An Informatics Based Approach to Respiratory Healthcare

A thesis submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Alan Charles Barber
B.Eng.(Hons), MIET

School of Health and Social Sciences
Middlesex University

July 2009

Abstract

By 2005 one person in every five UK households suffered with asthma. Research has shown that episodes of poor air quality can have a negative effect on respiratory health and is a growing concern for the asthmatic. To better inform clinical staff and patients to the contribution of poor air quality on patient health, this thesis defines an IT architecture that can be used by systems to identify environmental predictors leading to a decline in respiratory health of an individual patient.

Personal environmental predictors of asthma exacerbation are identified by validating the delay between environmental predictors and decline in respiratory health. The concept is demonstrated using prototype software, and indicates that the analytical methods provide a mechanism to produce an early warning of impending asthma exacerbation due to poor air quality. The author has introduced the term ***enviromedics*** to describe this new field of research.

Pattern recognition techniques are used to analyse *patient-specific* environments, and extract meaningful health predictors from the large quantities of data involved (often in the region of ¼ million data points).

This research proposes a suitable architecture that defines processes and techniques that enable the validation of patient-specific environmental predictors of respiratory decline. The design of the architecture was validated by implementing prototype applications that demonstrate, through hospital admissions data and personal lung function monitoring, that air quality can be used as a predictor of patient-specific health. The refined techniques developed during the research (such as Feature Detection Analysis) were also validated by the application prototypes.

This thesis makes several contributions to knowledge, including: the *process architecture*; ***Feature Detection Analysis*** (FDA) that automates the detection of trend reversals within time series data; validation of the ***delay characteristic*** using a Self-organising Map (SOM) that is used as an unsupervised method of pattern recognition; ***Frequency, Boundary and Cluster Analysis*** (FBCA), an additional technique developed by this research to refine the SOM.

Acknowledgements

This thesis would not be complete without mention of the many people who provided their encouragement, support, guidance, and quite often the odd word of wisdom, throughout the course of this research.

Thanks are due to Prof. Richard Bayford and Prof. Ron Hamilton, my supervisors, for guiding my work, and Prof. Ray Iles for providing advice. Thanks also to Helen Crabbe who influenced my early work and helped form ideas and direction, and John Watt who aided my writing style. I would like to thank all participants of the Medicate research project, during which my early concepts first originated, also Dr. Robert Colson who offered advice at an early proof reading stage, and those who kindly proof read early drafts from cover to cover. Thanks are also due to my examiners: Prof. Mark Woodman, Dr. Panos Liatis, and Dr. John Crowe, whose input has enhanced the readability and holistic nature of this research.

Special thanks go to my Mum, and especially my Dad who probably deserves his own PhD for coping with (and probably enjoying?) all the long discussions over the merits of my ideas and asking probing questions, and who certainly helped completion of the thesis during the final proof reading stages. A special thanks is owed to Anna, who volunteered this research with six months of her respiratory readings, without whom the work would not have been tested with such an extensive data set.

I would like to thank those at Monarch Charts and FastObjects who diligently corresponded with me in the development of their graphing package, and in their database discussion forum.

Thanks also, to all my former colleagues at Sun Microsystems, and my year spent there before this project began. Thank you for shaping my early interest in the use of technology, to solve real world problems.

Acknowledgements would not be complete without thanks to Rebecca. Thank you for your support, over the many years you gave it. Thank you – I appreciated it.

To all my friends: THANK YOU!

To all who asked, "is it finished yet", I thank you, and – Yes.

'Good intelligence is nine-tenths of any battle'. (Napoleon)

Contents

Figures

Tables

Chapter 1

Introduction

This chapter sets out the objectives, scope of thesis, and methodology used by this research, to identify processes capable of validating predictors of patient-specific asthma exacerbation.

1.1. Purpose of the Research

This work progresses research in the area of health informatics, in particular how clinicians and researchers use IT systems to investigate the affect of patient-specific environments on respiratory health (particularly periods of asthma exacerbation). The research embodies the improvements in a new process architecture that allows clinical staff and researchers to take the work forward.

Prior work has many limitations (correlation is discussed in Section 1.1.2), the complexity of interacting parameters means that the development of large-scale systems capable of collecting and analysing correlations between environmental factors and asthma can be difficult to achieve. In this thesis a new system to identify and monitor environmental predictors of respiratory decline for large scale studies is proposed.

There is evidence that environmental factors contribute to the decline in respiratory health, and sometimes death of a patient (Rabinovitch *et al.*, 2004). Uncertainty still remains as to the causes of some respiratory diseases, but the link between changing environmental factors as the possible trigger to a decline in respiratory health has been shown (Chin-Shen *et al.*, 2007; Stedman, 2001). Desensitisation to environmental conditions may render one person quite able to live a perfectly normal life without experiencing the symptoms of asthma, while another is hospitalised. Environmental factors, such as air quality, and their effect on people with respiratory disease have been a particular target for research in past years (Blanc *et al.*, 2005; Kim *et al.*, 2004; Kim J H *et al.*, 2005). Amongst the many respiratory conditions, asthma effects 300 million people worldwide and in 2005, asthma

contributed to approximately 255,000 deaths (WHO, 2007). The World Health Organisation states that asthma deaths will increase by almost 20% in the next 10 years if urgent action is not taken.

1.1.1 The Respiratory System

Respiration is the process by which all living organisms obtain oxygen, which is required to convert fuel into energy. The millions of cells in the human body facilitate the release of energy by a chemical reaction involving glucose and oxygen, so a supply of these ingredients must be maintained to them through the bloodstream. Oxygen is absorbed into the blood in large volumes via an efficient gaseous exchange surface, which is provided by an intricate structure of air-sacs (alveoli) within the lungs. These alveoli form an interface between the respiratory system and the rest of the body. Other components of the respiratory system provide a mechanism by which air may be inhaled and expired, and all these essential apparatus function as one distinct unit.

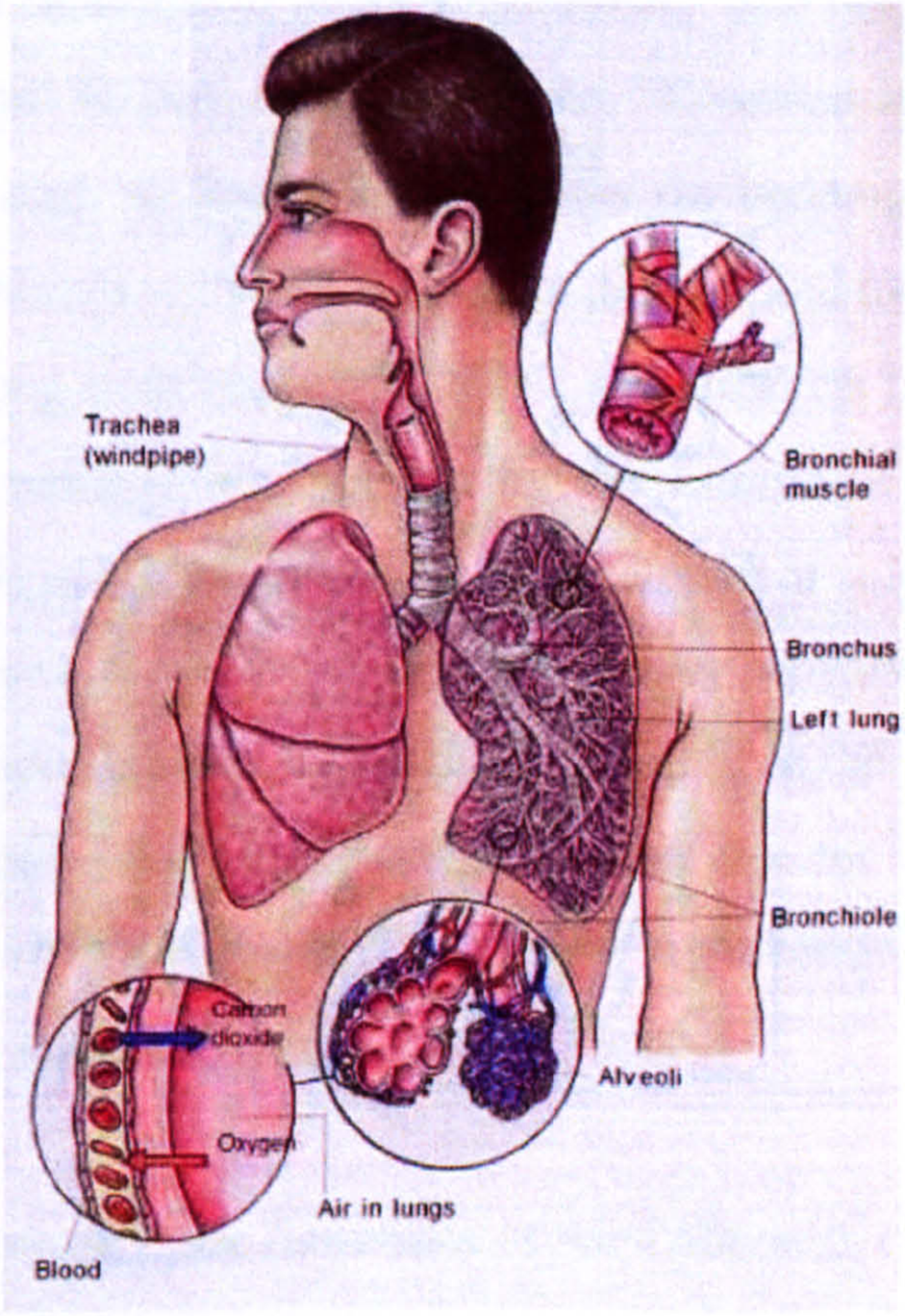


Figure 1 Anatomy of the respiratory system.

(Ayres J, 2005)

There are a variety of diseases which may impair the flow of air through the respiratory

system. *Breathlessness* can be caused by blockage or inflammation of the various tubes, or of the alveoli. Diseases such as: Bronchitis, Emphysema and Pneumonia, involve increasing difficulty of breathing over time due to gradual inflammation, excessive production of mucus, or the collapse of the lung tissue. However, one respiratory condition, Asthma, manifests itself in sporadic attacks of *wheezing*, caused by the narrowing of the bronchial passages when triggered by some agent or agents.

Lane (1996) states that the Greek word, Asthma, which literally means *panting* has been used for many years to describe clearly recognisable attacks of breathlessness. In order to illustrate early confusion when attempting to study the causes of asthma, he cites a twelfth century physician (Maimonides) who wrote that, "This disease has many aetiological aspects ... it cannot be managed without a full knowledge of the patient's constitution as a whole ... furthermore I have no magic cure to report". Lane (1996) uses the inconclusive outcome of an attempt by a panel of experts to define asthma, as an indication of the very complicated nature of this disease.

Many factors influence the condition of an asthma patient. "The Pocket Guide for Asthma Management and Prevention" (GINA, 2006) states that, "Common asthma triggers include viral infections; allergens such as domestic dust mites (in bedding, carpets, and fabric-upholstered furnishings), animals with fur, cockroach, pollens, and moulds; tobacco smoke; **air pollution**; exercise; strong emotional expressions; and chemical irritants." The effect of air pollution in asthmatic patients is supported by the Journal of Allergy and Clinical Immunology (JCAAI, 1995) which suggests that the inhalation of sulphur dioxide, nitrogen dioxide, or ozone is capable of inducing bronchospasm in patients with asthma. Stocks (1996) also indicates that there has been a significant increase in the prevalence, morbidity and mortality of asthma during the past 20 years, the reasons for which remain unclear although increasing levels of environmental pollutants and changes in housing customs (e.g. central heating, fitted carpets, etc.) have been implicated.

However according to the report, "UK Emissions of Air Pollutants 1970 to 2001" by Dore *et al.* (2003), atmospheric pollution has declined over the last 30 years. This trend indicates that there is more to the problem of identifying asthma triggers than simply analysing air quality data.

The National Asthma Audit (NAA, 2000) shows that in the UK, approximately one in

seven children aged 2 to 15 (over 1.5 million people) and at least 1 in 25 adults over the age of 16 (1.9 million people), have asthma-like symptoms which require treatment. The Global Initiative for Asthma (GINA, 2006) estimates that there are 300 million asthma sufferers world-wide. Some researchers (Cochrane *et al.*, 1996) estimate that 20 per cent of people with asthma can be described as having a "severe or very severe" condition. This means that they might have daily symptoms, frequent trips to the hospital, miss time from work or school and have a poorer quality of life. In 1997 the United Kingdom recorded 1584 deaths, which were directly attributed to asthma (ONS, 1997). The Office for National Statistics (2007) found that between 1993 and 2005 there was a steady decrease (by 27 per cent) in the number of avoidable male deaths due to respiratory disease. Amongst females there is a more complex pattern, with the rate per 100,000 population fluctuating between 15 and 18 across the entire period, suggesting no clear trend (ONS, 2007).

AsthmaUK estimates that the total cost of asthma to the UK is now in excess of £2000 million a year, calculated from estimated figures for National Health Service (NHS) expenditure, lost productivity and Department of Social Security (DSS) Sickness and Invalidity Benefits. In 2001, it was estimated that asthma cost the NHS £889 million (AsthmaUKa, 2007). Of this, £49 million (5.5%) was spent on hospital admissions for asthma. Since then, costs have risen. It is estimated that 75% of emergency admissions for asthma could be avoided with more appropriate and timely care. These figures are formulated from the cost of GP consultations, prescriptions for asthma medication, hospital in-patient and out-patient care and referrals to A&E departments. There were 59,859 hospital admissions for asthma in England in 2003, rising to 67,713 in 2004 (DH, 2004). In 2005, a hospital stay for asthma cost an average of £860.89 per patient, ranging from £781 for each uncomplicated hospital admission to £1,218 for those people experiencing an asthma attack with complications (AsthmaUKa, 2007). According to the Department of Health (DH, 2004), based on hospital admissions for 2004, that makes an estimated £58.3 million for hospital management of asthma in England each year. Caring for people who experience an asthma attack costs 3.5 times more than caring for those whose asthma is well managed (Hoskins *et al.*, 2000). This huge personal and economic cost means that any successful improvement in asthma management will help to increase the quality of life for asthma patients whilst reducing costs to healthcare providers.

Exposure of sensitive patients to inhalant allergens has been shown to increase airway

inflammation, airway hyper-responsiveness (the occurrence of wheezing and dyspnoea after exposure to allergens, environmental irritants, viral infections, cold air, or exercise), asthma symptoms, need for medication, and death due to asthma. Respiratory diseases such as asthma have for a long time been associated with influencing factors in the environment. Substantially reducing environmental exposures, significantly reduces these outcomes (NHLBI, 1997a). Conditions such as pollution and concentrations of pollen or dust are among the most common irritants. The term *environment* is used to encompass many different areas, with meteorology, air quality, particulate matter and occupational and domestic environments being the main ones. Asthma triggers are commonly found in these areas but have also been known to stem from strong emotional expressions, personality and inherited factors (Lane, 1996). Some patients experience asthma symptoms only in relationship to certain pollens and moulds. If the patient has seasonal asthma on a predictable basis, daily long-term medication should be initiated prior to the anticipated onset of symptoms and continued through the season.

Lane (1996) writes that asthmatics have irritable airways due to inflammation caused by allergy, infection and the effect of air pollution, "but rarely does one act alone". The combination of root causes such as allergens with pollutants or the smoker with an occupational hazard are documented by Lane to show this effect.

A key issue in the management of (respiratory) disease is the collection and interpretation of data (van den Hazel, 2007). The use of portable electronic monitoring devices carried by the ambulatory patient to record their respiratory condition improves the detail and accuracy of associated data over mechanical and manual means of taking and recording lung function data. Data collected as a patient moves from location to location provides a basis on which to probe for patterns or direct correlations between the respiratory data set and monitored environmental influencing factors (Chin-Shen *et al.*, 2007). A detailed picture of a patient's condition can be recorded and analysed. Currently, respiratory data (for an asthmatic) is used by clinicians to give an indication of how well a particular patient manages their asthma against standards set by research councils (BTS, 1995). The data used in the process however, does not give an indication of the influencing factors of respiratory episodes unless it is used in conjunction with a health diary kept by the patient (Reznik *et al.*, 2005) or analysed against real-time monitoring information (Cobern *et al.*, 2005; Crabbe *et al.*, 2004). It is not common practice to monitor the patient's environment on a continual basis, therefore short term events that continually impact the health of the

asthmatic are not validated, or at worst go undetected. The identification of factors contributing to a decline in respiratory health enable both the clinician and patient to better manage the condition. Knowledge is increased, quality of life improved through avoidance of detected influencing factors, and cost of treatment decreased by reducing the frequency of emergency cases. This thesis develops the processes necessary to define and construct a prototype software system, capable of facilitating the identification of factors that predict a decline in respiratory health.

1.1.2 Issues with Correlation as a Technique

During the research for this thesis, a number of techniques were explored. An early technique involved the use of correlation (Crabbe *et al.*, 2004). The technique analysed daily average and daily minimum lung function readings for a patient against the daily maximum for an environmental pollutant. The work concluded that there was some correlation between data set characteristics, but that further research was required, due to the limited size of data sets to ascertain a precise result.

An automated correlation model was used to explore the technique further during research for this thesis. The model analysed raw environmental and lung function data, and aimed to provide a tool which would assist in the identification of environmental time-series segments having a high correlation to low lung function measurements (shown in Figure 2). Correlation requires data at the same time interval, to enable a comparison between data sets. The sampling rate of the raw lung function data is irregular (reliant on the patient), while air quality sampling can be automated and recorded each hour for example. This makes the introduction of an interpolation technique necessary to estimate data values between actual readings. Interpolation techniques are discussed in Section 5.4.3. The use of an interpolation technique is necessary to synchronise the raw data sets, and allows the direct relationship between environmental and medical data sets for any given section to be analysed, (shown between the red start and finish lines in the top half of the Figure 2). The data period is set arbitrarily, and this example does not include a lag between the data sets. The second half of the figure, marked "correlation plot" shows the correlation of interpolated data values (every 12 minutes) between the start and finish markers in the top half of the figure. The advantage of this technique over interpreting daily averages or minimum/maximum readings is that it has an improved time

resolution, allowing individual exposure to environmental factors to be analysed if data is available.

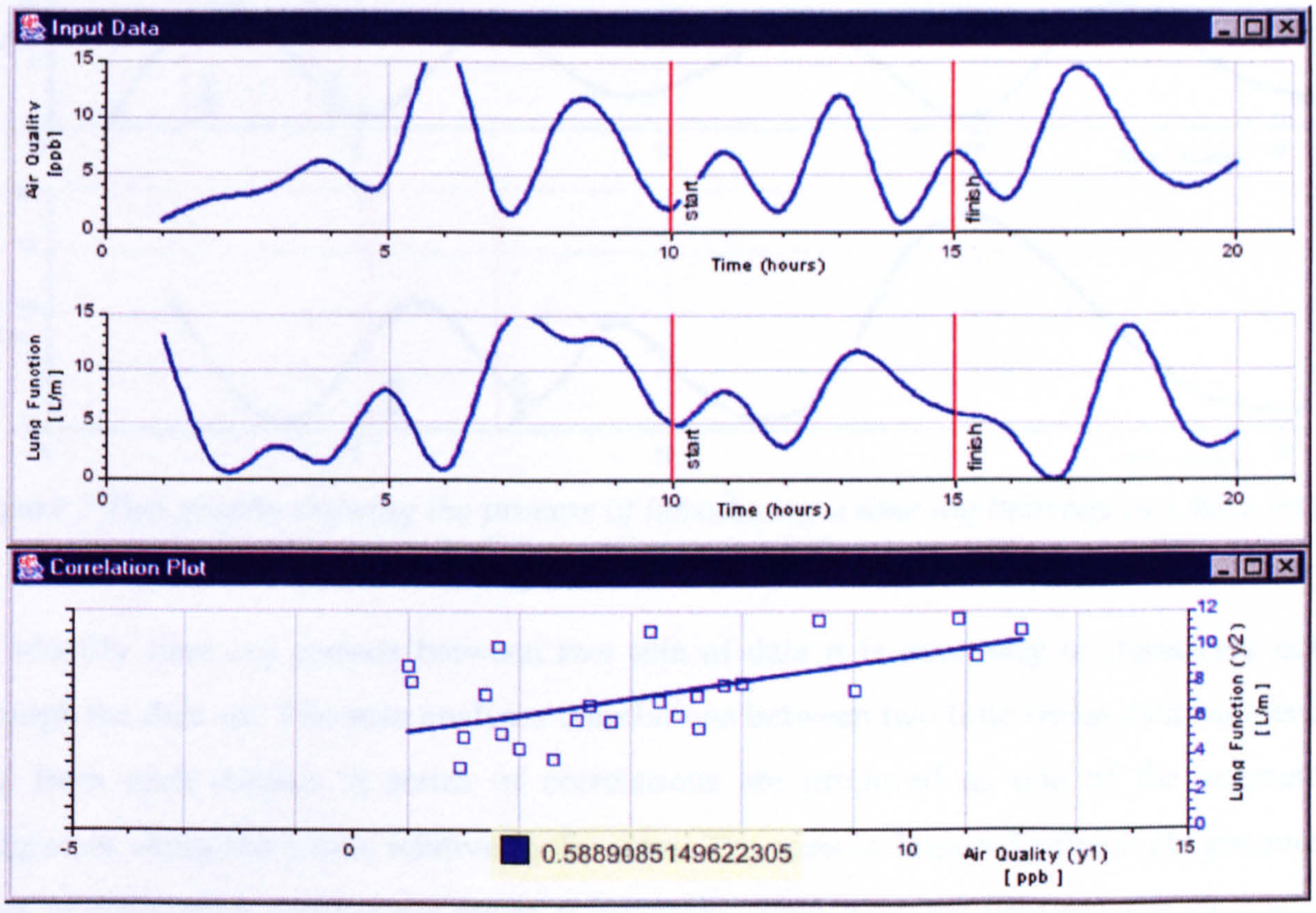


Figure 2 Correlation example, showing the xy plot of the two data sets (y values between the red start and finish lines) with the regression line of best fit and the correlation coefficient (0.589 to 3dp).

By using B-spline interpolation between raw data points, the technique is susceptible to noise if damping coefficients are not used. The use of damping coefficients reduces the volatility of the interpolated data points. An accurate fit (or estimation) of data is also dependant on the regularity of data readings. The air quality data in the figure is recorded regularly with an hourly interval, which reduces the likelihood of fluctuations in the estimation. However, the regularity of lung function readings is not guaranteed. For this reason, it may be more prudent to use straight line interpolation between raw data points rather than assuming that the smoothed curve represents the true characteristics of the underlying data. However, this requires further investigation.

A further technique, known as *time lag analysis* was integrated into the analysis to identify whether correlations existed between the data sets, but with a shift in time. For example, if a critically low peak expiratory flow (PEF) reading was given at 1pm in the afternoon and a very high sulphur dioxide reading was recorded at 11am, then it may be reasonable to assume a 2 hour time lag period connected the two pieces of data. An example is shown in

Figure 3.

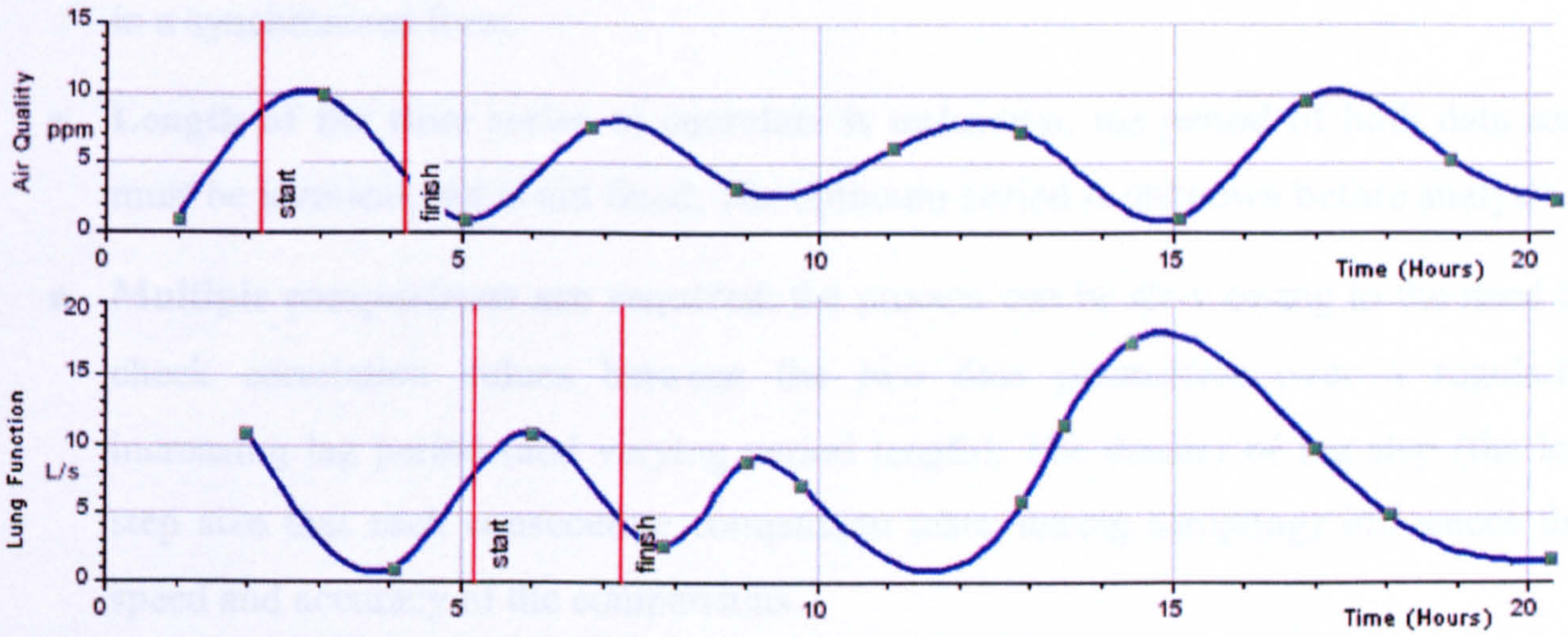


Figure 3 Two graphs showing the process of introducing a time lag between two data sets before correlation analysis

To identify time lag periods between two sets of data it is necessary to iteratively scan through the data set. The scan analyses correlations between two time series data segments, one from each dataset. A series of correlations are produced as one of the segments progresses along the x-axis relative to the other. The scan in Figure 3 effectively produces a set of correlation coefficients made at increasing time intervals between the air quality and lung function data sets, producing a graph similar to that in Figure 4.

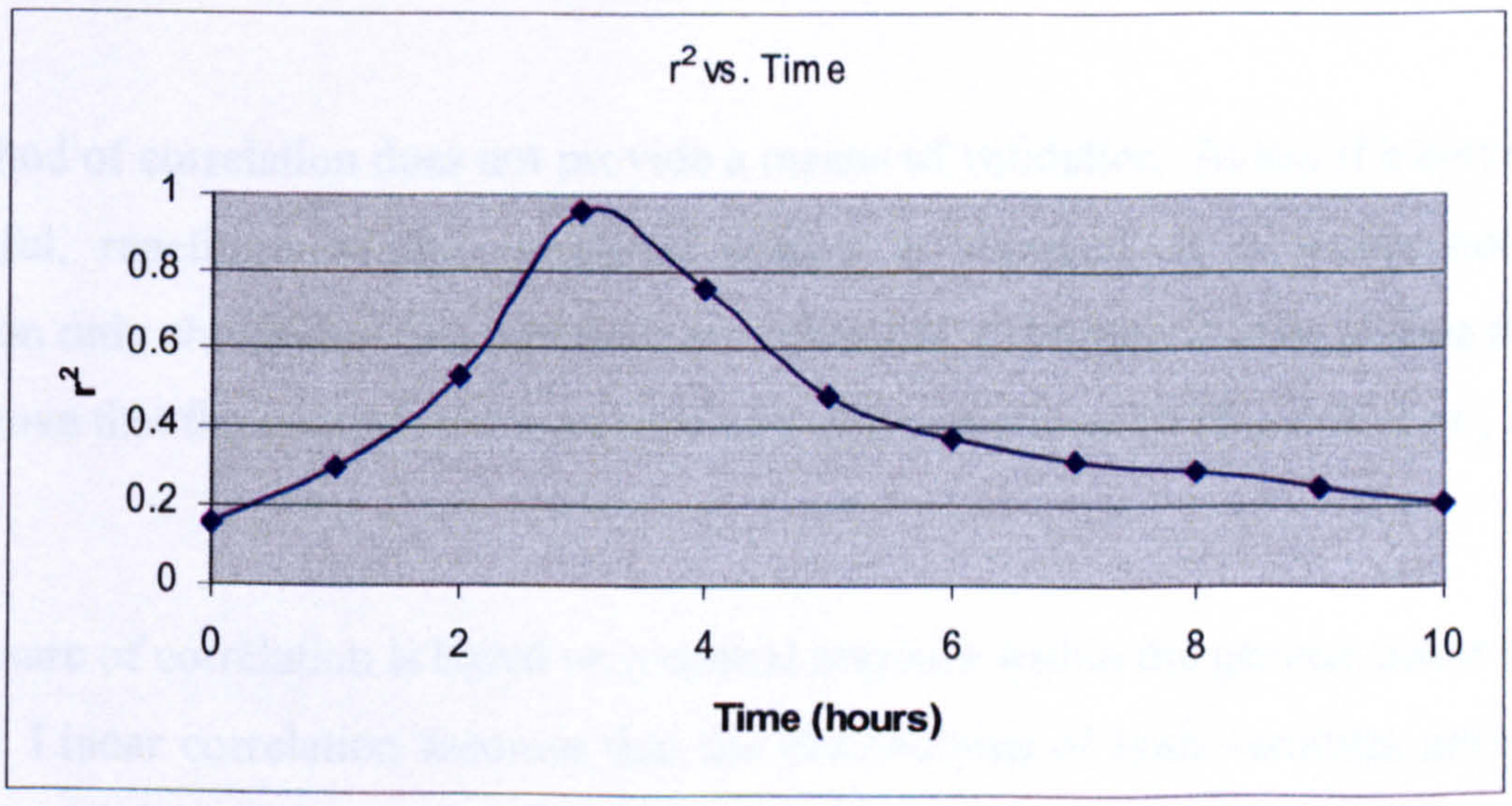


Figure 4 Graph showing an example of how the correlation coefficient can be plotted against lag time (lag between datasets)

In this particular example the graph indicates that there is a highly correlated time series section at a time lag of 3 hours. A number of issues were found implementing this technique, these included:

- **Raw data sampling is required**; correlation requires data at identical intervals between data sets. Interpolation methods can be used for this if data is not available in a synchronised form.
- **Length of the time series to correlate is unknown**; the period of both data sets must be identical, but is not fixed. The optimum period is unknown before analysis.
- **Multiple comparisons are required**; the process can be slow owing to the need to check correlation values between the two data parameters over a regularly increasing lag period (and varying period length). The density of lag step (the lag step size that each consecutive comparison takes during sampling) influences the speed and accuracy of the comparisons.
- **Multiple parameters can be involved**; correlation is traditionally undertaken between two data sets, analysis between multiple parameters would create many possible combinations, leading to a large increase in computation time.

In searching for periods of correlation, a more meaningful result is found if the correlation can be achieved over a longer period of time. In practice it is comparatively simple to find small periods over which data is highly correlated. However, the longer the time period of correlation, the more significant the overall result will be. *Appendix B* shows a correlation example using data from the Great London Smog (1952) to demonstrate this.

The method of correlation does not provide a means of validation. To test if a correlation is meaningful, repetition of the correlated pattern is required. It is worth noting that correlation only shows that two variables are related in a systematic way, it does not prove nor disprove that the relationship is a cause-and-effect relationship (Bewick *et al.*, 2003).

The measure of correlation is based on a central measure within the general linear model of statistics. Linear correlation assumes that the distributions of both variables are normally distributed, that the relationship is linear, and that the relationship is consistent throughout the sample. When these assumptions are violated, correlation becomes inadequate as an analytical technique.

Within the context of this thesis, the technique of correlation does not provide enough

information to form an alert of an impending decline in lung function. To provide a warning, an alert has to be generated before the occurrence of the decline in lung function. A new technique is required that provides analysis so an alert can be generated. In the development of a new technique it is useful to observe, that the only commonality between each parameter used in correlation analysis (in this instance) is the time component (or x-axis of the data). Characteristics of the environmental and medical data sets are none linear, for example, the analysis of a gradual build up of CO_2 in the atmosphere could be attributable to a sudden decline in respiratory function for some patients, but as the time frames of analysis are different between the two data sets (gradual vs. sudden), correlation would not identify this trait. The *correlation* between the two data sets may also posses more complex dependencies. It should be considered that perhaps detection of a gradual build up of atmospheric pollution is required, in addition to a threshold being exceeded.

1.1.3 Health Informatics

The phrase *medical informatics* is used to describe the application of information technology within healthcare. It was first coined in 1973 (Protti D J, 1995). Since then it has grown considerably as a medical discipline encompassing fields as varied as medical imaging, clinical decision support, health records, hospital information and financial systems. The use of *medical informatics* and *health informatics* as phrases are used interchangeably.

The use of informatics by the healthcare community over the last couple of decades has shown its ability to offer additional services in line with current treatment patterns in areas such as anaesthetics or remote surgery (Janetschek, 1998). In some situations the use of informatics has led to new health management techniques (Pande *et al.*, 2003). The variety of uses and benefits of informatics within the medical discipline has led to its widespread adoption by the healthcare community (Beuchat *et al.*, 2005).

Systems have gradually become more complex in design, driven forward by the needs of clinicians or by technological advances that enable old problems to be addressed in new ways. Early systems were essentially about providing communication links between medical experts and remote locations, and largely centred around the organisation of hospital schedules or supporting clinicians through protocol-based medicine (the treatment

of a patient to a medically predefined pattern).

Chapter 2 covers areas of informatics, and their benefits to patient care in more detail. This thesis uses knowledge from the area of informatics to analyse, and identify verifiable patterns between respiratory and environmental data sets.

1.2. Scope of Thesis

This thesis aims to identify a process architecture capable of identifying environmental predictors of patient-specific asthma exacerbation, and provide an application architecture, required for the automation of these processes.

By fulfilling this research aim, the application architecture can be used to gain a greater depth of knowledge regarding the relationships between a patient's environment and their health. Patients will always experience different environmental conditions (due to people's individuality) and react differently to allergens, particles, gases and other environmental parameters. The research questions are: *what is an individual patient susceptible to, and when will their susceptibility lead to a decline in health?* This thesis attempts to devise an architecture for a software application that will assist in providing an answer to these questions.

Clinical awareness of a patient's environment is vital for the successful management of respiratory health. Environmental factors related to the health of a patient must be correctly interpreted in order to react quickly and effectively to ensure the patient's wellbeing. Reacting to unforeseen changes in patient or environmental conditions requires a knowledge of both medical and environmental data sets. The analysis also requires that both environmental and medical data sets are available, so that possible relationships between asthma episodes and environmental triggers can be identified.

The ability to detect potential triggers or situations leading to an asthma attack is an assistive step in the management and treatment of asthma. Information indicating patient-specific issues or trends, that arrive in a timely manner can be used to improve the quality of care given to a patient.

1.2.1 Thesis Objectives

To meet the aim of detecting environmental predictors that are relevant to a decline in patient health, this thesis has several objectives:

1. Research into an appropriate analytical technique capable of attributing the decline in patient lung function to environmental *predictors*.
2. Identification of the necessary process architecture that enables the triggering of alerts when environmental predictors are encountered.
3. Design of an application architecture that enables the analysis of large scale datasets.
4. A proof of concept consisting of a set of prototyped software components demonstrating the derived process architecture.

