: 200 cases and Test 200 cases.

7.3 KOHONEN NETWORKS

Whereas MLP networks are designed for supervised learning task and are therefore adapted to regression studies, Kohonen Networks are primarily designed for unsupervised learning. Kohonen Networks do not use outputs, but instead learn the structure of the data. Consequently they are generally used in exploratory data analysis to recognise clusters of data or to perform classification tasks. Despite not being able to infer a correlation between the sensor responses and TOC concentrations, KNN may still present an interest for on-line monitoring of wastewater quality. For instance, as different classes of data are recognised, they can be labeled, so that the network becomes capable of differentiating between different levels of organic pollution (e.g. Low, Medium, High). Another potential use for KNN is novelty detection. For example, should new or unusual profiles be recorded in the case of a pollution incident or other anomalies, the network would fail to recognise it, and this could trigger an alarm or isolate the new data for further examination.

In the dataset studied here (subset 6), we can split the data into two distinct levels: weekend conditions (TOC < 150 mg/L) and normal weekdays conditions (TOC > 150 mg/L). As with MLP, the same training set, verification set and test set were used, with the 8 sensor responses as inputs. In effect, KNN have two layers, the input layer and the two-dimensional topological map (output layer) where related clusters are grouped together. A KNN with 50 units in the output layer gave relatively good classification results with 89%, 75% and 75% of the cases correctly identified for the training set, verification set and test set respectively. We then tried to refine the analysis by separating the data in an increasing number of subsets corresponding to different TOC concentration ranges. When divided into 7 classes, only 58% of the cases were correctly classified on the training set. As a rule, performances decreased with increasing number of classes. This trend was expected as narrowing the width of the clusters while at the same time increasing the number of "border areas"

between these groups increases the potential for misclassification. Other limitations with this approach may come from possible effects of the distribution of the training data. In a normally distributed dataset, the clusters or classes further away from the median will have less cases available for training which could lead to a higher recognition error than on those classes with a higher number of classes. For these reasons investigations on the use of KNN to differentiate between different TOC levels was not pursued. However the knowledge that KNN can distinguish between two different types of wastewater according to their organic strength is of real interest for pollution detection studies.

7.4 BASELINES AND RAW SENSOR RESPONSES

In all studies discussed earlier, we used the relative change in the sensors' resistance ($\Delta R/R$) to characterise the samples headspace. These were computed using the instrument's default extraction feature as recommended by the manufacturer. This approach allows for intercomparison of the sensor profiles obtained for different samples and compensate for the effect of drift, but does not provide any information on the drift itself, nor on the state of the sensors just before the acquisition period. Observation of the sensor baselines (resistance at 0 sec.) give us an indication of the evolution of the sensor condition (ageing, poisoning) with time. Figure 7.16 (a, b) shows the sensors baselines and corresponding extracted relative sensor responses ($\Delta R/R$ at 60 sec) obtained for a two month period of wastewater monitoring (23.01.01 to 15.03.2001).

Figure 7.16 (a), illustrates how sensors 1, 3 and 8 are particularly affected by drift with a 13% change in resistance for sensors 1 and 3 (this reached 20% after 6 months). The interest of using the relative sensor response to deal with such drift is demonstrated in Figure 7.16 (b). However the fact that sensors 5, 6, 7 and 8 baselines display the same diurnal pattern observed in Figure 7.16 (a) is a cause for concern. It indicates that the sensors did not fully return to their original baseline value after the clean up (depurge) period. Figure 7.17 shows a closer view of the sensor baseline

changes and clearly suggest the influence of the previous examples on measurements. In fact the effect on the sensor baseline is not systematic but occurs if a threshold concentration level of RH, global organic pollution or individual substance (or a combination thereof) is reached. Sensors 1, 2 3 and 4 (CR3's) appear to be generally less affected which shows that the type of sensor is also an important factor.

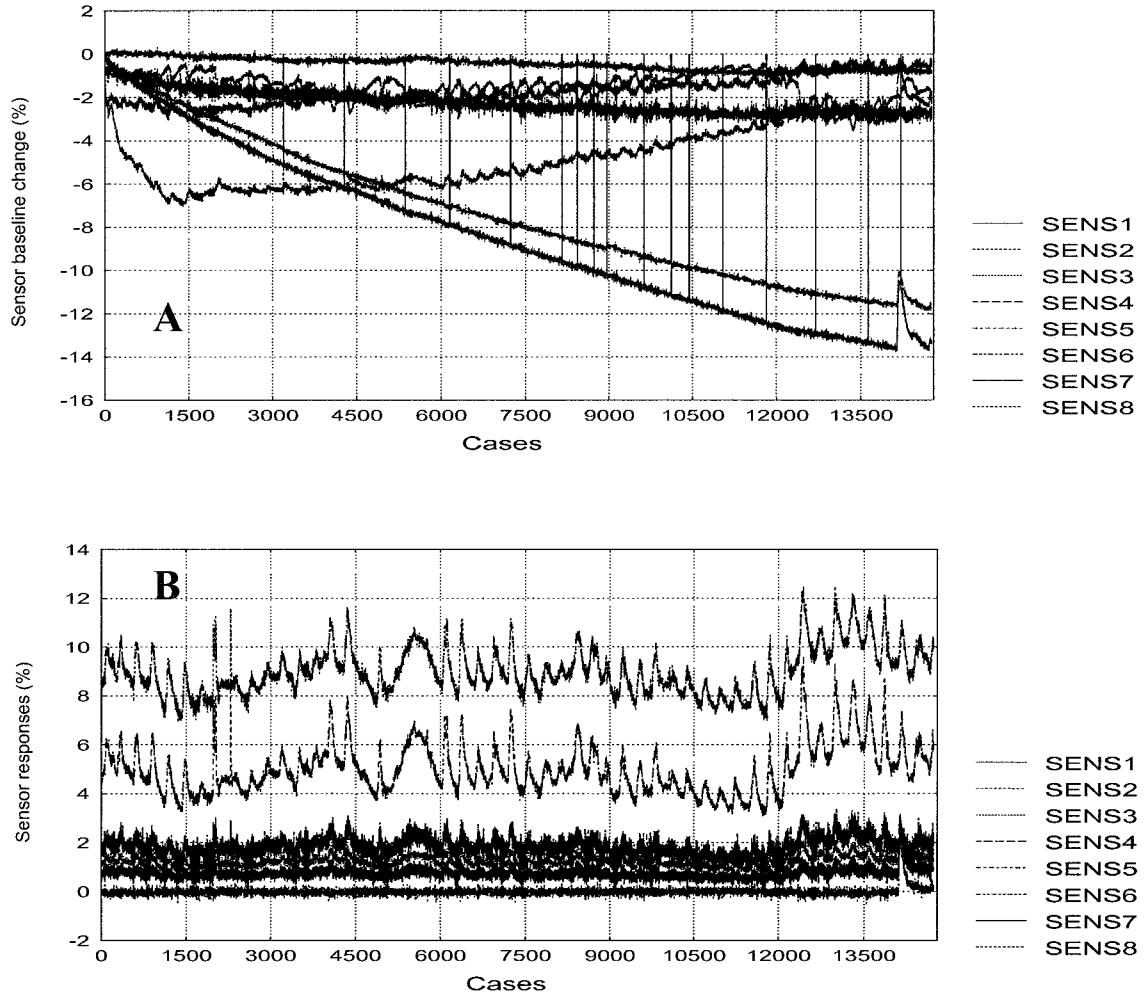


Figure 7.16: Plot of sensor baselines (A) and relative sensor responses at 1 min (B) from 23.01.01 to 15.03.01 (wastewater ring main)

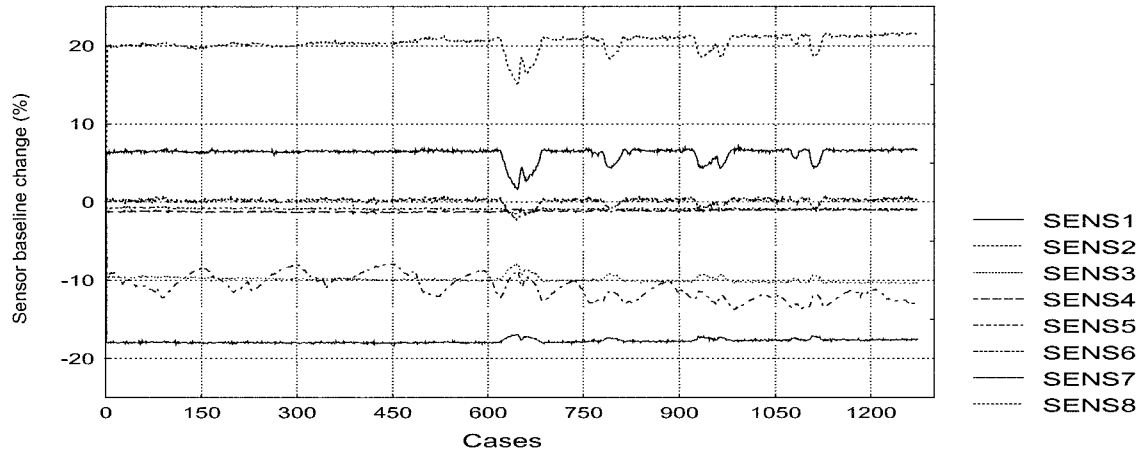


Figure 7.17: Plot of sensor baselines (20.06.01 to 28.06.01, wastewater ring main)

With regard to data analysis, these observations may have a major significance on the validity of the results presented in the previous section and could have contributed to the poor performance of some of the regression models that have been discussed. Because calculation of $\Delta R/R$ is directly affected by this phenomenon, a study was carried out that used the measured sensor resistance at 1 min (unsmoothed) as inputs to an 8-4-1 MLP. Following the same procedure as previously, the sensor responses were standardised using Statistica's Minimax conversion feature. Figure 7.18 shows the standardised response for sensors 1, 3, 4 and 5 (dataset 6) with the effect of drift on sensors 1 and 3 clearly visible. The results of the prediction of TOC with the 8-4-1 MLP (quasi-Newton) are presented in Figure 7.19. A comparison with Figure 7.2 and Table 7.1 where the relative sensor responses were used as inputs shows a major improvement. This significantly better performance suggests the importance of the information that is lost in the calculation of $\Delta R/R$ because of irregularities in the baselines. The effect of drift however can strongly limit the long-term validity of the model if using the raw sensor resistances as input.

Futures studies should therefore use the recommended $\Delta R/R$ but ensure that the sampling procedure is adequate for the type of application and sensors being used. It

is expected that this carry-over effect can be reduced by increasing the sensor clean-up stage and/ or decreasing the duration of the acquisition period which may require reducing the sampling frequency.

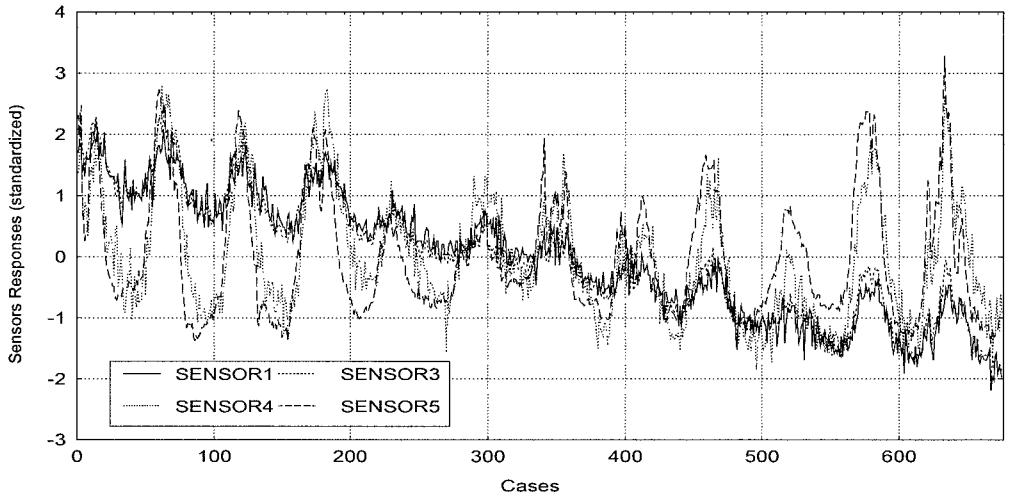
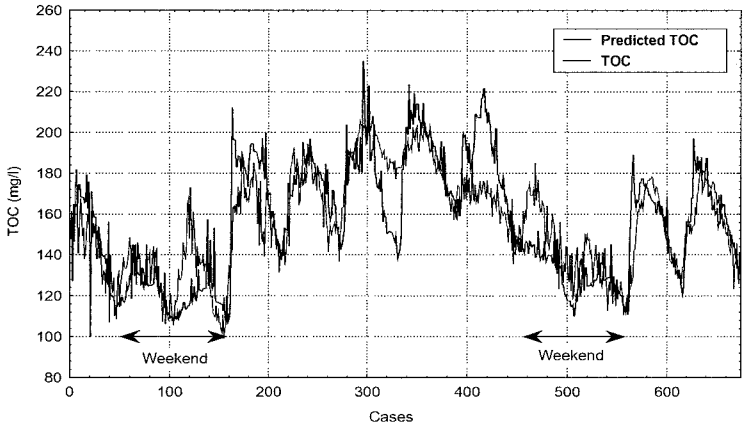


Figure 7.18: Standardised resistance of sensor 1, 3, 4 and 5 at 1 min.



Average RAE	8.4%
Std Dev	7.3%
Min RAE	0.02%
Max RAE	47.6%

<10%	67.1%
<20%	91.5%
<30%	98.5%

R = 0.80

Figure 7.19: TOC prediction with a 8-4-1 MLP and standardised sensor resistances (at 1min) as inputs.

7.5 WATER TEMPERATURE CONTROL AND CALIBRATION FIELD EXPERIMENT

Results in previous sections all pointed to RH and temperature variations as the major limiting factor in establishing a relation between sensor array data obtained in the field and wastewater organic load measurements such as TOC. Despite our attempts to numerically compensate for the effect of these factors using multivariate statistics or using RH and temperature as inputs to artificial neural networks, we have not been able to achieve performances comparable to those obtained with laboratory experiments. This demonstrates that physical control of experimental conditions is a more effective way to achieving satisfying results.

However, the principal objective of this research being to evaluate the application of an on-line sensor array system in the field, such control is highly impractical and cannot be considered as an option. In addition to changes in environmental conditions, the distribution of the data and the limited concentration range within which TOC variations occur, also limit the accurate prediction of values outside this range. This was illustrated in Figure 6.5 where the predictions are mostly centered around the average TOC concentration of 145 mg/L. As seen previously all parameters (RH, Temp. and TOC) show the same simultaneous diurnal pattern and are also strongly correlated with the sensor response. This particularity in the data does not allow us to separate the effect of each individual parameter on the sensors.

With these considerations in mind, a new in-situ experiment was carried out in order to acquire data that could be used for building a robust calibration model. Such model could then be evaluated against previous and subsequent data from the ring main. The experiment was statistically designed following a similar approach to the one described in Section 4.4. The aim was to study and quantify the individual effects of wastewater temperature, N_2 carrier gas temperature and organic load on the sensor responses as well as the interactions between one another. A 300 litres insulated tank was filled with raw sewage from the primary settling tank and diluted

during the course of the experiment to 50% and 25% with tap water (previously sparged with air for 48 hours to remove chlorine). A blank consisting of tap water only was also carried out. Water temperature was kept at 8 °C and 18 °C (+/- 1 deg. C) with a submerged heater cooler system. Two separate N_2 gas cylinders were used for rapid change of carrier gas temperature: cold (ambient) and warm (cylinder wrapped with electrically heated tapes and insulation). Gas temperature was recorded as usual by the PROSAT during acquisitions. At all times the water was re-circulated through the tank and aerated to avoid sedimentation and anoxia. Sensor responses, TOC, RACOD and manual COD and BOD5 were measured throughout the experiment. The design matrix of the experiment is presented in Table 7.2

Table 7.2: Design matrix for temperature control experiment in the field

Time	WW dilution	Water temperature	N_2 Temperature
12:00	100%	18°C	Ambient
18:00	100%	18°C	Warm
24:00	100%	08°C	Warm
06:00	100%	08°C	Ambient
12:00	50%	08°C	Ambient
18:00	50%	08°C	Warm
24:00	50%	18°C	Warm
06:00	50%	18°C	Ambient
12:00	25%	18°C	Ambient
18:00	25%	18°C	Warm
24:00	25%	08°C	Warm
06:00	25%	08°C	Ambient

A rapid on-site examination of the data showed that humidity levels were strongly affected by the water and gas temperatures. More interesting was the observation that changes in the organic load could be achieved independently of RH variations. Unfortunately, a few days after the end of the experiment a series of power cuts during a storm caused the Prosat instrument to crash, resulting in physical damage of the PC's hard drive and a loss of all unsaved data. Delays in obtaining a replacement instrument and logistic problems (availability of water coolers and re-circulation pumps) did not allow us to repeat this experiment in the field.

7.6 SUMMARY

- An 8-4-1 MLP gave better results than the traditional statistical techniques but failed to accurately predict lower weekend values.
- Exponential smoothing of the sensor array data significantly improved the models' performances and demonstrated the negative effect of noise on the prediction.
- Reducing the number of imputs to the less RH sensitive sensors worsened the prediction, suggesting that important information is present in the most RH dependent sensor responses.
- Calibration with clean water data could not be carried out due to the different levels in water temperatures.
- The use of PCA prior to MLP training improved the predictions. The first extracted component was strongly correlated with RH but was essential as an imput to the model.
- Reducing the dimensionality of the data was not crutial. Simple noise reduction via exponential smoothing was sufficient and recommended.
- Kohonen networks could differentiate between two different TOC concentration ranges (WE/weekdays) and may be of interest for pollution detection studies.
- An important drift was observed for sensors 1,3 and 8 but did not affect the analysis.
- Poisoning of sensor 5, 6, 7 and 8 had a major effect on the calculated $\Delta R/R$ and should be prevented in future experiments

Chapter 8: A MODEL FOR THE DETECTION OF UPSET EVENTS AND PROCESS CONTROL ANOMALIES

CHAPTER 8: A MODEL FOR THE DETECTION OF UPSET EVENTS AND PROCESS CONTROL ANOMALIES

8.1 INTRODUCTION

Wastewater that arrives at a municipal sewage works is highly variable in nature and the influent to be treated can be of different origins such as domestic and industrial sewage and surface run-off. Intermittent or accidental discharge of chemical pollutants and toxic substances into the sewers can have a damaging effect on the bioprocesses involved in treating wastewater. Consequently, polluted waters have the potential to pass through a treatment works untreated and reach the receiving waters where they can have a harmful effect on the environment and threaten drinking water abstraction points down the river. Traditional monitoring techniques are resource consuming and do not facilitate early warning of process failure. Additionally, they cannot provide a high-resolution picture of the nature and variations in wastewater quality and expose companies to the risk of undetected incidents. Results from the field are presented that show the effect of unknown acute pollution events and controlled diesel injections on the sensor response. A simple data mining approach for the rapid on-line detection and identification of anomalies is proposed as described in Bourgeois *et al.* (2002, b).

8.2 THE EFFECT OF UNKNOWN POLLUTION EVENTS

Over the course of our continuous monitoring programme a range of anomalies have been recorded but formal identification of the cause of a particular sudden change in the sensor profiles is generally difficult. Such identification relies mainly on manual logs of events and experience gained in the field over long periods of time. Figure 8.1 shows the sensor responses of an unknown pollutant being detected in the wastewater influent at the sewage work on the 13th of March. The rapid increase in the sensor profiles of 4 of the 8 sensors (sensors 1, 2, 3 and 4) corresponded with reports of a petroleum smell at the sewage works and with results from the on-line TOC analyser. Sensors 5 to 8 did not show significant changes in their response as a result of this incident but followed the usual diurnal patterns described earlier. This behaviour demonstrates the interest of using an array of broad selectivity for this type of application. The combination of sensors allow for the detection of a pollution incident (sensors 1-4) while providing simultaneous information on the normal change of the influent quality (sensor 5 and 8).

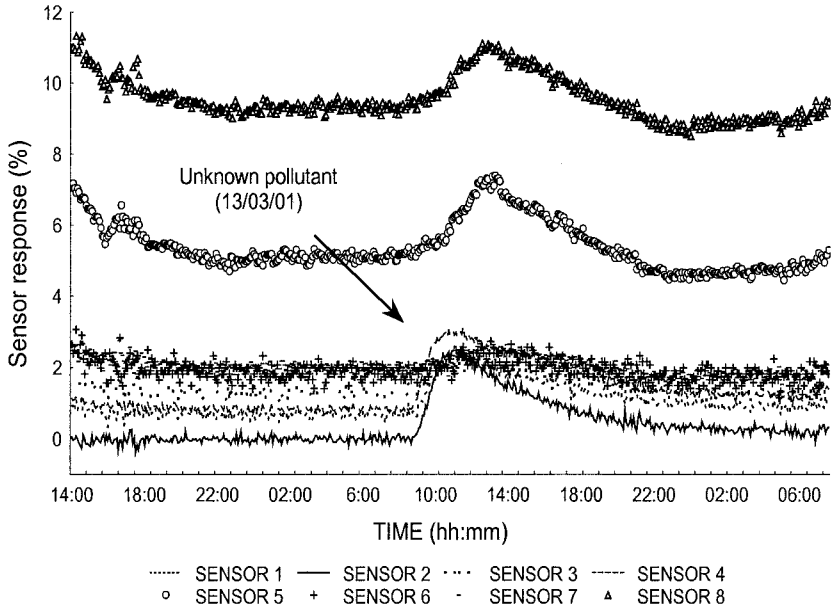


Figure 8.1: Plot of sensor responses showing the detection of an unknown discharge in the wastewater influent.