The research does not attempt to develop an application suitable for use in clinical diagnostics, or develop issues relating to the design of the user interface and the presentation of information and results to the end user in a graphical form, where the requirements for such features are beyond the scope of this thesis. However, in achieving the objectives listed above, this research provides the basis for a tool useful for researchers to identify predictors of change events. A number of software prototypes are used to test the architecture developed by this work.

The first objective of identifying an appropriate data analysis technique is given focus by the aims of this thesis; where it is proposed that a decline in patient lung function can be predicted from a change in the environment. More specifically, this thesis suggests that the detection of environmental events, leading to a reduced period of lung function can be identified with three steps:

1. **Identification of the asthma episode:** Achieved through the observation of a worsening trend, a sudden drop in lung function value, or a re-occurring pattern.

2. **Identification of the environmental predictor:** For the purpose of this thesis it is hypothesised that factors in the environment must be having an effect on the patient's health prior to the time that the patient's lung function trend first begins to decline.

3. **Validation of the environmental predictor:** Once an asthmatic episode has been identified it increases the probability that a predictor of that episode can be found. The time period preceding the asthma episode from which to extract possible predictors is arbitrary at this stage in the research, but has been taken to be anything between a matter of hours to a few days and is supported by general research in the area (Lebowitz, 1996).

These three steps are used in defining the application architecture, and are shown by this thesis to corroborate current research, at a patient-specific level. The steps are aimed at creating an alert from the identification of environmental predictors that are validated as contributing to a period of lung function decline. It is necessary to define a method for detecting adverse environmental conditions that are attributable to the decline in a patient's respiratory health, if an automated pattern identification application is to be created.

The second objective, of identifying a process architecture that would enable the triggering of patient alerts by implementing applications, is only made possible once a process of validating identified environmental predictors is found. The ability of the monitoring application to respond to environmental changes can be simplified through the choice of an appropriate analytical technique. The objective of alerting patients as the end point of the architecture has a defining role.

To support the automatic discovery of patterns an appropriate application architecture is required. A high level analysis is undertaken so that an architectural description of the software system, appropriate to meeting the research objectives can be developed. This includes appropriate pattern recognition components that remain flexible to adapt to future developments, and changes in use.

Typically it is *normal* to analyse data from a sample drawn from the population, rather than to take the approach of identifying factors related to an individual patient. Identification at a patient-specific level, combined with the task of identifying relatively rare environmental events, means that a greater quantity of data is required than finding correlations within population data. To identify patient-specific susceptibility to environmental factors, automatic discovery of patterns becomes necessary to link *factors* to the individual patient for several reasons:

- **Validation** of rare events is required.
- Collection of raw data is **continuous.**
- Environmental data requires **relating to the patient.**
- **Particular environmental factors** that are having an effect on the patient are unknown.
- Patients react in different ways to their **personal environments.**

Research into appropriate techniques for inclusion within this thesis means that software prototypes developed during this research have several purposes: to develop the application architecture using an iterative design process; to test the derived set of architectural processes; and to offer further detail as to how the application architecture could be applied in practice. The prototypes demonstrate positive results, in line with current respiratory research, and serve to provide supporting evidence that the hypotheses of the thesis can be proven.

1.2.2 Contribution to Knowledge

The research undertaken by this thesis has drawn out several key aspects that extend the work of others. The concepts specifically developed during this thesis as contributions to knowledge are:

1. A set of **processes** that focus on the delivery of **patient-specific analysis**, providing the ability to relate environmental data to ambulatory patients.
2. A method to recognise significant **changes in data trend**; this technique is applied to patient lung function and air quality data.
3. The *delay characteristic*, which defines the concept and processes involved in identifying patterns from the **time between environmental predictor and decline** in lung function.
4. A neural model that recognises **significant and repeatable** events, and
5. A clustering technique, to aid the performance of the neural model.

These contributions are discussed during Chapters 3, 4 and 5.

1.3 Research Methodology

The research and development of the system presented by this thesis, is divided into two deliverables:

1. Research into the concepts listed above, with the derivation of a set of processes that outline the requirements of the system.
2. Development of a set of *prototypes* known as the Environmental Monitoring System (EMS) in order to provide a proof of concept for the identification of predictive patterns (found between patient and air quality data sets), and the problem's automation.

The applied research methodology begins with a review of a broad number of subject areas including: the use of lung function measurements, environmental factors that influence the condition of asthma, system architecture, the role of health informatics, and analytical techniques. Correlation techniques that extend the methods used by the Medicate (2000) project are explored using a software prototype. Prototyping is then used in an iterative process to define a system architecture capable of recognising possible environmental predictors of patient lung function decline, through experimentation.

This thesis draws on research data obtained during the Medicate project (Crabbe *et al.*, 2004) described in *Appendix A*, additional hospital admissions data obtained from *The Information Centre for health and social care*, and a six month patient-specific data sample of lung function and air quality, collected during the course of this research. The prototypes that provide a proof of concept use these data sets, and additional information regarding typical characteristics of the data types to validate the process architecture.

The prototypes have three purposes: to aid the investigation of an appropriate analytical process, to help prove the thesis methodology, and provide a useful guide for an architectural implementation of the system. These three uses evolve chronologically during the research, and can be seen in Chapters 3 through to 5.

The objective, to warn patients of an impending asthma exacerbation guides this investigative methodology. Changes in environmental conditions are validated to ascertain if they are significant predictors for the onset of a patient's decline in respiratory health, and whether the environmental change provides an indication of how long the patient has before their health begins to deteriorate.

1.4 Summary

The objectives of this research address the creation of a new system architecture able to **identify interrelationships** between different large scale data sets, and capable of alerting both clinical staff and patients to factors in the environment that will adversely affect patient health. A research prototype is presented as a proof of concept. The prototype identifies patterns and relationships between patient lung function, and environmental factors identified as possible predictors to a decline in health. It is not the purpose of this thesis to provide a system that will identify the cause of asthma, but rather to provide a system architecture that will aid the **identification of when an attack** is likely to occur, by identifying **environmental factors** which can be used to predict the onset of asthma exacerbations.

Both environmental and respiratory readings have a location and time element to the reading (whether this is recorded or not). Environmental data tends to be collected on a regular basis with varying degrees of granularity from half an hour, to an interval of days. Lung function readings are taken wherever the patient happens to be at the time of the reading, and environmental data wherever an environmental monitor (or monitoring station) exists. The treatment of a patient's condition can be enhanced by understanding the relationship between environmental and medically related data. Treatment can be tailored on a patient-by-patient basis. This thesis seeks an approach to specifically identify air quality characteristics that act as predictors of patient-specific asthma episodes.

The next chapter provides an overview of the general research areas including the role of lung function in respiratory healthcare, the effect of environmental pollutants on asthma, health informatics, and system architecture.

Chapter 2

Overview: Respiratory Healthcare, the Environment, and Aspects of Informatics

This chapter begins by reviewing the areas of respiratory healthcare where the management of asthma is described. Then reviews how asthmatics are affected by their environment, before explaining the role of system architecture and areas of informatics (including pattern recognition) that are relevant to the work.

2.1 Lung Function

Clinicians use a regular cycle of monitoring and treatment to measure the *progress* of asthmatic patients. Set patterns of treatment follow guidelines that outline standard practices used in the medical profession, using what is known as a *predicted best value* as a reference point. Predicted best values are based on standard data tables (Nunn & Gregg, 1989) and depend on age, height, weight and the sex of a person. The predicted best value for a patient represents the maximum value of expiratory air flow that a patient is likely to achieve. The values are not always the best to use as an individual patient's benchmark as they are standard, and differences may occur on an individual bases. Therefore asthma patients can be given an individual best value by their doctor which is based around the predicted value but allows for a patient's individual characteristics.

Assessment of lung function is performed over one of two time scales. One is relatively long, involving discrete observations, usually in the form of *pulmonary function tests* (PFT) where a patient's parameters are compared to the set standard, at intervals in the order of days to years. The second time scale over which lung function is assessed is very short, observations are made either continuously or at intervals in the order of minutes to hours. This activity comes under the heading of patient monitoring (Webster, 1998).

The approach that clinicians take in the treatment of asthma is defined by the British Thoracic Society and published in the "British Guideline on the Management of Asthma" (BTS, 2004) and "The British Guidelines on Asthma Management" (BTS, 1995) in which

a 5-step scale (also referred to as the *stepwise* approach) is used as the basis for the management of patient care (definitions of *reliever* and *preventer* are given on the next page):

- Step 1 – *Mild Intermittent Asthma*. Recommended to use a *reliever* as and when required.
- Step 2 – *Regular Preventer Therapy*. Recommended to start using a steroid based *preventer* on a daily basis.
- Step 3 – *Add-on Therapy*. Addition of a long-acting $ß_2$ agonist to the therapy.
- Step 4 – *Persistent Poor Control*. An increased steroid dose (as prescribed by a doctor).
- Step 5 – *Continuous or Frequent Use of Oral Steroids*. The additional use of a daily steroid tablet, and consideration of alternative treatments (as prescribed by a doctor).

The US National Heart, Lung and Blood Institute (NHLBI, 1998) defines a number of methods that clinicians use to achieve control of asthma:

- Select appropriate medications.
- Manage asthma long term.
- Treat asthma attacks.
- Identify and avoid triggers that make asthma worse.
- Educate patients to manage their condition.
- Monitor and modify asthma care for effective long-term control.

A tool used for the management and diagnosis of asthma is the *spirometer*, a device which measures the volume of air leaving a patient's lung. Using this device, diseases of airflow obstruction and lung stiffening can be detected (NHLBI, 1997) and patient progress monitored. When the results are plotted on a regular basis they show a trend that is useful in determining whether or not a patient is responding to treatment. The graphs which are plotted also help to show any difficulties a patient may be experiencing, but may not necessarily be aware of. Velsor-Friedrich *et al.* (2005) study the effect that an intervention program has on a group of asthmatic students. The students were given appropriate knowledge of, and the ability to self-monitor their condition, which led to a general improvement of their respiratory condition.

The primary method of treatment involves the use of two different types of medicine called

relievers and preventers. Relievers are designed to relieve breathing difficulties as they happen by quickly relaxing the muscles which exist around the airways. This allows the airways to open wider making it easier to breathe, although airway swelling is not reduced. Preventers reduce the chance of asthma symptoms by protecting the lining of the airways, and have the added effect of reducing the inflammation of airways thus reducing their responsiveness to asthma triggers.

2.1.1 Peak Expiratory Flow (PEF)

There are a number of useful measurements that can be recorded to determine the severity of asthma and to monitor changes in condition. The most common is *peak flow*, which is a measure of how fast air can be exhaled from the lungs. Primarily this monitors how well asthma is being controlled and is a good indication of how well medication is working. *Peak flow* is dependant on how wide the airways within the lungs are, and is not a measure of fitness or how strong chest muscles are. *Peak flow* is also called *peak expiratory flow* or PEF.

Peak expiratory flow readings depend on the age of the patient, their height, and sex. A predicted PEF value can be obtained from charts using these three parameters. The standard method used in the United Kingdom was developed by Nunn & Gregg (1989).According to the UK standard, a male aged 35 with a height of 174cm should have a PEF value of about 637 litres per minute (L/m). For a male, any reading equal or up to 100 L/m lower than the predicted PEF value would be considered normal. In a female with similar characteristics, the PEF value would be expected to be in the region of 497 L/m with a range of up to 85 L/m lower than the PEF value being considered normal.

The most reliable measurement that shows an assessment of an asthmatic's condition is a PEF reading taken in the morning. Morning measurements are usually the lowest (with the afternoon being the highest) due to night-time occurrence of asthmatic symptoms. Fluctuation between the minimum and maximum values is known as the *variability* and is presented as a percentage.

PEF Variability, is used to determine how well a patient's asthma is being managed. A low variability means that a patient is controlling their asthma well, and medication is working satisfactorily. The formula is given below;

$$PEF\ Variability(\%) = \frac{(Max.PEF\ value\ per\ day - Min.PEF\ value\ per\ day)}{(Max.PEF\ value\ per\ day)} \times 100 \qquad Eq.\ 2.1$$

The graph shown in Figure 5 (showing a patient's data during the Medicate trial described later) can be used to plot Peak Expiratory Flow (PEF) readings for a patient, the graph can be used to monitor a patient's progress. The graph is banded into three sections: green, yellow and red to denote the 'seriousness' of reading level. A red reading would mean that the patient needed urgent treatment.

Figure 5 Peak flow graph showing PEF readings of a hospitalised patient (Jaeger, 1998).

In this example (Figure 5) of a patient recovering in hospital after a severe asthma attack the three colour coded bands can be seen. Any reading above 480 L/m is a good reading, between 300 and 480 L/m an indication that the patient is having difficulty breathing, and below 300 L/m the patient requires immediate hospitalisation or treatment.

The threshold bands are derived from the predicted peak expiratory flow rate a patient is

expected to obtain. This is usually defined after an examination by a doctor, although tables exist with suggested values. The general regions for the bands are shown in the table below, obtained from "Guidelines for the Diagnosis and Management of Asthma" (NHLBI , 1997a).

Table 1 The peak flow zone system

Zone	***% Predicted Peak Expiratory Flow Rate***
Green	80 – 100
Yellow	50 – 80
Red	< 50

The peak expiratory flow rate is generally defined using the tables (Nunn & Gregg, 1989) reading off the values using the patient's age and height.

2.1.2 Forced Vital Capacity (FVC)

Forced Vital Capacity is the maximum forced volume of air that can be expired from the lungs. FVC is measured using spirometry which is the name given to the measurement of change in lung volume for the testing of pulmonary function (regardless of the technique used).

A small number of parameters are used to describe the forced expiratory record:

- Forced Vital Capacity (FVC)
- Forced Expiratory Volume in one second (FEV_1)
- Mean Forced Expiratory Flow (FEF) during the middle half of FVC ($FEF_{25\text{-}75\%}$). This can also be explained as the average flow of air leaving the lung during the middle portion of the expiration (measured by volume).

FVC is recorded during a spirometry test starting with a patient's Total Lung Capacity (TLC) and ending with their Residual Volume (RV). There are two commonly used methods for displaying flow limitation during a FVC reading (Webster, 1998). Figure 6 shows graphs of the two methods, ***a)*** showing volume of air flow against change in

volume, and ***b)*** showing the change of air remaining in the lung over time.

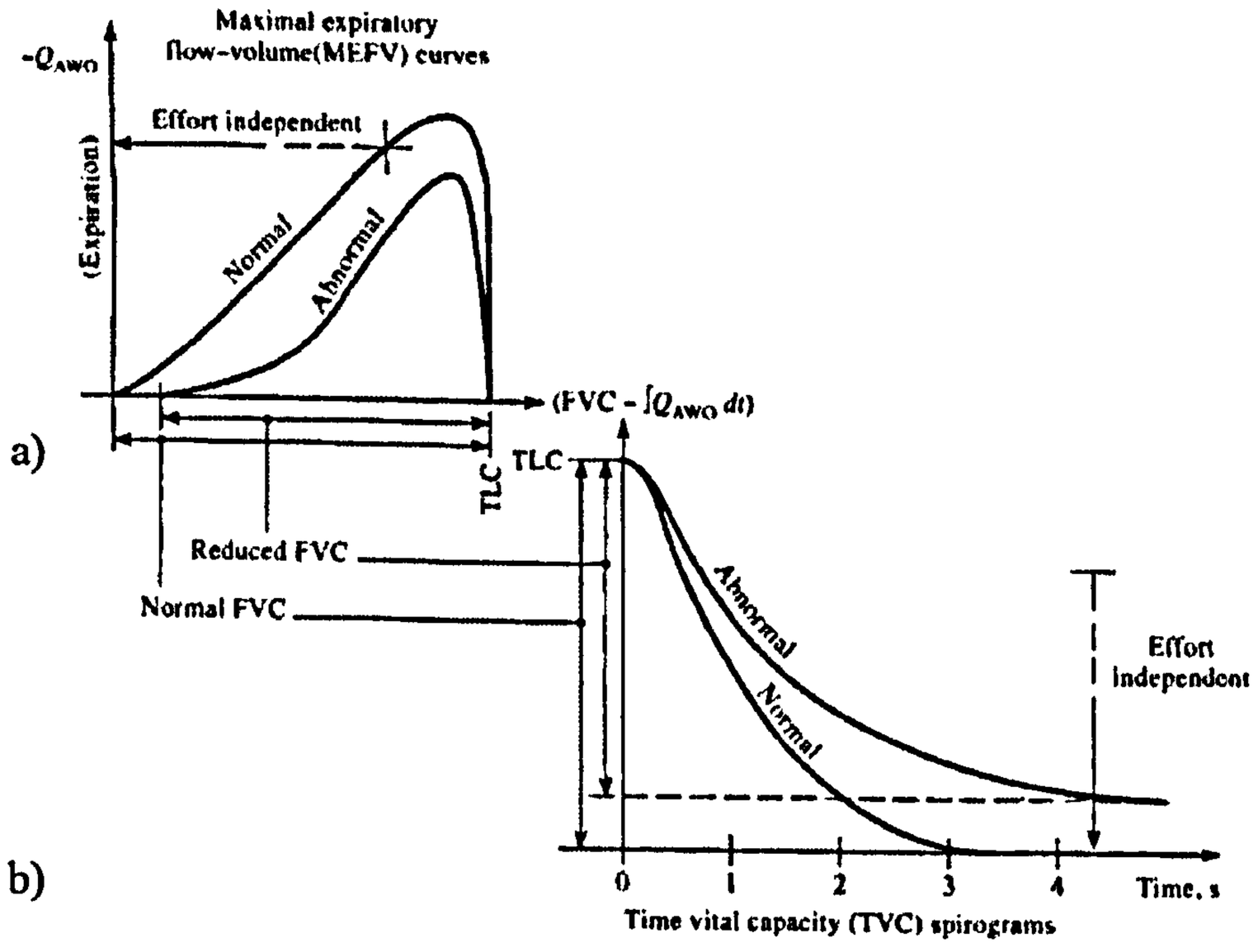


Figure 6 MEFV Curves and Timed Vital Capacity Spirograms (Webster, 1998)

Method 1

Plotting the volume flow of gas at the airway opening (Q_{AWO}) against volume change (or the integral of Q_{AWO}) subtracted from FVC against time.

This method produces a *maximal expiratory flow volume* (MEFV) curve shown in Figure 6a. The MEFV curve represents the relationship between variable $FVC - \int Q_{AWO}\,dt$ and $-Q_{AWO}$ its derivative. When this relationship is a straight line through the origin, it represents a homogeneous, linear, first-order differential equation (Webster, 1998).

$$Q_{AWO} = -K(FVC - \int Q_{AWO}\,dt) \qquad \textit{Eq. 2.2}$$

Method 2

Plotting against time (Figure 6b), gives a graph (or spirogram) of timed vital capacity (TVC). TVC is the difference between the total lung capacity (TLC) and the residual volume (RV), the smallest volume to which the lungs can be deflated due to a slow expiration. The reading or graph is shown over time and is the equivalent value to the FVC

value.

The two curves in Figure 6 provide information about the condition of the airways of the patient. During a FVC manoeuvre (or test) the respiratory system can be described as functioning in two ways: independently from the effort exerted by the patient, and effort-dependent. The dependent part provides information about the larger, upper airways and extrapulmonary parts of the respiratory system and is shown on the graphs as the plot between TLC and approximately 25% of the FVC below TLC. The independent section reflects the condition of the smaller airways of the lungs. Using the MEFV curves and TVC spirograms, the condition of the respiratory system can be explored.

2.1.3 Forced Expiratory Volume in one second (FEV_1)

Primiano (1998) suggests that FEV_1 is not a good index for small airways disease due to its partial dependency on effort-related data. Lane (1996) agrees but states that it is still useful if it is quoted as a percentage of the forced vital capacity (FVC) as the two variables are linked. Guidance from the US National Heart, Lung, and Blood Institute (NHLBI, 1997) goes further and says that the ratio of FEV_1 to FVC is often used to assess patients for airflow obstruction. A reduced ratio of less than 65 percent (FEV_1 / FVC) indicates obstruction to the flow of air from the lungs.

FEV_1 and FVC readings given by patients' both differ according to height, sex, age and race, and appropriate reference values should be used for their interpretation. An increase of 15% in FEV_1 is significant. Values should be expressed as absolute figures and also as percentage predicted, based on patients' age, height and sex.

2.1.4 Mean Forced Expiratory Flow ($FEF_{25-75\%}$)

$FEF_{25-75\%}$ represents an average flow produced during the middle half of the forced vital capacity expiration. So that comparisons of expiratory flow can be made between individuals of different sizes the parameter can be normalised. The technique used to compensate for differences in individual size (normalise) is to divide $FEF_{25-75\%}$ by FVC.

$$FEF_{25\%-75\%}=\frac{0.5\text{FVC}}{(t_{25\%FVC}-t_{75\%FVC})}$$

Eq. 2.3

2.2 Environmental Factors

The wide range of variables makes asthma a difficult condition to predict and monitor. Exacerbations of asthma are characterised by periods of increased symptoms and reduced lung function, which may result in diminished ability to perform usual activities. Exacerbations may be brought on by exposures to irritants or sensitisers in the home, work, or general environment. Education is needed to help the patient recognise both immediate and delayed reactions to these environmental exacerbations. Exposure of asthma patients to irritants or allergens to which they are sensitive has been shown to increase asthma symptoms and precipitate asthma exacerbations (NHLBI, 1997a). Asthma patients are known to react to a variety of triggers, such as allergens or a combination of allergens, exercise, pollutants or emotional stress. This complex nature can lead to a time consuming task of detailing and then analysing a patient's history. The complex nature of asthma is a major healthcare challenge. The management of the disease in asthmatics is particularly difficult without sufficient insight into the patient's environment.

2.2.1 Air Quality

An asthmatic's condition is affected by the specific environmental conditions that surround the individual. Peled *et al.* (2005), Chauhan *et al.* (2003), Moshammer & Neuberger (2003), Kehrl *et al.* (1999) and Lane (1996) consider the following pollutants to be of concern to asthmatics:

Table 2 Pollutants of concern to asthmatics

Pollutant	Description / Source
Particulate Matter (PM)	Particles which have an aerodynamic diameter of 10 microns or less are classified as PM_{10}. Smoke and particles from diesel engines would both fall into this category. Other common sub-categories include particles which have an aerodynamic diameter of 2.5 microns or less ($PM_{2.5}$) and $PM_{0.1}$ which has an aerodynamic diameter of 0.1 microns or less.
Sulphur Dioxide (SO_2)	The most important natural sources of sulphur dioxide, (and of other sulphur compounds) are volcanoes, during both active and dormant periods. Globally, these contribute perhaps 20% of the world's total sulphur emissions. However, in both developed and less-developed countries, particularly in urban areas, emissions that arise from the combustion of solid fossil fuels are of prime concern. Coal and oil both contain sulphur in varying amounts, and both therefore produce sulphur dioxide when burnt.
Nitrogen Dioxide (NO_2)	Vehicles produce nitric oxide, this changes to nitrogen dioxide in the atmosphere. Recent research has highlighted the increasing importance of directly emitted NO_2 primary emissions. There is evidence for significant amounts of NO_2 emitted directly from the tailpipes of diesel vehicles, especially when slow moving, with levels possibly as high as 25% of total NOx emissions in mass terms. These primary emissions have a significant impact on roadside NO_2 concentrations in areas where there is a large proportion of diesel vehicles (DEFRA, 2004).
Ozone (O_3)	Ground level ozone (O_3) is a secondary pollutant; it is generated from the reactions of primary pollutants in the atmosphere. It is important to note that ozone at ground level is classified as a pollutant, in stark difference to stratospheric ozone and the ozone layer occurring at around 11 to 15 km above ground level. Stratospheric ozone is naturally occurring and critically beneficial, as it protects us from harmful levels of UV radiation from the sun. Tropospheric or ground level ozone is formed mainly from the breakdown of NOx into NO_2 in the presence of UV (sunlight).

Air quality and other environmental influences affect people's respiratory system (NHLBI, 1997). Particles of diameter 10μm or less (also know as PM_{10}) and diameter of 2.5μm or

less (known as $PM_{2.5}$), along with irritant gases, are all noted to have an effect on health (COMEAP, 2006). According to Stedman (2001), irritant gases are those such as sulphur dioxide (SO_2) and nitrogen dioxide (NO_2) which enhance response to allergens and may increase the prevalence of respiratory infections in children, and other gases such as ozone (O_3) and carbon monoxide (CO) (Lee *et al.*, 2002). The health effects of both long- and short-term exposures to O_3 are shown by pulmonary function decrements (AQMD, 2003). nitrogen dioxide (NO_2) also has the potential to aggravate chronic respiratory disease and respiratory symptoms in sensitive groups (AQMD, 2003).

Sulphur dioxide arises from the sulphur present in most fuels, but its presence can also be attributed to volcanic activity where about $5x10^6$ tonne sulphur yr^{-1} (Halmer *et al.*, 2002), and even larger amounts are emitted by sea spray (as sulphate), which contribute in the region of $44x10^6$ tonne yr^{-1} (Hester, 1986), (NEGTAP, 2001) and is known to aggravate asthmatic symptoms. Exposure to sulphur dioxide can lead to bronchoconstriction for persons with asthma (AQMD, 2003).

It is extremely difficult to monitor the environmental conditions experienced by an individual as they move from location to location. Most air quality data in the United Kingdom originates from fixed location monitoring stations situated around the country, each of which monitors a set of pollutants (not always the same) including; nitrogen dioxide, ozone, carbon monoxide, sulphur dioxide and particulate matter $PM_{2.5}$ and PM_{10} which represent fine airborne particles of 2.5 and 10 microns diameter or less respectively. This provides a fairly detailed, although not ideal, data source for environmental data. The data is usually collected on an hourly basis but this can vary depending on the location and facilities at the monitoring station, and is not always reliable; missing or abnormal values pay a large part in the data sets. The data is also not ideal as it may not portray the actual levels of pollutant that the patient experiences due to the static location of the sites.

In the United Kingdom the monitoring stations are run by AEA Technologies, National Environment Technology Centre (NETCEN) monitoring network. The network is run on behalf of the UK Department for Environment, Food & Rural Affairs (DEFRA) and the Devolved Administrations. Details of the monitoring network and results can be found in an annual report from AEA Technology (AEAT, 2006). The network provides a reliable and consistent source of data which is available via the Internet (TAQA, 2008).

Data obtained from the monitoring stations can be used in validated *dispersion models*, which have the capability to estimate levels of pollution between known point sources. Although this research is not focused on the use of dispersion modelling techniques, a utility for the inclusion of such models would be a useful addition to the extensibility of the system.

In the United Kingdom air pollutant levels are banded into 4 categories: low, moderate, high, and very high. Each category describes a warning level according to health impact. The bands relating to each category and pollutant are shown by Table 3 below. The abbreviations *ppb* (*parts per billion*) and *ppm* (*parts per million*) are used.

Table 3 Air quality bandings (COMEAP, 1998)

Pollutant	**Pollutant Band**			
	LOW	**MODERATE**	**HIGH**	**VERY HIGH**
Sulphur Dioxide	Below 100 ppb (266μg/m^3), averaged over 15 minutes	100 ppb (266μg/m^3), averaged over 15 minutes	200 ppb (532μg/m^3), averaged over 15 minutes	400 ppb (1064μg/m^3), averaged over 15 minutes
Ozone	Below 50 ppb (100μg/m^3), as an 8 hour running average	50 ppb (100μg/m^3), as an 8 hour running average or averaged over one hour	90 ppb (180μg/m^3) averaged over one hour	180 ppb (360μg/m^3), averaged over on hour
Carbon Monoxide	Below 10 ppm (11.6mg/m^3), as an 8 hour running average	10 ppm (11.6mg/m^3), as an 8 hour running average	15 ppm (17.4mg/m^3), as an 8 hour running average	20 ppm (23.2mg/m^3), as an 8 hour running average
Nitrogen Dioxide	Below 150 ppb (287μg/m^3) averaged over one hour	150 ppb (287μg/m^3), averaged over one hour	300 ppb (573μg/m^3), averaged over one hour	400 ppb (764μg/m^3), averaged over one hour
Fine Particles	Below 50 μg/m^3, as a 24 hour running average	50 μg/m^3, as a 24 hour running average	75 μg/m^3, as a 24 hour running average	100 μg/m^3, as a 24 hour running average

Low indicates a pollutant level below the threshold of the National Air Quality Standard. *Moderate* indicates a level at the threshold. *High* represents a threshold set using advice from the Committee on Medical Effects of Air Pollution (COMEAP, 2002) that indicate that some discomfort may be experienced by those susceptible to air pollutants as a result of the levels, and *Very High*, also a threshold set by COMEAP, goes one step further and represents levels where considerable discomfort may be experienced.

The US National Heart, Lung, and Blood Institutes (NHLBI, 1997a) recommend that patients with any level of severity should avoid, exposure to allergens to which they are sensitive, exposure to environmental tobacco smoke (which counts as a type of particulate matter), and exertion when levels of air pollution are *high*. For successful long term asthma management, it is essential to identify and reduce exposures to relevant allergens and irritants and to control other factors that have been shown to increase asthma symptoms and/or precipitate asthma exacerbations. For example, Dominici *et al.* (2003) found that daily variations of PM_{10} are positively associated with daily variations of mortality, using a 24 hour PM_{10} average. However, the results also suggested that different regions within the geographical study all experienced differing pollutant levels, and that further detailed investigation was necessary to identify the effect of particular particle types.

2.2.2 Pollen

Pollen is a fine powder produced by the anthers of seed-bearing plants. It is particularly troublesome to asthmatics as it is carried by the atmosphere. Particular focus can be given to pollen when the weather is dry, sunny and moderately windy. Jacobson *et al.* (2007) found that there was a strong correlation between six separate pollen spikes and six peaks in emergency department and urgent care admissions during the summer months of May and June 2006. Grasses tend to shed their pollens during the morning, but when the weather is hot the process can extend into late afternoon. Symptoms due to pollen are considerably reduced when the weather is damp as dampness reduces the spread of pollens.

2.2.3 Health-Related Quality of Life

The World Health Organisation defines *health-related quality of life* as, "the individuals' perceptions of their positions of life in the context of the culture and value systems in which they live and in relation to their goals, expectations, and concerns" (Mohangoo *et al.*, 2007). While the American Thoracic Society Quality of Life Resource (ATS, 1997) states that "*health-related quality of life* is an individual's satisfaction or happiness with areas of life that are directly affected by *health*". It is generally recognised that the burden of illness is far reaching and can be measured in terms of the financial burden, discomfort, restricting ability and apprehension caused by the illness. Measurements determining the quality of life in studies of clinical evaluation are commonplace and are sometimes used to

offset higher costs of treatment. Chen *et al.* (2007) find that there is an inverse relationship between the number of asthma control problems and quality of life, while Mohangoo *et al.* (2007) find that the presence of at least four wheezing attacks during a year reduced perceived quality of life.

Over the last decade quantitative relationships between air pollution and adverse health effects have been studied (Taggart *et al.*, 1996; Howel *et al.*, 2001; Brunekreef and Holgate, 2002; Ho *et al.*, 2007). The effect of particulates, particularly PM_{10}, were studied by all four of the studies. Ozone effects were studied by Taggart *et al.*, Ho *et al.*, and discussed by Brunekreef and Holgate. Brunekreef and Holgate also discussed the affect of nitric oxides, which were additionally studied by Ho *et al.* Health effects of sulphur and nitrogen dioxides were studied by Taggart *et al.* (1996). The parameters and related studies are summarised by the table below.

Table 4 Study vs parameter summary

Parameter	***Study***
Particulate matter	Taggart *et al.* (1996); Howel *et al.* (2001); Brunekreef and Holgate, (2002); Ho *et al.*, (2007).
Ozone	Taggart *et al.* (1996), Ho *et al.* (2007) and discussed by Brunekreef and Holgate.
Nitric oxides	Ho *et al.* (2007), and discussed by Brunekreef and Holgate (2002).
Sulphur and nitrogen dioxide	Taggart *et al.* (1996).

Brunekreef and Holgate (2002) highlight the work of several others (including Samet *et al.*, 1998 and Pope and Kalkstein, 1996) for considering cofounders, that included weather variables. Ho *et al.* (2007) factored the effect of age, rhinitis, eczema, urban birth location, parental education level, exercise, cigarette smoking, environmental tobacco smoking, alcohol beverage consumption, and weather factors into their study. They analysed questionnaires from a screened sample of 64,660 students who displayed signs of having asthma. A repeat measurement regression model was used to examine the relationship between monthly asthma attack rate among asthma patients, and air pollution (particulates, nitric oxides and ozone). The model used a stratified random sample of students, and demonstrated that air pollution is related to asthma attack rate. Howel *et al.* (2001) investigated the association between the acute respiratory health of children, and

particulate levels over a six week period. Diaries of respiratory events were collected for 1405 children, along with concurrent monitoring of particulate levels over a six week period. It was found that frequently small and positive associations existed between PM_{10}, and respiratory symptoms, which were varied between communities.

Taggart *et al.* (1996) studied the relationship between asthmatic bronchial hyper-responsiveness and pulmonary function (represented by daily lung function tests of FVC and FEV_1) to ambient levels of summertime air pollution during 1993. The study of 38 subjects suggested that changes in the concentrations of traffic-related air pollution depicted by O_3, SO_2, NO_2 and smoke, were capable of potentiating airway inflammation. The study identified a log linear relationship between all lung function measures and pollutant levels, grass pollen concentrations and temperature. Correlation between levels of 24 hour mean SO_2, NO_2 and smoke, 48 hour mean NO_2 and smoke, and 24h lag NO_2, and bronchial hyper-responsiveness were also found. After reviewing relevant literature Brunekreef and Holgate (2002) concluded that the evidence of adverse health effects from air pollution have been estimated to be higher than effects from a long list of other environmental factors.

2.2.4 Meteorological

Meteorological effects such as humidity, temperature, wind direction, and weather condition have an effect on asthma sufferers. Research has linked the increase in asthma exacerbation to thunder storms (Anderson *et al.*, 2001). Damp and cold are also known to be particularly prominent in triggering asthma attacks, although some asthmatics would say that for them, hot and humid weather is more troublesome. Records show (Lane, 1996) that there are more emergency cases of asthma when there is a sudden drop in atmospheric temperature combined with the formation of mist or fog. A connection also exists with damp autumn and winter months where there is a steady increase in emergency cases. The exact cause of the increase in cases is largely unknown and can not be attributed to just one event. Asthmatics react in different ways, and they often react to more than one trigger. Complications also occur when the location of the patient changes. Wetter conditions would tend to suggest that more time is spent indoors where dust mites and other allergens effect the respiratory system. Wet and damp weather also increase the occurrence of chest infections, an occurrence of which would have an effect on an asthmatic. Other occurrences of asthma can be attributed to sudden heavy rainfall. Rainfall breaks up pollen

grains and mould spores, allowing them to be carried in the air (Dales *et al.*, 2003). In the summer months asthma attacks can be attributed to hot days, low wind speed, low relative humidity and a drop in temperature at night. The condition leads to a steady build up of allergens and pollutants trapped in an almost motionless layer of air.

2.3 System Architecture: Definition

During the design of the system architecture a number of key stages that underpin the creation of the operational components are identified. The architecture combines the analysis of data from two areas of research: environmental and medical. First though, it is useful to define the meaning of *system architecture*.

Buschmann *et al.* (1996) define software architecture as, "a description of the subsystems and components of a software system and the relationships between them." Buschmann *et al.* go on to say, "the software architecture of a system is an artefact. It is the result of the software design activity". In this context, the word *system* implies a set of entities (real or abstract), that together make up a whole entity. The denotation of the word *system* changes, depending on its use. When a contained *system* becomes used in a *larger* system it could be referred to as a *subsystem* or *component*. However, a subsystem could also be built from a number of smaller systems (subsystems), or components. The number of subsystems described in a system will usually itself depend on the level of abstraction that the describing architecture prescribes. Each subsystem will interact or relate to at least one other component and serve the common objective of the system as a *whole*.

The organisation of these subsystems into a *whole* system is the activity of the architecture. The ANSI/IEEE Standard 1471-2000 specification (IEEE, 2000) states that *architecture* is "the fundamental organisation of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution." This statement provides a clear description of the role that architecture plays, namely defining components, and their relationships to one another and the wider system environment, and considering how they may evolve in use over time. The description supports the definition previously given by Buschmann *et al.* (1996). The text of the ISO standard (ISO, 2007) (a republished version of IEEE 1471:2000) will form the basis of future joint revisions by the ISO and IEEE bodies. The ISO standard states that "an

architecture exists to respond to specific *stakeholder* concerns about the system being described". It also states that "system architecture descriptions are inherently multi-view, as no single-view is capable of capturing all concerns". The capturing of rationale for inconsistencies or unresolved issues between views is promoted in the standard in order to explain those areas where incorrect assumptions could be made. This is an important point to note whilst defining the architecture for an environmental monitoring system, as disambiguation should be reduced to a minimum through clarification for each *stakeholder.* The stakeholders of an environmental monitoring system used in this field would most likely be the patients being monitored, clinical staff responsible for patient care, and researchers interested with the development of new treatments or studies.

Two further proposals for the meaning of architecture are given by The Open Group Architecture Framework (TOGAF, 2007). The two architectural definitions given by the group are:

- a "formal description of a system, or a detailed plan of the system at component level to guide its implementation", or
- as "the structure of components, their interrelationships, and the principles and guidelines governing their design and evolution over time."

These proposals support the view given previously. For an environmental monitoring system to be implemented successfully in various settings, some probably unthought-of at the time of design, it is important to consider these definitions of architecture and abstract the key elements.

2.3.1 Architectural Design Process

Malveau and Mowbray (2004) talk about a process for architecture quality improvement called *architectural iteration*, where the architecture is adapted at various project stages from practical feedback during the project life cycle. "At every step, the architects strive to improve the quality of the design; they use the lessons learned to make the design better and better". The process Malveau and Mowbray describe, contains a core process, known as *architecture with subprojects*. Architectural planning partitions the problem into subsystems with stable boundaries. Malveau and Mowbray go on to describe the process of system development and how it is analogous to architectural planning with four steps, the main two (and first two) described steps are:

1. Identify subsystems.
2. Define subsystem interfaces.

These two steps enforce the view of Buschmann *et al.* (1996) where components are used as building blocks (with defined interfaces between them). The final points they make, cover project planning and developing subsystems in parallel. This becomes possible as each subsystem exists as its own entity.

Some typical design activities that take place during architectural design are component design, data structure and algorithm design. Components act as building blocks for the structure of systems (Buschmann *et al.*, 1996) and have interfaces, enabling them to be used easily by other parts of the system. Data structure design starts to develop the detailed specifics of how the component will handle the necessary information in order to accomplish the component's task. Algorithm design then ties the data structure to the task that is given to the component.

When decomposing a sub-system into modules there are two processes that could be used, they are (Shaw & Garlan, 1996):

1. Object-oriented, where a system is decomposed into a set of communicating objects.
2. Data-flow, where functional modules are designed to accept input data and transform it in some way to output data. This is also known as the pipeline approach.

The two processes of Shaw and Galan (1996) are used to guide the underlying structure of an environmental monitoring system, while the process of *architectural iteration* was used during the research phases to provide feedback into the design process.

2.3.2 Architectural Views (or Models)

Definitions, given by both the IEEE (2000) and TOGAF (2007) earlier in the chapter can be used as a *tool-set* for defining the architecture of an environmental monitoring system. The definitions suggest that an architecture should describe an information system in terms of a set of building blocks, and show how these blocks fit together. This definition is taken forward and applied to the system and subsystems required to implement the

Environmental Monitoring System (EMS) during Chapters 4 and 5. In defining the architecture for the EMS a number of aspects require consideration. The Reference Model for Open Distributed Processing (RM-ODP) defines five viewpoints for understanding a system. Although these view points are designed for environments where distributed processing is used, it is still useful to consider the points in the design of the EMS to increase the system's flexibility. The view points are given by ISO (1998) and shown by the table below:

Table 5 System architecture viewpoints

Viewpoint	***Description***
Enterprise	How the specified system fits into the wider organisation.
Information	Constraints on the use and interpretation of data.
Computational	The functional decomposition of the system into a set of interacting objects/components.
Engineering	Matters relating to infrastructure required to support the system.
Technology	Concerned with the choice of technology.

The first three viewpoints: Enterprise, Information, and Computational, relate directly to the architecture, whilst the Engineering and Technical viewpoints support the implementation of the architecture, and are not prescribed by this thesis.

Two further aspects, promoted by The Open Distributed Processing Model (also known as ISO 10746) are, the use of abstraction, and precision in concept definitions. Consideration should be given to these aspects, together they reduce the number of assumptions that can be made. ISO 10746 also encourages the consideration of five key features within the design of software systems. They are listed here only to provide a context for an implementation of the architecture, and not to guide this research. The five key features of the model are:

1. Interoperability – the ability to link and reconfigure systems and services.
2. Heterogeneity – the ability to link across different platforms and protocols.
3. Transparency – the ability to hide complications from users.
4. Trading/Broking – the presence of intermediary agents, to promote and distribute software artifacts and services.
5. Federation – where focus is given to the lack of central authority over the software design or configuration.

These five aspects are useful to consider within the context of designing the implementation architecture for the EMS. Chapter 4 defines the architecture of the Environmental Monitoring System, and develops prototype implementations. Chapter 5 details key analytical components of the EMS.

2.4 Health Informatics

Wooten (2001) outlines some advantages of health informatics. In a trial using home video phones, electronic stethoscopes and digital blood pressure monitors. Patients with chronic conditions were given 17% less home visits than control patients (who were not using video phones or measuring equipment). The trial patients had more traditional telephone contact, in addition to video consultation with nursing staff. The quality of care between the two groups was measured to be similar. Over the trial period the average cost of care in the trial group was 27% less than that of the control group. Another study, this time lasting for 20 months in Finland (also mentioned by Wooten, 2001), found that 52% of referrals from general practitioners to *Peijas Hospital* in Helsinki were dealt with electronically. Hospital staff used either electronic messages or a video link to treat the patients. Against two control groups of similar patients from general practitioners, the cost was shown to be seven times greater for those not in the electronic referral group.

The delivery of healthcare in the home is continuing to grow as a proportion of total healthcare provision. With increasing bed costs, patients are being discharged earlier from hospital or having care which once took place in hospital extended to the home. This has advantages. Risks such as hospital-acquired infection are reduced and 'creature comforts' can be maintained. For some patients, care in the home can span many years whilst for others it will be very brief. Regardless of which category they fall into, patients have a need for information about their care, which might be factual information about their therapy, or assurance from healthcare workers that their own self management is working well.

National healthcare budgets are constantly under pressure to keep up essential services and increase the quality of care, whilst at the same time medical research creates more clinical findings than can be integrated into best practices. Often general practitioners and other

clinical staff find it difficult to keep up with the latest treatments, and budgets are stretched yet further.

The Connecting for Health NHS programme (previously known as *National Programme for IT*, NPfIT) which had its origins in 1998 (HSC, 1998) and came into operation on April 1st 2005 is creating a multi-billion pound infrastructure (NHS, 2003) which will attempt to improve patient care by enabling clinicians and other NHS staff to increase their efficiency and effectiveness through the use of IT. The program is outlined in the publication *Delivering 21st Century IT Support for the NHS* - National Strategic Programme (NHS, 2002). The NHS Executive states that it is committed to spending £12.4bn (at 2004-5 prices) over the ten year life of the main contracts, to 2013-14 to modernise the NHS with information technology (NAO, 2006). The focus of the contracts is to ensure that both quality information and information technology provide clinical services to patients and increases population health. This is seen as a significant shift from the previous emphasis on management information where information was used primarily as a tool to monitor cost and activity. The new strategy aims to ensure that health information is available to clinicians, and increasingly patients, who need it, when they need it.

Health informatics has the potential to improve quality, effectiveness and efficiency, if it is applied to the complete cycle of patient care and to the transfer of information related to a given health problem. For example, resources can be better managed. Once a suitable condition has been diagnosed, hospitalisation of patients can be kept to a minimum through continued monitoring at home. When a problem is identified the patient can be notified and corrective treatment arranged remotely. Information gathered in a setting away from a hospital is more natural for a patient, and leads to long-term datasets that may be used in further detailed analysis. Through the use of automated monitoring, clinicians are made available to address a patient's problems as they occur, making better use of the clinician's and patient's time. The work flow and efficiency of medical personnel can be improved, and the delivery of more personalised healthcare given to patients.

There are a number of methods already in use to transfer asthma data from the patient to clinician (Glykas & Chytas, 2004), one of these is the use of electronic devices that measure lung function, and other such clinical data, and then on a regular basis download these to a clinician. For example, over a standard telephone line (Medicate, 2000). Maglaveras *et al.* (2002) use technology to facilitate the transfer of data from both patient

and their immediate home environment. Electronic devices (used for the measurement of lung function in asthma) make the process of taking regular readings easier for the patient. Devices store the data from tests in memory until they can be transferred to the clinician. Using electronic devices relieves the patient from having to use more conventional methods of manual note taking and are known to increase the accuracy of the data being presented to the clinician. Connecting these devices via the Internet as a method of data transfer, enables the clinician to gain a greater granularity of data, closer to the time when the readings are taken. Clinicians are then able to use types of *trend analysis* to evaluate how well an asthmatic's condition is being controlled.

A review of literature between 1974 and 2004, by Sanders & Aronsky (2006) focusing on biomedical informatics applications for asthma care, found 64 papers; of which 13 papers were focused on asthma (disease) monitoring or prevention, and one (Crabbe *et al.*, 2004) was a retrospective study (the Medicate Project). The small number of papers and relevant literature found by Sanders & Aronsky gives weight to the novelty of this research.

The Medical Diagnosis, Communication and Analysis Throughout Europe (Medicate) project attempted to provide a proof of concept system, capable of taking ambulatory lung function readings via an electronic spirometry device and sending these to a *Disease Management System* for automatic review, at the same time allowing clinicians access to the respiratory data for their own individual clinical evaluation. Medicate (2000) achieved a simple alert mechanism that monitored patient data, looking for values that fell below a threshold set by a clinician and alerting the clinician to any unforeseen problems.

2.4.1 Ambulatory Monitoring

Ambulatory monitoring takes place whilst the patient is on the move (not bed ridden, and capable of walking), the advantage to this type of monitoring is that it allows the monitoring of conditions that are exacerbated by everyday life. This is especially significant were environmental factors might influence the health of a patient. Ambulatory monitoring lends itself to this application area. Telemetry can be used to send back data from a patient's ambulatory monitoring device to a base station or mobile system in *real-time* for further analysis. If monitoring takes place using sensors at some distance from the

subject under investigation the term *remote monitoring* is used to describe the activity.

There has been significant effort to allow primary care practitioners to manage patients who would normally be referred to specialist centres, by supporting them with remote specialist advice. One study (Shanit *et al.*, 1996) gave primary care practitioners direct access to a hospital-based cardiac monitoring centre. They were able to transmit a 12-lead ECG and consult a cardiologist on a 24 hour basis. Following the transmission of the signal, discussion with the cardiologist would reveal the outcome of the test. The process gave the benefit of reducing the amount of time spent travelling to appointments with specialists, and gave pre-warning to hospital medical teams if a patient was suspected of myocardial infarction. Another system, named *LifeShirt*, developed by Levy *et al.* (2004) was an ambulatory respiratory and cardiac function monitoring system. The system was used to detect respiratory function abnormalities in sleep apnea syndrome as well as other disorders. However the accuracy of the measurements can not be verified due to the small sample of subjects studied.

In the clinical environment where an increasing amount of care is given away from the hospital, this frequently refers to clinical monitoring within the home. Rialle *et al.* (2002) use monitoring equipment installed at patients' homes, measuring blood pressure and cardiogram data, transmitting the results back from locations remote from the clinical setting. Clinical monitoring however should not be limited to this fixed location. As patients should be able to move normally from location to location, leading as far as possible, comparatively normal lives. Engin *et al.* (2005) develop a telemedicine system that transfers human electrocardiogram (ECG) signals via mobile phone. The real-time data transmission via mobile phone allowed doctors to check the coronary care of patients in rural areas. More recently Cleland *et al.* (2007) used an electronic lung function device attached to a mobile phone to transmit respiratory data. The day's temperature, wind speed and pollen count, specific to the mobile phone were also transferred. Results from the clinical trial indicated an increased rate of poor asthma control identification, and better communication with healthcare professionals without the need for face-to-face consultation.

Remote monitoring can be used in conjunction with ambulatory devices. This is especially the case were environmental factors are monitored. Many environmental factors are difficult or impractical to be monitored with portable measuring devices. Therefore, fixed

location devices must be used to record data.

Personal monitoring in this thesis refers to the use of remote monitoring and electronic ambulatory devices capable of obtaining regular readings of a patient's lung function measurements and environmental variables. Personal monitoring allows better communication between the patient and clinician. Recognition of early warning signs of worsening asthma are noticed and prompt warnings can be given by a clinician to any serious deterioration in symptoms or peak flow. A warning could also be given to remove or withdraw from allergic or irritant precipitants in the environment that may be contributing to the exacerbation. Beginning treatment away from clinical contact avoids delays and reduces the severity of exacerbation, at the same time adding to the patients sense of control over their asthma.

2.4.2 Need for Intelligent Monitoring

Intelligent monitoring is the process where by the information or data provided to the user about the monitored subject is both timely and directly attributable to the problem being monitored. The monitoring process usually generates information when a certain event condition has been met. The use of intelligent monitoring often takes this one step further, where a number of event driven signals are analysed, and an appropriate response formulated. Dawant *et al.* (1993) present a distributed computer architecture for intelligent patient monitoring that introduces a number of relevant concepts. These include: data acquisition, data reduction, selective display of information, and the facilitation of these concepts through the use of asynchronous software modules. The central modules are responsible for feature extraction, modelling the patient, and displaying information. The semi-independent nature of these processes is particularly relevant to this work, and promotes scalability.

There are many reasons why intelligent monitoring is desirable for monitoring patients. Clinicians face an incredible amount of information, and when time is limited or there are numerous patients to monitor in parallel, observations can be missed. This is also known as cognitive overload (Coiera, 1997). Some monitoring devices present more information than can be absorbed by the clinician while others distract the user with false alarms. Using intelligent monitoring can reduce these issues through better interpretation of signals.

Vázquez *et al.* (2006) presents a distributed module, based on intelligent agent technology. The module is dedicated to the process management of networked medical devices, and provides real-time acquisition and analysis of physiological data. Systems capable of reasoning with medical knowledge are classed as intelligent, and come under the category *Artificial Intelligence in Medicine* (AIM). The most common types of AIM system in routine clinical use are expert or knowledge-based systems. These systems contain medical knowledge about a specific area and are able to reason with data from individual patients and form a conclusion. Expert systems are also capable of generating alerts and reminders to warn of changes in a patient's condition.

One of the driving factors of AIM is to create systems which are able to learn from experience. Techniques of machine learning have accomplished this objective to varying degrees of success. The way in which knowledge is represented within such systems is more advanced than standard statistical tests, which are applied to data through manual searches. Machine learning systems are capable of identifying complex relationships between data sets or individual parameters through the manipulation of raw data. Systems which use machine learning can be used to develop the knowledge bases used by expert systems. This is achieved via a systematic description of data features which uniquely characterise each pattern or condition, then through the transformation of these into simple rules.

Artificial Intelligence (AI) offers medicine a way of constructing computer systems that have a capacity to capture and then reason with medical knowledge. AI systems have two distinct capabilities:

1) To take new data and create knowledge from relationships that exist within it, and
2) Take medical knowledge and use it to reason with data.

These two categories of AI system can be more simply classified as *model generators* or *model users*.

2.4.3 Information Discovery

Information discovery is the process of looking into a large datastore and discovering knowledge in the form of significant patterns and relationships. As an example Keles and Keles (2006) develop an expert system using fuzzy logic to aid the diagnosis of thyroid diseases. Their system is capable of diagnosis with a 95% accuracy, compared with the actual diagnosis of clinical staff. Discovery is usually either guided by a user (supervised) or automatically using intelligent software (unsupervised). In these systems it is not necessary for the user to have an understanding of statistics, or skill in using a query language because the system output indicates the key factors which shape the data. According to Parsaye (1993) the three stages of information discovery are:

1. Understanding.
2. Improvement.
3. Prediction.

It is extremely difficult to manage or control a process that is not understood, and often the extent of understanding is not known until an error is made that shows the magnitude of the misunderstanding. Understanding is frequently the goal of information discovery and interpretation. There are many forms of understanding. Three types of understanding that are particularly important are:

1. Differences,
2. Trends, and
3. Relations.

The steps that a human analyst might take to explore a database are as follows:

- Form a hypothesis.
- Make some queries.
- Run a statistics program.
- View the results and perhaps modify the hypothesis.
- Continue this cycle until a pattern emerges.

These steps would become tedious if repetitively performed by a human, hence the use of information discovery algorithms which automate the process of pattern discovery. Information or rule discovery can be guided through the use of hypotheses.

2.4.4 Information Processing Issues

The advantages of technology are numerous; using computers to automate tasks allows analysis to be performed on batches of data on a continual basis, due to calculations being achieved at speed. Efficiency is greatly improved over the productivity of a human operator. The number of parameters analysed concurrently can be increased beyond the usual one or two, which facilitates analysis of groups of parameters to determine if there is a combined effect. This would not be easily achievable without the use of computing technology.

However, the use of technology introduces new issues to consider. As the processes are required to make a transformation of the data, some work flow coordinator is required to schedule and monitor the analysis to ensure that the desired result format is produced. This coordination requires software that is easily maintained, giving the user enough flexibility to perform their analysis with the minimum amount of effort or specialist knowledge. Software should have the ability for integration into other products in the event that the analysis requires further investigation using a new technique or dissemination in a way not identified during the software design process.

The methodology used for processing the data, and finding if any patient specific relationships exist requires consideration. Mather *et al.* (2004) discuss the advantages and limitations of several statistical methods for linking health, exposure and hazards. They split types of analysis into three groups:

1. Tracking and trend analysis,
2. Ecologic analysis, and
3. Etiologic research studies.

Each successive group generally becomes more specific to an individual patient, starting with trend analysis which focuses on a population, and useful for characterising the background or seasonal base line. Correlation methods have been used to identify general relationships between air quality and a patient's asthma condition (Crabbe *et al.*, 2004), however these were primarily using minimum, maximum and average values for the data on a daily basis, and produced a general correlation between air quality and the patient's respiratory condition. This thesis, however, is seeking to specifically identify air quality characteristics that are a particular predictor of a patient's asthma episode. The use of statistical techniques to locate specific correlations between lung function and

environmental (specifically air quality) factors can be difficult to achieve. Statistical methods, such as correlation are not sufficient on their own to recognise and classify important changes in the data. Molitor *et al.* (2007) examine the uncertainty in spatial exposure models aimed at the etiological level, where spatial effects such as the proximity of patients to pollutants sources (such as roads) are considered. They say that assessing pollution distributions at the intraurban scale has proved challenging because of the lack of routinely collected data, they go on to say that the use of geographic information systems (GIS) with existing information now shows promise in assigning exposure to an individual's microenvironment.

It cannot be assumed that a particular air quality period is attributed to the cause of an asthma attack without some validation. While it is not difficult to find a set of data which appears to support a given hypothesis, it is necessary to validate the results by repeated testing before a level of confidence can be established.

The introduction of time lag complicates the identification of relationships even further, as the number of possible relations increases infinitely. A method for controlling the identification process is required to ease computational demands. The use of time lag in the analysis goes some way to developing a method for detecting a relationship between air quality levels and deteriorating lung function.

The ability of the system to recognise reoccurring patterns through a means of validation is important. Moreover the system should be capable of achieving this recognition via a semi-guided or automatic means. Mueller and Lemke (2000) suggest a synthesis of models into a hybrid solution; collective solutions reflect reality more thoroughly than any single model.

There are three affiliated areas of research that attempt to identify and validate patterns within data. They are (Michie *et al.*, 1994):

- Traditional statistics,
- Machine learning, and
- Neural networks.

Statistical methods are generally used to summarise or describe a collection of data.

Statistics can also be used to model data, and then used to draw inference about the process or population being studied. Statistical techniques often use a probabilistic approach (Barber, 2006) to classification that leads to an indication of the likelihood that an event belongs to a certain class.

Machine learning is a broad sub-field of artificial intelligence and is concerned with the development of algorithms that allow computers to learn (Michie *et al.*, 1994). Machine learning generally encompasses automatic computing procedures that learn a task from a series of examples. Classifying expressions are produced that are easily understood by the ordinary person. The major focus for machine learning is to extract information from data automatically, by computational and statistical methods. Therefore, machine learning is an extension of the field of traditional statistics.

Neural network techniques offer the advantage that they facilitate automatic pattern recognition, through response to input data on a continual basis (Haykin,1999), and could allow patterns of environmental events causing asthma exacerbation to be built up. Neural networks are useful to give an arbitrary classification, and generally combine the complexity of statistical techniques with the machine learning objective of imitating human intelligence.

Neural networks are tools that can be used for non-linear statistical data modelling. They can be used to model complex relationships between inputs and outputs, or to find patterns in data. As an example of a non-linear application – the weather is famously non-linear, where simple changes in one part of the system produce complex effects throughout.

The basic issue based on rule discovery may be conceptually represented by the following equation (Parsaye, 1993):

$$\text{Rule Discovery} = \text{Generation} + \text{Filtering}$$

The constraints imposed on *rule discovery* control the generation component, while parameters such as the confidence level or length, control the filtering.

Rules can be used for prediction. Rules obtained from expert experience or advice can be

compiled into a comprehensive list of cause and effect descriptors, and then used by a computer or a less experienced human operator to predict an outcome from a set of input variables. Prediction is usually always accompanied by an uncertainty or confidence level from each of the rules.

The first, and probably the most important problem faced when trying to apply artificial intelligence in a practical setting, is selecting attributes for the data at hand. Witten & Frank (2000) suggest that if meaningful attributes are not chosen that together convey sufficient information to make learning possible any attempt to apply machine learning techniques is likely to fail. The choice of a learning scheme is usually far less important than the application of a suitable set of attributes.

2.5 Summary

This chapter gave a broad overview of a number of topics. The overview began by outlining the processes used in the treatment of an asthmatics condition, the parameters routinely involved, and their measurement using peak flow monitoring devices.

Environmental factors shown to have an adverse effect on the asthmatic were then discussed, including a description of the levels of environmental factors such as air quality, required to be problematic to an asthmatic. A number of researchers identified links between the quality of air in the environment and health of a subject. For example, Tamburlini *et al.* (2002) found that environmental influences have an effect on health conditions, and is particularly prevalent on respiratory related illness

The purpose of system architecture within system design, was defined to clarify the role of the concept, and its use by this thesis. The design processes, and architectural aspects to consider when defining a system were presented to give a foundation for research by following chapters.

The area of health informatics was then introduced, with a description of its history, and the benefits to which it can be used. Areas of intelligent monitoring, information discovery, and general processing issues were also given a broad overview. Issues such as the

reduction of information through feature extraction were raised.

The overview given by this chapter established that every patient experiences their own unique personal exposure to environmental influences, and that informatics can be used as a tool to identify the levels of exposure that may adversely affect the asthmatic. Recently Chin-Shen *et al.* (2007) studied the effects of particulate matter on the peak expiratory flow rate of asthmatic children, and concluded that personal air quality data is more suitable for the assessment of changes in lung function than ambient monitoring data. Other research by Cullen (1996) suggests that the primary goal of past epidemiology investigations had been the establishment of causal relationships between an environmental agent and a health outcome. He stated that there had been little theoretical work on models for evaluating environmental patterns, rather than average or cumulative dose, as predictors of risk. The next chapter introduces a new term; ***enviromedics***, which describes a new field of research developed by the author, to better define the problem domain, and progress research in the field.

Chapter 3

Key Requirements for Enviromedic Architecture

This chapter proposes a set of requirements that underpin the architecture realised later in this thesis. All applications built using the realised architecture would fulfill these requirements.

3.1 Introduction

The term enviromedics has been introduced by this author to encapsulate the use of both environmental and medically related data sets in analysis. Enviromedics combines the analysis of geographical and temporally related data, specific to the environment of a patient, with the aim of providing or enhancing medical care for that patient. Key components of the analysis are presented in this chapter, and shown to be applicable to respiratory healthcare through prototype testing during Chapter 6.

The thesis aim (presented during Chapter 1) was to identify processes capable of identifying predictors of patient-specific asthma exacerbation, and provide a system architecture, required for the automation of these processes. The advancement of such a detection system is in synergy with enviromedics. A new technique was required to overcome the issues traditionally associated with correlation and scale to analysing large data sets by identifying key features that could act as predictors of respiratory decline.

For predictors to be identified, first the point at which a healthy lung function signal begins to decline (termed a *peak* by this thesis), and a *change event* in the environment such as a peak air quality reading must be recognised. To begin to meet these requirements, and in order to formally identify these points within the analysis, the concept of the *reference datum* was introduced.

3.2 The Reference Datum

The concept of the *reference datum* is introduced by the author to describe points marking key features within the analysed data sets. The term is generally borrowed from the science of geodesy where datums are used to mark locations on the earths surface (Section 3.4 describes the use of a datum in the geodetic sense). The denotation of the term *reference datum* here is; an identified point of interest (represented with a value), measurable in time and by location.

The introduction of reference datums focuses further analysis between these *singular* points (environmental and respiratory), rather than analysing two sets of time series with correlation. The benefit to referencing (in particular time) between single points is that it eliminates the need for direct correlations between sets of time-series data. This reduces the number of analytical permutations that are available, such as length of data series to correlate, number of interpolated data points to calculate, length of time lag to introduce between the two data sets, among others. The removal of these options increases the scalability of analytical techniques. By focusing on the points identified by the reference datums the need for exhaustive time-series data searches is minimised, reducing time and computational power required.

The introduction of the reference datum concept into analytical techniques fundamentally alters the approach required to find enviromedic patterns. This is because the analysis is now focused between two data points, one from an environmental data set, and the other respiratory. Using a correlation technique with two data points always yields a 100% correlation between the data, so the use of correlation obviously imparts little knowledge using this approach. It becomes necessary to identify an alternative form of analysis. The study of aetiology aims to explore cause and effect relationships between differing, but inter-related data sets. This research applied the principle of aetiology to the problem of identifying patterns between environmental predictors and the decline in a patient's lung function through the hypothesis that an environmental event, that often leads to a decline in patient lung function can be used as a future predictor of a decline.

Combining reference datums to mark *interesting* features in respiratory data, with the desire to identify a predictor of the change event (also marked by a reference datum), an additional concept, the *delay characteristic* was developed and introduced to link the two disparate reference datums together during analysis.

3.3 The Delay Characteristic

Without a method for consistently relating possible environmental predictors to a decline in lung function it is not possible to make associations between the two data sets. A number of key parameters are required to enable the identification of a link between a patient's environment and their respiratory health. These include: the **value** attributed to the environmental data, the lag time between the (environmental) predictor and reference datum belonging to the decline in lung function, and the **time and date** at which the environmental reading is taken.

The information recorded within a *delay characteristic* is shown below in Figure 7.

Parameter (data type/name)	***Date***	***Value***	***Lag***

Figure 7 Format and parameters recorded within the delay characteristic.

The *delay characteristic* contains a name (or key) that records the data type, and then three parameters *Date (time and date)* of the environmental reference datum, *Value* of the data type, and *Lag* (time) between the environmental and respiratory reference datums. The relationships that are most important are the *lag* time and *value*. The delay characteristic allows an environmental predictor to be related to a decline in lung function.

The process for defining a delay characteristic can be summarised in three steps:

1) Identify a decline in lung function (use reference datum).
2) Identify a possible predictor of the decline (use reference datum).
3) Relate the two datums (in 1 & 2 above) together through the use of a delay characteristic.

This series of steps outlines a new method, that in reality focuses further data analysis onto a set of data *outliers*. Reference datums identify periods that naturally occur at the extremes of data sets, and these are converted into delay characteristics, making a new data set where outliers become part of the core analysis.

The time and date of the reading is required to enable a meaningful comparison to be made between the various data sets. Following experimentation with prototypes later in this thesis, it is considered that it may benefit the system's capabilities if the change in value direction was also recorded. For example, if the ozone level were falling or rising at the time of the decline in lung function. However, the importance of this additional information will need to be verified through further research, and is outside the scope of this thesis.

3.4 Patient Location

In order to analyse personal air quality exposure levels (the actual air quality that a patient experiences), the ideal location to take environmental measurements at, is that of the patient. The measurement could be made using some sort of portable and personal monitoring device. Devices such as these are not commonly available though, and it is likely that *their* cost would negate wide-spread uptake at this time. However, use of such devices should not be ruled out in the future.

The capabilities of devices (EE&S, 2006) (Air Monitors, 2006) (IS&S, 2006) (Topac, 2003) common for monitoring do not contain a location recording component, portable air quality devices also do not widely monitor a range of particulate matter or a large number of gasses. For these reasons it is a requirement of the system that readings from fixed monitoring stations are used and matched with patient readings according to time and location; so as the patient moves, tracking occurs. Once portable devices use an appropriate means of tracking a patient's movements, these reading can be incorporated into the system. The matching of time and location data is important in the analysis of patient specific information, as a deeper understanding of the patient's real environment can be obtained.

Environmental data is not usually available from an ambulatory source. NETCEN (National Environmental Technology Centre) is the common source for automated air quality data in the United Kingdom. NETCEN is responsible for most of the automatic air quality monitoring stations in the UK. Due to the current inconvenience in carrying portable air quality monitors, available data is restricted to these sites. With further

technological development air quality monitoring in the home (which is currently possible) or on the person would be used. Examples of portable equipment are available from RKI Instruments (RKI, 2004).

In anticipation of future advances, the system must be capable of facilitating analysis of this detailed information. Implementation of a tracking system requires that data be capable of being analysed so closest match data can be extracted from the database. A distance formula is needed to ensure this.

The Global Positioning system (GPS) was introduced in 1973 by the United States Department of Defense. The position of an object (marked by GPS) is computed from time signals sent from satellites carrying extremely accurate atomic clocks. Using lung function measurement devices that are capable of recording location (using GPS) in addition to lung function data would allow the matching of environmental data to a patient's activity. The ability to track a patient, recording location and lung function measurements in addition to personal air quality exposure are two requirements for an *enviromedic* system identifying air quality effects on respiratory patients.

Personal exposure to airborne triggers is obtained in two ways:

- Personal environmental monitoring using ambulatory devices.
- Dispersion models that use data from static air quality monitoring stations.

For the concept of the delay characteristic to be feasible, it is necessary to obtain environmental data relating to the movements of the monitored patient. The movements of the patient should be related to the closest actual (or modelled) and relevant air quality data.

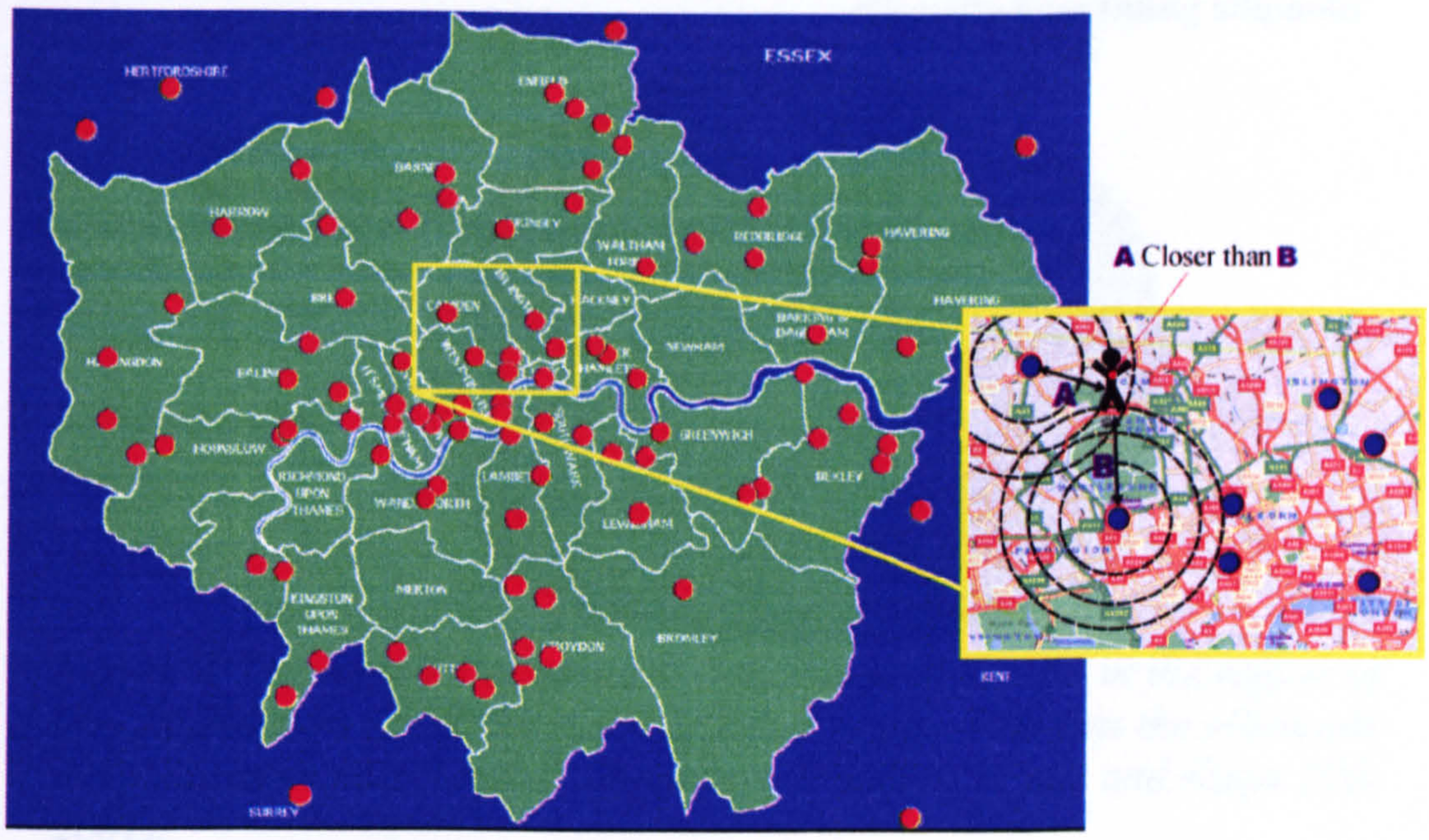


Figure 8 The decision between air quality monitoring sites

The measurement of location within the system is based on Longitude and Latitude co-ordinates which are spherical co-ordinates following the earth's approximately spherical shape. An example of latitude and longitude readings are, *30:16:28.82 N 97:44:25.19 W.* Location is referenced to a point defined when the measurement is taken. However there are a number of alternative reference systems that could be used to relate the position of the patient and their distance from the closest air quality data measurement point.