In Figure 8.2, a principal component analysis of the data presented in Figure 8.1 clearly shows the evolution in time of the intermittent discharge. This suggests that the unknown pollutant was present in the wastewater treatment plant for approximately 24 hours. These observations corroborate reports of a residual smell the following morning and demonstrate the sensitivity of the system. After reaching a peak within 2 to 4 hours, the polluted wastewater then gradually diluted and eventually reverted to its original quality at approximately 18.00 the following day.

This contrasts with the TOC measurements (Figure 8.3) which only show a significant difference in the organic load for a 6 hours period (TOC levels were down to 178mg/l at 17:23). TOC concentration increased from 147 mg/l to 240 mg/l between 09:53 and 10:53, which is significantly higher than the normal change (136 mg/l to 169 mg/l) recorded at the same time on the previous day.

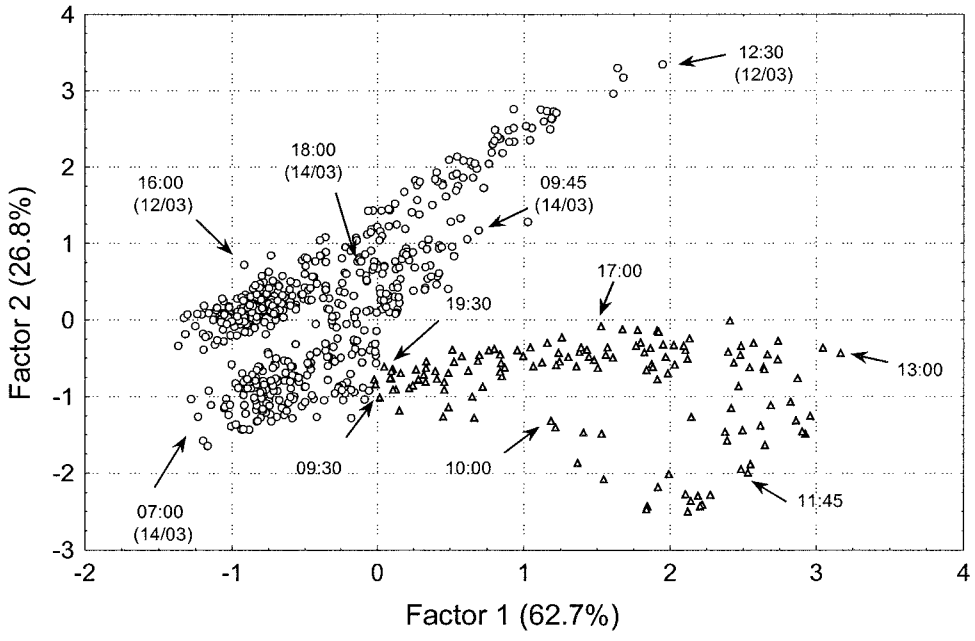


Figure 8.2: Plot of principal components showing the separation of an unknown discharge in the wastewater influent (represented by triangles), and a gradual return to its original quality.

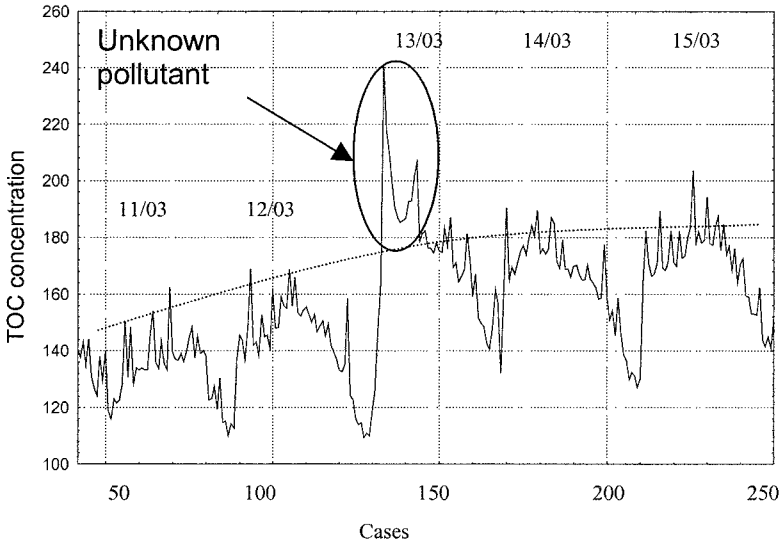


Figure 8.3: TOC profile (11/03/01 to 15/03/01) showing the effect of an unknown pollutant discharge in the wastewater collection system on the 13th of March 2001

The nature and origin of this pollutant remains unclear and a specific identification would require further investigations that are beyond the scope of this study. Nevertheless, a few possible sources can be listed

- Cranfield university has its own airfield with airplane servicing and refuelling areas and a range of compounds such as kerosene or antifreeze may accidentally find their way into the sewer system.
- Laboratories and workshops in many departments of the University also represent a risk of accidental discharge of toxic chemicals.
- Finally, road run-off from recent resurfacing work (tarmac) or unreported traffic incidents (leaks, crash..) is also a likely source of pollution which could upset the wastewater treatment processes.

8.3 SIMULATED INCIDENT

In an attempt to simulate similar pollution episodes, different concentrations of diesel were injected into the pre-sample vessel on April 17th and 18th (0.2% V/V and 0.4% V/V, respectively). Figure 8.4 shows the sensor responses for these additions. As in Figure 8.1 a very similar sensor response pattern to that of the unknown pollutant can be observed for sensors 1, 2, 3 and 4 showing some sensitivity. Given the small quantities of diesel added to the wastewater, a much more rapid return to normal could be observed within 15-20 minutes (3-4 acquisition cycles) for all sensors. The observed differences in the magnitude of the sensor response (on April 18th) shows the result of the different mixing time used in the pre-sample vessel prior to sampling and sensor array analysis and reflects the concentration effect of diesel in wastewater with time.

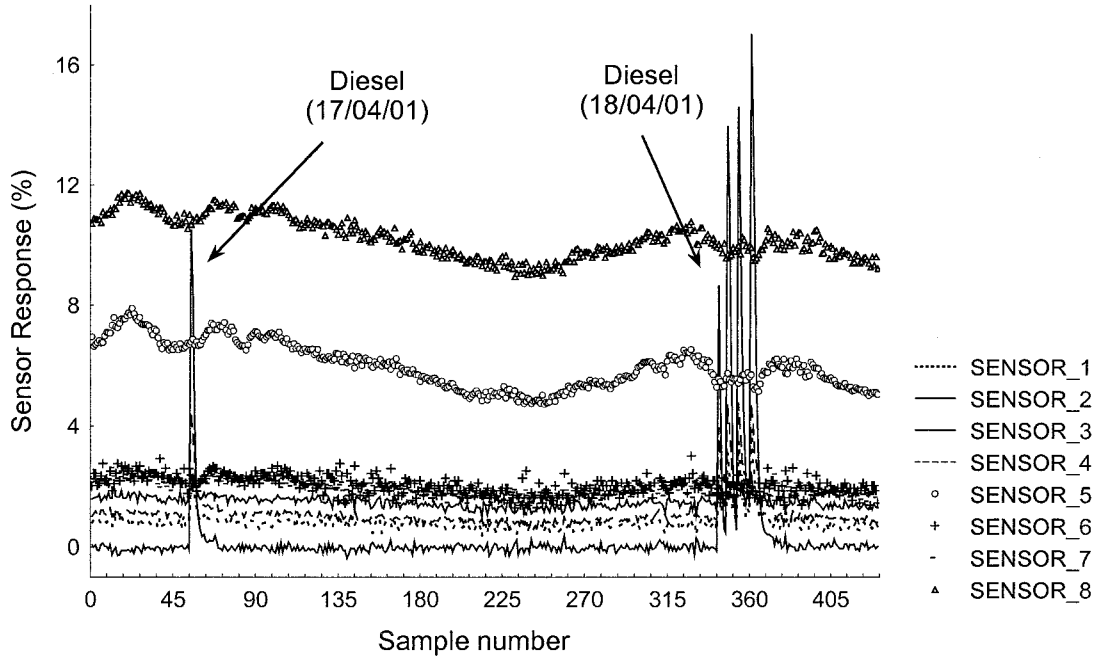


Figure 8.4: Plot of sensor responses showing the detection of diesel spikes (0.2% V/V and 0.4% VV) in the wastewater on two consecutive days (17.04.01 and 18.04.01 respectively).

8.4 DETECTION ALGORITHM

Over a 12-month sampling period, it is highly likely that the quality of wastewater will vary due to changing environmental conditions (such as rainfall) and intermittent industrial discharges. Figure 8.5 shows the relative responses of 3 sensors over the first 6 months of continuous monitoring. While a great number of anomalies can be observed in these sensor responses, such episodes may not always be detectable by simple visual examination of individual response profiles. Since manual screening of such large datasets would not be practical and extremely time consuming, the development a rapid screening method for the real-time detection of pollution incidents and operating anomalies is needed.

In Figure 8.6, a study of the sensor response (using data from Figure 8.5) versus relative humidity provides a simple but effective way to visually detect unusual patterns. These plots show that a large number of incidents can be detected over long periods of time independently of drift and diurnal variations. The marked points match incidents that were logged and identified by the plant operators. A more systematic approach to the detection of upset events within the gradual environmental changes in influent composition is also proposed which evolved from a macro originally developed for the removal of outliers from datasets prior to multivariate analysis (Section 5.4). A model is currently being developed to simulate on-line detection and is based on the comparison of the sensor's relative response to a moving average. The difference is then weighed against the standard deviation for that sensor (multiplied by a pre-defined coefficient) for each new data point. In principle, the sensitivity and selectivity of the model can be adapted by changing the size of the moving window (from a few minutes to a few days) and by selecting individual sensors as well as adjusting their respective threshold coefficient. Figure 8.7 shows an example of such an analysis where the model (24-hour moving window), successfully detected a whole range of pollution episodes and operating anomalies (diesel, gas failure, pump failure). The marked symbols show the points as identified by an integrated recognition algorithm that is called every time an outlier is detected. Pattern recognition techniques are being investigated for incorporation in an alarm generating software.

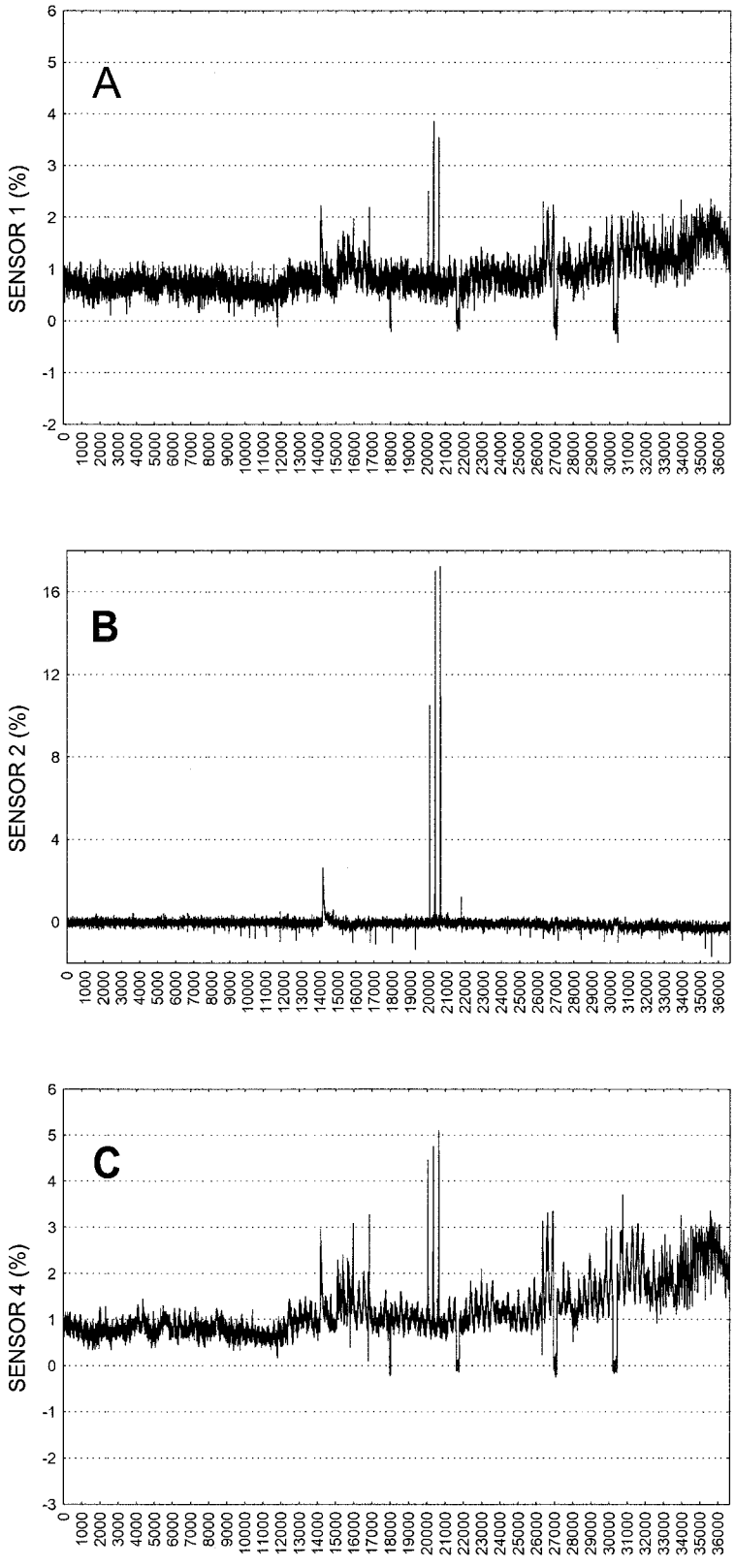


Figure 8.5: Examples of observed sensors responses over a 6-month period: Sensor 1 (A), Sensor 2 (B) and Sensor 4 (C)

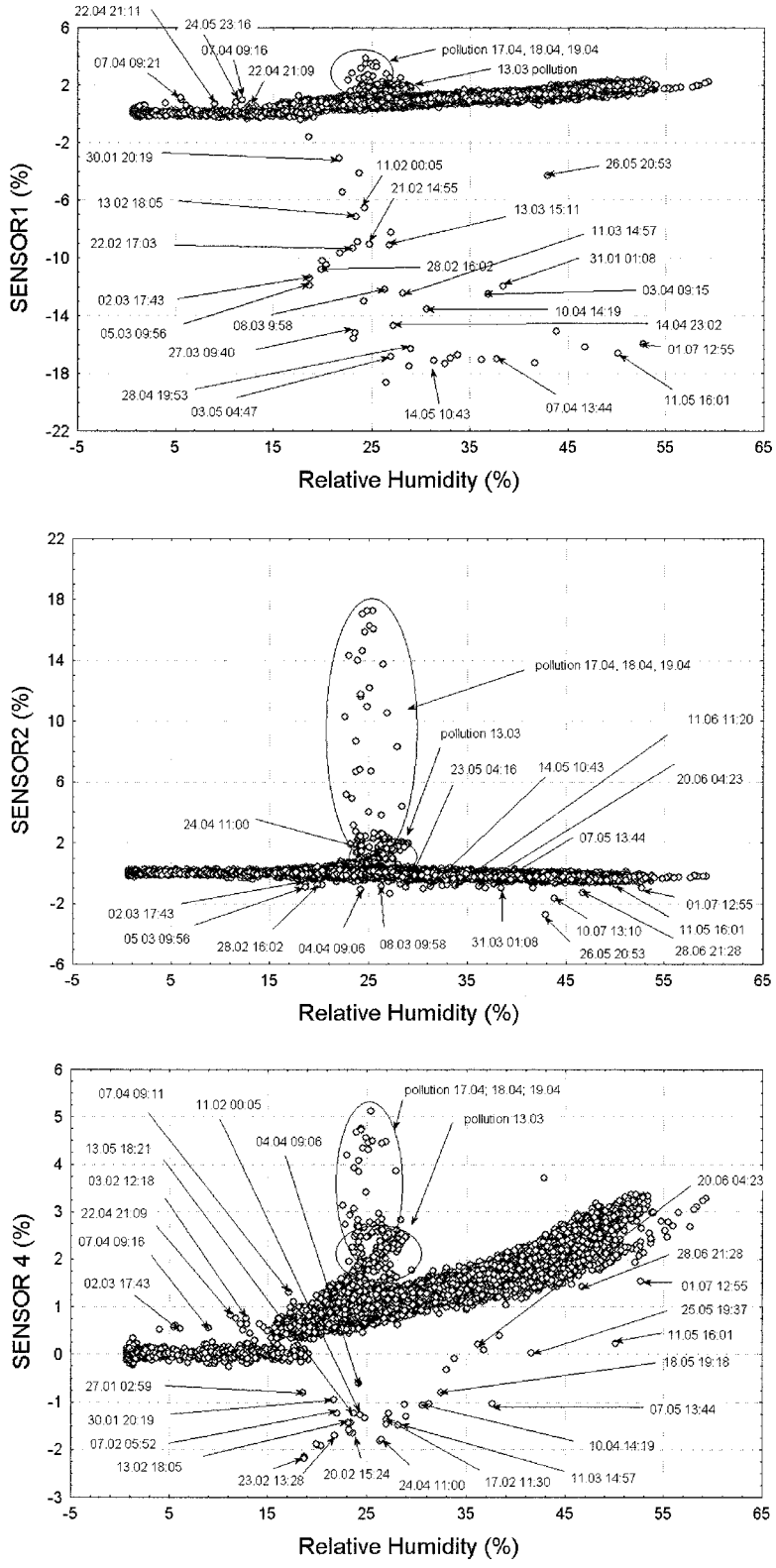


Figure 8.6: Plot of sensor sensors responses (sensors 1, 2 and 4) versus relative humidity (A, B and C respectively), showing unusual patterns. The marked points represent incidents identified by the plant operators.

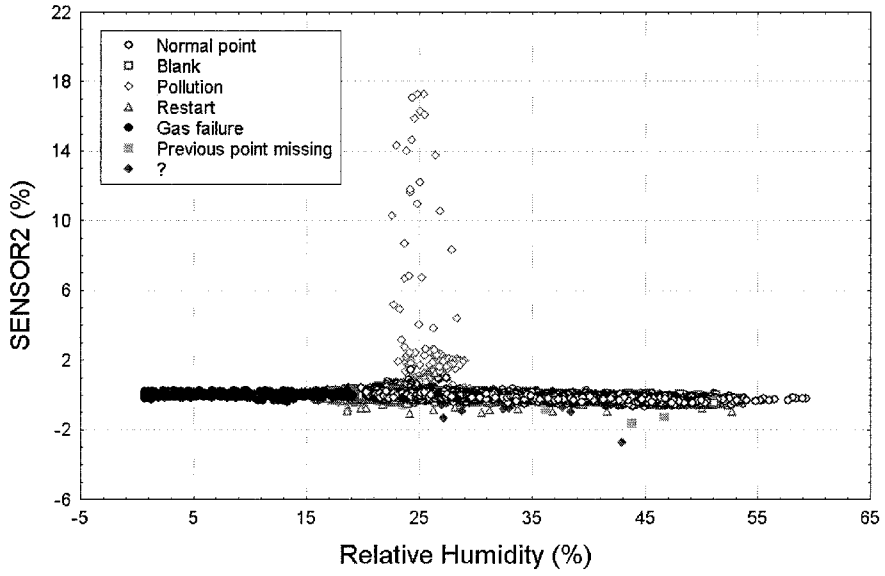


Figure 8.7: Example of simulated detection and identification of a range of incidents by the data mining algorithm using 6-months of continuous data.

The results have demonstrated how a chemical sensor array based system coupled with a simple data mining algorithm can be used for real time process monitoring and upset early warning at a wastewater treatment plant. Current work is focussing on the development of a non-invasive monitoring system that can be operated above a wastewater stream at the inlet of a sewage works. Such a system would provide early warning of process failure, and could be used to either bypass a pollutant to storm tanks or to alter the treatment process.

8.5 SUMMARY

- The effect of an unknown pollution event on sensors 1, 2, 3 and 4 was observed.
- The sensor array showed a better sensitivity to the pollutant than the TOC instrument. The pollution was shown to have remained within the plant for 24 hours.
- A simulated pollution experiment with diesel was successfully carried out and gave similar profiles to that of the unknown incident
- The amplitude of the response change was dependent on concentration as well as mixing time.
- A simple data mining algorithm based on a moving window was successfully used to detect and identify pollution events as well as operating anomalies independently of drift, diurnal variations or changes in temperature and humidity.

Chapter 9: DISCUSSION

CHAPTER 9: DISCUSSION

9.1 OVERVIEW

On-line monitoring of wastewater quality has become increasingly important in the last few years, but remains an unresolved issue to the water industry. International and national regulations still rely on the traditional measurements of global organic load parameters such as BOD_5, COD and TOC to assess whether a wastewater treatment plant meets its discharge requirements.