Due to the earth's irregular shape, and a legacy of methods for recording location, there are many ways to identify and record location. The number of mapping systems in existence is a result of historical (and localised) mapping methods (Dana, 1999). Since the increase in globalisation and the introduction of Global Positioning Systems (GPS) there has been a greater requirement to record position in a uniform manner. To achieve this, a number of assumptions have been made and tested. As the earth is not uniform in shape, a best fitting reference model is required from which to take measurements.

Figure 9 shows an example of two ellipsoids that are used to model the shape of the whole, and a segment of the Earth respectively.

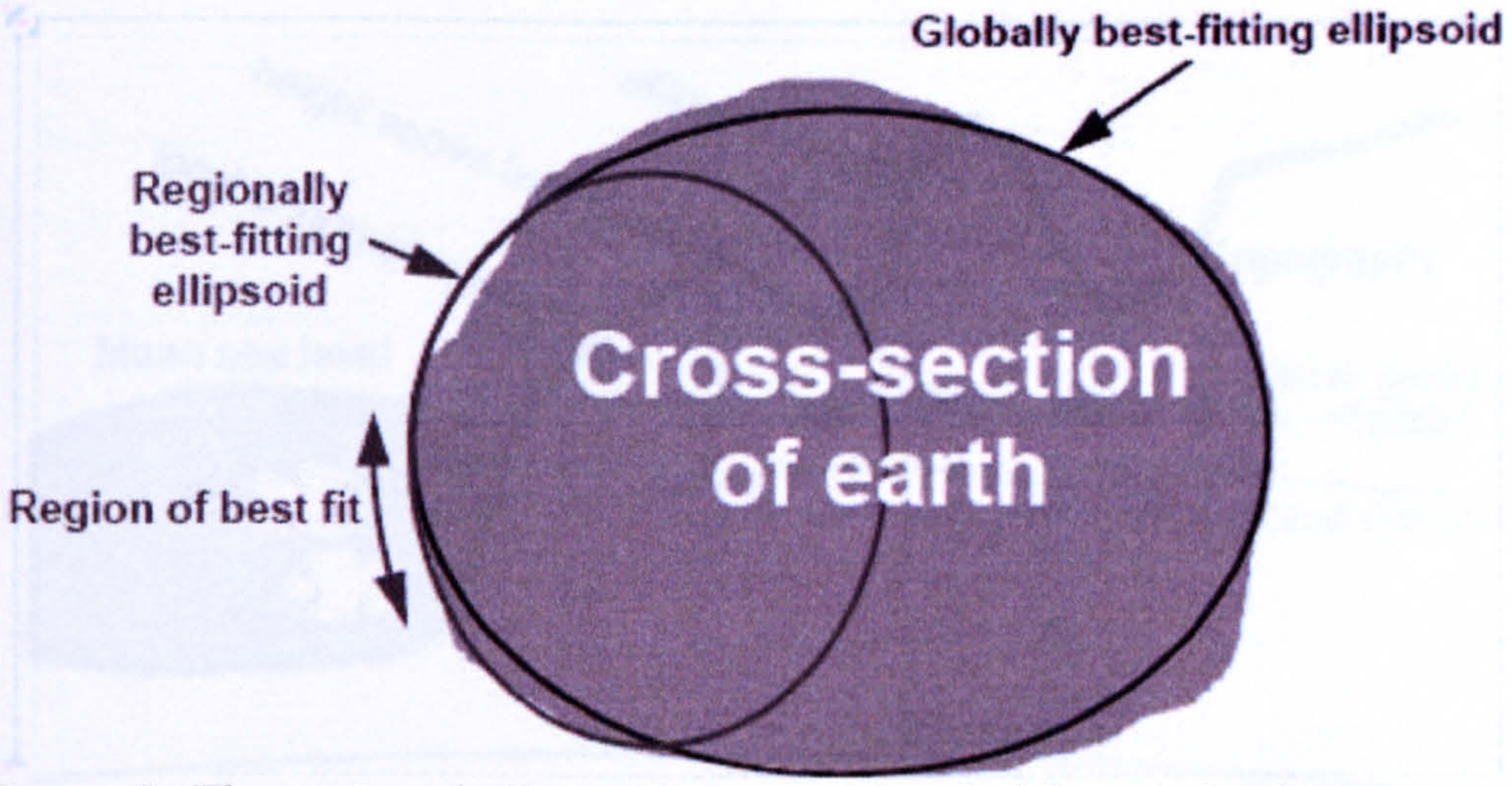


Figure 9 The regional ellipsoid is only intended for use in the region of best fit and does not fit the Earth in other areas. Note that the ellipsoids differ in centre position and orientation as well as in size and shape (OS, 2001).

The globally best-fitting ellipsoid does not fit the Earth perfectly, so there are many different ellipsoids in use. Some are designed to best fit the whole earth, and others designed to best fit one region. Although the modern trend is to use global coordinate systems, even for local applications, it is important to realise that in a global coordinate system, the ground on which we stand is constantly moving. This leads to subtleties in coordinate system definition and use. Therefore, to use latitudes and longitudes with any degree of certainty, the ellipsoid used for recording the location must be known and referenced to a *geodetic datum*. The term *geodetic datum* is usually taken to mean the ellipsoid and datum: a set of 3-D Cartesian axes plus an ellipsoid, which allows positions to be equivalently described in 3-D Cartesian coordinates or as latitude, longitude and ellipsoid height. The measurement of height is similarly complicated. The *Geoid* is a single unique surface and is the only level surface which best-fits the average surface of the oceans over the whole Earth. Figure 10 shows the variability in recording a location's height attribute.

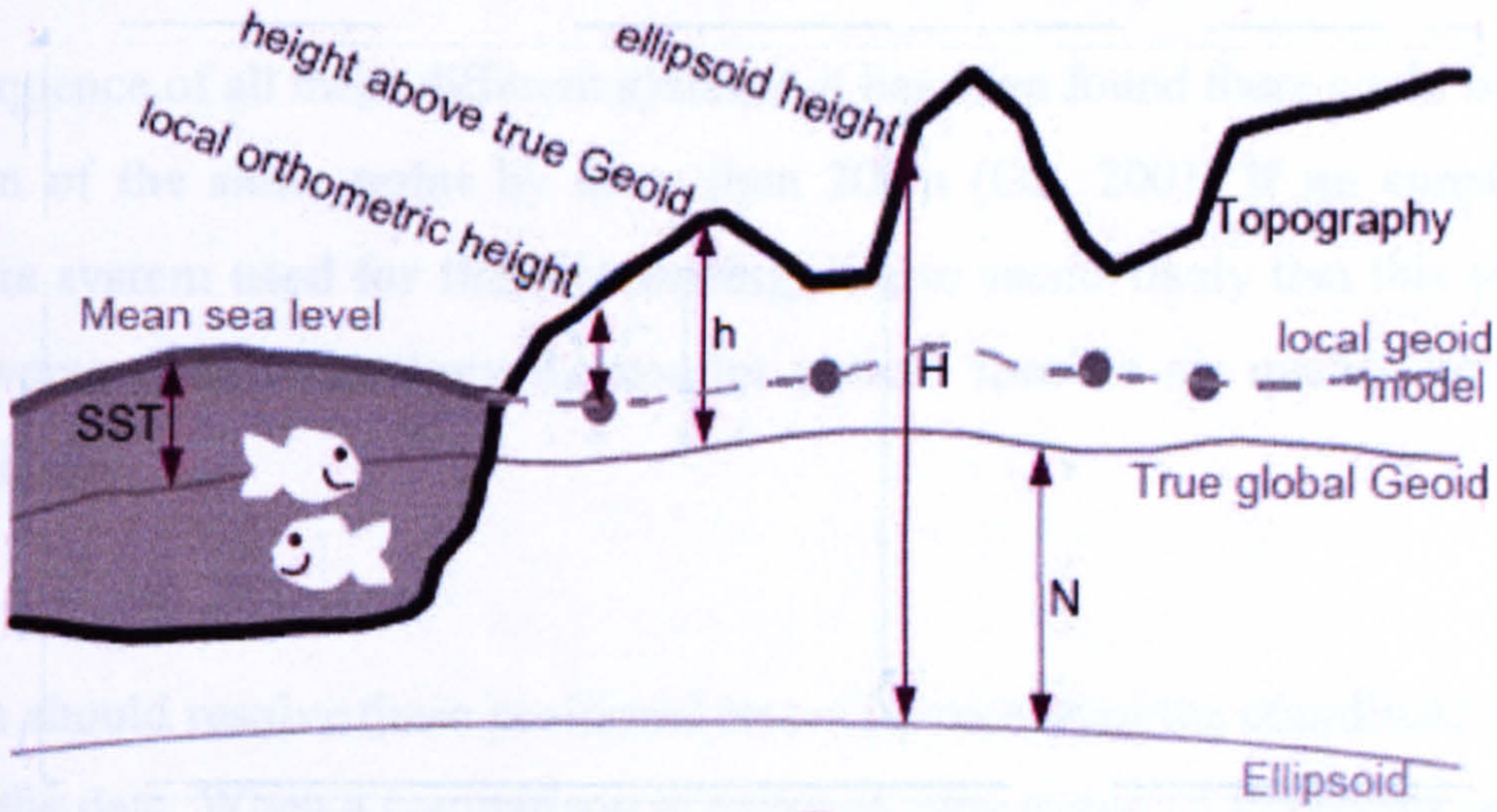


Figure 10 The relationship between the Geoid, a local geoid model (based on a tide-gauge datum), mean sea level, and a reference ellipsoid. The ODN geoid model is an example of a local geoid model (OS,2001).

The figure above shows the different measurements of height with regards to the various model geoids, and reference ellipsoid. The *true global geoid* is a reference model derived from mathematical calculation, and is the level surface which best fits global mean sea level (MSL). The local geoid deviates from the MSL due to water currents and variations in temperature, pressure and density. These produce watery "hills and valleys" in the average sea surface. This phenomenon is known as sea surface topography (SST) (OS, 2008). The *local geoid model* is the gravimetric surface of the earth, and defined through gravitational measurements. The *ellipsoid* model is the geometric idealised surface of the earth (Li & Gotze, 2001).

Once location and height have been recorded relative to a known ellipsoid, the measurement can be transformed from radial measurements and projected onto flat surfaces (maps) or converted into other coordinate systems. A map's projection is a way of depicting the spherical surface of the earth on a flat piece of paper.

The datum used for GPS positioning is called WGS84 (World Geodetic System 1984). It consists of a three-dimensional Cartesian coordinate system and an associated ellipsoid so that WGS84 positions can be described as either XYZ Cartesian coordinates or latitude, longitude and ellipsoid height coordinates. The origin of the datum is the Geocentre (the Earth's centre of mass) and it is designed for positioning anywhere on Earth (OS, 2001).

As a consequence of all these different systems, it has been found there could be an error in the location of **the same point** by more than 200m (OS, 2001) **if no consideration is made of the system used for their recording.** It also seems likely that this situation will be made worse once technology to monitor patient specific air quality becomes more readily available.

The system should resolve these positional issues by recording the coordinate system used to provide the data. When a comparison is required, conversion of the points into the same coordinate system is necessary to enable an accurate comparison. Coordinate systems use projections to convert angular measurements in degrees to linear measurements such as meters in order to give a length along the curvature of the earth. The conversion is performed in three steps (Mentor, 1999):

1. Convert the source coordinate to geographic form, latitudes and longitudes, using projection algorithms.
2. Apply a datum shift to the resulting latitudes and longitudes to convert them from the datum of the source coordinate system to the datum referenced by the target coordinate system.
3. Convert the resulting geographic coordinates back to Cartesian form using the projection algorithms.

A review of formulae available for measuring distance is provided in *Appendix O.* The Haversine Formula, discussed by Sinnott (1984) produces a mathematically and computationally exact result, the formula is widely used by graphical packages to plot distances. A standard way of storing location details is through the use of spherical coordinates (longitude and latitude), and the Haversine Formula is particularly suited to this. However, the Haversine Formula is not good at making calculations where the distance between locations is large (for example where the two points are either side of the earth) where there can be an error of up to 2km. This error is not considered to be a problem for the system, where calculations will be over small distances of up to approximately 5 miles.

3.5 Identification of Respiratory and Environmental Change Predictors

The US National Heart, Lung, and Blood Institute (NHLBI, 1997a) emphasised the need for patient-tailored monitoring and treatment in their Second Expert Panel Report, stating that "Asthma self management education should be tailored to the needs of each patient". A long term aim of this work is to provide both patient and clinician with additional patient-specific information. The need for the patient-specific approach was confirmed through discussions at the Healthcare 2002 conference and exhibition (Barber *et al.*, 2002). Further research into personal exposure of patients to air quality levels was also recommended by COMEAP (2006) with regard to cardiovascular effects resulting from poor air quality. Further discussion with respiratory researchers at the Whittington Hospital London helped identify that having a system that identified patient-specific allergens capable of exacerbating asthma would be useful. Carrer *et al.* (2001) suggest a number of measurements that would facilitate this in an indoor environment.

With the capability to collate and relate a patient's experience of their environment to their respiratory health, a system would have the processes required to facilitate personal enviromedic pattern recognition with asthma patients. The process can be described in three steps:

1. Identification of the asthma episode.
2. Identification of the environmental predictor.
3. Monitoring for the environmental predictor.

Step 1 – Identification of Asthma episode

The first step is to detect adverse patterns depicting signs of an asthma attack within the lung function data. It is important to identify these existing traits so that environmental factors useful for predicting the asthma episode can be found using the reference point for investigation, described in *Step 2*.

Step 2 – Identification of the Environmental Predictor

This involves the identification of environmental influences which appear to be related to adverse lung function patterns. The method behind the identification of the patterns should keep the identification process flexible and open to new influences. There are generally two types of process that can be used (Witten & Frank, 2000): *supervised*, and

unsupervised. When pre-analytical knowledge of the data is known such that classifications of the underlying data can be made, the process can use ***supervised learning***; whilst ***unsupervised learning*** should be used when data classifications are unknown.

As exploration of data should be left as far as possible unguided by an operator, the process should be focused on *unsupervised learning*. Investigation should be concerned with the identification and validation of patterns. Recurring patterns are an indication that a particular environmental characteristic (pattern) is more likely to be a good predictor of lung function decline, and learned by the identification process.

Step 3 – Monitoring for the Environmental Predictor

Once relations have been found among enviromedic data, the results should be made available within the system for the purpose of monitoring incoming data. Incoming patterns that follow the same characteristics as a learned pattern in the system, should trigger an alert to clinical staff prompting them to assist in patient care.

3.5.1 Identification of Asthma Episodes

The method used for identification of periods of asthma exacerbation outlined in clinical guidelines (NHLBI, 1997a) uses a threshold individually calculated by clinical staff for each patient. Lung function is then monitored, and if using ambulatory electronic devices, clinical staff can be alerted when the set threshold value is exceeded.

So that an element of prediction can be integrated into patient monitoring, it is necessary to extend this process by monitoring the trend of the patient's respiratory condition, in addition to the individually calculated threshold value. The technique developed by this research is shown in Section 5.2. The patient's lung function trend is monitored using a fitted regression line to calculate the rate of change in the patient's condition. The fitted (regression) line can then be used to identify points where the trend changes direction, creating *peak* and *trough* points.

Monitoring the trend of a patient's respiratory condition minimises the effect inaccuracies have, as the analysis relies on a number of readings. The technique is also applicable to

other data types, such as air quality, pollen, atmospheric, and other data sets where trends can be found.

3.5.2 Environmental Predictor Identification

Patterns that are hidden within large data sets, particularly patterns between environmental and respiratory data are difficult to verify without an automated process. There are almost an infinite number of relationships that could exist among the data sets. In addition there are many environmental factors which are monitored, but play no part in the patient's decline in lung function. For example, a high level of PM2.5 that has occurred (possibly with time lag) at the same time as a trigger point of an asthma attack, can not be assumed to be attributable to the patient's condition. Validation over a significant number of observations is required before confidence can be established.

There are two stages in identifying environmental patterns that can be used as reliable predictors of patient lung function decline. First, a probable relationship between the data sets is required. Second, the relationship has to be validated as similar relationships appear in future data sets.

3.5.3 Predictor Monitoring

Once environmental predictors have been recognised, a mechanism to facilitate the monitoring of real-time data (for similar patterns causing a problem for the patient) is required. The outcome of the matching process should provide an indication of the seriousness of the impending adverse event, and either give advice to increase medication, or for the patient to decide if an alteration in lifestyle for the problematic period should be considered.

Monitoring in a real-time environment requires the patterns being monitored to be stored using a method that is capable of matching against a stream of input data. Data from monitoring devices often contain spurious readings and general measurement noise. As a

result of this, and other factors, often an exact match is not possible. The system has to determine whether the input is sufficiently close to a stored trigger for it to raise an alert.

A recognised method for matching patterns is to use a *distance metric* (deSmith *et al.*, 2007), a method for testing if an input pattern is physically similar to the pattern being monitored. For mathematical purposes the input pattern is converted to take the form of a *vector.* The vector is then compared against a *model vector* using a *metric.* The most common form of metric is the *euclidean distance,* and one model type using this is a neuron belonging to a *neural network.*

Neural networks generally use a system of *weights* from which comparisons are made to the input data often using euclidean distances, in this case the distance between the weight vector and the input vector is analysed for a match. If the distance between the input and the weight is below a threshold then the neural network would recognise the match and activate a response to the pattern.

Certain types of neural network, such as networks based on the Radial Basis Function (RBF) are adaptable over time to change in the underlying monitored pattern. RBF networks employ a radial function, commonly based on the Gaussian distribution. The function is used to determine if an input pattern should be recognised by the neural network. The decision is derived by calculating the activation level of the radial function, with respect to the given input.

When an environmental *predictor* is identified, the characteristics of the predictor can be stored as a vector, and a neuron within the neural network adapted to represent it. The neuron determines how close a match it is, to its internal weight it is monitoring. If the metric is within the limits defined by the matching function, the neuron activates and triggers a response. Neural networks are particularly well suited to this role because they are capable of providing both the functions of adapting to changes in the underlying data, and providing alerts when the monitored pattern is encountered.

3.6 Overview of Architecture

The architecture used to build applications suitable for enviromedic analysis incorporate the major concepts outlined from Section 3.2 through to Section 3.5, which include: the reference datum; patient location, and matching patient-specific environmental data to the patient; the delay characteristic; and validation of the delay characteristic. The architecture in which these key concepts are combined, is shown by the enviromedic architecture in Figure 11.

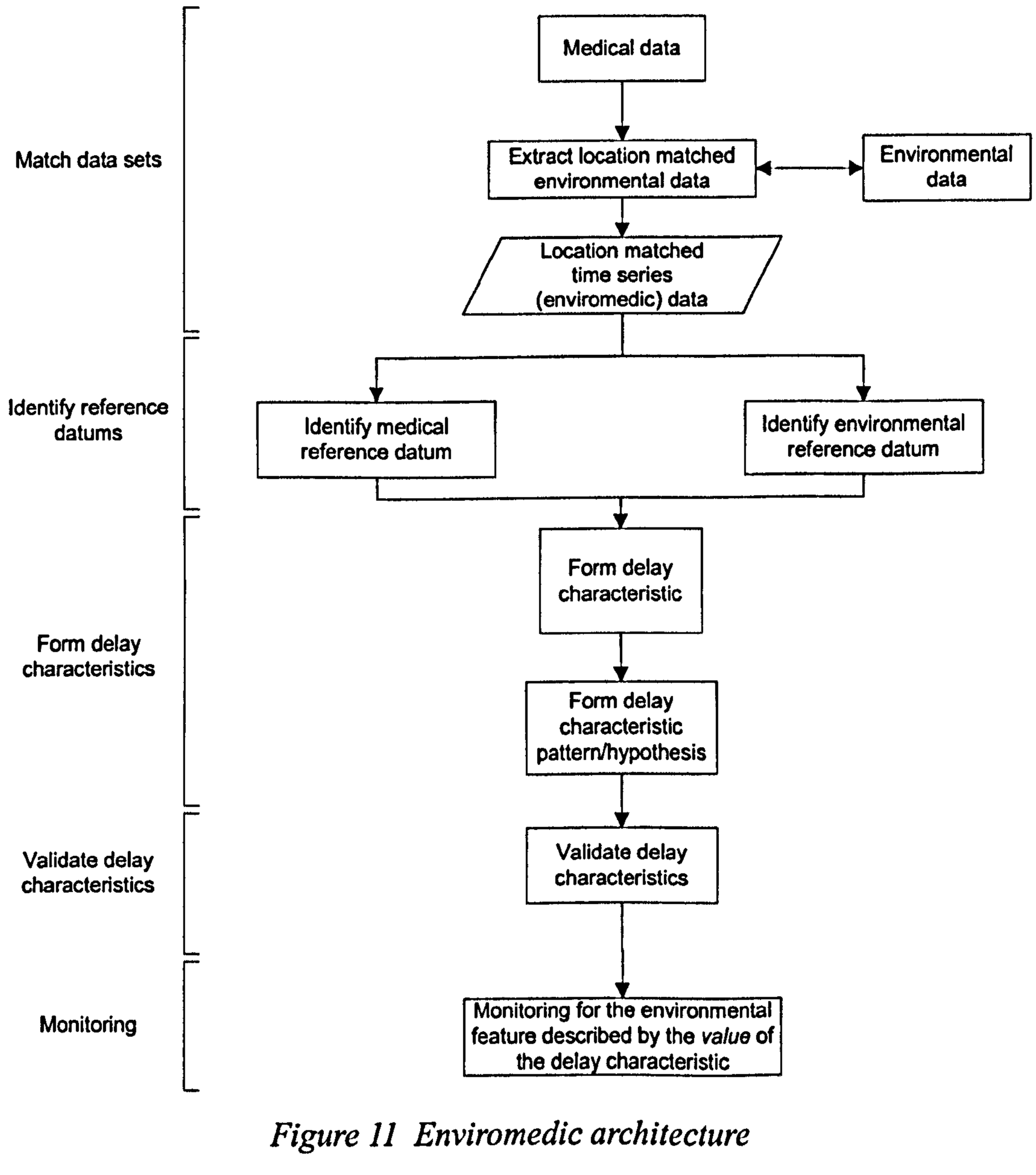


Figure 11 Enviromedic architecture

The extraction of medical and environmental data matched to the patient is the first process defined by the architecture. It is worth noting that matched data is not required at the same physical time or the exact physical location, but should be the best match possible. The second process identifies the reference datums in the medical and environmental data sets; the identification process for each data set is independent of the other. Once reference

datums have been identified from each data set, delay characteristics can be formed by extracting the time element between the two datums, and the value of the environmental datum. The next process outlined by the architecture combines the delay characteristics into sets, dependent on the hypothesis being tested. The sets can contain anything from a single delay characteristics, to a set of delay characteristics extracted from an environmental time series (discussed further in Section 5.4). The final sections of the architecture prescribe the use of validation, where delay characteristics are validated over time by experimentation. This process validates the characteristics that occur more frequently, and therefore those characteristics that can be used as reliable predictors of the *medical event*. The validated delay characteristics are then made available for monitoring purposes.

3.7 Summary

This chapter encompassed research into a number of issues associated with pattern identification between environmental and health related data. The research led to the construction of a number of steps which provided an outline for enviromedic analysis, as used by this thesis, the steps are:

1. Identification of the asthma episode.
2. Identification of the environmental predictor.
3. Monitoring for the environmental predictor.

The steps outlined a process, and relevant techniques were introduced to relate environmental data, specific to the health of individual patients. The work developed the concept of the ***reference datum***, marking trend reversals in analysed data sets, and the use of ***delay characteristics*** to relate pairs of reference datums. The method provides a way to relate two potentially related events in time and space, whilst not limiting the relationship to a direct correlation (in the traditional statistical sense).

Enviromedic analysis is used by this thesis to associate periods of patient-specific asthma exacerbation with validated environmental predictors, and is shown to be suitable for developing information for generating patient alerts, a concept which is developed over the next chapters.

Chapter 4

Environmental Monitoring System (EMS)

This chapter outlines the underlying architecture for a new system named the *Environmental Monitoring System (EMS).* The EMS uses environmental information to enhance respiratory healthcare, and is aimed at improving the quality of life for asthmatics through the identification of predictors which lead to asthma episodes.

4.1 Introduction

The EMS is developed primarily as a research tool to find relationships between data collected from electronic respiratory measuring devices, and environmental data from automatic air quality monitoring stations. The tool facilitates the identification of these relations through analysis using statistical and neural network techniques.

Key architectural components are identified, in particular the analytical requirements of the system. The architecture of the EMS is designed so that large data sets inherent in this field of research can be handled. System components forming the architecture of the Environmental Monitoring System (EMS) are described. A number of prototypes are developed to test the architecture during Chapter 6 (Results).

4.2 Development of a System Architecture

Turner *et al.* (1999) suggest that the structuring of a software system should be defined by its functionality from the users perspective, in the form of *use cases*. Where *use cases* are descriptions of a system's behaviour, when responding to a request that originates from outside the system. Toledano (2004) supports this view, but suggests that from an architectural point of view, not every use case can have the same importance. Aspects of

the architecture representing the basic functionality of the system, and reducing risk are more important. Fielding (2000) extends this hypothesis, by stating that the overall description of a system architecture must be capable of describing not only the operational behaviour of the architecture during each phase, but also the architecture of transitions between phases. Fielding uses the term *phase* in describing the possible need for several architectures for the same system; to fully describe phases of system operation. He states, "A system may be composed of many levels of abstraction and many phases of operation, each with its own software architecture". DeMarco (1995) supports this view, and states that "the ability to adapt to changing needs or a problem where the solution is unknown is paramount. The design of the system architecture needs to reflect the nature of the problem it is setting out to solve".

System architecture is generally described using perspectives called *styles* or *patterns*. Nitto and Rosenblum (1999) suggest that a given architecture may be composed from multiple styles, since architectural styles may address different aspects of a software's architecture. However, Shaw (1995) gives the opinion that some architectural styles are often portrayed as *silver bullet* solutions for all forms of software, and that a good designer should select a style that matches the needs of the particular problem being solved. In developing the system architecture of the EMS there are a number of generic aspects to consider. Avgeriou and Zdun (2005) specify the aspects as *modifiability*, *reusability*, and to a lesser extent *scalability*. Fielding (2000) introduces the term *modifiability*, as the ease with which a change can be made to an application architecture. An example of *modifiability* is the dynamic creation of class objects, through the use of XML property schemas and reflection. Objects created like this can be defined, and deployed within applications, without stopping and restarting the entire system. Modifiability can be further broken down into: evolvability, extensibility, customisability, configurability, and reusability (Fielding, 2000). Expanding the meaning of these terms:

- **Evolvability,** represents the degree to which a component implementation can be changed without negatively impacting other components.
- **Extensibility,** defined as the ability to add functionality to a system (Pountain and Szyperski, 1994)
- **Customisability,** refers to the ability to temporarily specialise the behaviour of an architectural element.

- **Configurability,** means for example, that components are capable of using a new service or data element type.
- **Reusability,** ability to be used in other parts of the system.

The development of the EMS adopts a process specified by Cheesman and Daniels (2001) where the use of component based software is promoted to develop flexible systems. The described process starts with a requirements description and produces an architecture showing the components to be developed, their interfaces, and their dependencies. For each interface operation, a specification is developed, consisting of a precondition, a postcondition, and additional information as required. The process does not consider the mapping of the developed specification to a model implementation. As a research tool, the EMS requires flexibility and the option to extend and modify its analytical capability. The approach taken by Cheesman and Daniels, and the adoption by the EMS of the approach, allows additional components to join and participate in the architecture as required.

The use of a component based approach to developing health information systems is promoted by Schlesinger *et al.* (1997) where they conclude that component architectures offer the potential to improve functionality, while simplifying the software development process. This point can be related to the EMS architecture by ensuring that distinct areas of functionality are segregated into components to allow their use by different parts of the system.

Tables 6 and 7 provide a summary of the key concepts and their reason for adoption into the EMS architecture. The concepts are covered by two views:

1. Elements that the architecture should incorporate.
2. Issues to consider in the design of the architecture.

Table 6 Elements that the EMS Architecture Should Incorporate

Architectural Element	***Function***	***Advantages***	***Disadvantages***	***Supported by***
Operational behaviour and transition between phases.	To abstract behaviour of each component, and describe evolution over time.	Promotes a system able to adopt to new tasks, with extensibility, and re-usability.	May use a component not specifically designed for a task	Fielding (2000) Garlan (2003) DeMarco (1995)
Modifiability, re-usability, and scalability	Saving of implementation effort, cost and time.	Flexibility, cost-effective.	Design can lead to a general solution.	Avgeriou and Zdun (2005) Fielding (2000)
Extensibility	To allow future extension of the system.	Not limited to providing the current functionality.	May extend a sub-standard component.	Pountain and Szyperski (1994)
Well defined interfaces	Structures principle components and interfaces for communication.	Can lead to extensibility, re-use, and scalability.	Badly designed interfaces can lead to operational restrictions.	Souquieres and Heisel (2004) Garlan (2003)
Interoperability between components	To promote common methods of communication between components.	System functionality can be modified.	Can limit communication of information.	Souquieres and Heisel (2004)

Table 7 Summary of Issues to Consider

Function	***Description***	***Advantages***	***Disadvantages***	***Supported by***
Use cases	To ensure development of a product meeting the end user's requirements.	The end product will meet the end users expectations.	Methods used to provide the result may not always be the best.	Toledano (2004)
Design constraints	Constraints implied by the architectural style. Specifying what should not change over time.	Core architecture remains consistent and designed for the purpose.	Could lead to an obsolete system.	Garlan (2003)
Implementation of multiple styles	Development of an architecture that aids the design of a suitable system.	Can pick style attributes that best fit the requirements of the EMS.	The architecture could become incoherent.	Nitto and Rosenblum (1999)
Structuring of the software system	Design of the system architecture by considering Its intended functionality.	The system's purpose is clearly defined, and developed components specified succinct.	Possibility for over engineering.	Turner *et al.* (1999)

The architecture must allow the flexible addition of new data types for analysis by the system's analytical components. This is because the precise parameters to monitor, and techniques to use in analysis are unknown, and will be subject to change as new knowledge

is gained of the patient's condition.

The use of the EMS as a research tool provides useful constraints for the architecture of the system. Used to identify and explore relationships between a patient's environment and their respiratory health, the interchange of analytical components must also be allowed by the architecture of the EMS.

4.2.1 Architectural Patterns

Avgeriou and Zdun (2005) discuss two different views that are expressed in literature. They observe that the term architectural pattern, and architectural style both refer to recurring solutions that solve problems at the architectural design level. They also observe that they have key differences. Patterns not only document how, but why, while styles describe components, connectors, and issues related to control and data flow.

Garlan and Shaw (1994) outline a number of architectural styles and show how they can be applied and adopted to specific software systems. Many of the styles are listed in *Appendix C*. Buschmann *et al.* (1996) also collate a number of architectural patterns that express a structural organisation for software systems. They say that architectural patterns provide a set of predefined subsystems that specify their responsibilities, and include rules and guidelines for organising relationships between them.

Validation of the EMS is dependent on implementation of the architecture, therefore consideration is required as to how the architectural implementation could be achieved. For architectures to be built in practice at a component level, design patterns are required to structure application functionality. The EMS implementation uses a mix of *Whole-Part,* and *Publisher-Subscriber* design patterns (expanded in *Appendix D*) to achieve a working system prototype.

Validation is shown through the architectural implementations described in Chapter 5, and subsequent results, in Chapter 6.

4.2.2 System Specification for the EMS

System specification provides a high level overview of the EMS describing: inputs, outputs, and what is required by the user. Four major aspects were considered as a framework during the design of the EMS architecture:

1. **User interface requirements;** to obtain settings to facilitate the running of the underlying system. The minimum sub-set of information required to run the EMS includes, a patient identifier, and lung function *deterioration* threshold. Ability to identify environmental conditions acting as predictors of patient asthma exacerbation.

2. **Output of the system;** alerts a member of clinical staff and/or patient to the onset of an asthma attack, based on the previous identification of environmental triggers and how they affect the patient.

   A research clinician and their patient require notification as to when the patient is likely to experience the onset of an asthma attack. The aim is to avoid adverse environmental conditions, or indicate when medication should be increased. Asthmatics are *hypersensitive* to varying *stimuli*. Early in diagnosis, most patients and clinical staff are unaware of which stimuli may affect the patient. Historically, clinical staff have been unaware of the patients' condition when not under their direct care. Patients also require pertinent information about their condition to aid the management of their condition. Therefore a system capable of delivering timely alerts to both patient and clinical staff with impending adverse environmental conditions is required to support the decision making process.

3. **Processes between the input and output** stages of the system, to enable the desired functionality, including the predictor identification process, and component to match real time data with the stored predictors. Taking into consideration performance, usability, recycling, economic and technological restrictions.

4. **To ensure appropriate alerts are triggered** when environmental conditions are experienced by a patient, first the affecting environmental condition must be identified

and verified. Personal lung function monitoring is required to provide personal patient data for the system, and used to detect a decline in respiratory condition. Data is required from environmental monitoring equipment local to the patient so their personal environment can be analysed for repeated triggers of asthma exacerbation.

The key information required in an alert created by the EMS, is the characteristic of the identified environmental condition, capable of acting as a verifiable predictor of respiratory decline. Therefore, the delay between the environmental condition and the expected decline in respiratory condition and the environmental parameters value should be given as information by the system.

Patterns of environmental conditions verifiable as leading to a decline in lung function, require recording in a form they are easily accessible. The verified patterns would then be used to provide alerts when the conditions are recognised in real-time.

Toledano (2004) suggests the use of the following framework (Table 8) to help detail an architecture that can be used to build an architectural implementation:

Table 8 Framework for Architecture Development

Framework Characteristic	***Example***
Basic characteristics	What defines and differentiates the architecture. For example: "the system is characterised by..."
Definition of the main actors	Definition of the main actors that participate in the system, as well as the basic use cases.
Main functional components	Specification of the main functional components of the system, as well as the relationships between them.
Logic architecture	Logic architecture and information flows, the exchange of their information, and so on.
Component architecture	Component architecture. This consists of mapping the functional components in the logic architecture of the application.
Physical architecture	Physical architecture. Specification of the deployment of the components.

Basic characteristics are described throughout this thesis, especially in earlier chapters, as are the *definitions of the main actors*, and the *main functional components*. The first part of this chapter considers aspects of the *component architecture*, while the later sections cover the areas of *logic* and *physical architecture*.

The aim is to provide a framework to facilitate the creation of relevant patient alerts when an environmental condition identified as affecting patient health is recognised. For this to occur, several distinct elements are required. Data storage, is necessary to act as a buffer between the real-time raw data and the identification of reference datums and delay characteristics by the *pattern identification* component. The pattern recognition component then requires data storage to match known patterns and create alerts. This is shown in Figure 12, where the fundamental roles of *Pattern Identification*, and *Pattern Recognition* are central to the operation of the system. The arrows in the figure represent the transfer of information, from raw sensor signals (delivered in real-time) on the left, to a patient specific alert on the right, produced when environmental *predictors* are first validated by the system, and then recognised in *real-time*.

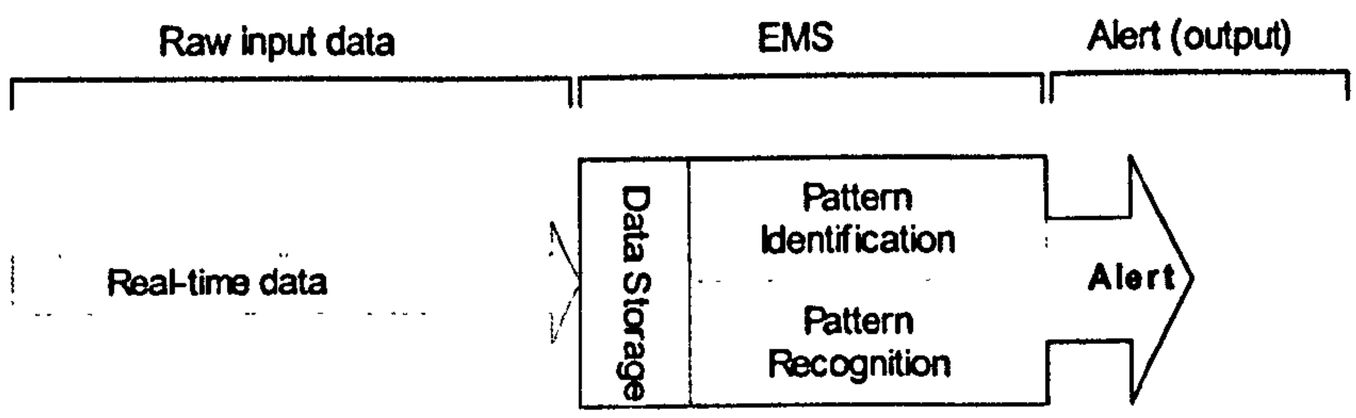


Figure 12 Basic Architecture of the EMS

4.2.3 Data Process Architecture

From the users perspective, the EMS can be broken into a number of high level subsystems consisting of: a *data interface* primarily focused on the autonomous collection of data, *data storage*, *data handling*, where the data is manipulated in some way to provide useful information, and *data dissemination*, where the useful information is finally presented to the user.

An architectural view showing the four distinct data processing areas is shown in

Figure 13. This chapter, and Chapter 5 concentrate on the work needed to fulfill the requirements for the *Data Handling* process, with consideration given to the surrounding processing layers.

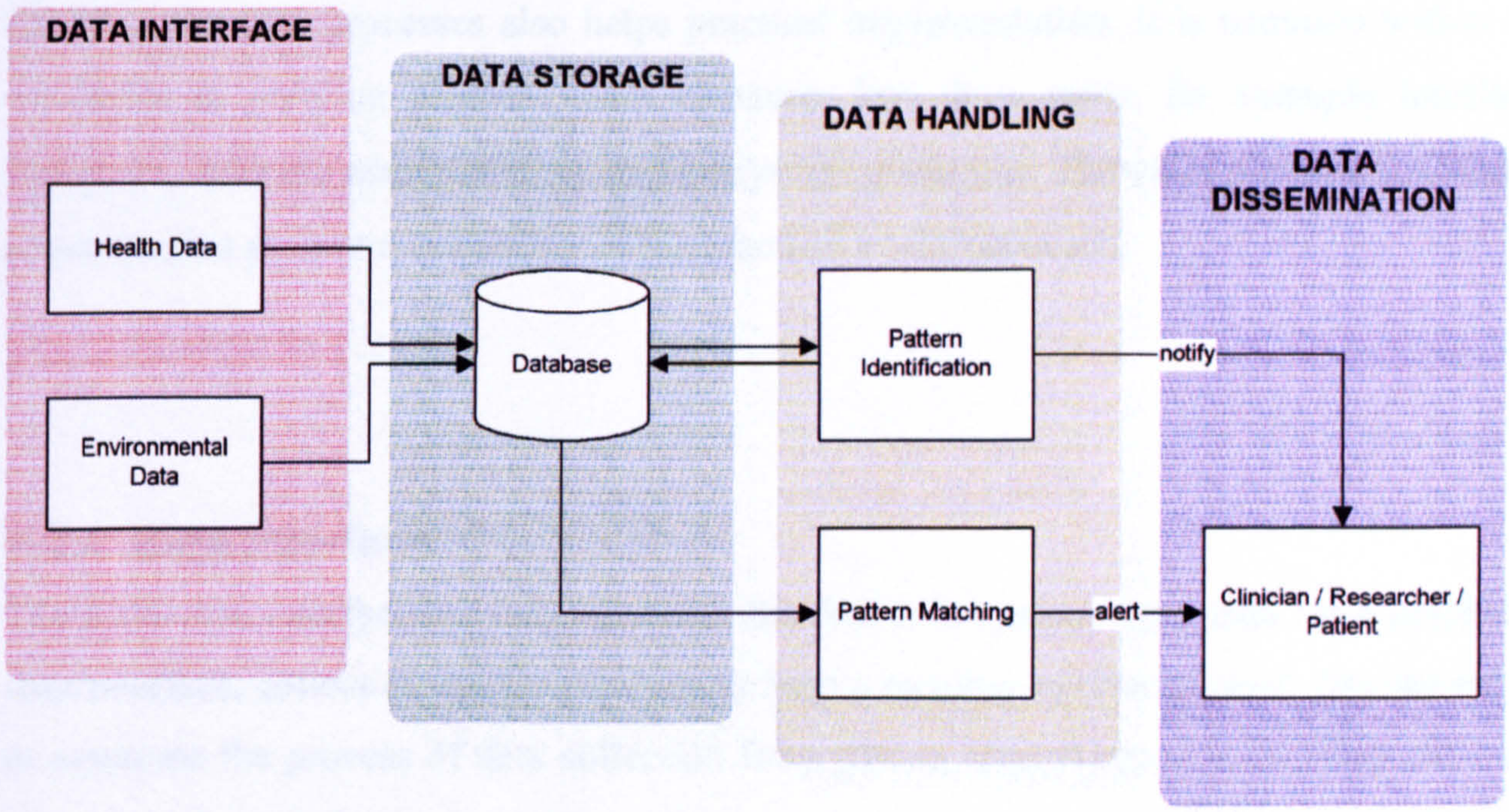


Figure 13 An architectural view of the data processing layers.

The processes described by the architecture manipulate raw input data into a form that is suitable for data mining, eventually allowing the triggering of alerts. The EMS performs the following varying functions:

- *Data Storage and pre-processing* – a generic means to store data before analysis.
- *Pattern Identification* – analysing trends in the data.
- *Validation/Realisation* – sorting and marking interesting and repeating relationships.
- *Recognition* – matching an incoming pattern against a previously found predictor.
- *Reaction* – of the system to the potential predictor pattern.

Partitioning the data processes into layers provides a degree of separation from any evolving technology used in any implemented part of the architecture (Avgeriou and Zdun, 2005). Separation extends the architecture's scalability, as each domain becomes self-contained and can operate independently. Processes can be staged across multiple servers if need dictates. The architecture also allows for the use of *sharding* data across multiple

databases, as each patient's analysis can be stored and accessed individually. Sharding is a process whereby system data is physically segregated to enable scaling of very large databases.

Dividing the data processes also helps practical implementation. It is common within an enterprise to partition physical team structures into these areas, for example interface designers, database administration, and analyst programmers. Therefore separation of these concerns also promotes scalability of the practical implementation.

4.2.4 Data Interface

The EMS data interface has two functions, the first is to receive input data via a graphical user interface, collecting system parameters from a member of clinical staff. The second is to automate the process of data collection from patient lung function monitoring devices, and air quality monitoring stations.

The Medicate (2000) project used a number of interface options (allowing the choice and control of the analysis). The options extended during research for this thesis include:

- The period of lag time in which to analyse, between lung function decline and possible environmental predictor, and
- Trend information relevant to the patient, set by clinical staff.

The EMS was designed with the capability of monitoring in *real-time,* for predictors using these options, satisfying an architectural requirement. With the addition of a patient identifier (complying with the need to keep patient details anonymous), identification of emerging environmental predictors of the patient's asthma exacerbations is made possible.

The EMS also has the capability of using common standards, for example XML (eXtensible Markup Language) (Bray *et al.*, 2006) to collect data. XML allows the structured transfer of data, and is able to ensure integrity before manipulation by other parts of the EMS architecture. However, although the capability was implemented, a comma delimited file format was adopted within the prototypes as this was the most common format used in both clinical and environmental data sets.

4.2.5 Sources of Data

The first function of the EMS is to collect the relevant data from environmental and clinical sources. This requires flexible interfaces to the EMS. There is no one standard for the transfer of air quality or lung function data. Comma delimited files used by the London Air Quality Network (LAQN, 2007), and shown in Figure 14 are typical of the de-facto standard for data storage. The aim for an EMS is to relate the location of the patient with the available environmental data.

Air quality data is disseminated across the Internet, either via direct feeds from monitoring stations via the Air Quality Monitoring Network (LAQN, 2007), or from standard web pages updated by the hour.

The structure of the raw lung function data is defined as follows (in Figure 14).

```
PatientID,     Date/Time     ,  PEF , FEV1 ,  FVC , FEF25, FEF50, FEF75,FEF2575,Type
    1     ,26/11/99 11:50:00,9.8333,3.5160,3.5160,9.4689,6.9019,0.0509,7.4879 ,0
    1     ,26/11/99 11:50:00,6.5999,4.1279,4.1279,6.7319,5.7800,3.8759,5.7280 ,1
    2     ,26/11/99 08:03:00,6.0833,2.3099,3.1559,5.1680,2.3629,0.5780,1.6799 ,0
    2     ,26/11/99 19:14:00,5.3666,1.8910,2.5759,4.6750,1.6319,0.4760,1.2799 ,0
    2     ,27/11/99 09:27:00,6.0333,2.4470,3.1549,5.6609,2.8389,0.8330,2.2880 ,0
    2     ,27/11/99 19:12:00,5.1666,1.9129,2.6319,4.2160,1.5640,0.5270,1.3120 ,0
    2     ,28/11/99 09:11:00,5.2833,2.0109,2.6210,4.7940,1.7680,0.6290,1.5520 ,0
    2     ,28/11/99 19:39:00,5.2833,1.8029,2.4930,3.6210,1.4620,0.4930,1.1679 ,0
    2     ,29/11/99 07:31:00,5.8166,2.1259,3.4119,4.0970,1.2580,0.2549,0.7200 ,0
    2     ,29/11/99 19:25:00,5.3666,1.9759,2.6540,4.5050,1.7000,0.6290,1.5679 ,0
    2     ,30/11/99 08:37:00,5.4666,1.9259,2.6159,4.3860,1.5809,0.5950,1.3919 ,0
    2     ,30/11/99 19:48:00,4.6833,1.9270,2.4519,4.4200,1.8190,0.5950,1.5679 ,0
    4     ,03/02/00 15:33:00,3.6833,1.6590,1.8439,2.5840,1.7000,1.1900,1.8400 ,0
    4     ,03/02/00 16:53:00,3.0000,1.5349,1.9309,1.7510,1.3940,0.9179,1.3919 ,0
    4     ,03/02/00 16:54:00,2.9166,1.4210,1.8459,1.6490,1.2580,0.8159,1.2799 ,1
    4     ,03/02/00 17:28:00,2.3833,1.8170,2.3410,1.2070,1.5299,1.1219,1.4880 ,0
```

Figure 14 Example of a raw Lung Function data file (Medicate, 2000).

Typically lung function data is organised in a matrix format where a data set is represented by a row and data type or value by columns. Each row contains a value which identifies a particular patient, a date/time stamp which gives the exact date and the time a data set was recorded (to the nearest minute), and then the data values including PEF (Peak Expiratory Flow), FEV_1 (Forced Expiratory Volume in 1 second), FVC (Forced Vital Capacity), $FEF_{25\%}$ (Forced Expiratory Flow at 25 percent), $FEF_{50\%}$ (Forced Expiratory Flow at 50 percent), $FEF_{75\%}$ (Forced Expiratory Flow at 75 percent), $FEF_{25\text{-}75\%}$ (Mean Forced Expiratory Flow). This particular lung function data set does not contain information about the location of readings or patient attributes such as weight and height; these are contained in other data sets. It is not common at this time for monitoring devices to incorporate a

GPS receiver to record location.

Neither format of data (air quality or lung function) is particularly suitable for facilitating access to large sets of information at optimum speeds in the raw form. An implementation of the architecture would require a database to support efficient querying of information.

4.2.6 Data Handling

A core decision algorithm capable of identifying the time at which a patient might experience the onset of an asthma attack was required. During the Medicate (2000) project, a simple level trigger had been applied such that if a patient's lung function fell below a predefined level, it would trigger an alert to their clinician. While these types of alerts proved useful in alerting clinical staff to a potential problem it was recognised that a patient's lung function usually started to decline some time before a fixed threshold was reached. It was therefore hypothesised that the onset of an asthma attack began, once the patients' lung function signal had reached its highest level. The process created and used by the EMS for identifying the change in direction of lung function trend has been named *Feature Detection Analysis* (FDA) and is presented during Section 5.2.

A technique was also required to select the types of environmental parameters included in the enviromedic analysis. As there is variability in asthma attack exacerbants from patient to patient, it is important for the EMS to adapt to these changes and use a mechanism for choosing if the monitored parameter actually has any effect or not. This is a complex task and is not covered explicitly within this thesis, although the clustering technique shown in Section 5.6 does provide a useful method for determining if a parameter is relevant to a particular patient's condition.

Another characteristic of the decision algorithm is to identify relationships that exist between changes in the environment and the asthma patient's period of declining lung function. The important aspect of this characteristic from the patient's point of view is: *how long do I have, before I experience discomfort?* Clearly the setting of lung function threshold levels, close, but prior to the level at which an individual patient experiences discomfort, are critical to the functioning of the EMS, and must be decided by discussion between clinical staff and the patient. Setting these levels is a type of prediction (from past

knowledge of influences), more importantly it is influenced by the recognition of the time interval between the environmental predictor and the asthma attack trigger-point.

These requirements form the basis of this research and guide the prototyping of an early warning system for asthmatics. For an asthma early warning system to be effective, it must possess the following features:

a) It must be patient-specific.

b) It must be capable of relating environmental data, gathered from fixed location or ambulatory monitoring systems, to patients who are mobile.

c) It must be capable of identifying significant and repeatable environmental events, which take place at a time sufficiently in advance of an asthma attack (deterioration in lung function) to allow corrective action to be taken.

4.3 Summary of Identification Architecture

The previous five sections have outlined the fundamental processes required for the identification of environmental predictors indicating a future decline in respiratory health. The figure below summarises the processes.

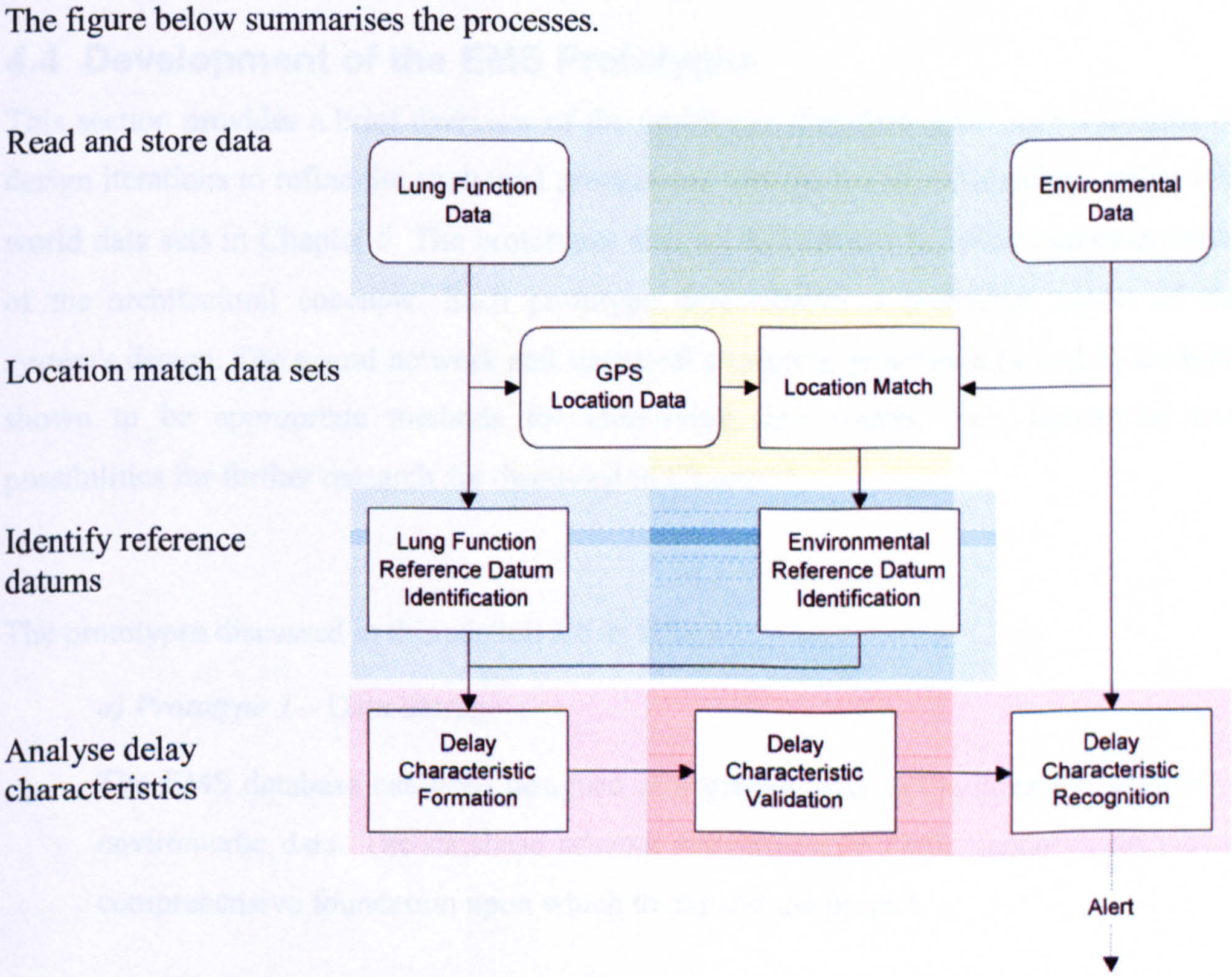


Figure 15 Summary of the Identification Architecture

The identification architecture (Figure 15) shows the process for determining an environmental predictor of lung function decline. The process begins with the collation of data from lung function devices, with a location reading. It is worth noting that there are no lung function devices currently available capable of measuring location (using a GPS receiver), so an additional device would be necessary to facilitate this measurement. The tool most capable of achieving the result of location based measurement is currently based on a lung function device that connects to a mobile phone.

Location is used to identify the appropriate environmental readings, represented by the *Location match data sets* process. Once both lung function and environmental data sets match the movements of a particular patient, analysis can begin with the identification of reference datums (from each data set). Following this stage, the pattern validation, and recognition process identify and then create an alert of impending lung function decline from the change in environmental condition.

4.4 Development of the EMS Prototypes

This section provides a brief overview of the prototypes that were used over a number of design iterations to refine the analytical process and test the thesis methodology using real world data sets in Chapter 6. The prototypes also act to illustrate possible implementation of the architectural concepts. Each prototype demonstrates a particular aspect of the system's design. The neural network and statistical clustering prototypes (4 and 5) are both shown to be appropriate methods for identifying data trends, their limitations and possibilities for further research are discussed in Chapter 7.

The prototypes discussed in this section are as follows:

*a) **Prototype 1*** – Data Storage

The EMS database has been designed in a generic way to facilitate the storage of enviromedic data. The database schema and prototype implementation provide a comprehensive foundation upon which to expand the project.

b) Prototype 2 – Feature Detection Analysis

A proprietary component capable of identifying significant peaks and trough data points, defined by the user.

c) Prototype 3 – Input Data Modifier & Organiser (Hypothesis Builder)

A prototype to arrange data into an appropriate format (vectorised) for analysis.

d) Prototype 4 – Statistical Clustering (FBCA)

Demonstrates a proprietary technique to achieve a frequency analysis of the input data and the evaluation of existing clusters.

e) Prototype 5 – Neural Network (SOM)

A neural network technique driven by the Self Organising Map (SOM) algorithm.

f) Prototype 6 – Overall Demonstrator

The five prototype modules were combined into one overall demonstrator to demonstrate the work flow from one end of the system to the other. The results presented during Chapter 6 use this prototype.

The following sections give an overview of the six prototypes and presents issues encountered.

4.4.1 Data Storage Implementation (*Prototype 1)*

The EMS prototype implementation of the database made use of *FastObjects* now owned by the Versant Corporation (Versant, 2008). FastObjects J1 (the database implementation used) is a 100% pure Java implementation of the JDO (Java Data Objects) (Roos, 2003) specification and an evaluation copy of the professional version of the software. Java has some advantages over other types of programing language, specifically portability between

operating systems and handling of garbage. A good evaluation of JDO is given by Srdanovic *et al.* (2005).

The use of an object oriented database allowed the object model of the system to be directly mapped to persistent storage. Integrating the system with the database was simplified as objects did not require translation into tables (like relational databases). The implementation database complied to the JDO specification, meaning that interfaces could be used to communicate with the database, providing separation so the type of database used for storage could be interchanged easily. Initially PSEPro Java edition (from ObjectStore) was used as the implementation database. The use of database indexes made querying more efficient, and the ability to internally filter queries was also available, further reducing the time taken to return a data set.

An example of the database indexes used can be found in *Appendix Q*. It was thought initially that it would be sufficient to index on data *type*, for example retrieving records associated with particulate matter. However, it became obvious that to search the records efficiently, additional indexes would have to be created, such as time and location in order to make the tracking of patients possible.

JDO supplies two methods for querying a database, a query interface which includes basic methods for extracting records from a database, and support for the Java Data Objects Query Language (JDOQL). It was found that computational speed and memory constraints were an issue with early prototypes. This problem was overcome by returning the ObjectID of records meeting the query criteria before extracting the associated records.

Initially the data sets used for testing the prototypes typically covered in the region of 5000 records extracted from a database with close to ½ million air quality and lung function recordings. These initial test data sets were then increased in size and databases created that contained the necessary data for each test, to speed the testing process.

4.4.2 Feature Detection Analysis (Prototype 2)

Feature Detection Analysis (FDA) is explained fully during Section 5.2. The core analytical components of the FDA prototype are the Pearson correlation, and regression algorithms. Analysis using FDA identifies a set of peak and trough points depending on the coefficient of determination. The prototype had the facility to control the algorithm, eliminating points that did not meet the requirement of the descent or ascent gradient of the regression line, or fell below a predefined threshold value. Figure 16 shows the user interface of the prototype.

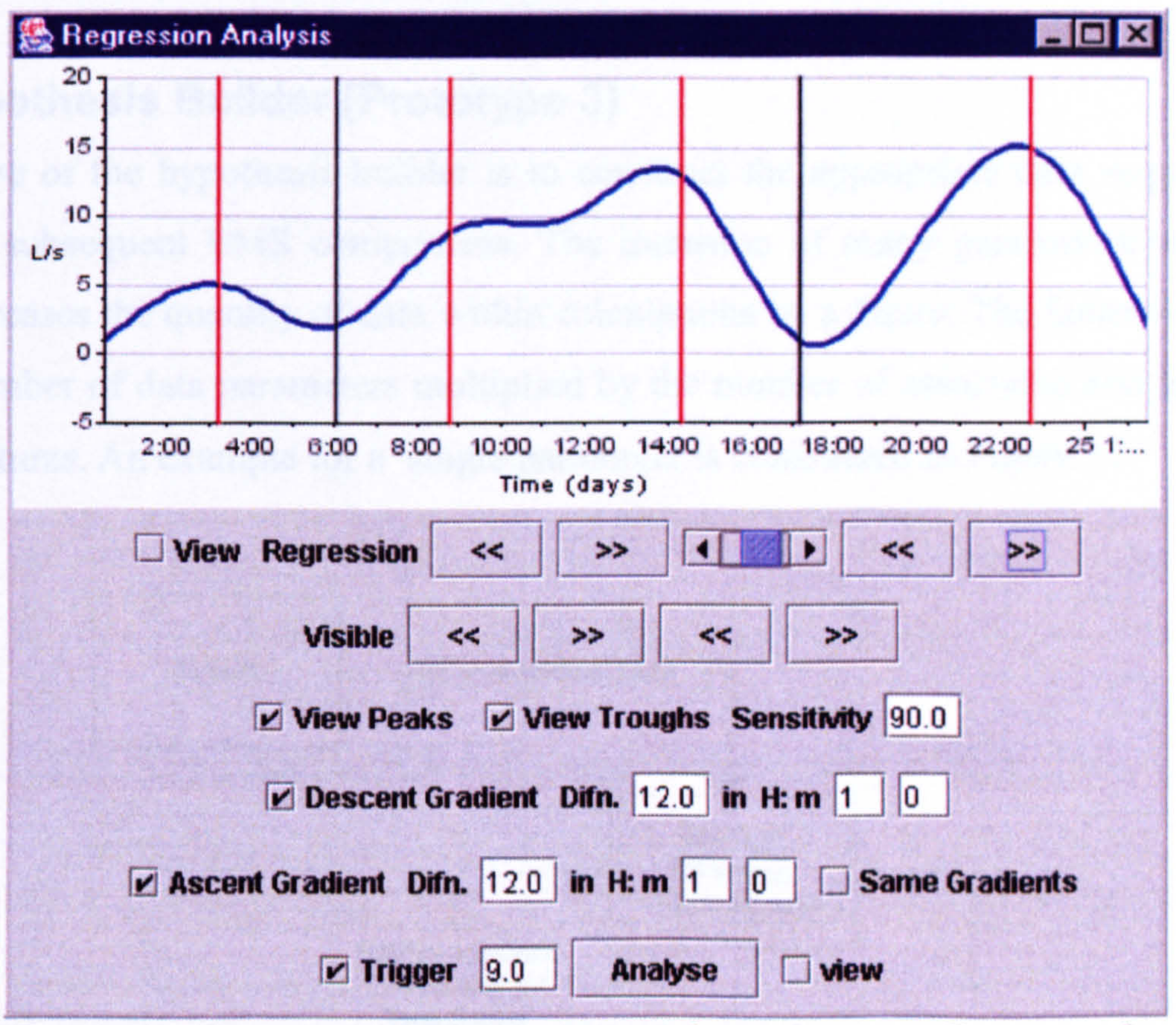


Figure 16 Prototype 2 (FDA) user interface, showing all options the prototype presented to the user.

The prototype user interface was iteratively produced during testing. There are two major functions:

1) Control over the visualisation aspect of the prototype.

2) Control of the analysis.

The visualisation aspect of the prototype, starting at the top left of the grey area (Figure 16) with the *View Regression* check box, and ending with the *View Troughs* check box, allows the user to see the regression line through any section of the data series, how much of the

x-axis is visible and if the identified peak/trough reference datums are displayed on the chart.

The analytical section starts with the *Sensitivity* setting, where the percentage corresponding to the coefficient of determination can be set. The gradient (descent/ascent) that the regression line must satisfy for a reference datum to be valid is displayed next on the interface. The last option given to the user is to set a threshold level, this option disregards all identified points if the regression trend does not fall below this value.

4.4.3 Hypothesis Builder (Prototype 3)

The objective of the hypothesis builder is to construct the appropriate data sequence for analysis by subsequent EMS components. The inclusion of many parameters within an analysis increases the quantity of data within calculations by a *factor*. The factor is derived from the number of data parameters multiplied by the number of associated and identified reference datums. An example for a single parameter is considered in Figure 17.

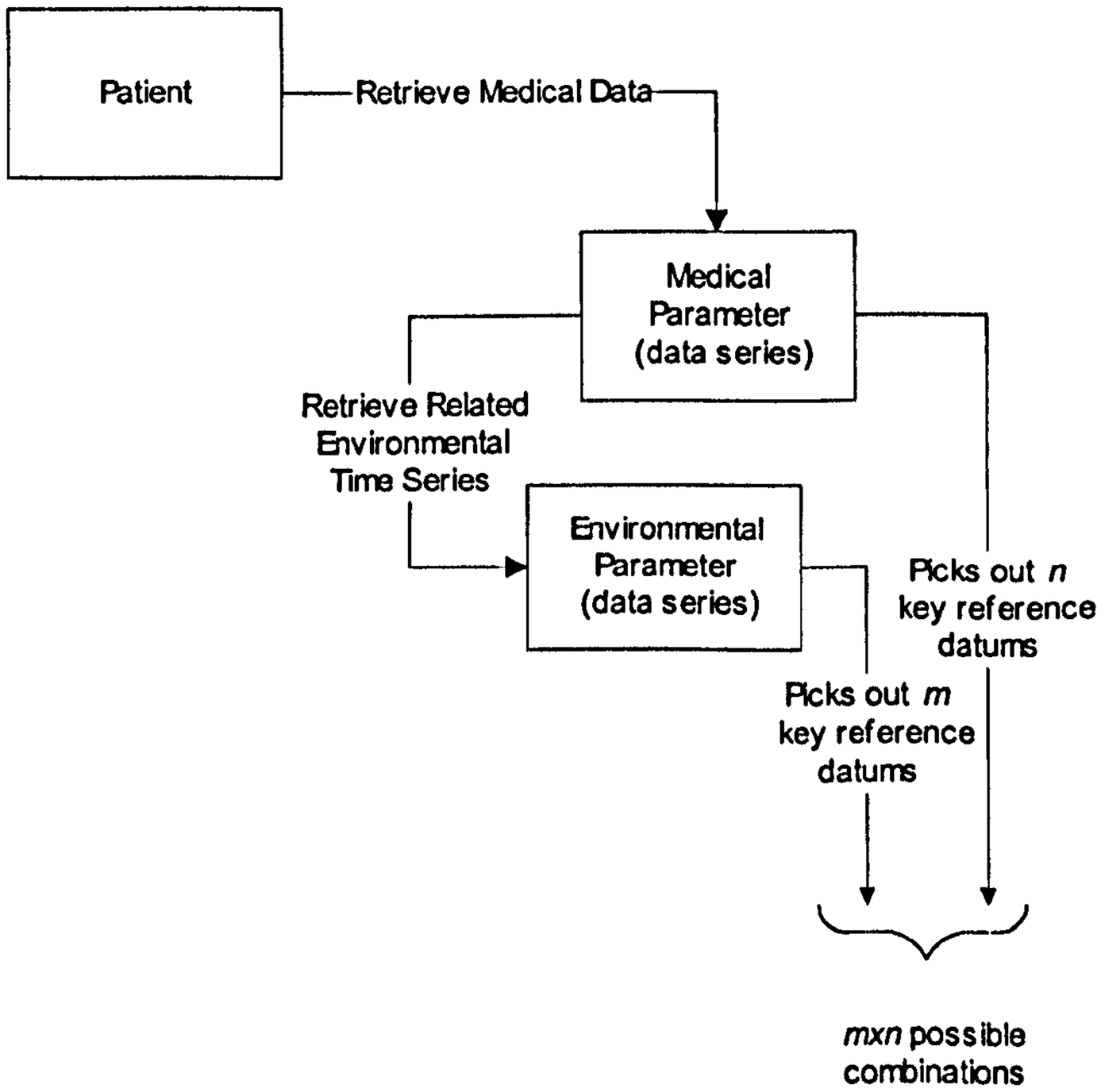


Figure 17 Handling a single parameter from each enviromedic data set.

As the number of parameters are increased, the number of possible combinations also increases, as shown by Figure 18.

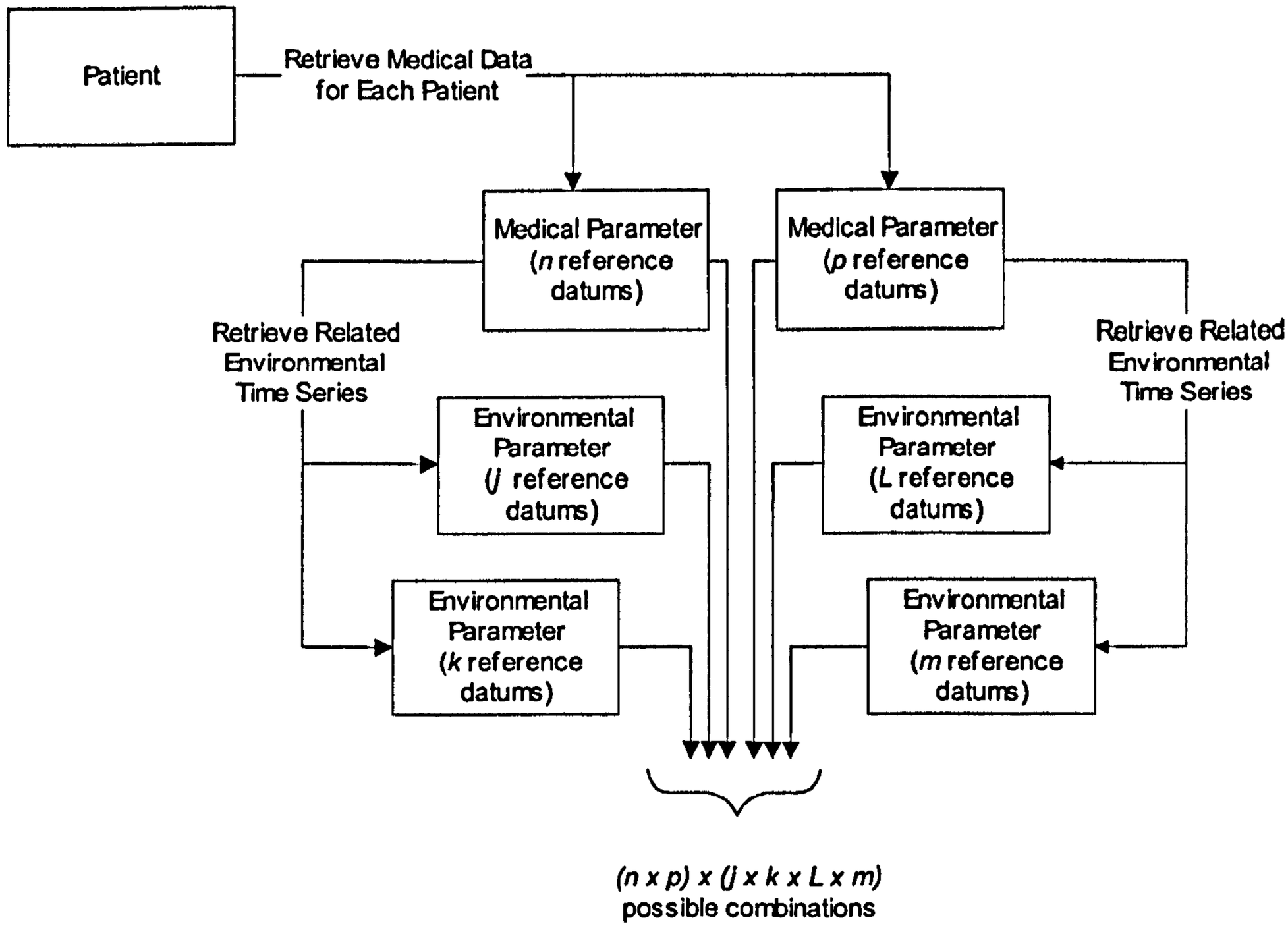


Figure 18 Handling multiple parameters.

Additional parameters such as patient selection, medical, environmental and time associated, adds significant complexity to the process. The need for computational memory increases as the number of parameters grow.