A rapid and non-invasive on-line device is needed that would allow plant operators to take immediate action against sudden pollution events, or could be used as a tool for real time process control and eventually reduce operating costs. The range of newly available instrumentation is quite varied, but only a few show some real potential. Among them, fluorescence based instruments are in their early days and still being developed. The more common TOC analysers on the other hand require the use of chemicals and frequent calibrations. However both are limited by their need for some form of sample handling and/or contact with the wastewater, and therefore prone to fouling.

Arrays of non-specific sensors may provide an alternative. In particular, conducting polymer-based systems may have shown promises for the measurement of volatiles present in pig slurry such as indole, skatole, ammonia, etc. (Persaud *et al.*, 1996). As demonstrated by Persaud *et al.* (1992), Persaud *et al.* (1996) and Pelosi and Persaud (1998), a clear quantitative relationship exists, since the change in conductance measured is almost proportional to the concentration of adsorbed molecules and their binding affinity to the conducting polymer.

Indeed such sensors are highly sensitive to polar molecules which correlates well with the range of volatile chemicals that are present in wastewater (see section 2.3), and therefore can be related to the organic content of these waters. Recent studies of the headspace of domestic wastewater samples, (Stuetz *et al.*, 1999a, b; Stuetz *et al.*, 2000) have confirmed that such a quantitative relationship between the sensor responses and global organic parameters such as BOD_5, COD and TOC could be established. However, the technology is itself very new and there are few reports of its application to environmental monitoring in the field. In addition, preliminary work has been mostly concerned with the analysis of static, quiescent samples and has been carried out under carefully controlled conditions.

Therefore one of the main challenges of this study was to evaluate and transfer the technology to the field using a simple and robust approach. A commercial CP based sensor array was tested and its potential for wastewater organic load monitoring studied.

9.2 SYSTEM DEVELOPMENT

The first obstacle when dealing with liquid samples is to generate or collect a headspace gas that can be reliably and safely measured by the sensor array. With parameters such as flow, temperature, or suspended solids constantly changing, the major difficulty lies in drawing a headspace sample that is sufficiently representative of the liquid phase. Currently available sensor array systems generally have a limited sampling frequency. Therefore an original approach to generating a gaseous sample from a continuously flowing liquid was developed. Between each acquisition period, wastewater is pumped through a temperature controlled flow-cell where it is then sparged with zero grade N_2, thus generating the headspace gas for sensor array analysis. Preliminary evaluation confirmed previous observations (Gardner & Bartlett, 1999) that the conducting polymer sensors are sensitive to changes in humidity levels. The flow cell was designed so as to allow a tight control over parameters such as temperature, gas flow rate as well as sparger porosity (i.e. number

and size of the bubbles generated). Results showed that it could be used to limit RH variations and produce a reproducible headspace from a liquid sample (Bourgeois and Stuetz, 2000; Bourgeois *et al.*, 2002).

Principal component analysis of the sensor array data demonstrated that the technique can be used to differentiate between RO water, raw sewage and final effluent (Section 4.3.2). Further tests and modifications were carried out in the lab and an optimum sampling methodology for on-line monitoring of a continuously flowing liquid sample was selected using a chemometrics approach (Section 4.4). When combined with multiple linear regression, the technique proved successful in predicting the dilution rate of known wastewater samples as well as the organic strength of unknown samples over a period of a few days only. We discussed in Section 4.5.5 how the model rapidly collapsed when trying to extrapolate the model over periods of more than a week. Samples collected a week later are very likely to have a different odour characteristic than those used for training or collected only a few days after. These findings support previous observations by Stuetz *et al.* (1999b, c) that a better correlation exists between the BOD and the sensor response over shorter periods of time. This "drift" makes the generalisation of a linear model more difficult and becomes more prevalent if limited amount of data is available for training.

Laboratory experiments (Section 4.5) relied on the manual collection of wastewater samples whose biochemical and physical properties rapidly deteriorate due to biological respiration, oxygen depletion and sedimentation. Storage at 4-5°C as well as aeration and mixing may limit these changes but are highly impractical and do not allow for a high sample throughput and the collection of sufficient data for subsequent multivariate analysis. In addition it is thought that prolonged aeration would eventually result in a significant proportion of volatile compounds being stripped off and lost from the samples.

A portable monitoring system based on an array of 8 CP sensors was used to continuously monitor the headspace of wastewater from the primary settlement tank at the university's wastewater treatment plant. Although environmental variables such as temperature cannot be realistically controlled in the field, the same

methodology as previously selected under controlled laboratory conditions is used here. We assume that despite the inevitable variations, the use of the temperature-controlled flow cell, can act as a buffer and limit the amplitude of changes in RH. Since the aim is to eventually move on to a simpler and truly non-invasive system, this approach was judged as good as any for investigating the potential of the technique.

A comparison of this approach to other headspace extraction procedures developed for (HS)-GC analysis (Meyer *et al.*, 2001; Cruwys *et al.*, 2002) underlines the interest of this sparging approach in terms of its relative simplicity and rapidity. Although HS-GC methods are a significant improvement to the analysis of volatile organic compounds from wastewater sample, they remain static and generally rely on the addition of solvents or salts such as methanol, NaCl, $NaHSO_4$, and can also require heating up and relatively long equilibration time. Heating samples to promote volatilisation of VOC's was also conducted by Kaipainen *et al.* (1997) who heated up sugar crystals to 80°C in their research on aroma using electronic noses. Other studies of wastewater by Dewettinck *et al.* (2000) using an array of 12 MOS sensors, relied on heating up spot sample to 90°C for at least 1 hour. A practical limitation with this approach (other than the long equilibration time and energy requirements) was the occurrence of condensation in the tubing of the instrument and the relatively poor repeatability and reproducibility of the system despite the fixed experimental conditions.

Although it has been demonstrated that careful system design and sample pre-conditioning can help minimize changes in relative humidity (Bourgeois and Stuetz, 2000; Ogawa and Sugimoto, 2001; Ueyama *et al.*, 2001), this can make the instrument more complex and expensive and also limit sample throughput. The alternative approach has been favored in a number of studies (Faglia *et al*,1997; Orts *et al*, 1999; Dickert *et al,* 2000; Hierlemann *et al,* 1995) where RH is measured and used as an input to artificial neural networks for parametric compensation. Yet, despite the relative success, these lab-based experiments do not reflect the reality of environmental monitoring where complex mixture must be analysed in rapidly changing background conditions.

Observation of the response profiles obtained from the field (Section 5.3) show diurnal variations that closely match those of the recorded TOC. Multiple correlation studies confirmed previous observations that a good correlation exists between the changes in sensor response and the organic load of wastewater which improves over shorter periods of time (Stuetz *et al.,* 1999b,c). Here, the possibility to use the measured RH for parametric compensation and reduce its effect on the sensors is suggested. The method used consisted in substracting the responses obtained when the sensors are exposed to RO water, from the signal obtained with wastewater samples in the same RH conditions. In principle, the fraction of the response that corresponds to the effect of humidity is eliminated and the remaining fraction can in theory be associated with changes in wastewater quality.

The detection of operating anomalies and the visible dilution effect of heavy rainfall on the profiles confirmed that the selected methodology was appropriate for this study. Thus, the apparatus was left to run continuously for a 12-month period so as to generate data for a more comprehensive investigation of the relationship between the sensor response and the organic content of wastewater.

TOC is used as the parameter of choice because of the instrument's high sampling frequency and reproducibility. Despite the principles of TOC analysis being fundamentally different to that of BOD (Section 2.3.4), it generally provides an accurate and high-resolution record of changes in the organic load. It is also more likely to show some similarity with the sensor array data which depends solely on physicochemical interactions rather than on a biological process. The use of Racod data is voluntarily limited because of concerns of the impact of temperature and toxic chemicals on the results. Furthermore, the Racod instrument continuously pumps wastewater into a bioreactor at a very low flow rate, so measurements are not representative of the ring main BOD at any particular time because of the "Buffer" effect (continuous dilution) that results from this sampling strategy.

9.3 MULTIVARIATE ANALYSIS

On-line monitoring of industrial processes and the analysis of sensor array data have a number or requirements which generally justify the use of multivariate statistics. As noted by Kresta *et al.* (1991), the selected method must be able to deal with collinear data of high dimension, reduce the dimensionality of the problem and give a good prediction of the dependent variable. Rosen *et al.* (2002) lists additional challenges that are specific to wastewater monitoring:

- Poor data quality and reliability due to the hostile environment in which the instruments have to function.
- Non-linear relationships between variables.
- Changing conditions due to diurnal and seasonal changes, which cause deviations from the assumption that the data is stationary in the timescale of interest.
- Changes in the relationship between the variables (Dynamic covariance structure).

Multivariate analysis represents an essential part of the sensor array concept. Arguably the most basic and fundamental aspect, is the need to carefully consider the quality of the data. After removing outliers such as operating anomalies and pollution incidents as well as dealing with missing points, a preliminary examination of the data reveals that a number of parameters may potentially affect multivariate models for the prediction of wastewater organic load from sensor array data (Section 5.5.3). These are mainly the sensor's sensitivity to RH, a slight deviation from the assumption of normality, possible non-linearities and synchronous RH and TOC diurnal variations, which must all be carefully taken into account during the analysis in order to avoid misinterpretation of the results. Among these, RH in particular is a well-known limiting factor that affects most chemical sensors (Gardner and Bartlett, 1999).

Various approaches to the analysis of sensor array data are reported in the literature (Gardner and Bartlett, 1999; Jurs *et al.*, 2000) and the most commonly encountered techniques are reviewed in Section 2.1.4. These can be divided in two main

categories; the traditional multivariate statistics and the more recent artificial neural network models. MLR, PLS, PCR, CC and MLP with BP and RBF respectively are amongst the most frequently used regression methods. However, the choice of a technique varies widely from one application to another as well as from one user to another. With the lack of a standard or recommended approach, and in the absence of reported studies of wastewater organic load using a similar system, a trial and error approach had to be adopted.

The problem at hand is two-fold: a) find a regression model in order to establish that a quantifiable and generalisable relationship exists between the sensor array data and the primary effluent TOC; b) identify and develop a dynamic processing approach suitable for real-time monitoring.

Traditional linear and non-linear multivariate regression were our first choice since these techniques may give a better understanding of the relationships observed – MLR, PLS, Polynomial Regression and Factorial Regression are initially investigated. In practice, most multivariate techniques can tolerate minor deviations from the normality and linearity assumptions discussed above. Nevertheless pre-processing of the data is a crucial part of the analysis. Each variable must be re-scaled so as to have the same minimum and maximum values and to avoid large sensor responses from dominating the analysis. This provides a balanced input to the model and ensures that sensors with smaller responses, but which may contain important information, are equally considered. MLR gave promising results when applied to data obtained under controlled conditions over short periods of time and was therefore our first obvious choice (Section 6.2).

Clearly the models trained on 11 different datasets (each 1 to 2 weeks long) show a poor generalisation when applied to unknown data obtained over a 6-month period (average prediction error <19% for training data and <40% for unknowns). In every case, MLR models appear to be sensitive to the observed diurnal variations in the data but the predictions remain centered around the mean with little deviation from this value. Analysis of the residuals suggests a slight possible influence of the non-normality of the data as well as unequal variances. In a quantitative analysis of VFA's in wastewater headspace samples using HS-GC, Cruwys *et al.*, 2002, also

reported how replicate measurements taken at various concentration can show increasing variance. In our study, this heteroscedasticity is attributed to the fact that most variables are skewed and others (sensor 2, TOC) are not. Although a simple transformation of the variable can reduce or eliminate this effect, it can also make the results more difficult to interpret. Since heteroscedasticity does not invalidate the analysis, the data was left untransformed.

A study of the respective effects of the duration and amount of data used in training confirms previous observations of a time dependent relationship (Section 6.2.2). This shows the relatively short-term validity of MLR based models as well as the need for a relatively long training period. In contrast the number of cases used in training has little influence.

With reference to previous work in which the effect of RH could be numerically compensated (Section 5.3.2), the same approach used on field data showed no improvement (Section 6.2.3). Indeed this procedure assumes that the responses fall within the linear range of the sensors and the poor results may be explained by the fact that most sensors exhibit a non-linear response with RH as seen in Section 5.5.2. In addition, in the same way that CP sensors are sensitive to RH, the RH sensor itself can be sensitive to the presence of VOC's in the headspace gas because of the similar nature of the sensors used in this study. This makes the distinction between RH and VOC's more difficult and, in retrospect, the use of a dew-point humidity sensor or condensation hygrometer is recommended and would allow to objectively quantify the effect of water vapor on the sensors

Despite the different attempts to improve the models, MLR did not perform as well as in our preliminary investigations. The effect of humidity and the longer periods of time considered here for the purpose of generalisation undoubtedly play a significant role. Despite this, the performances of the model on the training sets only remain unsatisfactory and the better predictions appear to be more strongly correlated with RH.

Although PLS is also based on a linear approach, it usually achieves better results than MLR and it is often successfully used in sensor array applications (Jurs *et al.*,

2000). In a comparative study, Nicolas *et al.* (2001), used PLS scores plots to show that data from an array of 6 tin oxide sensors could be used to measure odour intensity around a landfill site and noted that it gave slightly better prediction than both PCR and MLR. However, our prediction results and correlations using PLS on our data showed no difference to that of MLR. It is therefore suggested that the poor predictions may be the result of non-linear relationships that are too important for MLR and PLS.

Interestingly, the application of non-linear multivariate regression models to sensor array data has not been reported by any of the authors listed in Table 2.4. We therefore implemented and evaluated such models, but this proved unsuccessful. A second-degree polynomial regression shows no improvement over the linear methods in the same conditions and, as with factorial regression, increasing the degree and complexity of the models only improve the performances on the training sets. This is a typical demonstration of the adaptability of these methods, which increases with the number of coefficients. Unfortunately, the poor generalisation is a clear indication of the model's propensity for overfitting and their strong dependence on the training set. Another negative aspect is the relatively complex and slow computation, which make these techniques inadequate for real-time monitoring applications.

In both sensor array applications (e.g. Nicolas *et al.*, 2000; Wilson *et al.*, 2000; Capone *et al.*, 2001) and wastewater monitoring studies (Rosen & Olsson, 1998; Rosen *et al.*, 2002; Lennox, 2001), authors have reported the interest of using principal component analysis, simply as a tool to demonstrate the classification ability of a particular system, or as a pre-processing step to reduce the dimensionality of the data (and computational time) and cope with highly collinear variables. In regression studies, PCA can also be useful in removing process or measurement noise from the data. Therefore, principal components were extracted and the scores used as an input to both linear (PCR) and non-linear (polynomial) regression models. No difference was observed between the three different models tested and the results were very similar to those of PLS, MLR or polynomial regression alone. These observations show that PCA brings no benefit as a first step prior to regression

analysis and also indicates that it cannot be used to elucidate the finer features that depend on non-linear relationships.

Artificial neural networks are free of the traditional assumptions and have become increasingly popular because of their great flexibility and ability to learn both linear and non-linear relationships. Although their use is not always clearly justified and the choice of ANN may sometimes appear arbitrary, it is clear that the technique is very attractive to most authors involved in sensor array research. ANN have indeed proved to be a successful and powerful tool in a number of regression studies. Hierlemann *et al.* (1995) showed the advantage in using ANN rather than PLS for complex ternary mixtures and long-term measurement. Consequently, we also investigated the potential of ANN in finding a relationship between the sensor array and corresponding TOC data. From our previous observations, a successful model would be expected to cope with some or most of the limitations associated with RH, time-related changes, non-linearities and noise in the data, in order to outperform the statistical methods. This is supported by Hierlemann *et al.*, 1995; Orts *et al.*, 1999 and Dickert *et al.*, 1999 who showed the interest of using RH as an input to ANN.

A 3-layer (8-4-1) MLP gives an average error inferior to 14% and is able to predict more than 91% of the data with an error <30%. Although this is a better performance than previously achieved with statistical regression techniques, the predictions are still limited around the mean TOC value and do not improve when increasing the complexity of the model. On the other hand, significant improvements can be achieved by first applying a smoothing function such as exponential smoothing to reduce the noise from the sensor array data. The network is able to predict both diurnal variations as well as differences between weekend and weekday levels (R = 0.71; Average relative error = 11% and 96.1% of cases predicted with an error <30%). This demonstrated the negative effect of noise which is particularly prevalent with on-line data obtained in the field. Consequently, the application of a low pass filter or other smoothing method as a standard pre-processing step prior to the analysis is highly recommended. Indeed, systematic noise reduction has been suggested by a number of authors, both in sensor array applications and on-line multivariate analysis of wastewater systems (Rosen, 2000; Cremancini *et al.*, 2001; Pardo and Sberveglieri, 2002).

A comparison with the use of PCA as a pre-processing technique demonstrates that the dimensionality of the data is not so much of an issue here. Any improvements observed with PCA are mostly due to the fact that the last extracted components are generally associated with noise in the original data. Removing these components also removes the noise. This is confirmed by the fact that no improvement is seen when a PCA is performed on smoothed data since most of the noise has already been removed.

In our effort to determine whether a sensor array can be used for wastewater organic load monitoring, the combined specificities of the technology and of wastewater systems have led us to an extensive multivariate analysis of the data. The outcome of the above study is two-fold:

- it has shown that ANN performed better than traditional MVS and that noise is a limiting factor in establishing solid regression models.
- it provided some original information on the nature of the data and its characteristics which may be useful for the development of continuous data analysis algorithms

In this respect, our contribution to knowledge may be valuable to researchers since there is little comparative research that deals with the applicability of a particular statistical monitoring technique to wastewater treatment processes. This adds to the work carried out by Rosen (2001), Mounce *et al.* (2001), Yoo *et al.* (2001), and Beck and Lin (2003), who despite being mostly concerned with the detection and isolation of disturbances, also highlighted the difficulty in coming to terms with the complexity of the higher-order, multivariable and non-stationary character of the datasets, i.e., in interpreting the interactions among contemporarily measured entities.

9.4 A SENSORY UPSET EARLY WARNING DEVICE

The observation of the sensor array data before isolation and removal of outliers has shown that a whole range of anomalies in the profiles had been recorded. In particular, attention is drawn to the system's ability to detect pollution events as well as operating anomalies, and to follow their evolution in time. A simple data mining algorithm based on a moving window is presented for the detection of incidents and sudden changes in wastewater quality. With further development, such system could be used as an upset early warning device for process stream control. It is anticipated that a whole range of pattern recognition techniques could be incorporated into alarm generation software. For instance, we showed the ability of Kohonen networks to distinguish between wastewaters according to their organic strength. Other dynamic data analysis techniques for novelty detection include the newly developed adaptive PCA reported by a number of authors (Capone *et al.*, 2001; Rosen & Lennox, 2001) as well as modified dissimilarity index as described by Yoo *et al.* (2001). These techniques are all based on the principle of a moving window in which the variance is examined. Typically a window of a few hours or a few days is considered for processes that exhibit a diurnal or seasonal pattern since increasing the window size will increase variance and reduce sensitivity. Similarly important drift will limit the time-scale that can be applied. Still, these techniques have successfully been tested on real data from wastewater treatment plant for the detection of sudden and gradual changes in the process. Their association with sensor array systems could provide a simple non-invasive early warning device and present a real commercial interest in a number of industries.

9.5 OVERALL CONSIDERATIONS

It has been discussed that sensor array technology is a relatively new and strongly multidisciplinary area where continuous progress and developments are still being

made. With regard to environmental monitoring applications and with particular reference to wastewater organic load monitoring, this study clearly exposed a number of limitations that must be addressed before such systems reach an acceptable level of maturity.

Among these, sampling and conditioning of the headspace gas, the type of sensors being used, the quality of the data generated, the choice of data processing and pre-processing techniques are all critical and specific factors that require to be addressed by highly qualified personnel. Indeed, wastewater monitoring in itself represents a real challenge to the water industry which faces major difficulties associated with, for instance, the harsh environment and dynamic aspects of the processes involved.

Drift is also a major limiting factor. An examination of the sensor baselines reveals how sensors 1, 3 and 8 are particularly affected. A 20% change in the resistance prior to acquisition can be observed at the end of the 6-month monitoring period. Although important, the drift did not affect the sensors' ability to respond to changes in wastewater quality, and the relative change in sensor resistances (dR/R) when exposed to a sample remained consistent throughout that period. However, because of the significance of the drift, we would recommend to regularly (e.g. monthly or quarterly) test the performance of the system with a known standard solution particularly if it is used over long periods of time. This approach which is routinely used as a calibration step with other commercial instruments (ENose 4000, Marconi Applied Technologies; VOCmeter, AppliedSensors) could be easily automated and would also be of interest for intercomparison of results between instruments or if sensors have to be replaced.

In addition to drift, some irregularities in the baselines profiles were observed. A real cause for concern is the failure of sensors 5, 6, 7 and 8 to return to their original baseline levels after acquisition. This carry-over or poisoning of the sensors is not systematic and directly affects the calculation of the relative response change (dR/R). The good performance of an 8-4-1 MLP using the sensor resistances demonstrates that important information is lost when using dR/R because of the irregularities in the baseline. It is therefore strongly recommended to significantly reduce the acquisition/ de-purge time ratio in order to avoid cross contamination of the sensors.