Three types of analysis are discussed during the *Hypothesis Builder* section of the next chapter. They are *Point*, *Series of Points*, and *Series* analyses. Emphasis should be given to the need for consistency in preparation of the input pattern; each vector's dimensions must contain the same data type. Comparison is made between each vector by comparing the difference between each vector dimension. Examples of vector construction are provided in *Appendix J.*

Additional functionality within the Hypothesis Builder allows the user to select the parameters that will be included by the analysis (*Include Options* area of Figure 19). The

included parameters are shown in the *Vector Parameter Order* area of the interface; the order of the selected parameters can be re-arranged by dragging and dropping the *tags*. For example, if SO_2 – *Lag* was required before SO_2 – *Value*, it could be dragged to the earlier position in the vector, which the interface facilitates. This functionality becomes more useful when a large number of multiple parameters are analysed.

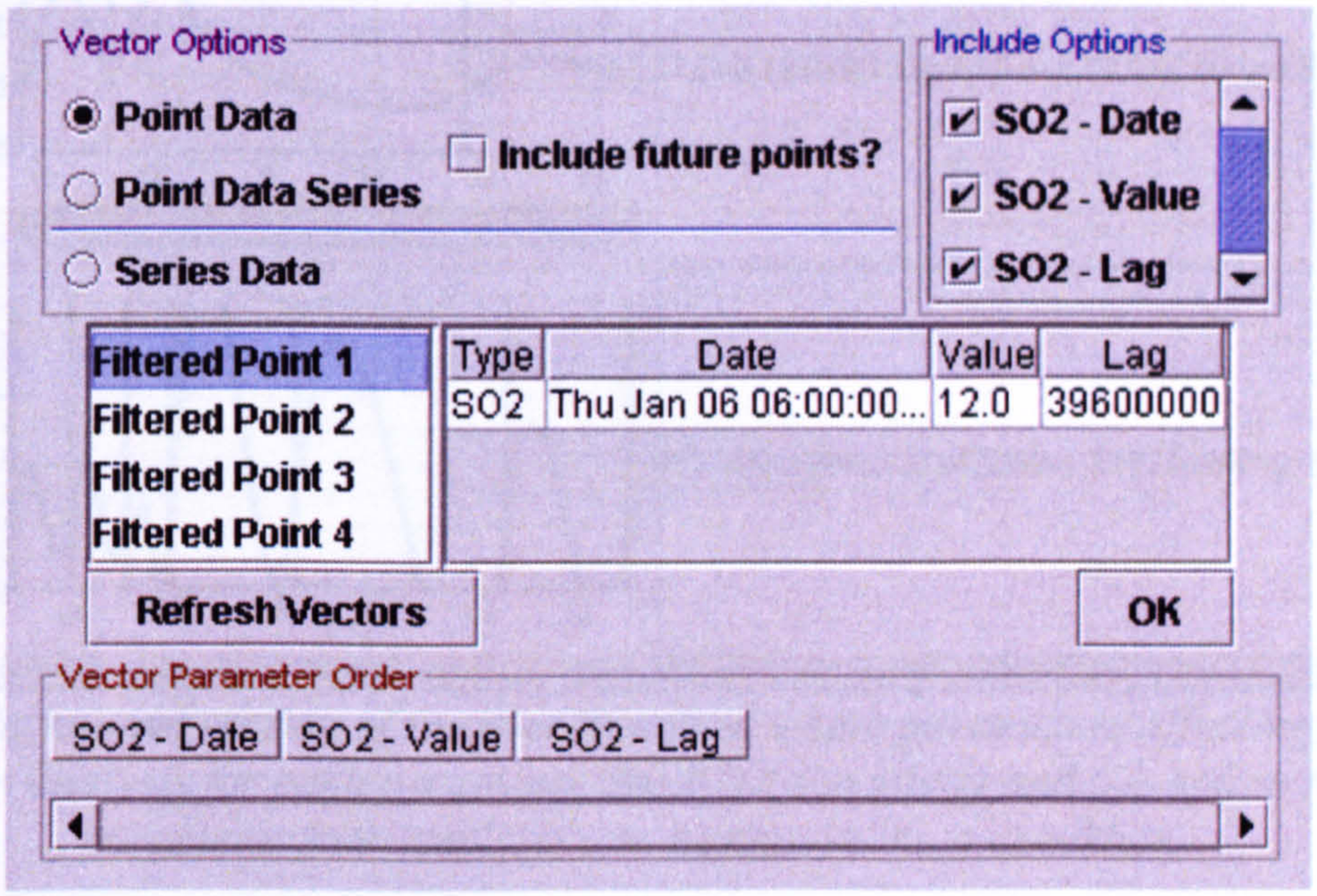


Figure 19 The hypothesis builder, including the area 'Vector Parameter Order' where the dimensions for each of the vectors can be ordered.

The component extracts the lag between the environmental and medical data sets, the value of the environmental parameter and date of reading, shown in Figure 19 as a 'Filtered Point'.

4.4.4 Statistical Clustering - FBCA (Prototype 4)

The statistical clustering prototype combines *Frequency*, *Boundary* and *Cluster Analysis* (FBCA) into one prototype (explained further in Section 5.6). The prototype provides a user interface where parameters controlling the size of the buckets used during *Frequency Analysis* are set. This could be automated through analysis of the signals content type; a topic for further research. Bucket sizes for *Lag* and *Date* are denoted in milliseconds.

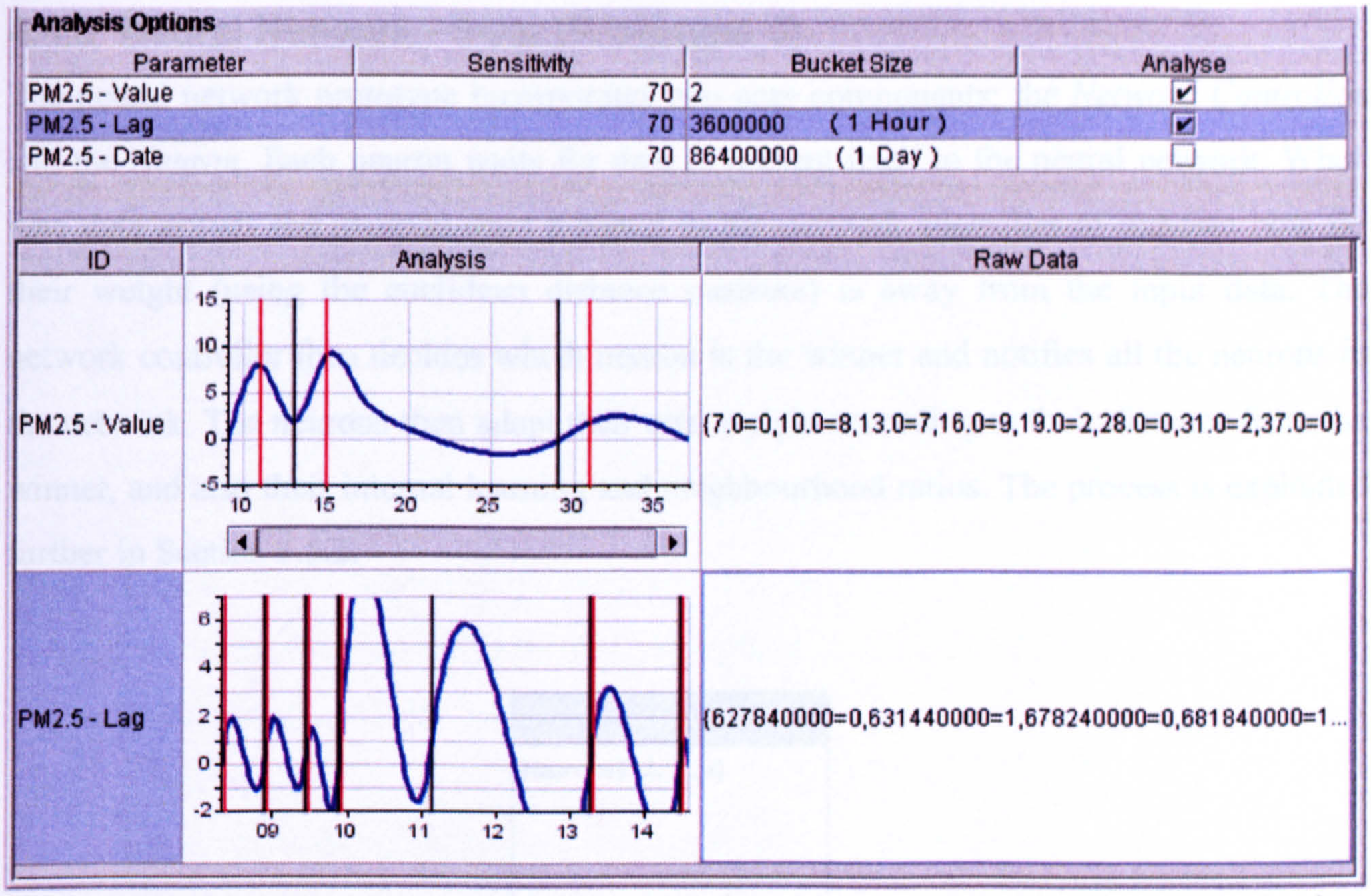


Figure 20 The control section of the user interface where parameters affecting the bucket size used in analysis for each parameter, the FDA sensitivity and the option to omit a parameter from analysis can be chosen via a checkbox.

Once the *Frequency Analysis* has been achieved by recording a tally of data falling into each bucket, *Boundary Analysis* utilises the FDA component in locating the distribution boundaries. Distributions normally cover more than one bucket and are not often unimodal. FDA provides a method through trough detection to define a point at which one distribution ends and another begins. Shown in Section 5.6.2.

Once the clusters are ready to receive input data the process of identifying which clusters are active can begin. The structure to record the cluster boundaries within the FBCA component, and monitor for matching patterns is shown in *Appendix I.*

As the cluster ranges are known, along with the likelihood (compared with the other identified clusters) of an input pattern matching a particular cluster, a probability indication can be given. This is achieved through the division of the number of hits a particular cluster has received by the total number of hits received by all clusters.

4.4.5 Neural Network - SOM (Prototype 5)

The neural network prototype incorporates two core components; the *Network Controller*, and the *Neuron*. Each neuron waits for data to present itself to the neural network. When new data arrives the neurons send a signal to the network controller to indicate how far their weight (using the euclidean distance measure) is away from the input data. The network controller then decides which neuron is the winner and notifies all the neurons in the network. The neurons then adapt their own weight according to their distance from the winner, and also their internal learning and neighbourhood ratios. The process is explained further in Section 5.5.2.

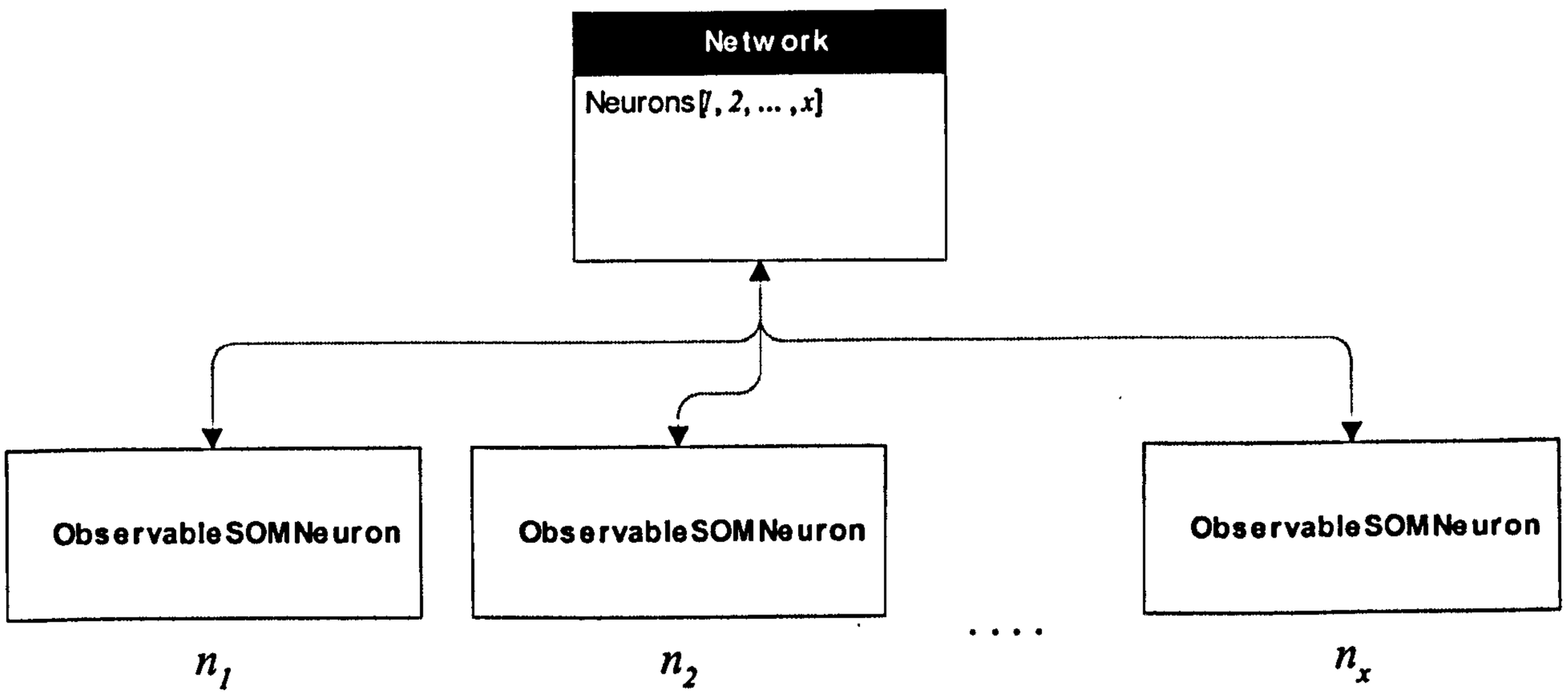


Figure 21 High level structure of the neural network. Showing x number of neurons (indicated by ObservableSOMNeuron), and Network that keeps a record of, and controls the number of neurons in the network.

Additional classes were written to control the behaviour of the network such as neuron splitting and visualisation functions. Communication between the *Network Controller* and *Neurons*, and components from outside the neural network use an *Observer* design pattern. This gives good separation of control between the components, making the network flexible. The use of the observer design pattern exchanges events between components for control and gives each neuron an amount of autonomy that could be used to advantage in scaling the network to a larger number of neurons.

4.4.6 EMS Architecture - Overall Demonstrator (Prototype 6)

Workflow coordinator

The analytical process begins within the *workflow coordinator* where general parameters outlining the analysis are first set. The graphical interface used for prototyping the EMS is shown in Figure 22 below. The figure shows the 'setup' panel on which the options for selecting the monitored patient (*owner* of the data) and environmental parameters associated with them can be chosen for analysis. It should be noted that as the data schema was designed to be flexible it is the user who defines which data is medical and which is environmental. Also available on the primary set up panel are the parameters for:

a) defining the analytical period of time (*Previous Time Period Interested In*),

b) the reading density for time series analysis. A reading every *x* hours/minutes is derived, and

c) the Lag used – if a lag is to be introduced into the time series analysis, between the data sets.

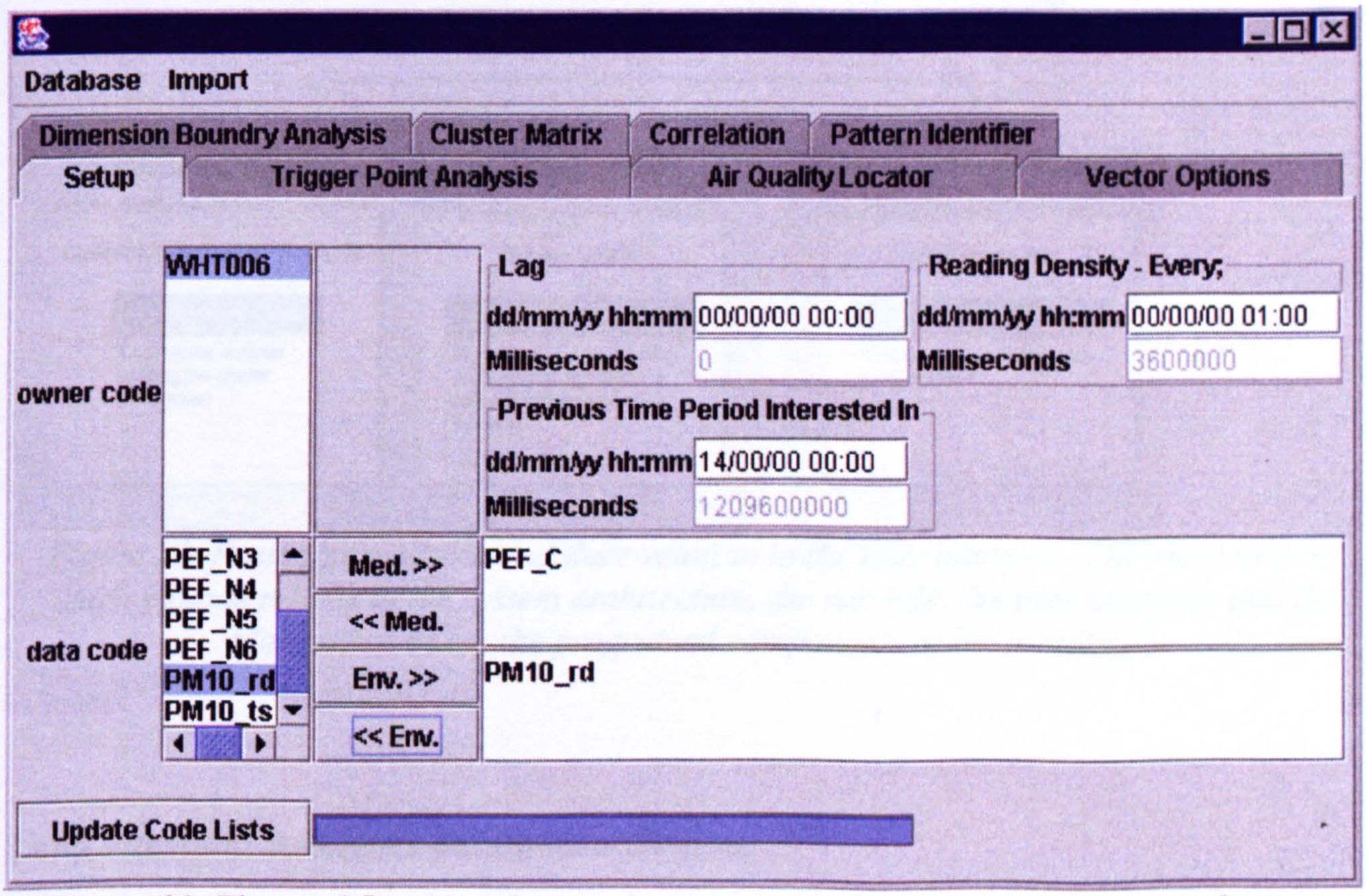


Figure 22 The workflow coordinator; prototype interface, showing the general setup panel.

It is worth emphasising that the user interface shown in Figure 22 evolved through the iterative EMS design process and as such is a testing prototype. Figure 23 relates the interface to the prototype modules developed as a proof of concept; the tab titles in Figure 22 relate to the sub titles in the diagram of Figure 23.

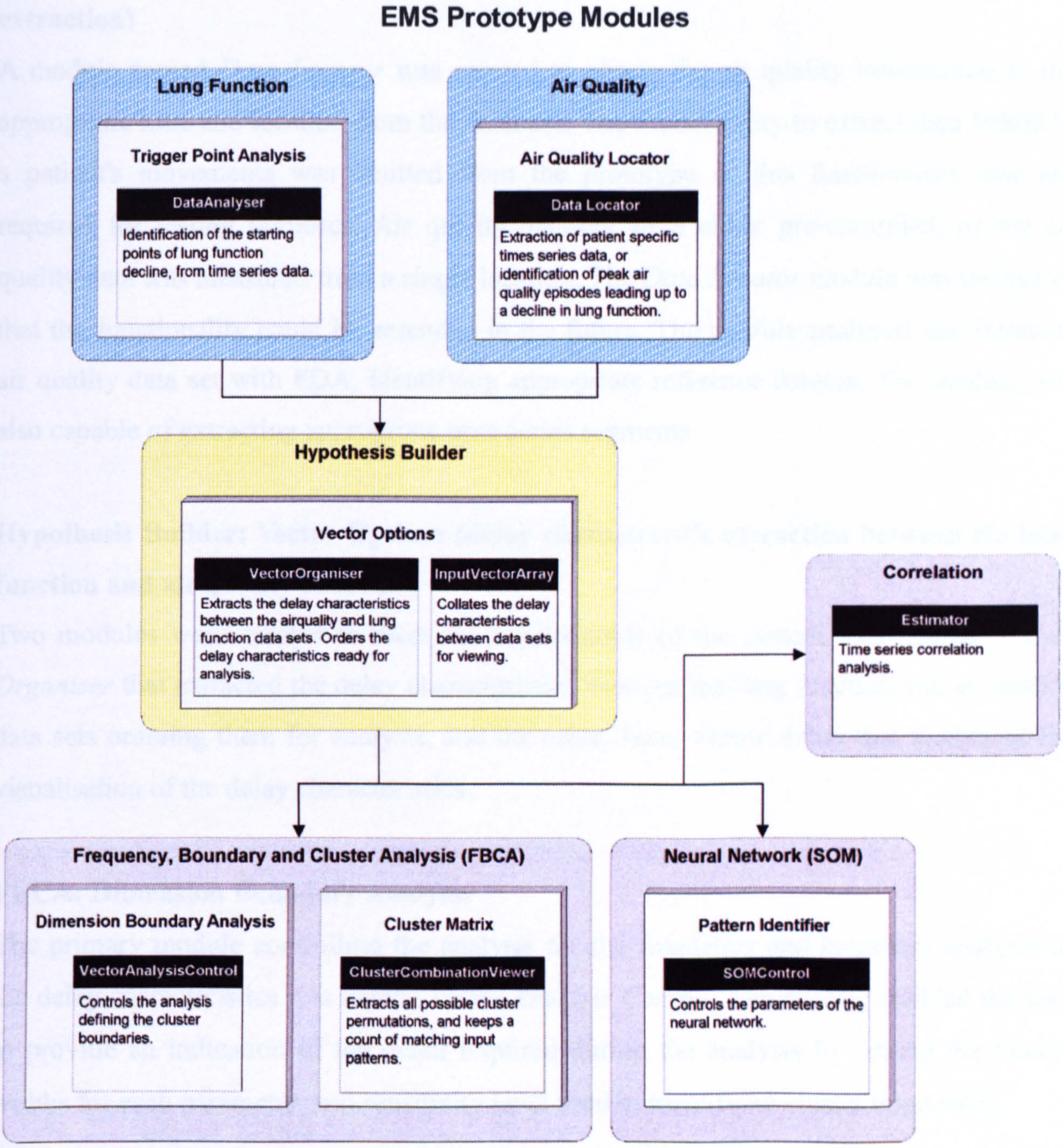


Figure 23 Prototype modules and their relation to the user interface. The main title of each module relates to the system architecture, the sub title, the user interface and the black titled boxes the programed components of the prototypes.

Lung Function: Reference datum identification

Using the information provided during the set up of the system, a module named *Data Analyser* is given control and extracts the patient's lung function records from the database over the relevant time period. Once obtained, the FDA module is used to identify the lung function reference datums. These datums are returned to the *workflow coordinator.*

Air Quality: Air Quality Locator (reference datum identification, or time series extraction)

A module named *Data Locator* was created to obtain the air quality information at the appropriate time and location from the database. The functionality to extract data linked to a patient's movements was omitted from the prototype as this functionality was not required for testing purposes. Air quality datasets were either pre-compiled, or the air quality data was measured from a single location. The *Data Locator* module was written so that the functionality could be extended in the future. The module analysed the extracted air quality data set with FDA, identifying appropriate reference datums. The module was also capable of extracting appropriate time series segments.

Hypothesis Builder: Vector Options (delay characteristic extraction between the lung function and air quality data sets)

Two modules were created to meet the requirements of the system. One named *Vector Organiser* that extracted the delay characteristics between the lung function and air quality data sets ordering them for analysis, and the other, *Input Vector Array* that facilitated the visualisation of the delay characteristics.

FBCA: Dimension Boundary Analysis

The primary module controlling the analysis for the frequency and boundary analysis of the delay characteristics was named *Vector Analysis Control*. The module enabled the user to provide an indication of the detail required during the analysis by setting the bucket widths for each parameter, and sensitivity level used in identifying cluster boundaries.

FBCA: Cluster Matrix

The second module (*Cluster Combination Viewer*) used within the FBCA component extracted all cluster permutations derived by the Dimension Boundary Analysis module. Incoming data was then monitored by each cluster permutation for a matching record, and when recognised, was stored.

Neural Network: Pattern Identifier

The neural network, based on a self-organising algorithm, was developed with a single point of access to its functionality. A module name *SOMControl* was used to control the network of neurons, and provided access points to visualise the process.

4.5 Summary

This chapter presented the way in which system architecture was used to develop the structure of the Environmental Monitoring System (EMS). A system specification giving a high level overview of key features, and elements that the system incorporates was presented. Research into the *data process architecture* developed a four layer model, where the third layer (data handling) is central to further research into analytical components in Chapter 5. Processes required for the identification of environmental predictors leading to asthma exacerbation were outlined, providing a summary of the identification architecture.

This chapter also presented a number of prototype components, and described aspects of an architectural implementation used for validation by this thesis. The implementation of the architecture through prototypes was not designed to undergo scrutiny in a clinical setting. However, the iterative design and verification process did reveal a number of real design issues which were used to guide the development of this research.

The identification of key communication aspects and data interchange between components led to the refinement of some of the fundamental concepts of this thesis, including refinement of the *reference datum* and *delay characteristic*.

Chapter 5

Analytical Process

This chapter presents research into specific analytical components, that when combined are capable of validating environmental predictors of asthma exacerbation. Techniques are developed to: identify reference datums within environmental and medically related data, find associations using delay characteristics, and validate the characteristics. The validated delay characteristics provide key information that enable the triggering of patient-specific alerts.

5.1 Introduction

Results obtained from the use of correlation techniques are unable to relate specific events in the environment to the decline of patient lung function, or provide a suitable mechanism to automatically detect adverse patterns or rules without further analysis. Correlation also does not help build a picture of particular attributes or characteristics that play the most significant roles in affecting a patient when the time frames of the data sets are acting differently. This is supported by research during the Medicate (2000) project and research in Section 1.1.2.

With the change in focus from identifying a correlation between two data series (lung function and environmental) to constructing verifiable delay characteristics, this research defined a method that indicated the time delay between the environmental predictor, and the asthma exacerbation.

The method starts by using a technique capable of detecting the point at which a healthy lung function signal begins to decline. The author has named the developed technique *Feature Detection Analysis* (FDA). This is of course a very loose description of what the analysis achieves, but is one that works for the purposes of this thesis. An explanation of the technique follows.

5.2 Feature Detection Analysis: Asthma Episode Detection

During the Medicate (2000) project clinical staff were alerted if a lung function **measurement** fell below a predefined threshold. The new technique of *Feature Detection Analysis* (FDA) automates the analysis of patient specific **trends** and includes the ability to: identify a rate of change in lung function trend over time, calculate if the **trend** falls below a threshold, and when the trend reverses.

The first purpose of FDA is to determine the point at which a trend reverses, and more importantly, does the trend meet the condition set by the clinician. For example;

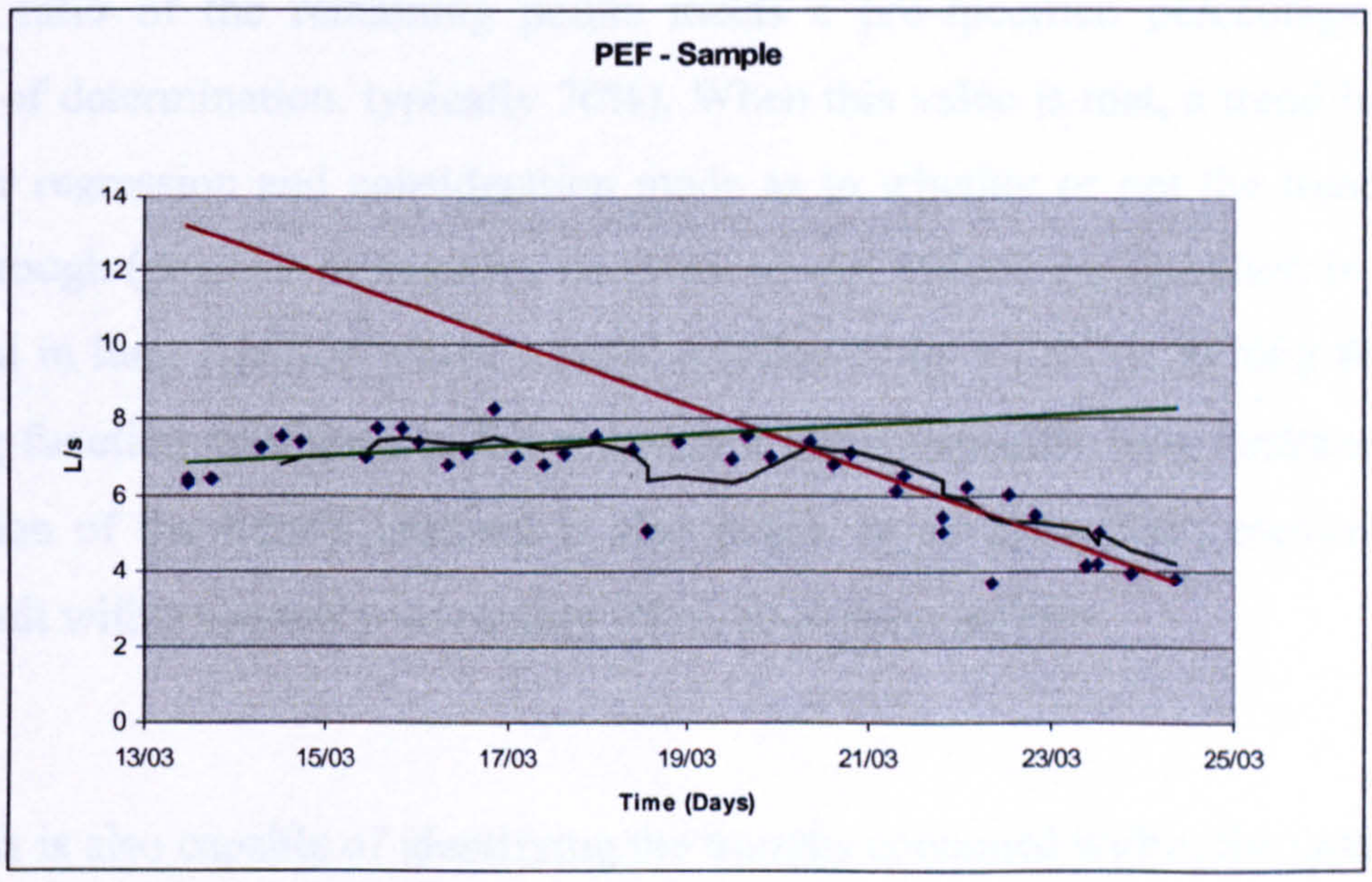


Figure 24 A sample of PEF data. The trend for the declining section of lung function is shown by the red linear regression line. The trend for the stable period of lung function is shown in green While a 1 day moving average is presented in black.

Figure 24 shows a sample of PEF data from a patient during the Medicate (2000) project. A regression line (shown in red) has been fitted to the declining period of lung function trend, beginning on the 20th March, after a period of stability. The objective for FDA is to determine if the declining trend passes through a threshold level set by a member of clinical staff. If the trend falls below this predefined threshold a reference datum is created to mark the change in trend, in this example, on the 20th March, where the two trend lines meet. FDA achieves these purposes with a combined correlation and linear regression technique to analyse the trend of the lung function data segment.

A regression model is constructed for a segment of data where the trend is being analysed. The length of segment is not critical to the analysis, but must contain a minimum of two data points to operate. The technique has been trialled on datasets containing up to 1 year of data (8,700 data points). The correlation model for the analysed segment is then used to obtain the goodness of fit of the model to the underlying data. The goodness of fit of the model is measured using the R-squared function (also known as the coefficient of determination) of the correlated of data points, within the analysed segment.

The technique is implemented by extracting the correlation ratio from an ever-decreasing subset of data. The number of points covered by the data set is reduced by one until the correlation ratio of the remaining points meets a pre-specified percentage value (the coefficient of determination, typically 70%). When this value is met, a trend line is drawn using linear regression and consideration made as to whether or not the trend leads to a peak or a trough (positive or negative correlation). An asthma exacerbation is represented by the trend in lung function where a peak is followed by a decline in lung function, and where lung function continues to fall to reach a patient-specific lung function threshold. Consideration of the trend's gradient is also made, as an asthmatic's recovery becomes more difficult with a suddenly worsening trend, so is more serious.

The analysis is also capable of identifying the troughs contained within the data set. This is useful when lung function reference datums are being analysed, as it is likely that if environmental (predictor) analysis is focused purely on peak reference datums, good predictors of the declining respiratory trend could be missed.

However, peak points are of particular interest due to their position in the data trend. Data following a peak represents a patient whose ability to breath normally is reducing, and the *peak* is the earliest indicating point of the start of this trend, and therefore the earliest point at which a patient can be warned of their condition. The method used to measure the ability to breath depends on the monitoring device used. The devices used during the Medicate (2000) project measured the average of three lung function tests, while a more recent device (Ferraris, 2008), used for the capture of patient lung function, records the highest peak value from readings taken over a three minute period. If the level of lung function falls below a pre-defined threshold (set by clinical staff), the time and location of the peak

reading is of specific interest. A reference datum is assigned to the point and used to extract possible (environmental) predictors from the database. The figure below shows the raw data points of a lung function time series.

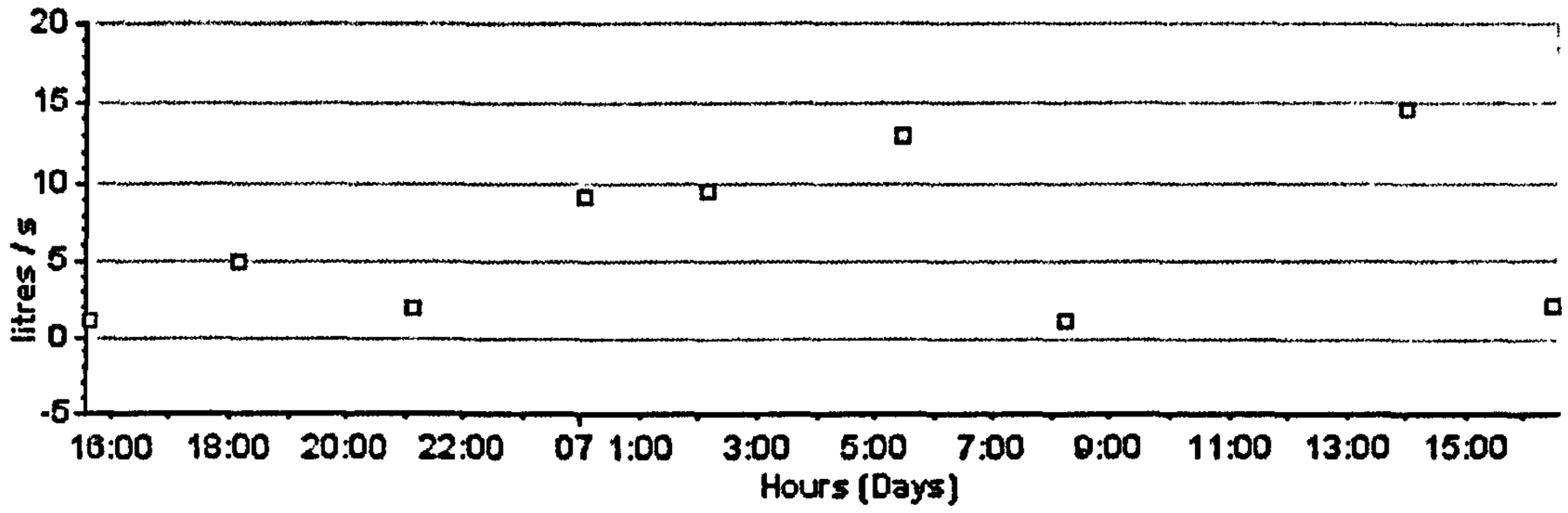


Figure 25 Lung function time series - data points

The objective is to locate peaks (trend reversals) automatically, rather than through visual inspection of the graphs, whilst fulfilling the requirements previously specified. The peak values of the data identified through visual inspection (but found automatically) are highlighted by the grey arrows in Figure 26.

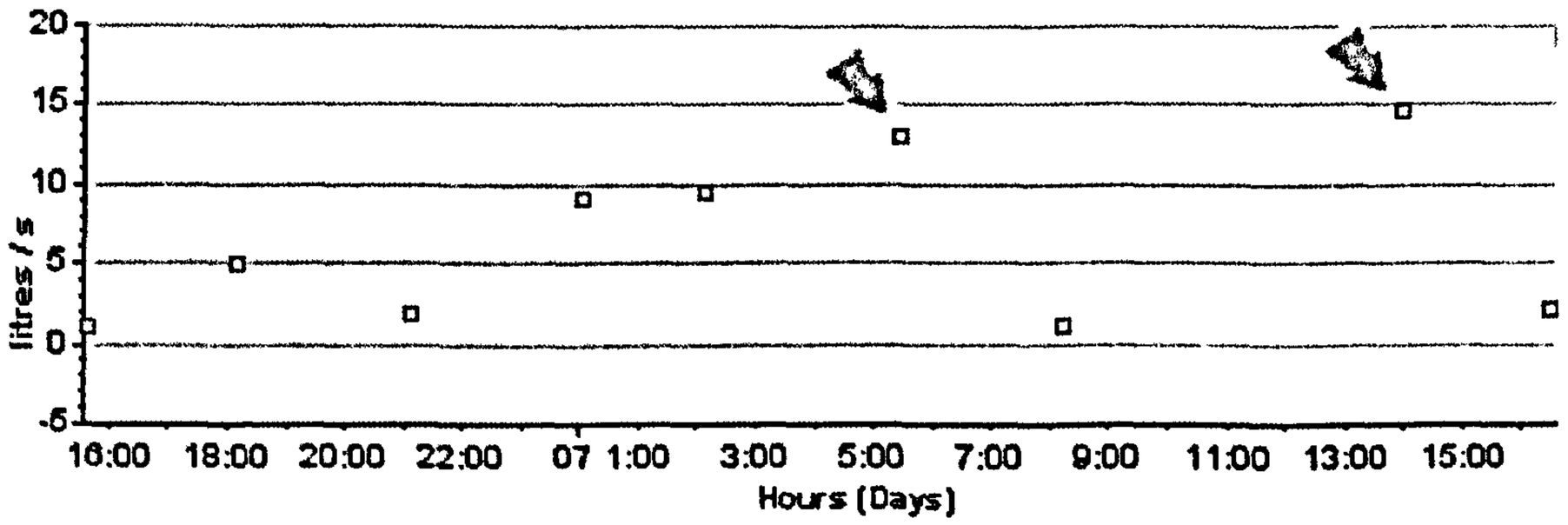


Figure 26 Lung function (Peak Expiratory Flow) data for 24 hours, with visually inspected peak values highlighted by arrows.

The analysis is initiated with all data points in the dataset (Figure 27). The analysis measures the square of the correlation coefficient (R^2), which is also equal to the coefficient of determination of the regression line. The length of data set is gradually reduced by one, until the coefficient value satisfies a pre-specified value (defined as a percentage). The specified value has been found to be adjustable depending on the length of time series being analysed, and the outcome that clinical staff are looking to obtain.

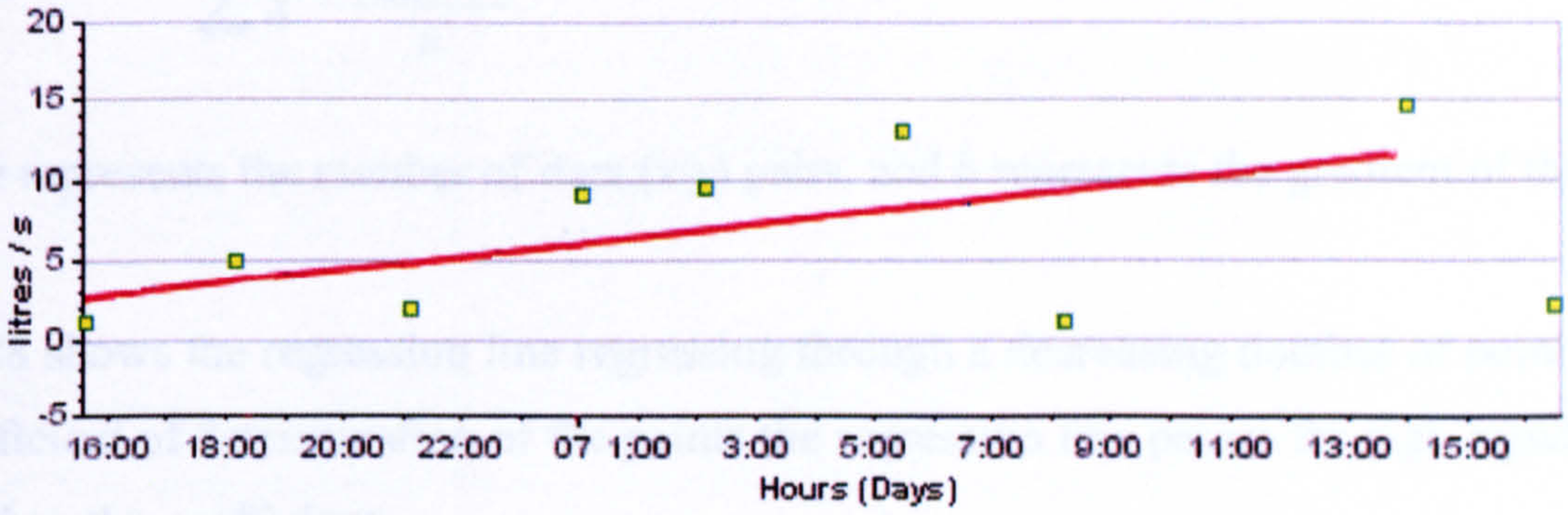


Figure 27 Reducing regression line on a complete dataset.

Figure 27 shows the first progression of the regression line (in red) through the data points. The regression coefficient in the figure equals 36% as this figure is not greater or equal to the pre-defined level the analysis continues to reduce the number of data points covered by the analysis.

$$r=\frac{\sum^{n} xy-\frac{\sum x\sum y}{n}}{\sqrt{\left(\sum^{n} x^2-\frac{(\sum^{n} x)^2}{n}\right)\left(\sum^{n} y^2-\frac{(\sum^{n} y)^2}{n}\right)}} \qquad \textit{Eq. 5.1}$$

Equation 5.1 is the formula for the Pearson product moment correlation coefficient. The technique uses this coefficient, predetermined by the system (or user) as a benchmark to set which points the regression line passes through. The value used is R^2 the coefficient of determination, which represents the percentage variance of y which is explained by the variable x, an indication of the importance of the correlation.

The equation of the regression line is given by (Pearson & Turton, 1993);

$$y=a+b\cdot x \qquad \textit{Eq. 5.2}$$

where a is given by Equation 5.3, and b by Equation 5.4.

$$a=\frac{\sum^{n} y}{n}-b\frac{\sum^{n} x}{n} \qquad \textit{Eq. 5.3}$$

$$b = \frac{\sum^{n} xy - \frac{\sum^{n} x \sum^{n} y}{n}}{\sum^{n} x^2 - \frac{\left(\sum^{n} x\right)^2}{n}} \qquad \textit{Eq. 5.4}$$

Where n represents the number of data (x/y) pairs, and b represents the gradient of the line.

Figure 28 shows the regression line regressing through a decreasing number of points until the coefficient of determination of the points the regression line passes through equals or is greater than the coefficient.

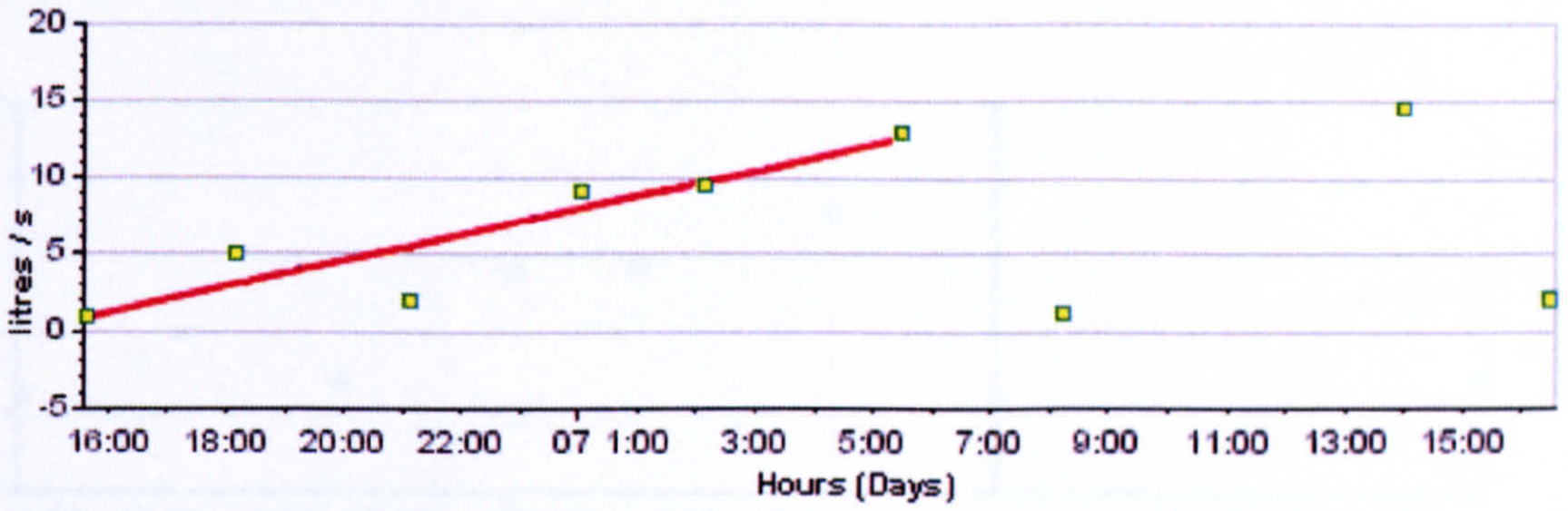


Figure 28 Regression analysis decreasing data set

The stage (shown above) has satisfied the trend fitting parameters (in this example set to 70%) with a coefficient of determination of 83%. The gradient of the line (given by ***b*** in Equation 5.2) is *0.8* which equates to a positive gradient of 0.8L/s per hour. These settings can be set by experienced clinical staff for each individual patient.

The method picks out *peaks* and *troughs* within the data series. It is also able to consider the gradients of +'ve or -'ve regression lines and whether or not the line crosses a *trigger* value before encountering the next identified data peak or trough. Regression line details that satisfy the set parameters are recorded for use in the next analytical process. Figure 29 shows the second pass through the data set to identify a *trough* reference datum.

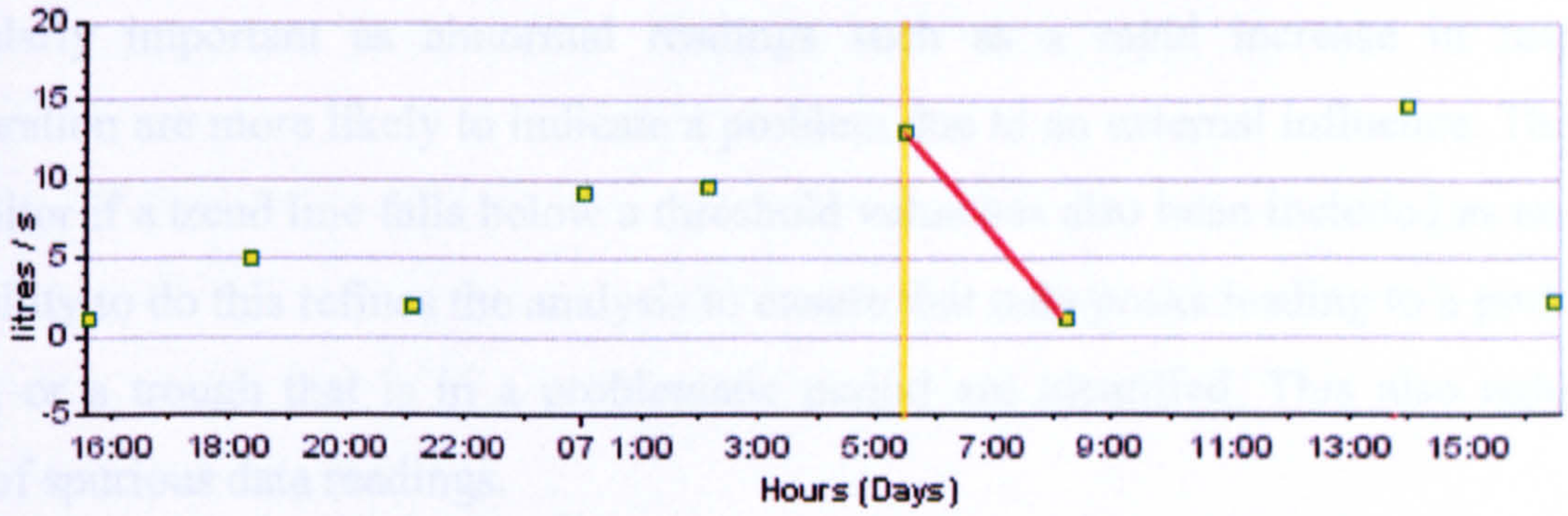


Figure 29 The analysed section leads to the identification of a trough.

Figure 30 shows the start of the third iteration through the data set. The yellow line represents the peak identified by the first pass through the data set. The data set starts from the last peak or trough that was found.

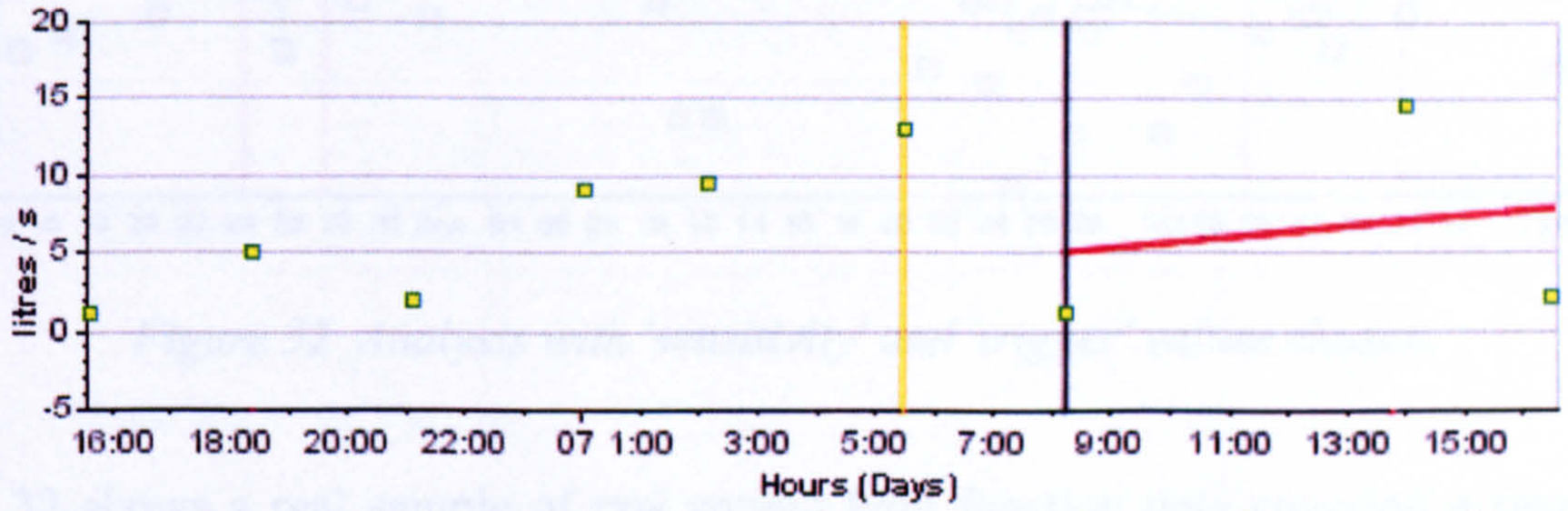


Figure 30 Regression analysis after the second reference datum, (in this case, a trough shown by the black vertical line) has been identified.

The third iteration through the data set locates a trend that satisfies the coefficient percentage between two data points (Figure 31).

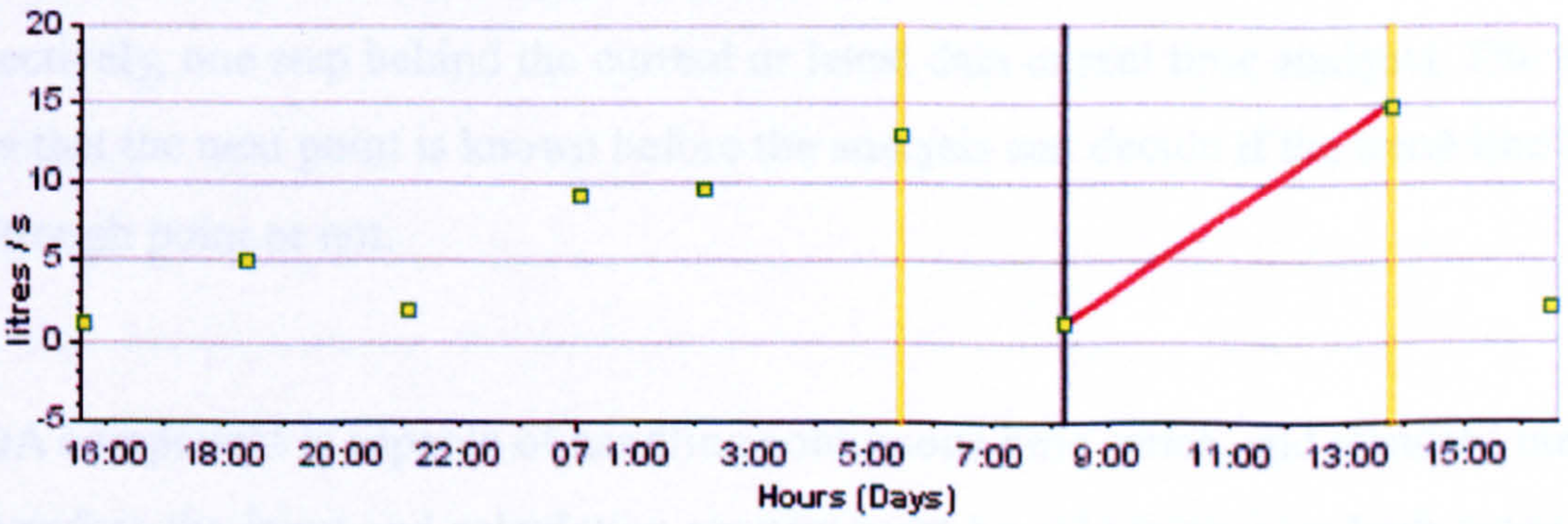


Figure 31 Reducing regression trend line (last iteration).

The user interface used in the prototype allows the user to set parameter values affecting the analysis. The capability to identify trends that descend or ascend over a certain time (monitoring the gradient of the trend) increase the usefulness of the technique. This is

particularly important as abnormal readings such as a rapid increase in respiratory deterioration are more likely to indicate a problem due to an external influence. The ability to monitor if a trend line falls below a threshold value has also been included as an option. The ability to do this refines the analysis to ensure that only peaks leading to a problematic period, or a trough that is in a problematic period are identified. This also reduces the effect of spurious data readings.

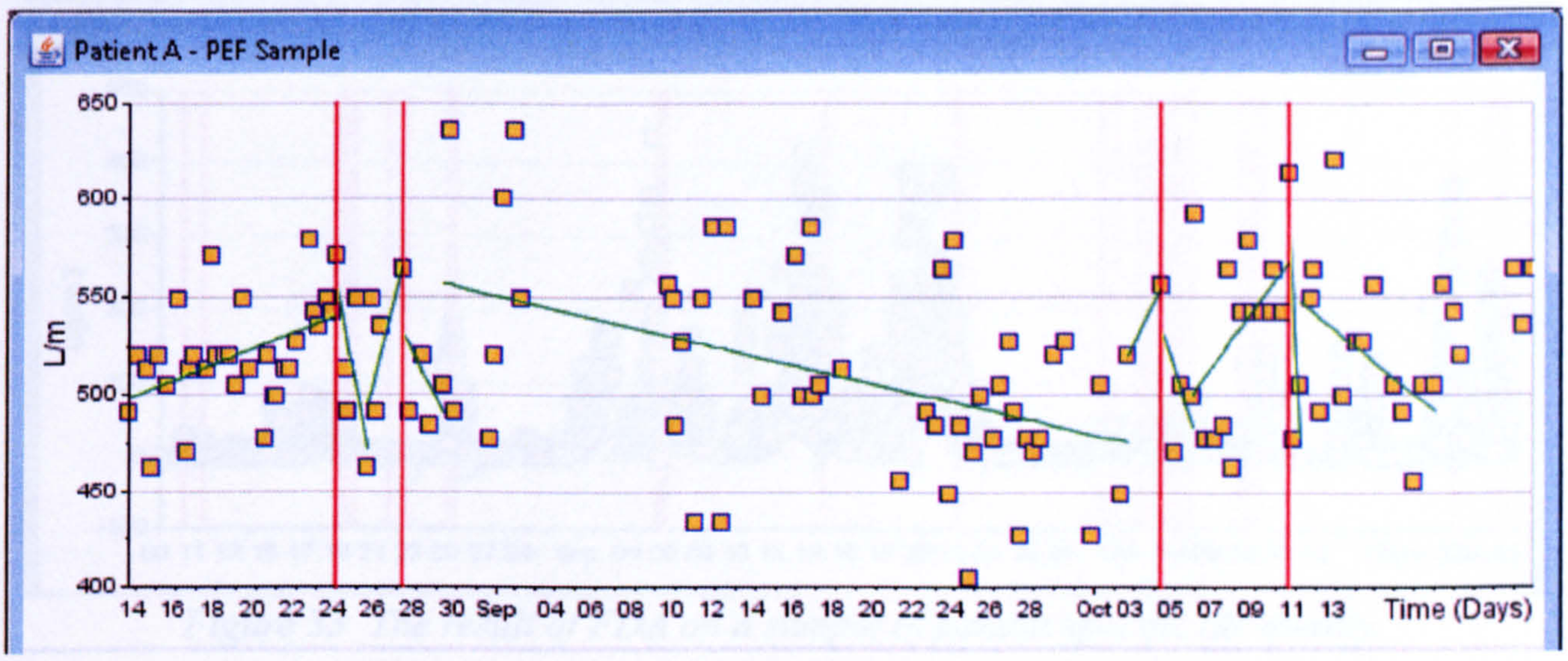


Figure 32 Analysis with 'sensitivity' and 'trigger' values chosen.

Figure 32 shows a real sample of raw patient lung function data covering a two month period. The analysis has identified four (peak) trend reversal points that lead to a decline in lung function (falling below the specified threshold). Each trend marked with a reference datum (represented by the vertical red lines), declines below a threshold of 480 L/m.

It is worth noting that the application of FDA requires the analysis to be used retrospectively, one step behind the current or latest data in real time analysis. The analysis requires that the next point is known before the analysis can decide if the trend line leads to a peak/trough point or not.

The FDA component is capable of handling continuous time series, and standard numerical data, therefore the input and calculation process must be able to process both types of data stream. Analysis of time series data is achieved by conversion of the date/time element of the data into a value. The value used by the EMS is called the *epoch*, and is the number of milliseconds since 1st January 1970. An example of FDA using a numerical x-axis can be seen in Section 4.4.4 (Frequency, Boundary and Cluster Analysis).

5.3 Feature Detection Analysis: Environmental Predictor Detection

Feature Detection Analysis is designed as a generic tool and can be applied to data sets other than lung function. The figure below shows a sample of nitric oxide data that is specific to a patient, with FDA analysis shown by the red vertical lines.

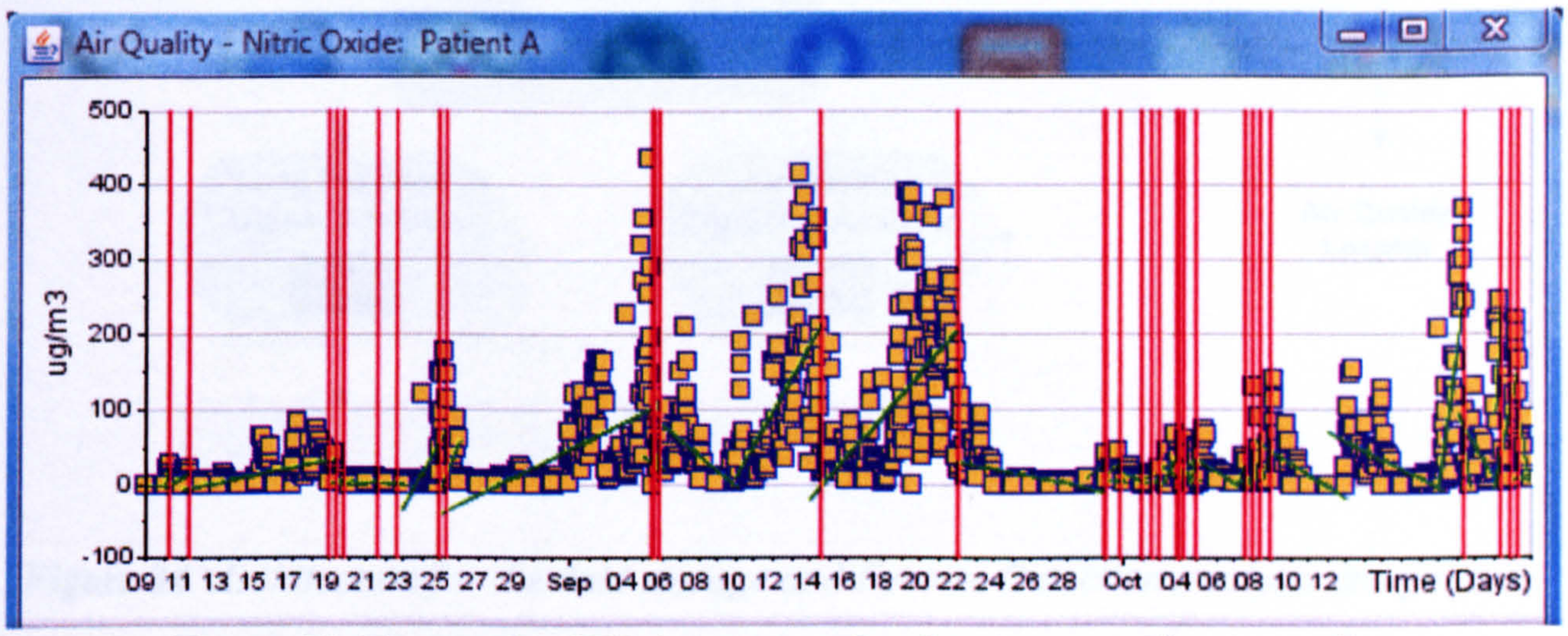


Figure 33 The result of FDA on a sample of patient specific air quality.

Figure 33 demonstrates the use of FDA on a sample of air quality data. The technique remains identical to the use of FDA with respiratory data. Although air quality is a highly variable data type (often in the region of ±40% at 1 standard deviation), the analysis is still capable of identifying features that after validation, could be predictors of asthma exacerbation.

Air pollutant levels are categorised into quality bands of *high*, *medium*, and *low*, as discussed in Section 2.2.1. These parameters are used internally within the EMS to identify air quality reference datums. As lung function FDA is guided by clinical staff, so analysis of air quality data is guided by the air quality banding thresholds. Figure 33 shows FDA without this option selected, therefore matching data by trend significance alone.

Feature Detection Analysis is also capable of analysing other sources of data. Section 6.7 demonstrates this through analysis of hospital admissions data; linking hospital admissions due to asthma exacerbations, and poor air quality.

Figure 34 shows a summary of the architecture presented during the previous sections of this thesis. Following the identification of reference datums from each of the data sets, a method is required that prepares them for further analysis. The next section explains the

processes involved.

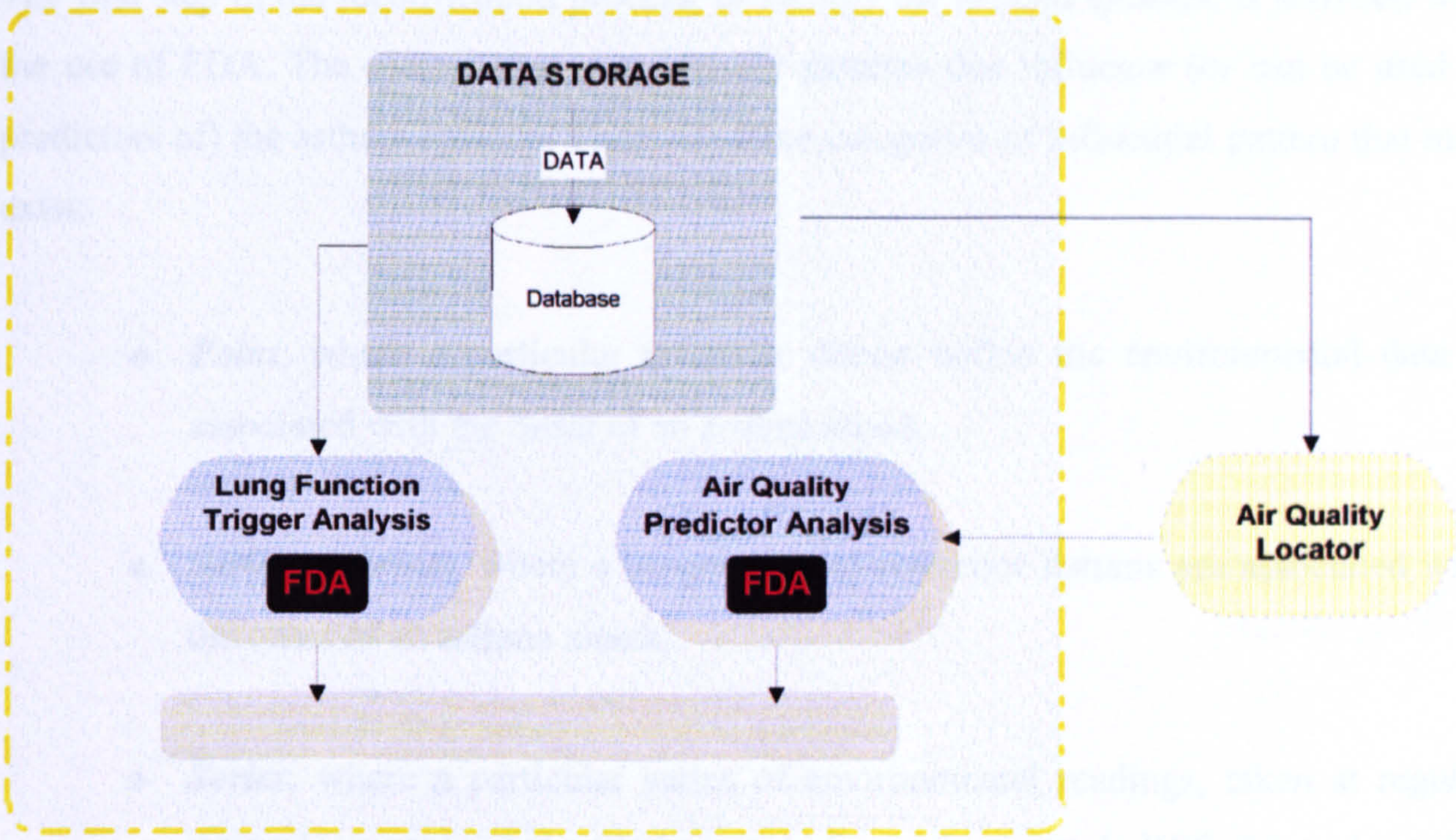


Figure 34 Architecture for the data storage and Feature Detection Analysis components.

Implementation of FDA as a distinct component creates scalability by enabling the instantiation of new FDA modules each time an additional parameter, or patient is *created* during analysis. The FDA component, and EMS architecture were designed with this consideration, and will spawn a new FDA component to analyse each additional data parameter.

The underlying requirement is for reference datums to be created from the analysis. The reference datum takes the following form;

Figure 35 Example of a reference datum required by the Hypothesis Builder to define the analysed data.

The patientID belongs to the patient regardless of whether the reference datum belongs to a respiratory, or environmental data set. When multiple FDA components are in use, the PatientID associates the datum to the appropriate Hypothesis Builder (described in the next section). Respiratory and environmental data sets both use the same type of reference datum and PatientID. If the analysis is being applied to a new domain then the patientID would be replaced with an alternative identifier, relevant to its domain.

5.4 Building a Hypothesis with the Hypothesis Builder

The first step in the identification process, *to identify the asthma episode*, is satisfied with the use of FDA. The second step is to identify patterns that influence (or can be used as predictors of) the asthma episode. There are three categories of influential pattern that may exist:

- ***Point***, where a particular reference datum within the environmental data is associated with the onset of an asthma attack.

- ***Series of points***, where a progression of reference datums are associated with the onset of an asthma attack.

- ***Series,*** where a particular series of environmental readings, taken at regular intervals, is associated with the onset of an asthma attack. With this technique it is necessary to estimate missing data values. This aspect is discussed later in Section 5.4.3.

The process of organising the (pre-)processed data is controlled through a user interface (the *Hypothesis Builder*) which is described below in relation to the software implementation of the prototype. The *Hypothesis Builder* is a module designed to provide an interface between the system and clinical staff. However the choice of analysis used by the *Hypothesis Builder* could be set before the analytical process begins, when patient details are selected by the user. The *Hypothesis Builder* was developed to test the functionality of system components, but helped identify and refine the methods of collating delay characteristics shown in this section. Environmental data is analysed according to the type of analysis chosen (listed above). Data sets derived from the pre-processing options are analysed further by the pattern identification components of the system.

The module collects the data produced by the initial FDAs and extrapolates all the potential permutations between them, forming a list of possible predictors and their relationships to each *outcome* reference datum. The prototype interface for this module (Figure 36), contains two sections that influence how the analysis can be taken forward.