However, a comparison of the sensor array technique with existing and recently improved technologies such as biosensors, UV-absorbance or TOC (section 2.2), clearly shows the advantage of using non-invasive instruments in wastewater treatment systems which limits problems associated with fouling, corrosion or maintenance. Although alternative new techniques such as fluorescence analysis may also have a potential for non-invasive measurements, they still require further research and development. For instance, Reynolds and Ahmad (1997) and Ahmad and Reynolds (1999), noted the effect of changes in temperature and pH on the relationship between the fluorescence spectra and the organic content of a wastewater. Similarly for sensor array analysis, these interferences together with the long-term changes in wastewater quality and the selection of a suitable data analysis protocol must all be carefully considered. We therefore anticipate that these techniques still face the same challenges with regard to organic load measurement. In the case of sensor arrays, one particular limitation lies in the current difficulty to establish unambiguous relationships and solid models that can be applied over extended periods of time (i.e. more than one week). Further improvement would require long-term expertise and the use of complex multivariate analysis which may be a significant barrier to both developers and end users. A summary table which compares sensor arrays versus other on-line wastewater measurement techniques is given in Table 9.1.

Although the ability of a commercial array of CP sensors to provide a measure of the organic strength of a wastewater has been demonstrated, it appears that the use of the technology as a surrogate analytical instrument for continuous organic load measurements in the field is still an ambitious task. In the current state of the art, we also argue that a sensible approach may be to move away from already complex applications. In this respect, the development of a sensor array based upset early warning system for the detection of pollution incidents or process anomalies could bring more immediate returns to both the end-user and the manufacturers. In addition, such on-line and real-time monitoring systems could be used as a stepping stone to generate revenues and acquire valuable experience and knowledge while more advanced systems (e.g. new sensors and pattern recognition) are being developed.

Table 9.1: Comparison of on-line wastewater organic load monitoring techniques

Principle	Variable	Sample treatment	Frequency	Interests	Limitations	References & examples
Biochemical	respirometry (respiration rate or % inhibition)	At line	variable	commercial instruments, characterisation of activated sludge kinetics, provides a measure of biodegradability and BOD_5 estimates	fouling, sensitive to toxics, Strongly depend on system used, need to specify biomass source, type of substrate and time aspect (continuous, batch..)	RODTOX, Varolleghem *et al.*, (1994) Spanjers *et al.*, (1998)
Electro-chemical	Dissoved Oxygen	On-line (immersed)	continuous	essential to activated sludge aeration control (cost savings), reliable and accurate, good indicator of oxygen uptake	fouling, location sensitive, provides limited information on water quality and disturbances	Widely used
Chemical	TOC	At line	5 minutes	commercial instrument, rapid, sensitive, accurate and reproducible	cost, no information on biodegradability, pre-filtration, calibration needed maintenance, fouling, high power demand	Widely used e.g. Shimadzu TOC 4100
	COD	At line	30 minutes to 2 hours	standardised technique, extensively used, not affected by toxics, good correlation with BOD	uses chemicals, hazardous wastes, clogging of instruments, no distinction between inert and biodegradable matter	Widespread use Ademoroti (1986) Korenaga *et al.*, (1990)
Physico-chemical	Sensor arrays	On-line (soon non-invasive)	less than 5 minutes	rapid, robust versatile, non-invasive, no reagents, good potential for early warning and detection of pollutants	selectivity, site specificity, time dependent relationship with organic load, data analysis intensive, development needed. RH & temperature sensitive.	Bourgeois and Stuets, (2002) Stuetz *et al*, (1999a,b)
	Fluorescence	Non-invasive	continuous	rapid, non-invasive, no reagents potential for screening for pollutants	immersed parts (beam terminator), data analysis intensive, More development needed. affected by pH and temp	Ahmad and Reynolds (1999)
	UV absorption	On-line (immersed)	few minutes	low cost, no chemicals, good indicator of organic load (COD, TOC)	poor sensitivity and selectivity, interference of particulate material, fouling (immersed sensor)	Brookman (1997) Matche and Stumwohrer (1996)

Alternatively, one could reason that sensor arrays or electronic noses should be considered and used as instruments per se and not as surrogates to other already existing techniques. One of the arguments is that the data generated is sufficiently original and relevant to be useful as such for wastewater treatment process control, as successfully demonstrated in a number of food and drink applications. This requires

the end user to fully embrace this philosophy, and revises the question of the value of the technique with regard to discharge requirement and other legal issues.

Chapter 10: CONCLUSIONS

CHAPTER 10: CONCLUSIONS

- This study shows that an array of non-specific conducting polymer sensors can be used for on-line monitoring of wastewater quality.

- A sampling system was developed and used to generate a reproducible headspace gas sample from a continuously flowing liquid. The flow cell allowed for the effects and interactions of temperature, gas flow rate and sparger porosity over RH, to be quantified and minimised using a chemometrics approach.

- Laboratory results show that a modified commercial instrument can differentiate between different types of wastewater (raw sewage, final effluent and clean water) and different dilutions of a single wastewater sample using the selected methodology.

- Examination of sensor array data from a wastewater treatment plant (primary settled effluent) also indicates that a strong relationship exists with global organic parameters. Diurnal variations and lower responses over the weekend were recorded that matched TOC measurements.

- Important limitations with regard to on-line monitoring in the field have been exposed that underlined the complexity of both wastewater monitoring and sensor array analysis in uncontrolled conditions. These are: a) effect of RH variations on the sensor profiles, b) effect of noise and quality of the data on the analysis, c) effect of time dependant relationship on the long term validity

of the models, d) erratic poisoning of the sensors which affects the relative sensor response.

- An 8-4-1 MLP neural network with quasi-Newton algorithm gave the best results for the prediction of on-line TOC using dR/R as input (R=0.49, Average RAE = 13.9%, and 91.2% of cases predicted with less than 30% RAE).

  The performance was significantly improved by pre-processing the data and applying an exponential smoothing function to remove the noise. The model's average prediction error was (R = 0.71; Average relative error = 11% and 96.1% of cases predicted with an error <30%).

- The use of PCA as a pre-processing step to extract PC's and use them as input also showed some clear benefits. The interest of PCA however, was associated with the removal of noise included in the last extracted PCs rather than to a reduction of the dimensionality of the data. In comparison, exponential smoothing proved a more straightforward and more efficient approach.

- Multivariate analysis using traditional linear and non-linear statistical techniques (MLR, PLS, Polynomial regression, Factorial regression) was in comparison less adapted to this application than the more flexible ANN. On the other hand, MVS provided vital information on the quality of the data and helped detect problem areas such as non-linearities, positive skewness and heteroscedasticity.

- Continuous monitoring of primary effluent in the field recorded a range of recurring patterns and incidents: diurnal variations, dilution effect of heavy rain, pollution incidents and operating anomalies.

- Unknown and simulated (diesel spikes) pollution events as well as operating anomalies (e.g. pump failures) were detected and identified regardless of humidity or temperature conditions using a moving time window and a simple data mining algorithm.

- It is argued that in this current state of development, sensor array systems cannot be used as an on-line instrument for real-time measurement of wastewater organic content. Instead, they may have a greater potential as non-invasive upset early warning devices for the protection of wastewater treatment plants and other process streams. Although not strictly of interest with respect to statutory discharge requirements, the implementation of sensor array based systems in wastewater treatment plants may yield significant benefits to water companies as it could help in reducing costs generated by intermittent or accidental discharges. Ultimately the experience gained from such systems may be built upon and used for more demanding applications.

Chapter 11: RECOMMENDATIONS FOR FUTURE RESEARCH

CHAPTER 11: RECOMMENDATIONS FOR FUTURE RESEARCH

A number of areas can be identified for future research work that continues the themes developed in this study. Having identified a number of problem areas as well as potentially promising applications for the technique, it would be particularly interesting to carry out a comparative study of different commercial or prototype instruments based on different types of sensors (MOS, SAW, QCM).

Arguably the most promising area for future research would be the development of a sensory upset early warning device at the inlet of a wastewater treatment plant.

Evaluation of such systems in different locations, under a wide range of operating conditions and over long periods of time would be very valuable.

Experiments aiming to identify and establish detection limits for a range of substances such as diesel, petrol and antifreeze may be of real interest to the water industry.

Comparison of sensor array versus fluorescence or GC data could help identify specific compounds that particularly affect the profiles. Similarly, a parallel examination of the suspended solids and volatile suspended solids content of wastewater would be highly beneficial.

With regard to organic load monitoring and process control applications, repeating the experiments described in Section 7.5 should yield valuable information on the effect and interaction of environmental variables.

In every case, such work may be divided into two main development stages:

1) Method Development

The most important issue, that is common to all collections of on-line measurements for analysis is the quality of the data. As with all new techniques there remain some basic problems and uncertainties which should be considered a priority for the successful application of sensor arrays in wastewater treatment plants. The following aspects are key factors which may be considered as research topics:

- Sampling and sample pre-treatment: for instance a truly non-invasive technique would be highly desirable. The interest of a drying mechanism (e.g. Nafion membrane) or humidity control device could be investigated.
- Calibration and drift compensation.
- Optimisation of the acquisition procedure: the acquisition phase may be reduced and the desorption phase extended.
- Feature selection: transient versus steady state sensor responses as well as desorption profiles may be studied.
- Pre-processing: noise reduction (smoothing, filtering), re-scaling and variable selection are straightforward yet powerful ways to improve the analysis.

2) Data Analysis

Chemometrics and multivariate data analysis are vast and complex disciplines which require careful considerations. The selection of an appropriate protocol is an essential component of both sensor array analysis and wastewater monitoring and will depend on the particular application (e.g. qualitative, quantitative or novelty detection). Future work in this field would be best carried out by experts and should put the emphasis on automated and real-time analysis protocols as well as on the overall accessibility to the end user. Techniques such as Independent Component Analysis (ICA), Dynamic PCA, Fourrier analysis and Wavelett transform have been successfully used by a few researchers and may also present an interest for wastewater applications.

REFERENCES

REFERENCES

Ademoroti, C. M. A. (1985). The effects of metallic toxicants on BOD measurements. *Environmental International Journal*, USA.

Ademoroti, C. M. A. (1986). Model to predict BOD from COD values. *Effluent & Water Treatment,* **26**, 80-84.

Ahmad, S. R. and Reynolds, D. M. (1999). Monitoring of water quality using fluorescence technique: prospect of on-line process control. *Water Res.,* **33**(9), 2069-2074.

Ahmad, S. R., Foster, V. G. and Reynolds, D. M. (1993). Laser scattering technique for the non-invasive analysis of wastewater. *Proc. Substance Detection Syst. (SPIE)*, **2092**, 353-359.

Aishima, T. (1991). Discrimination of liquor aromas by pattern recognition analysis of responses from a gas sensor array. *Analytica Chimica Acta*, **243**, 293-300.

Albert, K. J., Lewis, N. S., Schauer, C. L., Sotzing, G. A., Stitzel, S. E., Vaid, T. P. and Walt D. R. (2000). Cross-Reactive Chemical Sensor Arrays. *Chem. Rev.*, **100**, 2595-2626.

American Public Health Association (APHA) (1995). *Standard Methods for the Examination of Water and Wastewater,* 19th ed. American Water Works Association and Water Environment Federation, Washington DC, USA.

An L., Niu, H. and Zeng, H. (1998). A new biosensor for rapid oxygen demand measurement. *Water Env. Res.*, **70**, 1070-1074.

Baby R. E., Cabezas M. and Walsöe de Reca, N. E. (1999). Quantitative analysis with an electronic nose of lindane and nitrobenzene in water. In: *Proceedings of ISOEN 99*, Tübingen , 351-354.

Baby, R. E., Cabezas, M. and Walsoe de Reca, E. N. (2000). Electronic nose: a useful tool for monitoring environmental pollution. *Sensors and Actuators B*, **69**, 214-218.

Bachinger T., Lidén H., Mårtensson P., Mandenius C.-F. (1998). On-line estimation of state variables in Baker's yeast fermentation using an electronic nose. *Seminars in Food Analysis*, **3**, 85-91.

Bartlett, P. N., Archer, P. B. M. and Ling-Chung, S. K. (1989). Conducting polymer gas sensors. Part I: Fabrication. *Sensors and Actuators*, **19**, 125-140.

Bartlett, P. N. and Gardner, J. W. (1992). Odour sensors for an electronic nose. In: *Sensors and sensory systems for an electronic nose*, Gardner, J. W. and Bartlett, P. N. (eds.), **212**, pp. 31-51.

Bartlett, P. N. and Li-Chung, S. K. (1989a). Conducting polymer gas sensors. Part II: Response of polypyrrole to methanol vapour. *Sensors and Actuators*, **19**, 141-150.

Bartlett, P. N. and Li-Chung, S. K. (1989b). Conducting polymer gas sensors. Part III: Results for four different polymers and five different vapours. *Sensors and Actuators*, **20**, 287-292.

Beck, M.B. and Lin, Z. (2003). Transforming data into information. *Wat. Sci. Tech.*, **47** (2), 43-51.

Becker, T., Muhlberger, S., Bosh-v. Braunmuhl, C., Muller, G., Ziemann, T. and Hechtenberg, K.V. (2000). Air pollution monitoring using tin-oxyde-based microreactor systems. *Sensors and Actuators B*, **69**, 108-119.

Becker, T., Tomasi, L., Bosh-v. Braunmuhl, C., Muller, G., Sberveglieri, G., Faglia, G. and Comini, E. (1999). Ozone detection using low power-consumption metal-oxide gas sensors. *Sensors and Actuators A*, **74**, 229-232.

Bishop, C. M. (1995). *Neural networks for pattern recognition*, Oxford University Press, Oxford.

Bourgeois, W. and Stuetz, R. M. (2000). Measuring wastewater quality using a sensor array: prospects for real-time monitoring. *Water Science and Technology*, **41**(12), 107-112.

Bourgeois, W., Burgess, J. E. and Stuetz, R. M. (2001). On-line monitoring of wastewater quality: A Review. *J. Chem. Technol. Biotechnol.*, **76**, 337-348.

Bourgeois, W., Gaugler, M. and Stuetz, R. M. (2001). On-line evaluation of a sensor array for monitoring changes in effluent quality. In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp. 271-278.

Brezmes, J., Ferreras, B., Llobet, E., Vilanova, X. and Correig, X. (1997). Neural network based electronic nose for the classification of aromatic species. *Analytica Chimica Acta*, **348**, 503-509.

Brezmes, J., Llobet, E., Vilanova, J., Orts, J., Saiz, G. and Correig, X. (2001). Correlation between electronic nose signals and fruit quality indicators on shelf-life measurements with pinklady apples. *Sensors and Actuators B*, **80**, 41-50.

Brookman, S. K. E. (1997). Estimation of biochemical oxygen demand in slurry and effluent using ultra-violet spectrophotometry. *Water Res.*, **31**, 372-374.

Buck, T. M., Allan, F. G. and Dalton, M. (1965). Detection of chemical species by surface effects on metals and semiconductors. In: *Surface effect in detection*, Bregman, J. I. and Dravnicks, A. (eds.), pp. 147-163.

Buhlmann, K. Schlatt, B., Cammann, K. and Shulga, A. (1998). Plasticised polymeric electrolytes: new extremely versatile receptor materials for gas sensors (VOC monitoring) and electronic noses (odour identification/discrimination). *Sensors and Actuators B*, **49**, 156-165.

Byun, H. G., Persaud, K. C., Khaffef, S. M., Hobbs, P. J. and Misselbrook, T. H. (1997). Application of unsupervised clustering methods to assessment of malodours in agriculture using an array of conducting polymer odour sensors. *Comput. Electron. Agric.*, **17**, 233-247.

Capone, S., Epifani, M., Quaranta, F., Siciliano, P., Taurino, A. and Vasanelli, L. (2001). Monitoring of rancidity of milk by means of an electronic nose and a dynamic PCA analysis. *Sensors and Actuators B*, **78**, 174-179.

Carotta, M. C., Martinelli, G., Crema, L., Gallana, M., Merli, M., Ghiotti, G. and Traversa, E. (2000). Array of thick film sensors for atmospheric pollutant monitoring. *Sensors and Actuators B*, **68**, 1-8

Carotta, M. C., Martinelli, G., Crema, L., Malagu, C., Merli, M., Ghiotti, G. and Traversa, E. (2001). Nanostructured thick-film gas sensors for atmospheric pollutant monitoring: quantitative analysis on field tests. *Sensors and Actuators B*, **76**, 336-342.

Cecile, J. L. (1998). Needs and use of continuous monitoring equipment for wastewater treatment. Standardisation y/n? In: *Monitoring of Water Quality*, Elsevier, pp. 237-244.

Chee, G. J., Nomura, Y., Ikebukuro, K. and Karube, I. (1999). Development of highly sensitive BOD sensor and its evaluation using preozonation. *Anal. Chim. Acta*, **394**, 65-71.

Clark, D. W. (1992). BOD: The modern alchemy. In: *Public Works*, January, pp. 50-51, 86-89.

Colin and Quevauviller. (1998). Monitoring of Water Quality. In: *Proc. of the European workshop on standards, measurements and testing for the monitoring of water quality: the contribution of advanced technologies,* Nancy, France, May 29-31 1997, Elsevier, 269 pp.

Craven, M., Hines, E. L., Gardner, J. W., Morgan, D., Hogan, P. and Ene P. A. (1994). Bacteria detection and classification using artificial neural network in conjunction with an electronic nose. In: *International Conference on Neural*

Networks and Expert Systems in Medicine and Healthcare, Univ. of Plymouth, UK, August 23-26 1994.

Cremoncini, A., Di Francesco, F., Lazzerini, B., Marcelloni, F., Martin, T., McCoy, S. A., Sensi, L. and Tselentis, G. (2000). Electronic Noses using "Intelligent" Processing Technqiues. In: *Electronic Noses and Olfaction 2000*, Conference proceedings, Brighton, UK, pp. 97-106.

D'Amico, A. and DiNatale, C. (2000). A contribution on some basic definitions of sensor properties. *IEEE Sensor Journal*, **1**(3), 183-190.

DARPA (1998). *Neural Network Study*, AFCEA International press, p.60.

De Wit, M., Vanneste, E., Geise, H. J. and Nagels, L. J. (1998). Chemiresistive sensors of electrically conducting poly(2,5-thienylene vinylene) and copolymers: their response to nine organic vapors. *Sensors and Actuators B*, **50**, 164-172.

Dejous, C., Rebiere, D., Pistre, J., Tiret, C. and Planade, R. (1995). A surface acoustic wave gas sensor: Detection of organophosphorous compounds. *Sensors and Actuators B*, **24-25**, 58-61.

Delpha, C., Lumberas, M. and Siadat, M. (2001). Discrimination of Forane 134a and carbon dioxide concentrations in an air conditionned athmosphere with an electronic nose: influence of the relative humidity. *Sensors and Actuators B*, **80**, 59-67.

Dewettinck, T., Van Hege, K. and Verstraete, W. (2001). The electronic nose as a rapid sensor for volatile compounds in treated domestic wastewater. *Water Research*, **35**(10), 2475-2483.

Di Natale, C., Macagnano, A., Davide, F., D'Amico, A., Legin, A., Vlasov, Y., Rudnitskaya, A. and Seleznev, B. (1997). Multicomponent analysis on polluted waters by means of an electronic tongue. *Sensors and Actuators B*, **44**, 423-428.

Di Natale, C., Martinelli, E. and D'Amico, A. D. (2002). Counteraction of environmental disturbances of electronic nose data by independent component analysis. *Sensors and Actuators B*, **82**, 158-165.

Dickert, F. L., Hayden, O. and Zenkel, M. E. (1999). Detection of volatile compounds with mass-sensitive sensor arrays in the presence of variable ambient humidity. *Analytical Chemistry*, **71**, 1338-1341.

Dignac, M-F., Ginestet, P., Rybacki, D., Bruchet, A., Urbain, V. and Scribe, P. (2000). Fate of wastewater organic pollution during activated sludge treatment: nature of residual organic matter. *Water Research*, **34**, 4185-4194.

Dobbs, R. A., Wise, R. H. and Dean, R. B. (1972). The use of ultra-violet absorbance for monitoring the total organic carbon content of water and wastewater. *Water Res.*, **6**, 1173-1180.

Dravnieks, A. and Trotter, P. J. (1965). Polar vapor detection based on thermal modulation of contact potentials. *J. Sci. Instrum.*, **42**, 624-627.

Droste, R.L. (1997). Theory and Practice of Water and Wastewater Treatment. John Wiley & Sons Inc., Chichester.

Dunteman, G. H. (1984). *Introduction to multivariate analysis*, Sage Publications Inc.

Ekama, G. A., Marais, G. V. R. and Dold, P. L. (1986). Procedures for determining influent COD fractions and the maximum specific growth rate of heterotrophs in activated sludge systems. *Water Science and Technology*, **18**, 91-114.

Eklöv T., Johansson, G., Winquist, F. and Lundström, I. (1998). Monitoring Sausage Fermentation Using an Electronic Nose. *J. Sci. Food Agric.*, **76**, 525-532.

ETACS, European Testing and Assessment of Comparability of On-line Sensors/Analysers. Project 3256 – ETACS, Part of the Standards, Measurement & Testing Programme, Managed by DGXII.

Faglia, G., Bicelli, F., Sberveglieri, G., Maffezzoni, P. and Gubian, P. (1997). Identification and quantification of methane and ethyl alcohol in an environment at variable humidity by an hybrid array. *Sensors and Actuators B*, **44**, 517-520.

Fausset, L. (1994). *Fundamentals of neural networks*, Prentice Hall, New York.

Fayyad, M., Tutunji, M., Ramakrishna, R. S. and Taha, Z. (1987). Dissolved oxygen: Method comparison with potentiometric stripping analysis. *Anal. Lett.*, **20**, 529-535.

Fekadu, A. A., Hines, E. L. and Gardner, J. W. (1993). Generic algorithm design of Neural network based electronic nose. In: *Artificial neural networks and genetic algorithms*, edited by R. F. Albrecht, C. R. Reeves and N. C. Steele, Springer-Verlag, New York, pp. 691-698.

Fenner, R. A. and Stuetz, R. M. (1999). The application of electronic nose technology to environmental monitoring of water and wastewater treatment activities. *Water Env. Res.*, **71**, 282-289.

Fenner, R. A. and Stuetz, R. M. (1999). The application of electronic nose technology to environmental monitoring of water and wastewater treatment activities. *Water Env. Res.*, **71**, 282-289.

Furlong, C. and Stewart, J. R. (2000). Using a portable electronic nose for identification of odorous chemicals. In: *Electronic Noses and Olfaction 2000*, Conference proceedings, Brighton, UK, pp. 285-290.

Gardner J. W. (1991). Detection of Vapours and Odours from a Multisensor Array Using Pattern Recognition. Part 1: Principal Component and Cluster Analysis. *Sensors and Actuators B*, **4**, 109-115.

Gardner, J. W. and Bartlett, P. N. (1999). *Electronic Noses: Principles and Applications*, Oxford University Press, Oxford, 233 pp.

Gardner, J. W. and Hines, E. L. (1997). Pattern analysis techniques. In: *Handbook of biosensors and electronic noses: Medicine, food and the environment*, Kress-Rogers, E. (ed.), CRC Press, pp. 633-652.

Gardner, J. W., Craven, M., Dow, C. and Hines, E. L. (1998). The prediction of bacteria type and culture growth phase by an electronic nose with a multi-layer perceptron network. *Meas. Sci. Technol.*, **9**, 120-127.

Gardner, J. W., Hines, E. L. and Wilkinson, M. (1989). Application of artificial neural networks to an electronic olfactory system. *Meas. Sci. Technol.*, **1**, 446-451.

Gardner, J. W., Pearce, T. C., Friel, S., Bartlett, P. N. and Blair, N. (1994). A multisensor system for beer flavor monitoring using an array of conducting polymers and predictive classifiers. *Sensors and Actuators B*, **18**, 240-243.

Gardner, J. W., Shin, H. W., Hines, E. L. and Dow, C. S. (2000). An electronic nose system for monitoring the quality of potable water. *Sensors and Actuators B*, **69**, 336-341.

Germili, F., Orhon, D. and Artan, N. (1991). Assessment of the initial inert soluble COD in industrial wastewaters. *Water Science and Technology*, **23**, 1077-1086.

Gernaey, K., Petersen, P., Ottoy, J. P. and Vanrollenghem, P. (1999). Biosensing activated sludge. *WQI*, **May/June**.

Getino, J., Ares, L., Robla, J. I., Horrillo, M. C., Sayago, I., Fernandez, M. J., Rodrigo, J. and Gutierrez, J. (1999). Environmental applications of gas sensor arrays: combustion atmospheres and contaminated soils. *Sensors and Actuators B*, **59**, 249-254.

Gibson, T. D., Prosser, O., Hulbert, J. N., Marshall, R. W., Corcoran, P., Lowery, P., Ruck-Keene, E. A. and Heron, S. (1997). Detection and simultaneous identification of microorganisms from headspace samples using an electronic nose. *Sensors and Actuators B*, **44**, 413-422.

Gopel, W. (1998). Chemical imaging: I. Concepts and visions for bioelectronic noses. *Sensors and Actuators B*, **52**, 125-142.

Gostelow, P., Parsons, S. A., and Stuetz, R. M. (2001). Odour measurements for sewage treatment works: a review. *Water Research*, **35**(3), 579-597.

Grady, C. P. L. and Lim, H. C. (1980). *Biological wastewater treatment, Theory and Applications*, Marcel Dekker Inc., New York.

Green, S. A. and Blough, N. V. (1992). Absorption and fluorescence spectra of waters from the Gulf of Mexico and western coastal Florida. *Abst. Am. Chem. Soc.*, **203**, 247-252.

Grieve, I. C. (1985). Determination of dissolved organic matter in streamwater using visible spectrophotometry. *Earth Surf. Proc. Landf.*, **10**, 75-78.

Gustaffson, G. and Lundstrom, I. (1987). The effect of ammonia on the physical properties of polypyrolle. *Synth. Met.*, **21**, 203-208.

Guwy, A. J., Farley, L. A, Cunnah, P., Hawkes, F. R., Hawkes, D. L., Chase, M. and Buckland, H. (1999). An automated instrument for monitoring oxygen demand in polluted waters. *Water Research*, **33**, 3142-3148.

Hack, M. and Kohne, M. (1996). Estimation of wastewater process parameters using neural networks. *Water Science and Technology*, **33**, 101-115.

Hair, J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998). *Multivariate data analysis*, 5th ed. Prentice-Hall International Inc.

Hashem, S., Keller, P. E. and Kangas, L. J. (1996). Electronic noses and their applications in environmental monitoring. In: *Applications of Neural Networks in Environment, Energy and Health*, Ch. 9, pp. 74-81.

Hatfield, J. V., Hicks, P. J., James-Roxby, P., Neaves, P., Persaud, K. C. and Travers, P. J. (1994). Towards an integrated electronic nose using conducting polymer sensors. *Sensors and Actuators B*, **18-19**, 221-228.

Haugen, J.-E. and Kvaal, K. (1998). Electronic nose and artificial neural network. *Meat Science*, **49**, Suppl. 1, S273-S286.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*, McMillan Publishing, New York.

Hedges, J. I. and Lee, C. (1993). Measurement of dissolved organic carbon and nitrogen in natural waters. In: *Proceedings of NSF/NOAA/DOE Workshop*, Seattle, WA, USA, July 1991, *Marine Chemistry*, **41**, 1-290.

Hellinga, C., Vanrolleghem, P., Van Loosdrecht, M. C. M. and Heijen, J. J. (1996). The potential of off-gas analyses for monitoring wastewater treatment plants. *Water Science and Technology*, **33**, 13-23.

Henze, M. (1992). Characterization of wastewater for modelling of activated sludge processes. *Water Science and Technology*, **25**, 1-15.

Hierlemann, A. (2002). *1st NOSE II short course on artificial olfactory sensing lecture notes*. [WWW document: http://www.nose-network.org].

Hierlemann, A., Weimar, U., Kraus, G., Schweizer-Berberich, M. and Gopel, W. (1995). Polymer-based sensor arrays and multicomponent analysis for the detection of hazardous organic vapours in the environment. *Sensors and Actuators B*, **26-27**, 126-134.

Hobbs, P. J., Misselbrook, T. H. and Pain, B. F. (1995). Assessment of odours from livestock wastes by photoionistion detector, electronic nose, olfactometry and gas chromatography-mass spectrometry. *J. Agric, Eng. Res.*, **60**, 137-144.

Hodgins, D. (1995). The development of an electronic nose for industrial and environmental applications, *Sensors and Actuators B*, **26-27**, 255-258.

Hogben, P. and Stuetz, R.M., (2001). Use of a chemicalsensor array and an on-line flow cell for monitoring water quality. In: *Proceedings of AWA-WQT 2001*, Nashville, USA

Höger, S., Heckt, L., Keding, H.-J. and Borcherding, G. (1999). Fire and hazardous event detection – a new area of application for Electronic Nose systems with miniaturised sensor arrays. In: *Proceedings of ISOEN 99*, Tübingen, 387-390.

Holmberg, M., Gustafsson, F., Gunnar Hörnsten, E., Winquist, F., Nilsson, L. E., Ljung, L. and Lundström, I. (1998). Bacteria classification based on feature extraction from sensor data. *Biotechnology Techniques*, **12**, 4, 319-324.

Hyun, C. K., Tamiya, E., Takeuchi, T. and Karube, I. (1993). A novel BOD sensor based on bacterial luminescence. *Biotechnol. and Bioeng.*, **41**, 1107-1111.

Ide, J., Nakamoto, T. and Moriizumi, T. (1994). Development of odour-sensing system using an autosampling stage and identification of natural essential oils. In *Olfaction and Taste XI*, Springer-Verlag, Tokyo, pp. 727-730.

Ikegami, A. and Kaneyasu, M. (1985). Olfactory detection using integrated sensors. In: *Proceedings of the 3rd international conference on solid-state sensors and actuators*, IEEE Press, pp. 136-139.

Iranpour, R., Moghaddam, O., Bina, B., Abkian, V. and Vossoughi, M. (1999). Comment on "Response characteristic of a dead-cell BOD sensor" by Qian, Z. and Tan, T.C. *Water Research*, **33**, 595-596.

Iranpour, R., Straub, B. and Jugo, T. (1997). Real time BOD monitoring for wastewater process control. *J. Environ. Eng.*, **February**, 154-159.

Jacobsen, S. and Lynggaard-Jensen, A. (1998). On-line measurement in wastewater treatment plants: Sensor development and assessment of comparability of on-line sensors. In: *Monitoring of Water Quality*, Elsevier, pp. 89-102.

Janata, J. (1992). Microsensors based on modulation of work function. In: *Sensors and sensory systems for an electronic nose*, Gardner, J. W. and Bartlett, P. N. (eds.), pp. 103-116

Jurs, P. C., Bakken, G. A. and McClelland H. E. (2000). Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chemical Reviews*, **100**, 2649-2678.

Kaipainen A., Ylisuutari S., Lucas Q. and Moy L. (1997). A new approach to odour detection. *Int. Sugar Journal*, **99**, 403-408.

Kalabina, M. M. (1946). Effects of Copper and Lead bearing wastes on purification of sewage. *Water and Sewage works*, **93**, 30.

Kalman, E. L., Winquist, F. and Lundstrom, I. (1997). A new pollen detection method based on an electronic nose. *Atmospheric Environment*, **31**(11), 1715-1719.

Kaneyasu, M., Ikegami, A., Arima, H. and Iwanga, S. (1987). Smell identification using a thick-film hybrid gas sensor. *IEEE Trans. Comp. Hybrids Manufact. Technol.* CHMT-**10**, 267-273.

Khan, E., Babcock, R. W. Jr., Suffet, I. H. and Stenstrom, M. K. (1998a). Method development for measuring biodegradable organic carbon in reclaimed and treated wastewater. *Water. Env. Res.*, **70**, 1025-1032.

Khan, E., Babcock, R. W. Jr., Viriyavejakul, S., Suffet, I. H. and Stenstrom, M. K. (1998b). Biodegradable dissolved organic carbon for indicating wastewater reclamation plant performance and treated wastewater quality. *Water Env. Res.*, **70**, 1033-1040.

Konig, A., Bachmann, T. T., Metzger, J. W. and Schmid, R. D. (1999). Disposable sensor for measuring the biochemical oxygen demand for nitrification and inhibition of nitrification in wastewater. *Appl. Microbiol. Biotechnol.*, **51**, 112-117.

Lauf, R. J. and Hoffheins, B. S. (1991). Analysis of liquid fuels using a gas sensor array. *Fuel*, **70**, 935-940.
Lee, D. D. and Lee, D. S. (2001). Environmental gas sensors. *IEEE Sensors Journal*, **1**(3), 214-224.

Lee, D. S., Jung, H. Y., Lim, J. W., Huh, J. S. and Lee, D. D. (2000). Recognition of VOC gases using the nanochrystalline thick film SnO_2 gas sensor array and pattern recognition analysis. In: *Tech. Dig. 8th Int. Meeting Chem. Sens.*, Basel, Switzerland.

Li, X. M., Ruan, F. C., Ng, W. G. and Wong, K. Y. (1994). Scanning optical sensor for the measurement of dissolved oxygen and BOD. *Sensors and Actuators B*, **21**, 143-149.

Lidén, H., Mandenius, C., Gorton, L., Meinander, N. Q., Lundström, I. and Winquist, F. (1998). On-line monitoring of a cultivation using an electronic nose. *Analytica Chimica Acta*, **361**, 223-231.

Llobet, E., Ionescu, R., Al-Khalifa, S., Brezmes, J., Vilanova, X., Correig, X., Barsan, N. and Gardner, J. W. (2001). Multicomponent gas mixture analysis using a single tin oxide sensor and dynamic pattern recognition. *IEEE Sensors Journal*, **1 (3)**, 207-213.

Logan, B. E. and Patnaik, R. (1997). A gas chromatographic-based headspace biochemical oxygen demand test. *Water Env. Res.*, **69**, 206-214.

Logan, B. E. and Wagenseller, G. A. (1993). The HBOD test: a new method for determining biochemical oxygen demand. *Water Env. Res.*, **65**, 862-868.

Londong, J. and Wachtl, P. (1996). Six years of practical experience with the operation of on-line analysers. *Water Science and Technology*, **33**, 159-164.
Lowden, G. (1981). Tests for assessing the oxygen demand of effluents. *Wat. Res. Tkop.*, **1**, 142-147.

Lu, Y., Bian, L. and Yang, P. (2000). Quantitative artificial neural network for electronic noses. *Analytica Chimica Acta*, **417**, 101-110.

Lynggaard-Jensen, A. (1999). Trends in monitoring of waste water systems. *Talanta*, **50**, 707-716.

Martin, M. A., Santos, J. P. and Agapito, J. A. (2001). Application of artificial neural networks to calculate the partial gas concentrations in a mixture. *Sensors and Actuators B*, **77**, 468-471.

Masmoudi, R. A. (1999). Rapid prediction of effluent biochemical oxygen demand for improved environmental control. *Tappi J.*, **82**, 111-119.

Matsche, N. and Stumwohrer, K. (1996). UV absorption as a control-parameter for biological treatment plants. *Water Science and Technology*, **33**, 211-218.

Matzger, A. J., Lawrence, C. E., Grubbs, R. H. and Lewis, N. S. (2000). Combinatorial approaches to the synthesis of vapor detector arrays for use in an electronic nose. *J. Comb. Chem.*, **2**, 301-304.

McEntegart, C. M., Penrose, W. R., Strathmann, S. and Stetter, J. R. (2000). Detection and discrimination of coliform bacteria with gas sensor arrays. *Sensors and Actuators B*, **70**, 170-176.

McGill, R. E, Nguyen, V. K., Chung, R., Shaffer, R. E., DiLella, D., Stepnowski, J. L., Mlsna, T. E., Venezky, D. L. and Dominguez, D. (2000). The "NRL-SAWRHINO": a nose for toxic gases. *Sensors and Actuators B*, **65**, 10-13.

Menzel, R. and Goschnick, J. (2000). Gradient gas sensor microarrays for on-line process control: a new dynamic classification model for fast and reliable air quality assessment, *Sensors and Actuators B*, **68**, 115-122.

Metcalf and Eddy Inc. (1991). *Wastewater Engineering: Treatment, disposal and reuse*, McGraw Hill, New York.

Miasik, J. J., Hooper, A. and Toefield, B. C. (1986). Conducting polymer gas sensors. *J. Chem. Soc., Faraday trans., I*, **82**, 1117-1126.

Misselbrook, T. H , Hobbs, P. J. and Persaud, K. C. (1997). Use of an electronic nose to measure odour concentration following application of cattle slurry to grassland. *J. Agric. Eng. Res.*, **66**, 213-220.

Moncrieff, R. W. (1961). An instrument for measuring and classifying odours. *J. Appl. Physiol.*, **16**, 742-749.

Mopper, K. and Schultz, C. A. (1993). Fluorescence as a possible tool for studying the nature and water column distribution of DOC components. *Marine Chemistry*, **41**, 229-238.

Moriizumi, T., Nakamoto, T. and Sakubara, Y. (1992). Pattern recognition in electronic nose by artificial neural network models. In: *Sensors and sensory systems for an electronic nose*, edited by J. W. Gardner and P. N. Bartlett, Chapt. 14, Kluwer, Dordrecht.

Mounce, S.R., Day, A.J., Wood A.S., Khan, A., Widdop P.D. and Machel J. (2001). A neural network approach to burst detection. In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp. 349-356.

Moy, L., Tan, T. and Gardner, J. W. (1994). Monitoring the stability of perfume and body odours with an electronic nose. *Perfum Flavor*, **19**, 11-16.

Munch, E. V. and Pollard P. C. (1997). Measuring bacterial biomass-COD in wastewater containing particulate matter. *Water Research*, **31**, 2550-2556.

Nakamoto, T., Fukunishi, K. and Moriizumi, T. (1990). Identification capability of odor sensors using quartz-resonator array and neural network pattern recognition. *Sensors and Actuators B*, **1**, 473-476.

Namdev, P. K., Ahoy, Y. and Singh, V. (1998). Sniffing Out Trouble: Use of an Electronic Nose in Bioprocesses. *Biotechnol. Prog.* 14, 75-78.

Nayak, M. S., Dwivedi, R. and Srivastava, S. K. (1992). Transformed cluster analysis : an approach to the identification of gases / odours using an integrated gas-sensor array. *Sensors and Actuators B,* **12**, 103-110.

Nicolas, J., Romain, A.C. and Maternova, J. (2001). Chemometrics methods for the identification and the monitoring of an odour in the environment with an electronic nose. *Sensors and Chemometrics*, 75-90

Nicolas, J., Romain, A.C., Wiertz, V., Maternova, J. and Andre, P. (2000). Using the classification model of an electronic nose to assign unknown malodours to environmental sources and to monitor them continuously. *Sensors and Actuators B*, **69**, 366-371.

Niebling, G. (1994). Identification of gases with classical pattern-recognition methods and artificial neural networks. *Sensors and Actuators B*, **18-19**, 259-263.

Niebling, G. and Schlachter, A. (1995). Qualitative and quantitative gas analysis with non-linear interdigital sensor arrays and artificial neural networks. *Sensors and Actuators B*, **26-27**, 289-292.

Nikolaou, A.D., Gofinopoulos, S.K., Kostopoulou, M.N., Kolokythas, G.A. and Lekkas, T.D. (2002). Determination of volatile organic compounds in surface waters and treated wastewater in Greece. *Water Research*, **36**, 2883-2890.

Ogawa, S. and Sugimoto, I. (2001). Detecting odorous materials in water using quartz crystal microbalance sensors. In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp. 317-322.

Ortega, A., Marco, S., Sundic, T. and Samitier, J. (2000). New pattern recognition systems designed for electronic noses. *Sensors and Actuators B*, **69**, 302-307.

Orts, J., Llobet, E., Vilanova, X. Brezmes, J. and Correig, X. (1999). Selective methane detection under varying moisture conditions using static and dynamic sensor signals. *Sensors and Actuators B*, **60**, 106-117.

Osbild, D. and Vasseur, P. (1998). Microbiological sensors for the monitoring of water quality. In: *Monitoring of Water Quality*, Elsevier, pp. 37-49.

Pantsar-Kallio, M., Mujunen, S-P. Hatzimihalis, G., Koutoufides, P., Minkkinen, P., Wilkie, P.J. and Connor M.A. (1999). Multivariate data analysis of key pollutants in sewage samples: a case study. *Analytica Chimica Acta*, **393**, 181-191.

Pardo, A., Marco, S., Calaza, C., Ortega, A., Perera, A., Sundic, T. and Samitier, J. (2000). Methods for Sensors Selection in Pattern Recognition. In: *Electronic Noses and Olfaction 2000*, Conference proceedings, Brighton, UK, pp. 83-88.

Pardo, M. and Sberveglieri, G. (2002). *Learning from data: a tutorial with emphasis on modern pattern recognition methods*, Evaluation version, unpublished.

Parsons S. A. and Stephenson T. (2003). Introduction to Wastewater Treatment. In:. *Phosphorus in Environmental Technology - Removal, Recovery, Applications.* Valsami-Jones E (Ed)IWA, London.

Patterson, R.G., Jain, R.C. and Robinson, S. (1984). Odour controls for sewage treatment facilities. In: *Proceedings of the 77th annual meeting of the air pollution control association, San Francisco, June 1984.*

Pearce, T. C., Gardner, J. W., Friel, S., Bartlett, P. N. and Blair, N. (1993). An electronic nose for monitoring the flavours of beers, *Analyst*, **118**, 371-377.

Pelosi, P. and Persaud, K. C. (1988). Gas sensors: towards an artificial nose. In: *Sensors and sensory systems for advanced robots*, Dario, P. (ed.), NATO ASI series, **F43**, pp. 49-70.

Persaud, K. C. (1997). Arrays of broad specificity films for sensing volatile chemicals. In: *Handbook of biosensors and electronic noses: medicine, food and the environment*, Kress-Rogers, E. (ed.), CRC Press, pp. 563-592.

Persaud, K. C. and Dodd, J. H. (1982). Analysis of discrimination mechanisms of the mamalian olfactory system using a model nose. *Nature*, **299**, 352-355.

Persaud, K. C. and Pelosi, P. (1985). An approach to an artificial nose. *Trans. Am. Soc. Artif. Organs*, **31**, 297-300.

Persaud, K. C. and Pelosi, P. (1992). Sensor arrays using conducting polymers. In: *Sensors and sensory systems for an electronic nose*, Gardner, J. W. and Bartlett, P. N. (eds.), pp. 237-256.

Persaud, K. C., Khaffaf, S. M., Hobbs, P. J. and Sneath, R. W. (1996b). Assessment of conducting polymer odour sensors for agricultural malodour measurements. *Chem. Senses*, **21**, 495-505.

Persaud, K. C., Khaffaf, S. M., Payne, J. S., Pisanelli, A. M., Lee, D. H. and Byun, H. G. (1996a). Sensor array techniques for mimicking the mammalian olfactory system, *Sensors and actuators B*, **35-36**, 267-273.

Persaud, K. C., Pisanelli, A. M., Szysko, S., Reichl, M., Horner, G., Rakow, W., Keding, H.J. and Wessels, H. (1999). A smart gas sensor for monitoring environmental changes in closed systems: results from the MIR space station. *Sensors and Actuators B*, **55**, 118-126.

Ping, W. and Jun, X. (1996). A novel recognition method for electronic nose using artificial neural network and fuzzy recognition. *Sensors and Actuators B*, **37**, 169-174.

Pisanelli, A., Qutob, A. A., Travers, P., Szyszko, S. and Perseau, K. C. (1994). Applications of multi-array sensors to food industries. *Life Chem. Rep.*, **11**, 303-308.

Placak, O. R., Ruchhoft, C. C. and Snap, R. G. (1950). Copper and chromates ions in sewage dilutions. *Ind. Eng. Chem.*, **41**, 2238-2241.

Plumey, J. (1999). Quality monitoring for compliance assurance. In: *The Spencer's Guide to the UK Environmental Industry 1999*, B. M. Publishing, pp. 43-45.

Pouet, M. F., Thomas, P., Jacobsen, B. N., Lynggaard-Jensen, A. and Quevauviller, P. (1999). Conclusions of the workshop on methodologies for wastewater quality monitoring. *Talanta*, **50**, 759-762.

Preti, G., Gittleman, T. S., Staudte, P. B. and Luitweiler, P. (1993). Letting the nose lead the way – malodorous components in drinking water. *Analytical chemistry*, **65**, 699A-702A.

Qian, Z. and Tan, T. C. (1999). Response characteristics of a dead cell BOD sensor. *Water Research*, **32**, 801-807.

Rapp, M., Bob, B., Voigt, A., Gemmeke, H. and Ache, H. J. (1995). Development of an analytical microsystem for organic gas detection based on surface acoustic wave resonators. *Fres. J. Anal. Chem.*, **352**, 699-704.

Reddinger, J. L. and Reynolds, J. R. (1999). Molecular engineering of p-conjugated polymers. *Adv. Polym. Sci.*, **145**, 57-123.

Reynolds, D. (1999). Prospects for 'star wars' monitoring of water and wastewater quality. *WQI*, **Jan/Feb.,** 12-13.

Reynolds, D. M. and Ahmad, S. R. (1997). Rapid and direct determination of wastewater BOD values using a fluorescence technique. *Water Reseacrh*, **31**, 2012-2018.

Riedel, K., Lange, K. P., Stein, H. J., Khun, M., Ott, P. and Scheller, F. (1990). A microbial sensor for BOD. *Water Research*, **24**, 883-887.

Rigal S. (1995). Odour and flavour in wates: quantitative method for a new European standard. *Wat. Sci. Tech.*, **31** (11), 237-242.

Ripley, B. D. (1996). *Pattern recognition and neural networks*, Cambridge University Press, Cambridge.

Romain, A. C., Nicolas, J. and Andre, Ph. (1997). In-situ measurement of olfactive pollution with inorganic semiconductors: limitation due to the influence of humidity and temperature. *Semin. Food Anal.*, **2**, 283-296.

Romain, A. C., Nicolas, J., Wiertz, V., Maternova, J. and Andre, Ph. (2000). Use of a simple tin oxide sensor array to identify five malodours collected in the field. *Sensors and Actuators B*, **62**, 73-79.

Roncali, J. (1992). Conjugated poly(thiophenes): Synthesis, functionalisation and applications. *Chem. Rev.*, **92**, 711-738.

Rosen, C. (2001). A chemometric approach to process monitoring and control, with application to wastewater treatment operation. PhD thesis, Dept of industrial electrical engineering and automation, Lund University, Sweden, 278 pp.

Rosen, C. and Lennox, J. A. (2001). Multivariate and multiscale monitoring of wastewater treatment operation. *Water Research*, **35**(14), 3402-3410.

Rosen, C., Rottorp, J. and Jeppsson, U. (2003). Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. *Wat. Sci. Tech.*, **47** (2), 171-179.

Rose-Pehrsson, S. L., Shaffer, R. E. and Gottuk, D. T. (1999). Probabilistic Neural Network for Early Fire Detection Using an Electronic Nose. In: *Proceedings of ISOEN 99*, Tübingen, 195-198.

Rozzi, A., Ficara, E., Massone, A. and Verstraete, W. (2000). Titration biosensors for wastewater treatment process control. *Water21 Casebook*, **April,** 50-55.

Rozzi, A., Massone, A. and Alessandrini, A. (1997). Measurement of rbCOD as biological nitrate demand using a bisensor: Preliminary results. In: *3rd International Symposium Environmental Biotechnology*, Oostende, Belgium, April 21-23 1997.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning internal representations by error propagation. In: *Parallel Distributed Processing, Volume 1*, Rumelhart, D. E and McClelland, J. (eds), MIT Press, MA.

Ryman-Tubb, N. (1995). Computers learn to smell and taste. *Expert Systems*, **12**, 157-161.

Sawyer, N. C., McCarty, P. L. and Parkin, G. F. (1994). *Chemistry for Environmental Engineering*, 4th ed. McGraw Hill, New York, 658 pp.

Schroeder, E. D. (1977). *Water and wastewater treatment*, McGraw Hill, New York.

Schweizer-Berberich, M., Vaihinger, S., and Gopel, W. (1994). Characterisation of food freshness with sensor arrays. *Sensors and Actuators B*, **18,** 282-290.

Scully, P. (1998). Optical techniques for water monitoring. In: *Monitoring of Water Quality*, Elsevier, pp. 15-35.

Shaffer, R. E., Rose-Pehrsson, S. L. and McGill, R. A. (1999). A comparison study of chemical sensor array pattern recognition algorithms. *Analytica Chimica Acta,* **384**, 305-317.

Sharp, J. H., Suzuki, Y. and Munday, W. L. (1993). A comparison of dissolved organic carbon in North Atlantic Ocean nearshore waters by high temperature combustion and wet chemical oxidation. *Marine Chemistry*, **41**, 253-260.

Shepherd, A. J. (1997). *Second-order methods for neural networks*, Springer, New York.

Shurmer, H. V. (1990). The fifth sense: on the scent of the electronic nose. *IEE Review*, **March**, 95-98.

Singh, S., Hines, E. L. and Gardner, J. W. (1996). Fuzzy neural computing of coffee and tainted-water data from an electronic nose. *Sensors and Actuators B*, **30,** 185-190.

Skotheim, T. A., Elsenbaumer, R. L. and Reynolds, J. R. (1998). *Handbook of conducting polymers*, 2nd ed., Dekker, M. (ed.), **1**, 1097pp.

Spanjers, H., Vanrolleghem, P. A., Olsson, G. and Dold, P. (1998). Respirometry in control of the activated sludge process: Principles. *IAWQ scientific and technical report*, No. **7**.

Srivastava, A. K., Shukla, K. K. and Srivastava, S. K. (1998). Exploring neuro-genetic processing of electronic nose data. *Microelectronics Journal*, **29**, 921-931.

Statham, P. J. (1997). *Analytical Chemistry lecture notes (OC 603)*, M.Sc. in Oceanography, University of Southampton.

Statistica NN. (1998). *Statistica Neural Network user's manual.*

Stetter, J. R., Strathmann, S., McEntegart, C., Decastro, M. and Penrose, W. R. (2000). New sensor arrays and sampling systems for a modular electronic nose. *Sensors and Actuators B*, **69**, 410-419.

Stuetz R. M., Georges, S., Fenner, R. A. and Hall, S. J. (1999b). Monitoring wastewater BOD using a non-specific sensor array. *J. Chem. Technol. Biotechnol.*, **74**, 1069-1074.

Stuetz, R. M. and Fenner, R. A. (1998). Electronic nose technology: a new tool for odour management. *Water Quality Int.*, **July/Aug.**, 15-17.

Stuetz, R. M. and Nicolas, J. (2001). Sensor arrays: An inspired idea or an objective measurement of environmental odours? *Water Science and Technology*, **44**(9), 53-58.

Stuetz, R. M., Engin, G. and Fenner, R. A. (1998a). Sewage odour measurements using a sensory panel and an electronic nose. *Water Science and Technology*, **38**, 331-335.

Stuetz, R. M., Fenner, R. A. and Engin, G. (1999a). Characterisation of wastewater using an electronic nose. *Water Research*, **33**, 442-452.

Stuetz, R. M., Fenner, R. A. and Engin, G. (1999c). Assessment of odours from sewage treatment works by an electronic nose, H_2S analysis and olfactometry. *Water Research*, **33**, 453-461.

Stuetz, R. M., Fenner, R. A., Hall, S. J., Stratfull, and Loke, D. (2000). Monitoring of wastewater odours using an electronic nose. *Water Science and Technology*, **41**(6), 41-47.

Stuetz, R. M., White, M. and Fenner R. A. (1998b). Use of an electronic nose to detect tainting compounds in raw and treated potable water. *J. Water Services Res. Technol. Aqua*, **47**, 223-228.

Sugimoto, I., Seiyama, M. and Nakamura, M. (1999). Detection of petroleum hydrocarbons at low ppb levels using quartz resonator sensors and instrumentation of a smart environmental monitoring system. *Journal of Environmental Monitoring*, **1**(2), 135-142.

Szczurek, A., Szecowka, P. M. and Licznerki, B. W. (1999). Application of sensor array and neural networks for quantification of organic solvent vapors in air. *Sensors and Actuators B*, **58**, 427-432.

Tabachnick, B. G. and Fidell L. S. (1996). *Using multivariate statistics*. 3rd ed. Harper-Collins College Publishers.

Tan, T. C. and Qian, Z. (1999). Author's reply to "Comment on response characteristics of a dead cell BOD sensor". *Water Research*, **33**, 597-598.

Thomas, O., El Khorassani, H., Touraud, E. and Bitar, H. (1999). TOC versus UV spectrophotometry for wastewater quality monitoring. *Talanta*, **50**, 743-749.

Thomas, O., Theraulaz, F., Agnel, C. and Suryani, S. (1996). Advanced UV examination of wastewater. *Environ. Technol.*, **17**, 251-261.

Thomas, O., Theraulaz, F., Cerda, V., Constant, D. and Quevauviller, P. (1997). Wastewater quality monitoring. *Trends Anal. Chem.*, **16**, 419-424.

Topart, P. and Jocowicz, M. (1992). Characterisation of the interaction between (poly)pyrrole films and methanol vapour. *J. Phys. Chem.*, **96**, 7824-7830.

Traversa, E., Sadaoka, Y., Carotta, M. C. and Martinelli, G. (2000). Environmental monitoring field tests using screen-printed thick-film sensors based on semiconducting oxides. *Sensors and Actuators B*, **65**, 181-185.

Ueyama, S., Hijikata, K. and Hirotsuji, J. (2001). Water monitoring system for oil contamination using polymer-coated quartz crystal microbalance chemical sensor. In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp. 287-292.

Unistast 4.5. (1997). *Statistical package for Windows*: *User's Guide.*

Vaidyanathan, S., Macaloney, G., Vaughan, J., McNeil, B. and Harvey, L. M. (1999). Monitoring of submerged bioprocesses. *Crit. Rev. Biotechnol.*, **19**, 277-316.

Vig, J. R. (2001). Temperature-insensitive dual-mode resonant sensors – A review. *IEEE Sensors Journal*, **1**(1), 62-68.

Vincent, A. and Hobson, J. (1998). Odour Control, CIWEM "monographs on best practice" No2. Terence Dalton Publishing Ltd.

Viraghavan, T. (1976). Communication: Correlation of BOD, COD and soluble organic carbon. *J. Wat. Contr. Fedn.*, **48**, 2213-2214.

Wacheux, H. (1998). Sensors for waste water: many needs but financial and technical limitations. In: *Monitoring of Water Quality*, Elsevier, pp. 229-235.

Wilson, D. M., Hoyt, S., Janata, J., Booksh, K. and Obando, L. (2001). Chemical sensors for portable, handheld field instruments. *IEEE Sensors Journal*, **1**(4), 256-274.

Wilson, D., Dunman, K., Roppel, T. and Kalim, R. (2000). Rank extraction in tin-oxide sensor arrays. *Sensors and Actuators B*, **62**, 199-210.

Wilson, D., Roppel, T. and Kalim, R. (2000). Aggregation of sensory input for robust performance in chemical sensing Microsystems. *Sensors and Actuators B*, **64**, 107-117.

Wilson, F. (1997). Total Organic Carbon as a predictor of biological wastewater treatment efficiency and kinetic reaction rates. *Water Science and Technology*, **35**, 119-126.

Wolff, U., Dickert, F. L., Fischerauer, G. K., Greibl, W. and Ruppel, C. C. W. (2001). SAW sensors for harsh environments. *IEEE Sensors Journal*, **1**(1), 4-13.

Xu, S. and Hasselblad, S. (1996). A simple biological method to estimate the readily biodegradable organic matter in wastewater. *Water Research*, **30**, 1023-1025.

Yang, Z., Suzuki, H., Sasaki, S., McNiven, S. and Karube, I. (1997). Comparison of the dynamic transient and steady-state measuring methods in a batch type BOD sensing system. *Sensors and Actuators B*, **45**, 217-222.

Yoo, C.K., Choi, S.W. and Lee, I. (2001). Disturbance detection and isolation in the activated sludge process. In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp. 333-340.

Young, J. C. and Clark, J. W. (1965). History of the Biochemical Oxygen Demand Test. *Wat. Sew. Works*, **112**, 3.

[WWW document: http://www.appliedsensors.com], Date accessed: 08/10/2002

[WWW document: http://www.galactic.com/algoritms.html], Date accessed: 08/10/2002

[WWW document: http://www.nose-network.org], Date accessed: 08/10/2002

[WWW document: http://www.statsoft.com/textbook.html], Date accessed: 08/10/2002

PUBLICATIONS

PUBLICATIONS

Journals:

Bourgeois W, Gardey G., Servieres, M and Stuetz R.M. (2003) A chemical sensor array based system for protecting wastewater treatment plants. *Sensors and Actuators B* (in press)

Bourgeois W. and Stuetz R.M. (2002) Use of a chemical sensor array for detecting pollutants in domestic wastewaters. *Water Research,* 36, 4505-4512

Bourgeois W., Hogben P., Pike A. and Stuetz R.M. (2003) Development of a sensor array based measurement system for continuous monitoring of water and wastewater. *Sensors and Actuators B* (88), 312-319.

Bourgeois W., Burgess J. E. and Stuetz R.M. (2001) On-line monitoring of wastewater quality: a review. *Journal of Chemical Technology and Biotechnology,* 76, 1-12.

Bourgeois W. and Stuetz R.M. (2000) Measuring wastewater quality using a sensor array - prospects for real-time monitoring. *Water Science and Technology,* 41(12), 107-112

Bourgeois W., Romain, A.C., Nicolas J. and Stuetz R.M. (2002). The use of sensor arrays for environmental monitoring: interests and limitations. *Journal of Environmental monitoring.* (submitted)

Conferences and presentations:

ISOEN02 conference, September 2002, Rome, Italy: (oral presentation)
Bourgeois, W. and Stuetz, R.M. An upset early warning sensory device for process stream monitoring. (Paper submitted to IEEE.)

IMCS9 Conference, July 2002, Boston, USA: (oral presentation)
Bourgeois W., Gardey G., Servieres, M and Stuetz R.M. A chemical sensor array based monitoring system for the protection of wastewater treatment plants

3rd IWA Young Researcher Conference, April 2002, Nottingham, UK: (oral presentation)
Bourgeois W. Electronic nose technology for wastewater monitoring and inlet protection: Interests and limitations

School of Water Sciences Showcase, December 2001, Cranfield University, UK: (oral presentation)
Monitoring of wastewater quality using an electronic nose

IWA- ICA Conference, June 2001, Malmo, Sweeden: (oral presentation)
Bourgeois W., Gaugler M. and Stuetz R.M. On-line evaluation of chemical sensor arrays for monitoring changes in effluent quality In: *Instrumentation, Control and Automation, (IWA-ICA 2001),* Malmo, Sweden, preprints, pp 271-278.

NOSE I, 3rd Sch. June 2001, Sta Cesarea Therme, Lecce, Italy:
Application of sensor arrays for real-time monitoring of wastewater quality (poster + presentation)

2nd IWA Young Researcher Conference, April 2001, Cranfield University, UK:
Analysis of sensor array data for monitoring and predicting wastewater quality (oral presentation)

APPENDICES

APPENDIX A

Here follows some selected comments on the analysis of the data (Neil Collins, Marconi Applied Technologies). The prime purpose of these experiments was to optimise the system with respect to minimising the RH variation and, therefore, ensuring steady and consistent sensor responses.

All three reps were used in full which gives a very good error measure and ensures the model is as robust as possible, bearing in mind the limitations of squeezing the data into linearity.

Results:

ANOVA with respect to RH RSD%:

ANOVA; Var.:RH_SD%; R-sqr=**.65701**;
Adj:.53595
2**(3-0) design; MS
Residual=.1000886

DV: RH_SD% — **Analysis with respect to RH stability**

	SS	df	MS	F	p
(1)FLOW	0.91053199	1	0.91053199	9.09726194	0.00778168
(2)POROS	0.61616077	1	0.61616077	6.15615485	0.02384897
(3)TEMP	0.61017762	1	0.61017762	6.09637627	0.02444406
1 by 2	0.96731498	1	0.96731498	9.66458926	0.00638312
1 by 3	0.12041433	1	0.12041433	1.20307764	0.28800212
2 by 3	0.03465477	1	0.03465477	0.34624103	0.56398554
Error	1.7015058	17	0.10008858		
Total SS	4.96076026	23			

The prime effect is the flow/porosity interaction followed by the flow variable. Porosity and temperature are also very statistically significant.

The model fit is not very good, demonstrating that the system with respect to RH variability is highly non-linear in the experimental area measured. This immediately tells us two things:

1) Further experiments will be required in another region to examine the system and

2) Projections should NOT be made outside the measured experimental area as the model will break down rapidely.

Desirability plots were constructed from the models and are given below. The 1,2 interaction plot shows a complex shape and possibly two regions of maximum desirability: one with both variables at their HIGH values OR with both variables at their LOW value. It is not easy to see which is most appropriate in isolation. There are also signs that a 'saddle' is developing in the plot towards the bottom right. This

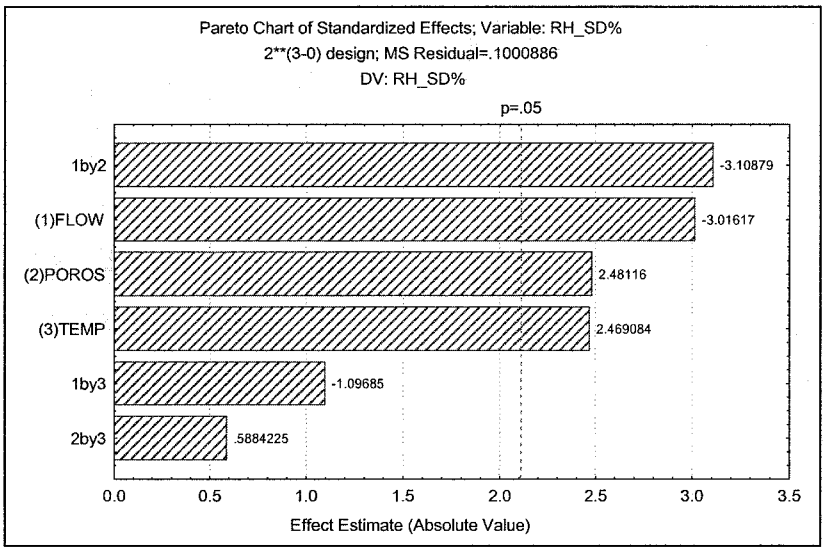
is a very amiguous result reflecting the linear approx. of a complex system. Further decisions cannot be made without infornation from the other plots. The 1,3 and 2,3 interaction plots are slightly more consistent and show that max. desirability is achieved with the variables at the following levels:

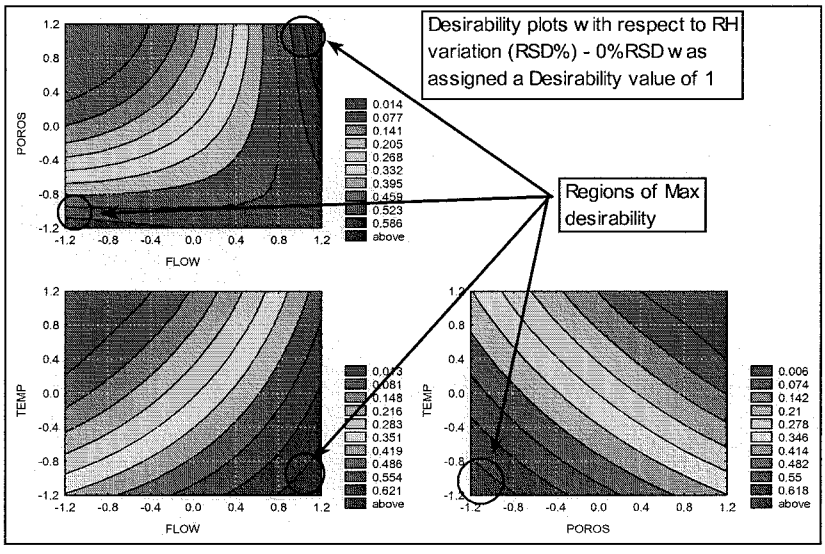
Flow	HIGH
Temp	LOW
Porosity	LOW

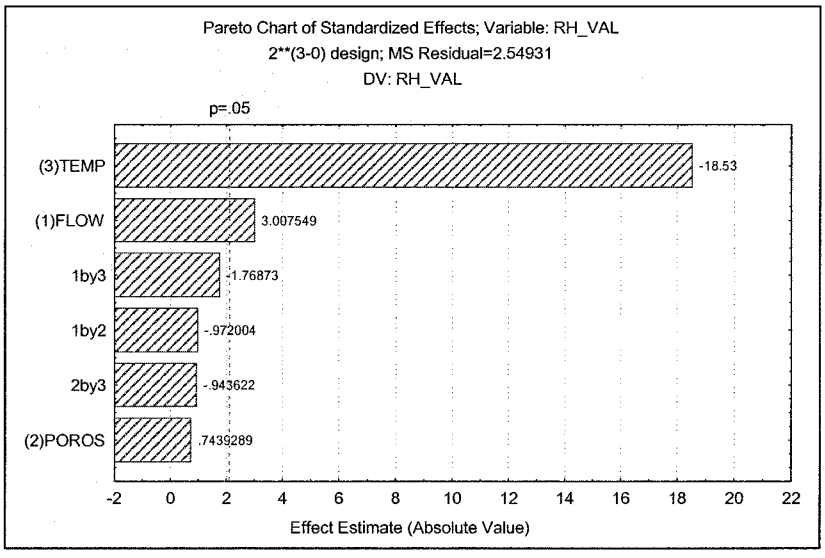
From this information it would appear that there may well be a saddle on the first plot and the two regions circled as possibles are not the way to proceed (too many countours giving a 'knife edge' system). To proceed, it is necessary to perform more experiments in a different area based on this information.

The following is suggested:
Flow rate: 130 and 160 ml/min (possibly higher but this could have bad additive effects on other parameters in the system because of the 1,2 interaction – Go carefully!)
Porosity: 0 and 1
Temp: two achievable temperatures below 30, the lower the better.

With reference to RH value obviously temperature is by far the biggest effect. The second variable of importance is flow rate. However, this is of secondary importance and use until the system is stabilised effectivley (once the system is under control in a robust manner, it will be possible to control and predict Rh with some degree of accuracy – But not until the system is robust)


Pareto Chart of Standardized Effects; Variable: RH_SD%
2**(3-0) design; MS Residual=.1000886
DV: RH_SD%
p=.05
1by2
(1)FLOW
(2)POROS
(3)TEMP
1by3
2by3
-3.10879
-3.01617
2.48116
2.469084
-1.09685
.5884225
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Effect Estimate (Absolute Value)


Desirability plots w ith respect to RH variation (RSD%) - 0%RSD w as assigned a Desirability value of 1
Regions of Max desirability
POROS
FLOW
TEMP
FLOW
TEMP
POROS
above


Pareto Chart of Standardized Effects; Variable: RH_VAL
2**(3-0) design; MS Residual=2.54931
DV: RH_VAL
p=.05
(3)TEMP
(1)FLOW
1by3
1by2
2by3
(2)POROS
-18.53
3.007549
-1.76873
-.972004
-.943622
.7439289
-2
0
2
4
6
8
10
12
14
16
18
20
22
Effect Estimate (Absolute Value)

APPENDIX B

Descriptive statistics for the TOC-Prosat subsets

Variable	Average	Std.Dev	Minimum	Maximum	cases
RH	34.00	8.25	20.70	55.50	2969
SENSOR1	1.09	0.37	0.34	2.27	2969
SENSOR2	-0.13	0.15	-0.65	0.60	2969
SENSOR3	2.20	0.81	0.93	4.43	2969
SENSOR4	1.53	0.59	0.62	3.21	2969
SENSOR5	10.36	4.63	4.49	23.73	2969
SENSOR6	2.93	0.88	1.21	5.28	2969
SENSOR7	2.87	0.92	1.30	5.06	2969
SENSOR8	13.13	2.76	8.62	19.31	2969
TOC	144.39	35.76	19.62	295.86	2969

Descriptive statistics for the Racod-Prosat subsets

Variable	Average	Std.Dev	Minimum	Maximum	Cases
RH	24.53	6.15	0.70	59.30	7598
SENSOR1	0.74	0.22	-0.22	2.20	7598
SENSOR2	-0.02	0.15	-0.52	1.99	7598
SENSOR3	1.37	0.42	-0.23	4.63	7598
SENSOR4	0.93	0.30	-0.23	3.29	7598
SENSOR5	5.80	2.04	-0.17	23.17	7598
SENSOR6	1.94	0.58	-0.56	4.58	7598
SENSOR7	1.86	0.53	-0.26	4.73	7598
SENSOR8	9.64	1.96	-0.15	18.47	7598
BOD	8.86	5.02	0	20	7598

Normality tests for each TOC-Prosat data sets

		S1	S2	S3	S4	S5	S6	S7	S8	Toc	Rh
Data set 1	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 2	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 3	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 4	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 5	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 6	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 7	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 8	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 9	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 10	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 11	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										
Data set 12	Shapiro-Wilks' w test										
	Lilliefors test										
	Kolmogorov-Smirnov										

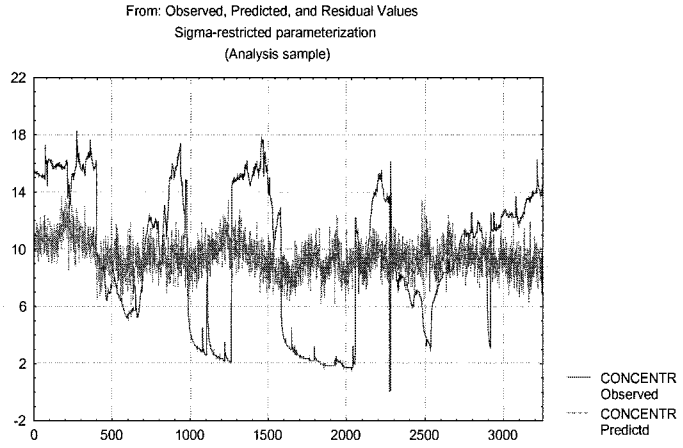
normal data distribution

APPENDIX C

RACOD- PROSAT
MLR sensors 1-8

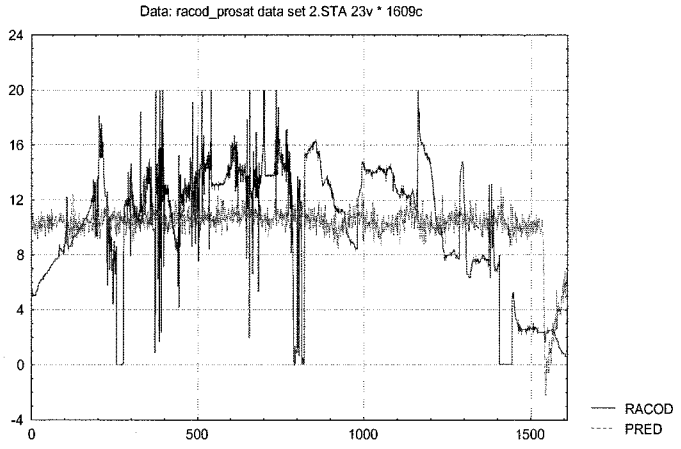
(Note: all raw data given in mA. Convertion to mg/l BOD: x*25)

Data set 1: 30.01 to 23.02.01



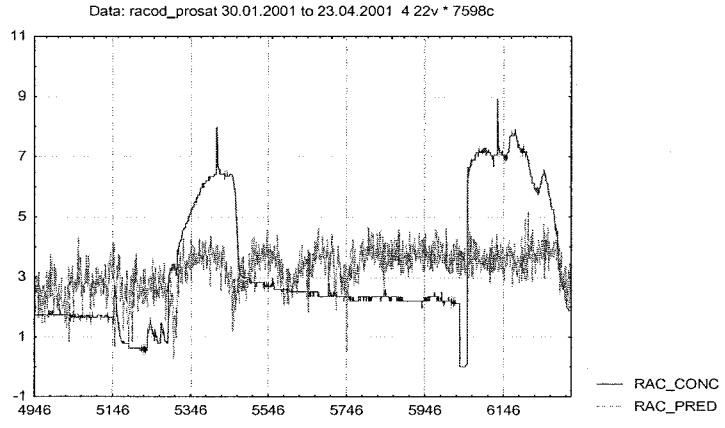
Error Min	0.00%
Error Max	12901%
Error Mean	107%
Error StDev	4.55

Data set 2: 27.02 to 13.03.01



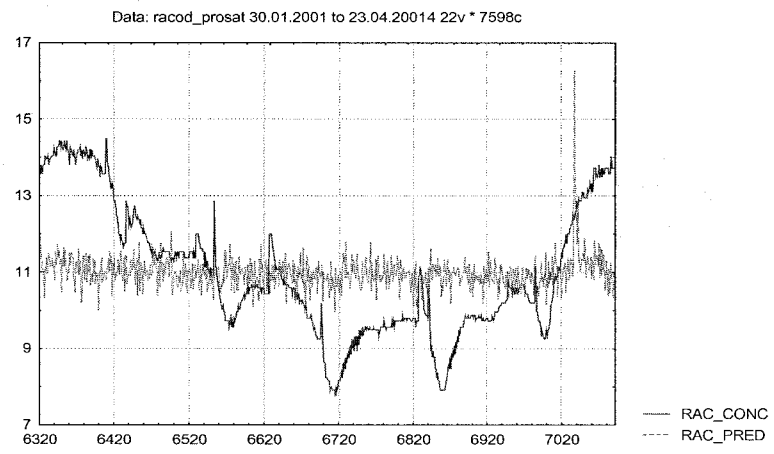
Error Min	0.00%
Error Max	16054%
Error Mean	466%
Error StDev	22.69

Data set 3: 28.03 to 07.04.01



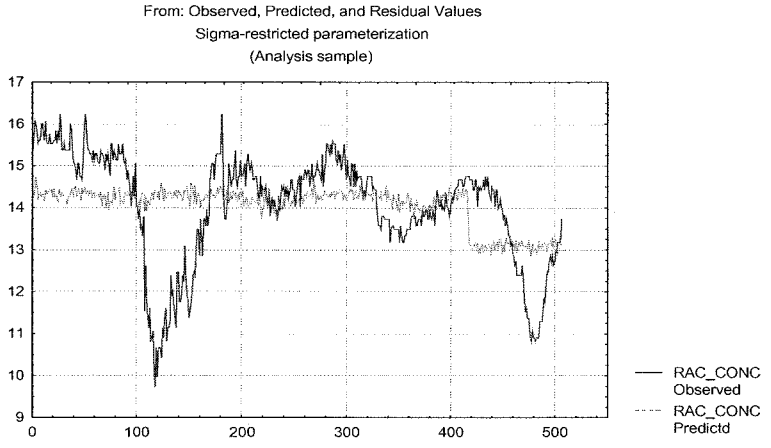
Error Min	0.00%
Error Max	4620%
Error Mean	76%
Error StDev	2.35

Data set 4: 12.04 to 17.04.01



Error Min	0.00%
Error Max	43%
Error Mean	13%
Error StDev	0.09

Data set 5: 20.04 to 23.04.01



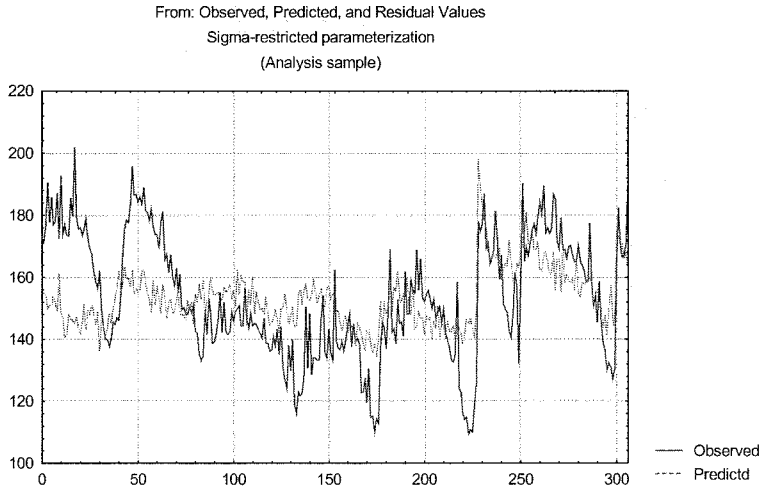
Error Min	0.00%
Error Max	48%
Error Mean	7%
Error StDev	0.07

APPENDIX D

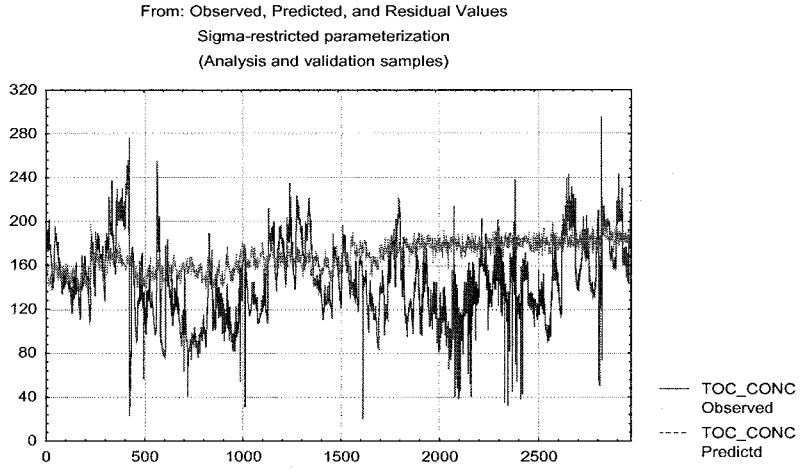
TOC - PROSAT
MLR sensors 1-8

Data set 1: 08.03 to 15.03.01

A) Training

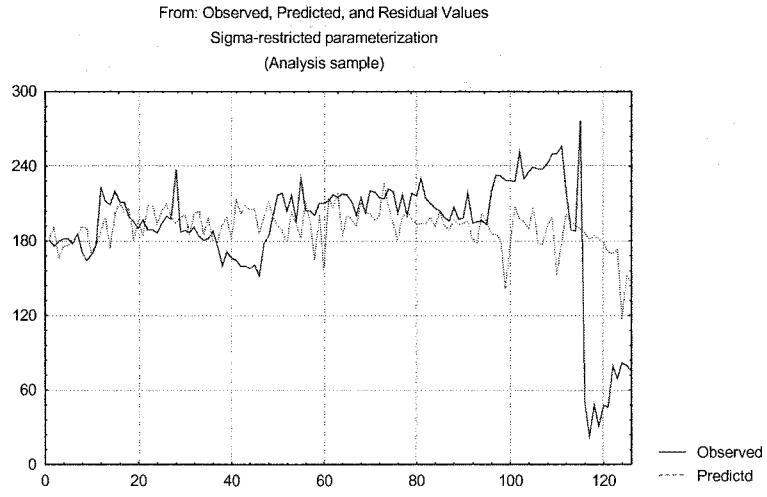


B) All data

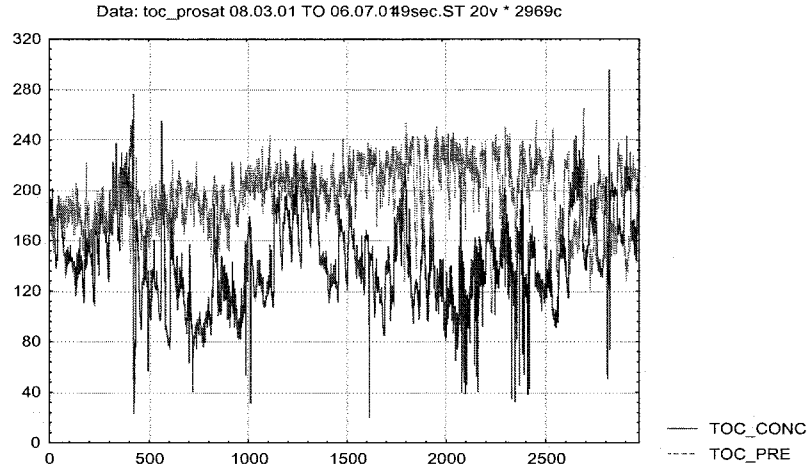


Data set 2: 28.03 to 31.03.01

A) Training

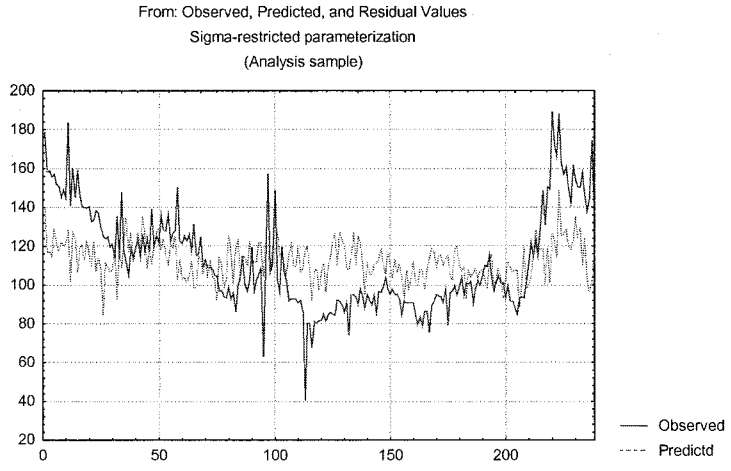


B) All data

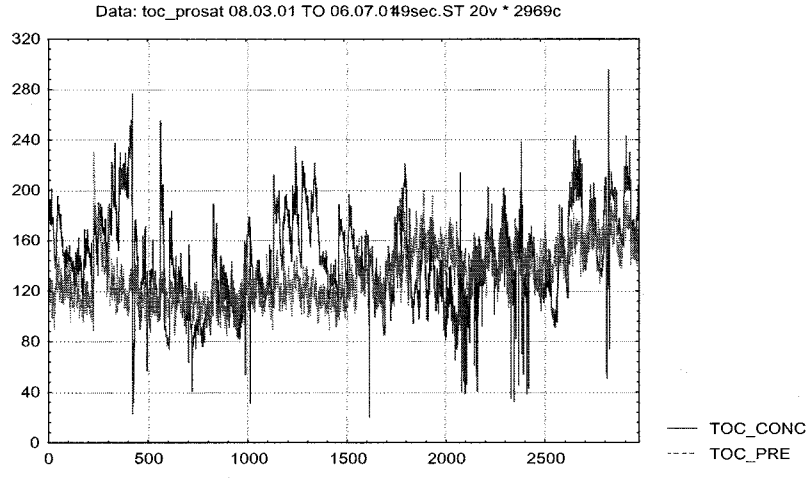


Data set 4: 12.04 to 17.04.01

A) Training

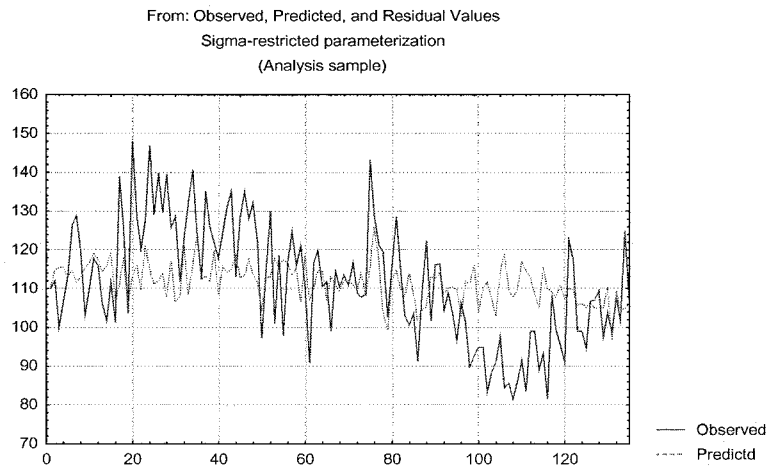


B) All data

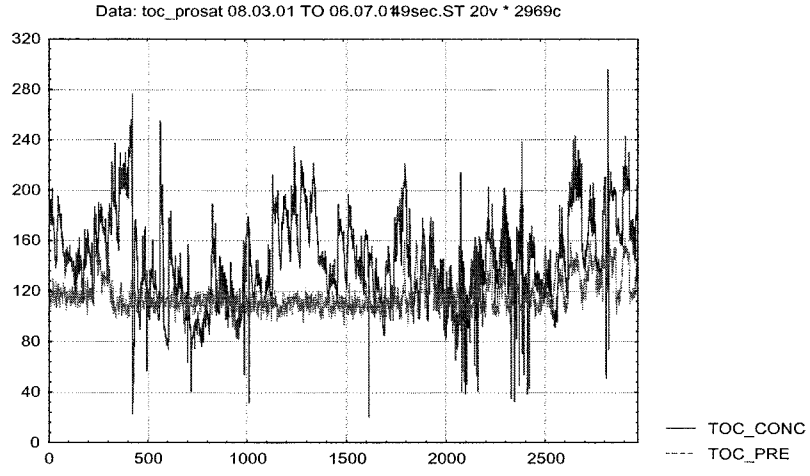


Data set 5: 20.04 to 22.04.01

A) Training

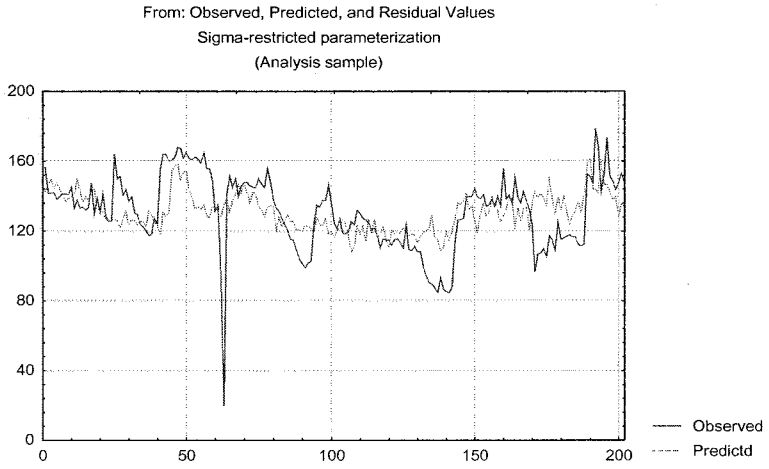


B) All data

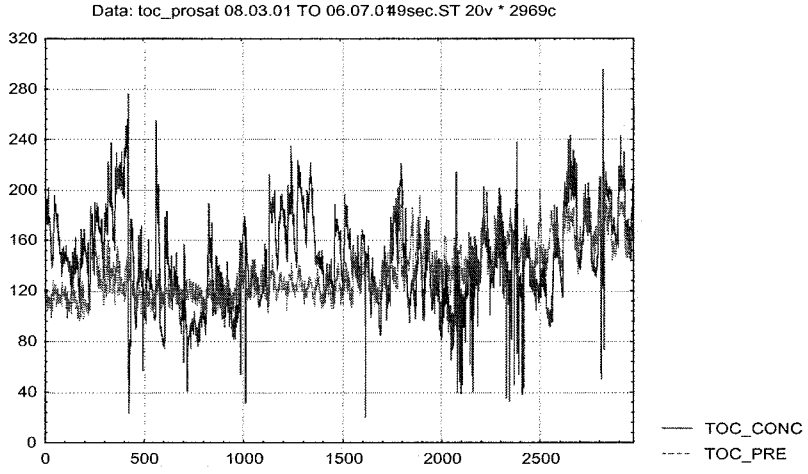


Data set 7: 14.05 to 20.05.01

A) Training

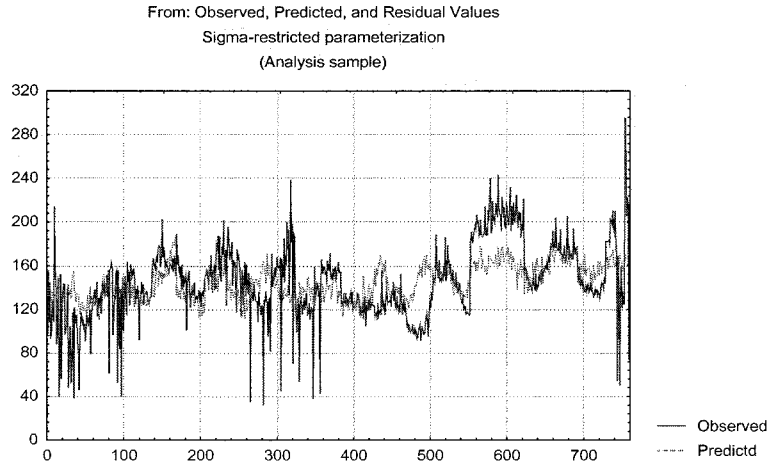


B) All data

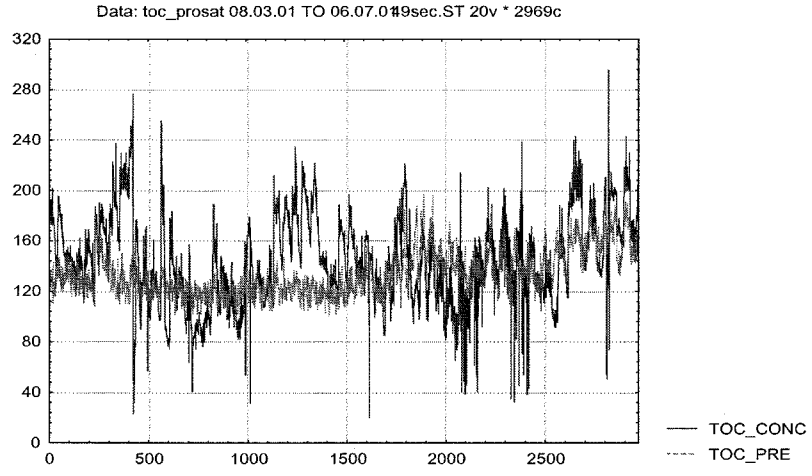


Data set 10: 14.06 to 29.06.01

A) Training

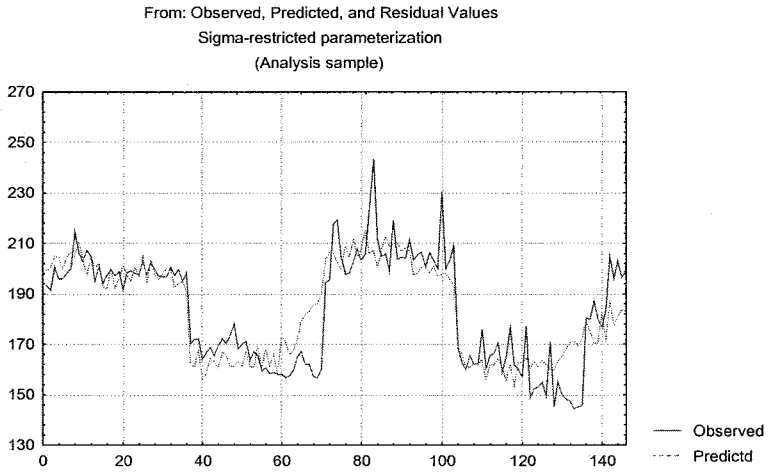


B) All data

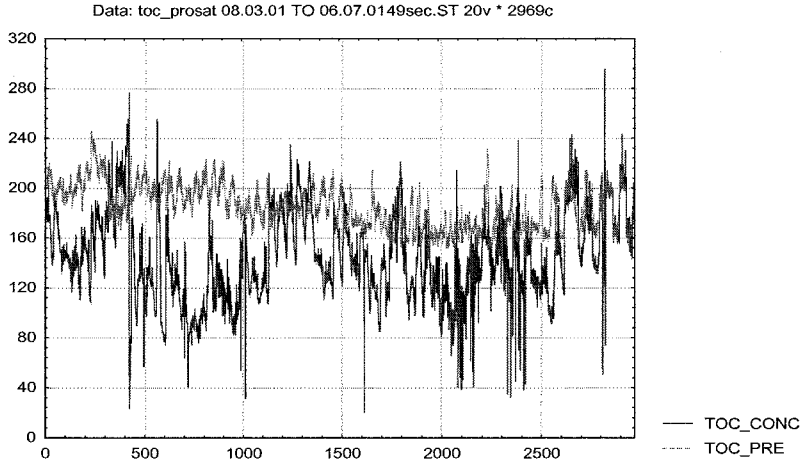


Data set 11: 04.07 to 06.07.01

A) Training



B) All data

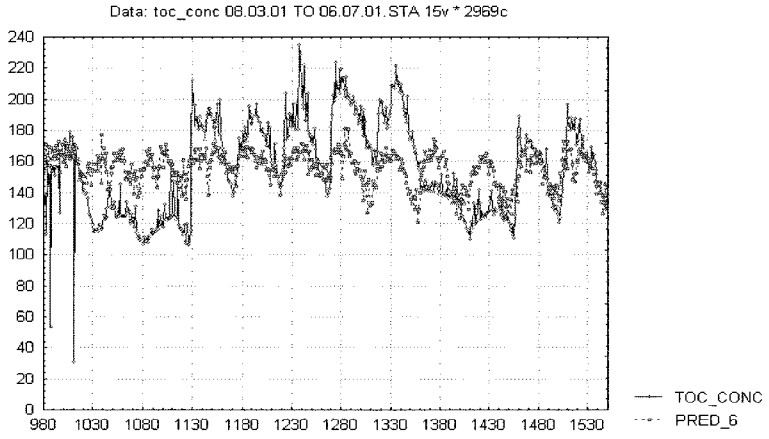


APPENDIX E

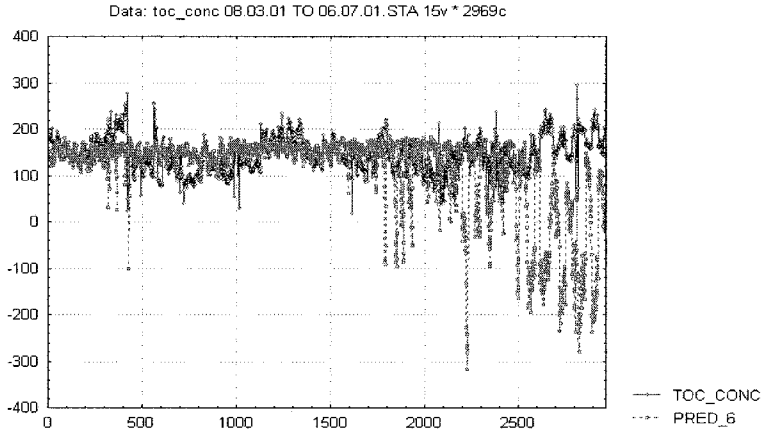
TOC - PROSAT
Polynomial Regression sensors 1-4

Data set 6: 26.04 to 09.05.01

A) Training

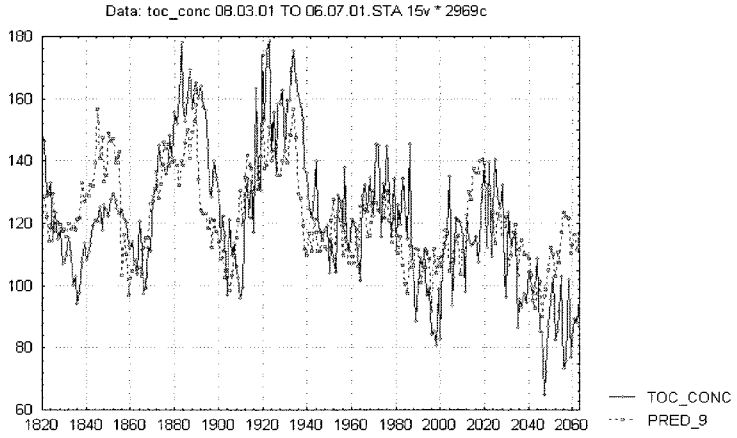


B) All data

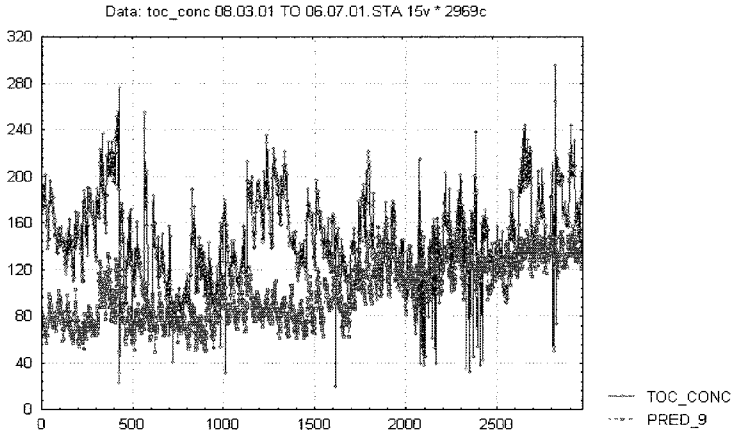


Data set 9: 27.05 to 02.06.01

A) Training



B) All data



2^{nd} degree polynomial regression predictions RAE and correlations on training sets (sensors 1-4)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
Min RAE (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE (%)	38	729	85	181	37	425	585	21	68	406	17
Mean RAE (%)	9	33	16	18	11	16	13	7	12	19	5
RAE StDev	0.07	0.91	0.14	0.16	0.08	0.22	0.42	0.05	0.1	0.35	0.04

Fraction of cases predicted with an RAE < x%, using 2^{nd} degree polynomial regression on training sets (sensors 1-4)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11
<20%	92.48	77.77	73.56	61.76	85.18	71.8	85.14	75.62	85.18	75.26	64.38
<30%	99.34	88.88	84.48	84.45	95.55	90.54	93.56	97.1	96.29	87.50	90.41
<50%	100	91.26	98.85	97.89	100	99.29	99.0	100	99.17	94.21	100

2^{nd} degree polynomial regression predictions RAE and correlations on all data (sensors 1-4)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
Min RAE (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max RAE (%)	748	856	1213	1255	470	755	585	825	471	585	802	634
Mean RAE (%)	67	48	71	55	25	45	22	29	33	21	56	21
RAE StDev	0.87	0.48	0.88	0.75	0.24	0.56	0.30	0.41	0.24	0.28	0.58	0.31

Fraction of cases predicted with an RAE < x%, using 2^{nd} degree polynomial regression on all data (sensors 1-4)

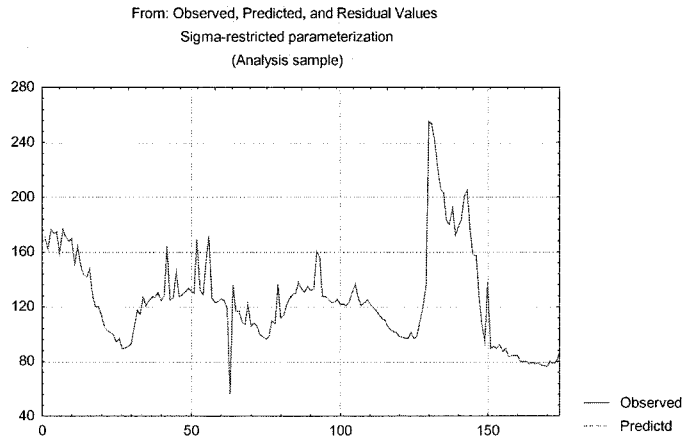
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
<20%	38.19	28.59	29.80	35.12	44.62	42.94	58.26	46.71	30.07	62.88	31.18	46.57
<30%	51.80	41.56	42.77	51.76	66.89	57.12	72.89	66.85	47.12	79.99	40.92	95.2
<50%	65.54	61.73	57.35	69.65	95.75	72.54	95.85	86.08	82.78	94.71	57.32	100

APPENDIX F

TOC - PROSAT
Factorial Regression sensors 1-8

Data set 3: 3.04 to 07.04.01

A) Training

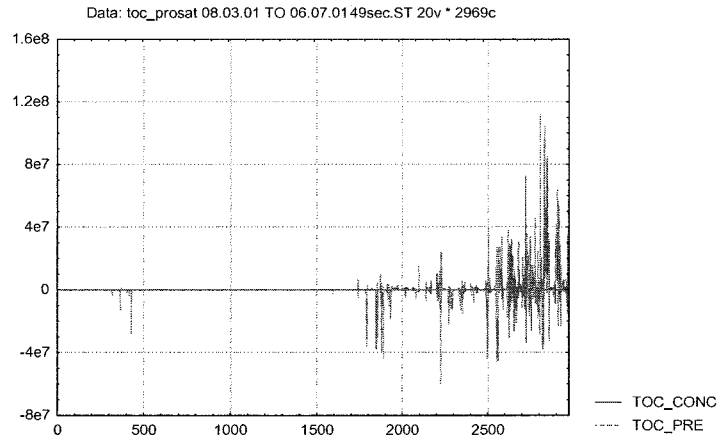


Error Min	0.00%
Error Max	0.00%
Error Mean	0.00%
Error StDev	0.00

<10%	100%
<20%	
<40%	

Correlation between:
Toc_pred and toc_conc: 1.00
Toc_pred and Rh: 0.67

B) All data

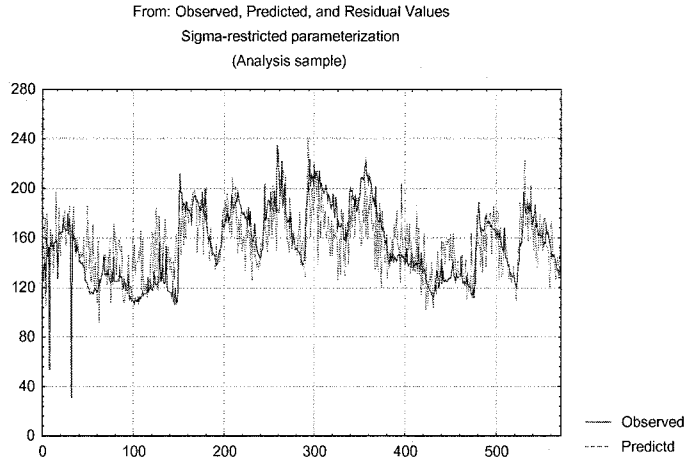


Error Min	0.00%
Error Max	64104610%
Error Mean	1657074%
Error StDev	51339.44

Correlation between:
Toc_pred and toc_conc: 0.26
Toc_pred and Rh: 0.29

Data set 6: 26.04 to 09.05.01

A) Training

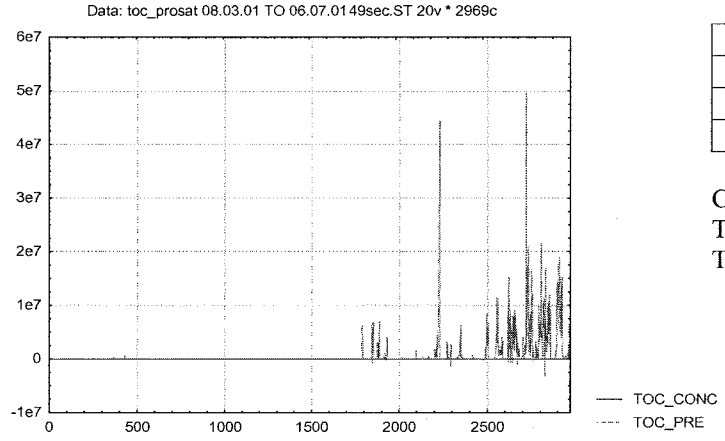


Error Min	0.00%
Error Max	419%
Error Mean	12%
Error StDev	0.20

<10%	57.26%
<20%	83.88%
<40%	97.72%

Correlation between:
Toc_pred and toc_conc: 0.68
Toc pred and Rh: 0.29

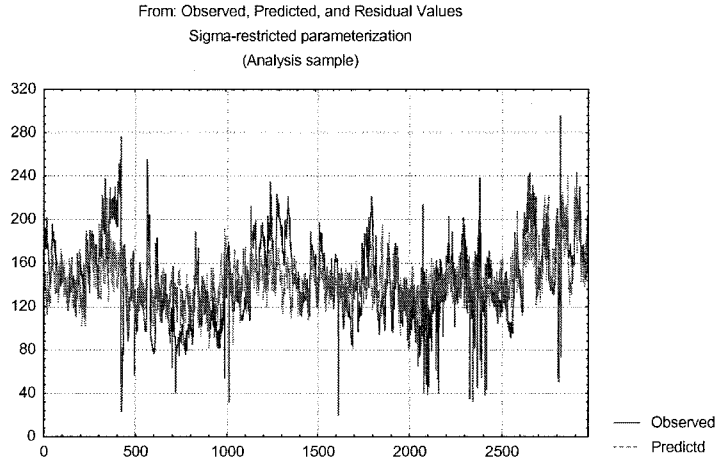
B) All data



Error Min	0.00%
Error Max	29555410%
Error Mean	418992%
Error StDev	16158.18

Correlation between:
Toc_pred and toc_conc: 0.24
Toc_pred and Rh: 0.44

Data set 12: 08.03 to 06.07.01



Error Min	0.00%
Error Max	681%
Error Mean	18%
Error StDev	0.30

<10%	40.08%
<20%	70.66%
<40%	92.93%

Correlation between:
Toc_pred and toc_conc: 0.56
Toc pred and Rh: 0.49