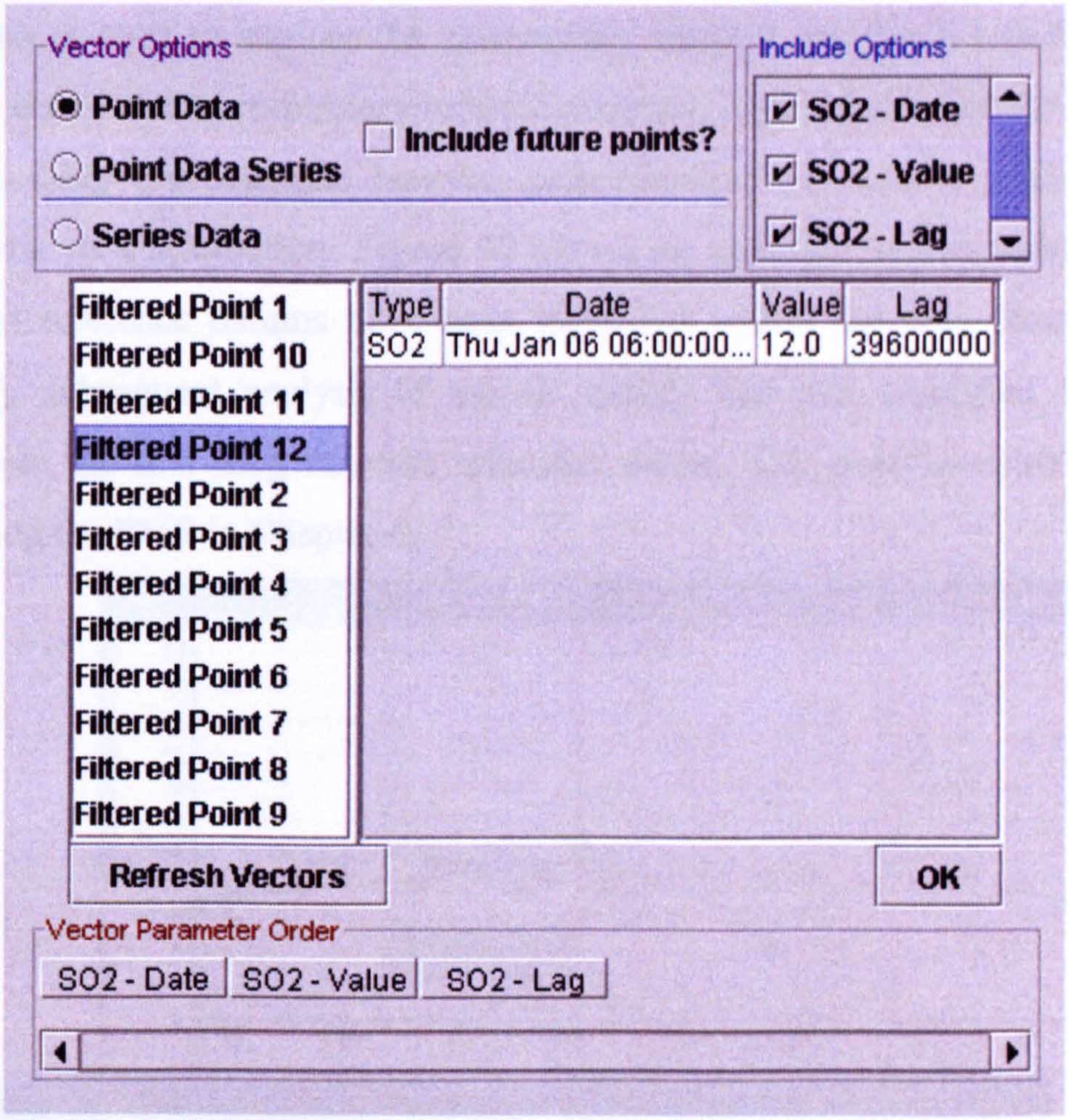


Figure 36 Prototype interface to facilitate the selection of analysis type and associated output options.

The ***first section***, shown on the left hand side of the prototype interface (Figure 36 '*Vector Options*') contains the *type* of analysis (described over the following sections), with the addition of an option that allows analysis of future environmental predictors (for *point data*). This works by allowing the inclusion of air quality reference datums occurring after the date/time of the lung function reference datum, which is possible when all permutations are being extracted between the two data sets.

The ***second section***, appearing on the right side of the prototype interface *Include Options* provides a choice to the user to select the data parameters taken forward for further analysis (SO_2 – *Date, Value,* and *Lag* in Figure 36).

5.4.1 Point Analysis

Point Analysis is used to explore the relationship between single air quality predictors, identified by the FDA and respiratory reference datums. The objective of the analysis is to generate one delay characteristic between each identified air quality predictor and the identified point of exacerbation. Figure 37 shows an example of this technique. In the example, two reference datums have been identified within the lung function data set (using FDA), subsequent analysis of the air quality data has identified four possible predictors from the first lung function reference datum. The *point analysis* technique is used for testing the EMS in Chapter 6.

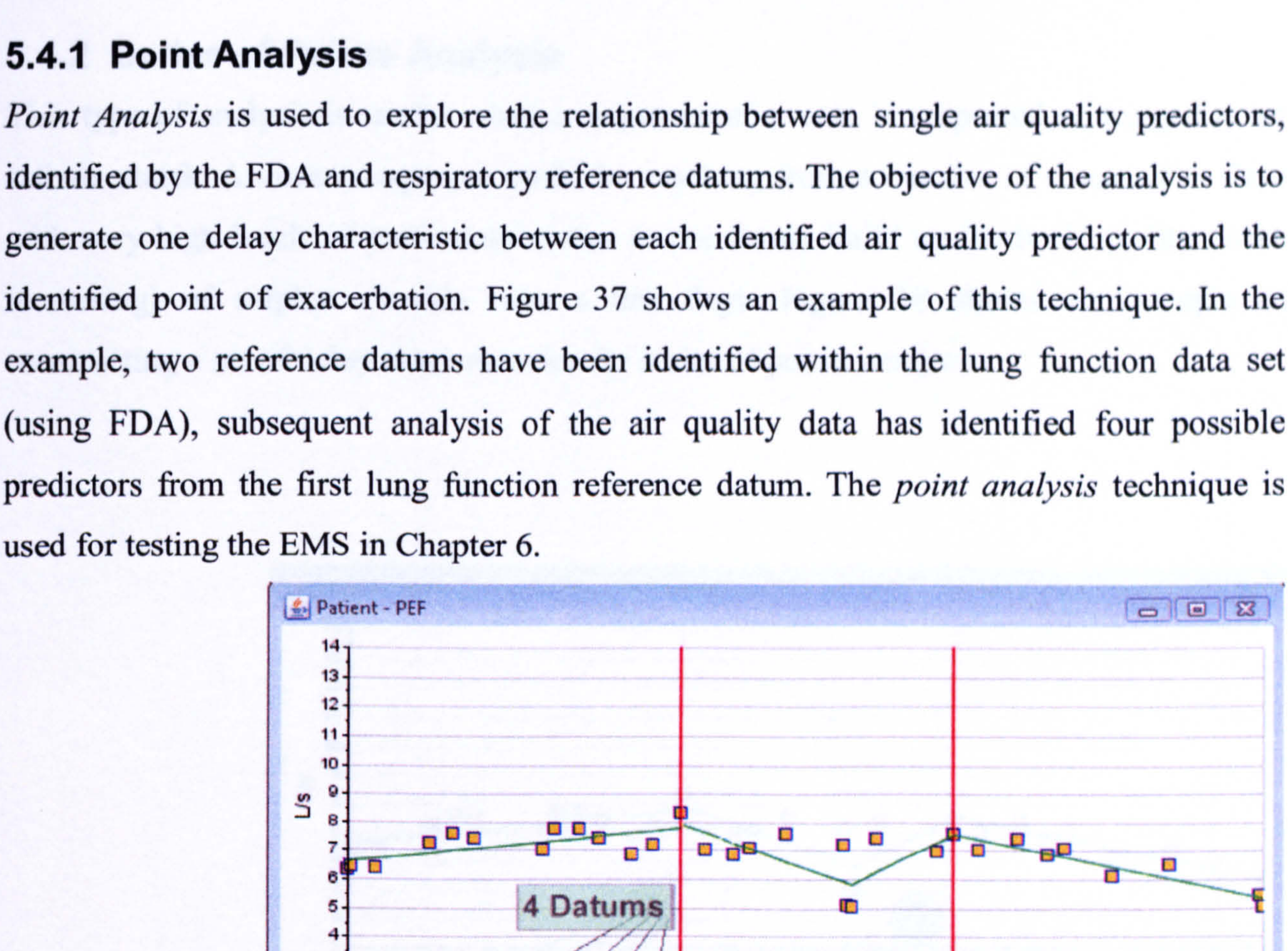


Figure 37 Showing four individual air quality reference datums that could each be a predictor of the decline in lung function.

In this example there are a total of ten possible combinations if the second identified lung function (PEF) datum is included within the analysis, and the time period for analysis is seven days prior to each PEF reference datum.

5.4.2 Series of Points Analysis

This type of analysis is useful when a sequence of events is suspected of triggering an asthma attack. An event sequence could be anything from a number of consecutive days with very high levels of particulate matter to the steady build up (with air quality peaks increasing) of sulphur dioxide over a few days. Figure 38 shows an example of constructing a set of delay characteristics for series of points analysis.

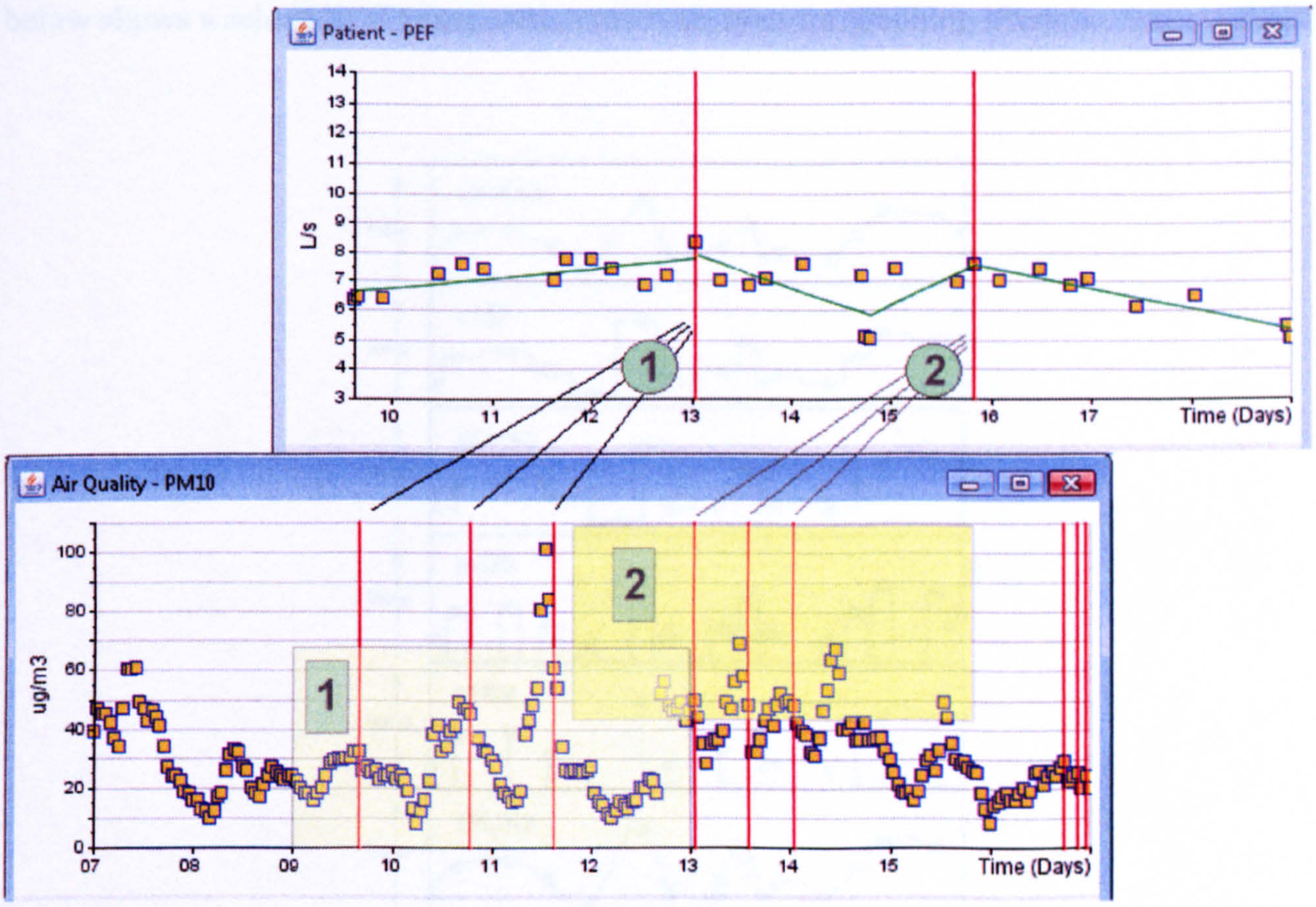


Figure 38 Shows the two combinations that are used during a series of points analysis. The three air quality reference datums would be used in series, against both lung function triggers.

The number of reference datums within each combination must be consistent for further analytical components in the EMS to analyse the data. This is due to the number of parameters (or dimensions) being presented to the components having to be of equal length. Components like the neural network must have the same dimensional vectors presented to them for each set of analyses. If additional forms of analysis are required using different length vectors, then an additional number of neural network components would be required. The number of reference datums identified within the **first period** sets the required number of reference datums in subsequent periods.

5.4.3 Series Analysis

Readings taken at environmental monitoring stations are taken at regular time intervals that are not controlled by the EMS, and bear no relationship to the timing of lung function readings. This creates a significant measurement difficulty. Analysis of enviromedic data requires that both sets of data (lung function and environmental data) are available as near as possible to the same time and location. For readings from both data sets to be recorded at the same time, a technique for estimating a reasonable value is required. The figure below shows a selection of interpolation methods used for graphing irregular data readings.

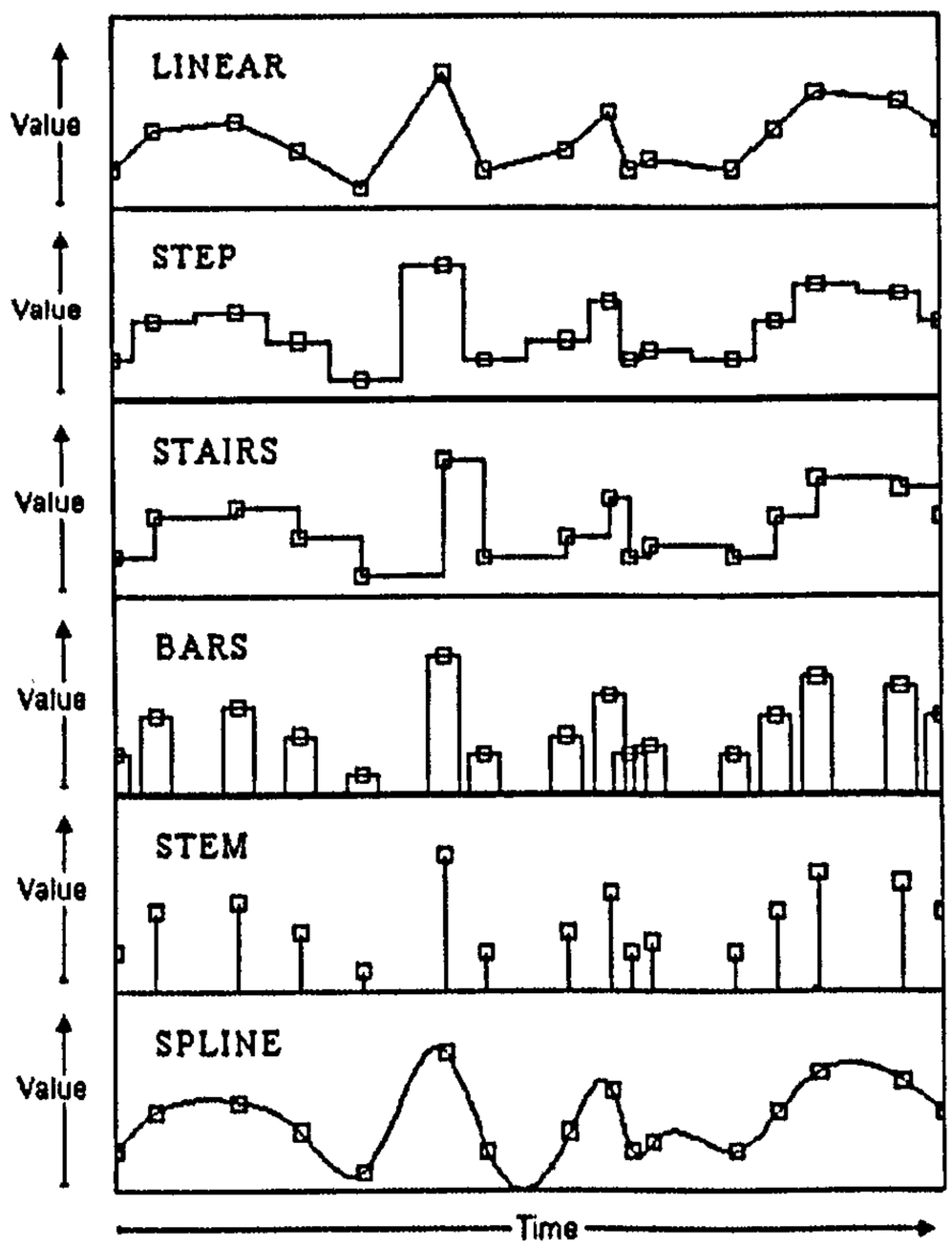


Figure 39 Interpolation methods

Irregular data intervals produce difficulties for many pattern recognition systems as trends become more difficult to identify. The method used by the EMS overcomes some of the problems associated with irregularity through the use of estimation. Estimation techniques such as the B-spline (Unser & Aldroubi, 1992) can be used within the EMS to estimate values between known data points. However it is not always suitable for time series data sets where data is irregularly spaced. A linear method of interpolation may be preferable in

this case. Interpolation techniques enable a set of environmental data to be estimated from data extracted from the database, as a series of readings taken at fixed time intervals.

Figure 40 shows two periods (1) & (2) highlighted in green, that are to be analysed further as a result of their identification by the FDA.

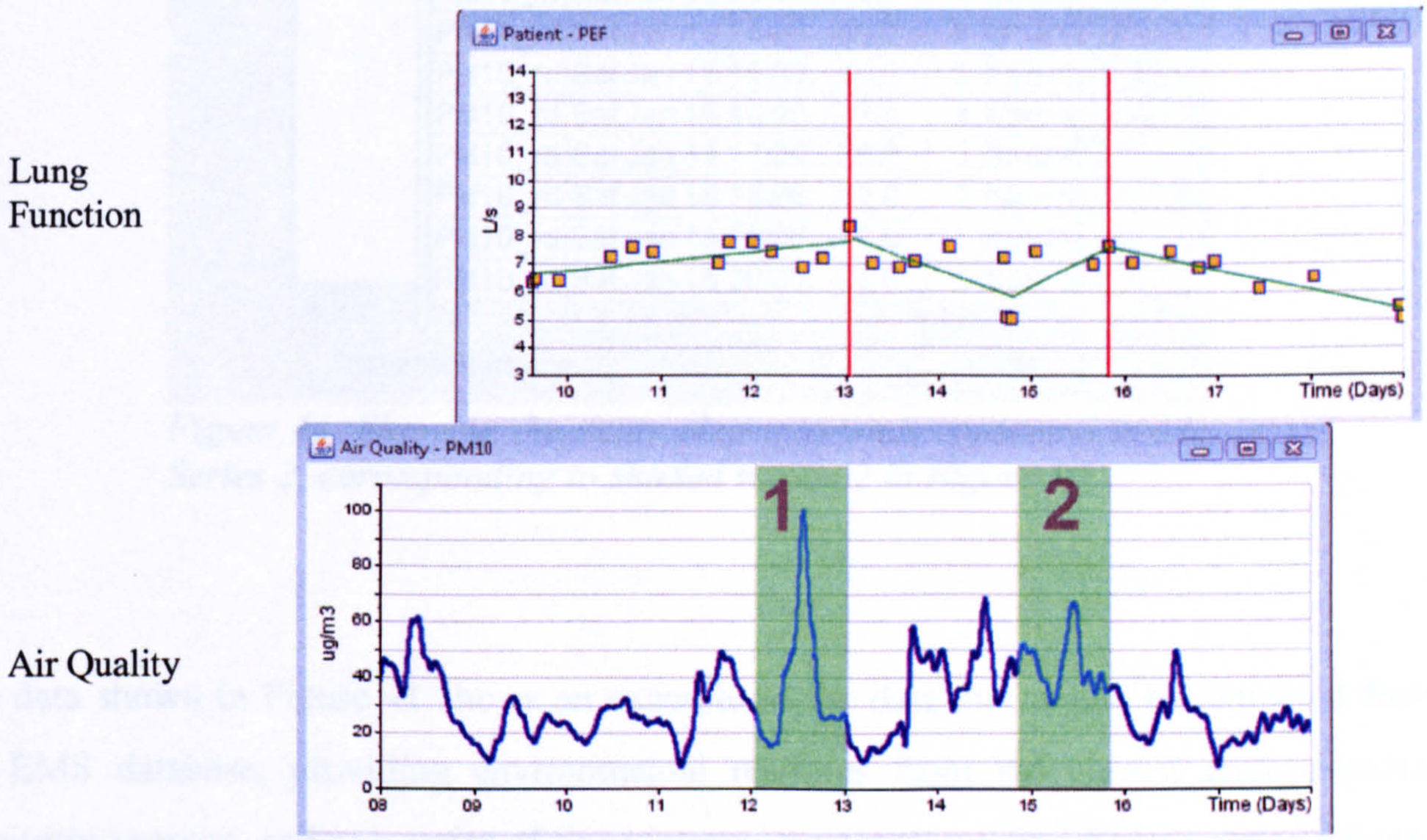


Figure 40 Data Series Analysis. The air quality period before each lung function peak is split into smaller time segments to analyse the characteristic of the air quality before the lung function reference datums (marked by the two shaded areas).

Following FDA the hypothesis builder extracts delay characteristics from relationships between each value of the series and the start point of lung function decline. The result of which is shown in Figure 41, where two time series can be seen, in this case two lung function peak points were identified after the 12th January (Figure 40) and one series has been extracted for each. *Series 2* corresponds to the second shaded region (marked 2) in Figure 40 and would be identical in length to that of *Series 1*.

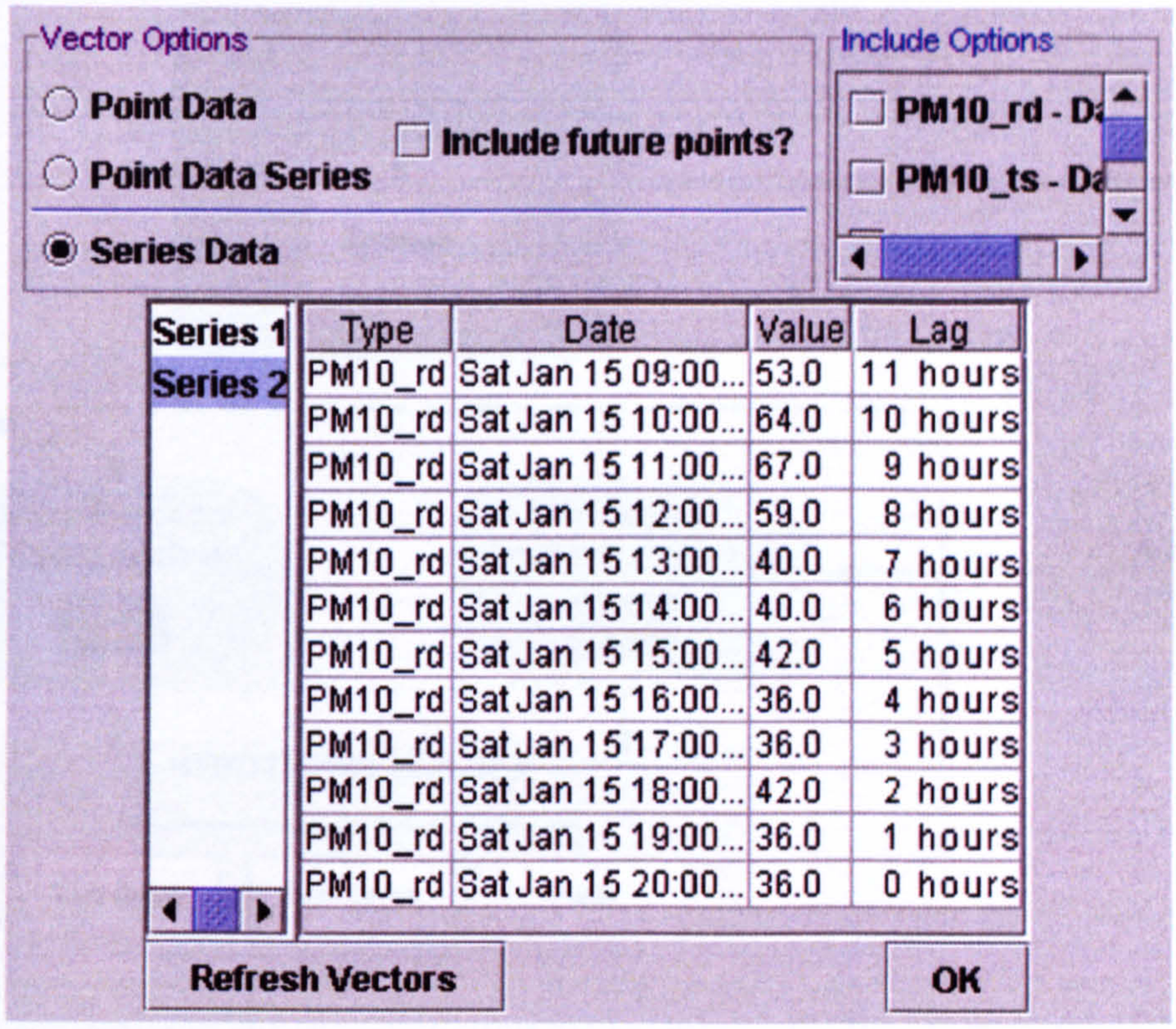


Figure 41 Showing the delay characteristics contained inside Series 2, corresponding to shaded region 2 in Figure 40.

The data shown in Figure 41 shows an example of the data that would be extracted from the EMS database, providing environmental readings from the closest environmental monitoring source, and as a series of fixed interval readings, running from a period prior to the reference lung function, and continuing up to the time of the asthma reference datum. These signals provide the basis for further analysis by the pattern recognition components of the system.

5.4.4 Operational Overview

Figure 42 shows the system architecture of the EMS, including the hypothesis builder. The hypothesis builder transforms reference datums identified by Feature Detection Analysis, into delay characteristics in preparation for further analysis.

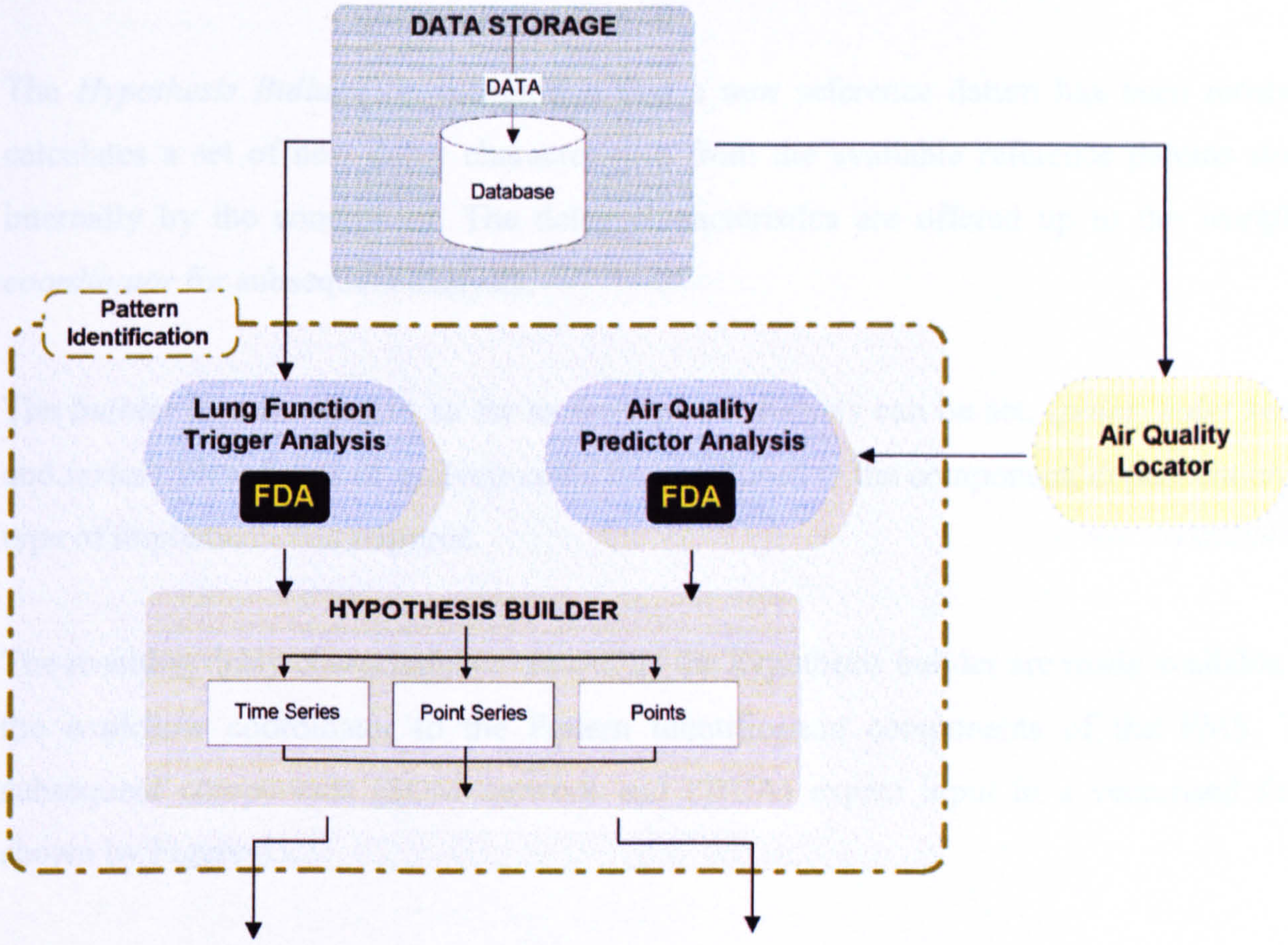


Figure 42 Architecture showing data storage, FDA, and the hypothesis builder components.

The Hypothesis Builder has been designed so that one *builder* is required per monitored patient. The input data presented to the component originates from air quality monitoring stations and lung function monitors, and is received by the builder via the FDA components as reference datums.

The Hypothesis Builder combines the *reference datums*, to form *delay characteristics* that are passed on to the subsequent pattern identification components. This is the Hypothesis Builder's primary purpose.

The Hypothesis Builder works by recording all the environmental reference datums that have been presented to the component over a set period of time. Using the work of Lebowitz (1996) this can be arbitrarily set to 10 days. Respiratory reference datums identified by the FDA component are also recorded and used as reference markers in the creation of subsequent delay characteristics. It should be noted that the FDA component could be interchanged for an alternative mechanism of identifying *interesting* features within the datasets. The underlying requirement is that delay characteristics must be created for further analysis.

The *Hypothesis Builder,* on recognition that a new reference datum has been received, calculates a set of new delay characteristics from the available reference datums stored internally by the component. The delay characteristics are offered up to the *workflow coordinator* for subsequent analysis.

The *builder* is controlled in so far as the type of analysis can be set, (*point*, *point series*, and *series*). New forms of analysis could be introduced to the component, dependent on the type of implementation required.

The resulting delay characteristics created by the hypothesis builder are made available via the workflow coordinator to the Pattern Identification components of the EMS. The subsequent components (SOM network and FBCA) expect input in a vectorised form, shown by Figure 43.

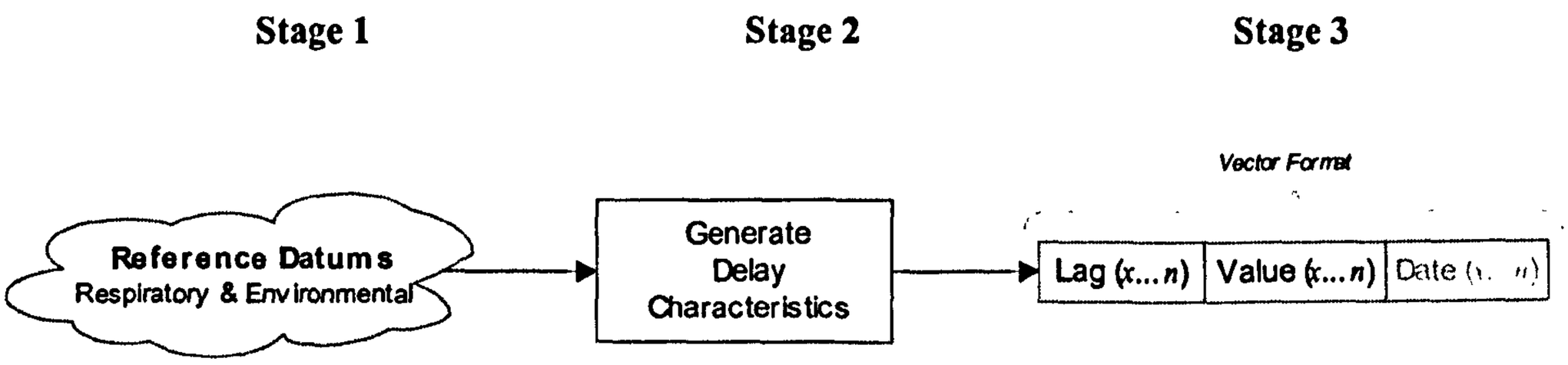


Figure 43 Process overview outlining the creation of vectors representing sets of delay characteristics, for further analysis by analytical components.

5.5 Pattern Recognition

Pattern recognition falls into the second step of the *identification process* (Section 3.5). However, it is useful to consider the output required from the EMS before the process of pattern recognition is discussed.

Monitoring for troublesome air quality patterns in real time is fundamental to the success of the system, and a technique capable of this is required for the research and development work of *step two* to be satisfied. Characteristics of the required technique are as follows:

- Real-time pattern matching capability

- Adaptable to new data types
- Capable of handling many parameters
- Capable of being applied to varying problems
- Ability to give a *strength* indication (or probability) of pattern match.
- Ability to pick out re-occurring data characteristics.

These characteristics enable a *picture* of real-time patient exacerbants to be developed, and stored for instant recognition as they occur in the environment. *Appendix M* outlines some algorithms that could be used within the EMS to assist with the task of recognising commonly occurring patterns. Systems that learn how to recognise patterns without being taught under supervision (usually by training data) are generally classified into the category of being unsupervised systems.

Neural networks have the ability to ***a)*** adapt to recognise *patterns* in real-time, and ***b)*** recognise similar traits when they happen in the future. These characteristics are complementary to the aims of the EMS.

5.5.1 Neural Networks

A neural network is a network of interconnected elements which learn by modifying the connection strengths between elements to match the inputs and outputs of the system being modelled. The basic element of a neural network is a neuron (or processing element). The basic functions of each element are (Parsaye, 1993):

1. Evaluate input signals and determine strength
2. Calculate a total for the combined input signals
3. Compare the total with a threshold value
4. Determine what its own output will be

Mathematically, the calculation of input strength is the dot (inner) product of two vectors, one vector being the input signals and the other being the *weights* that are assigned to each signal, presented in the figure below.

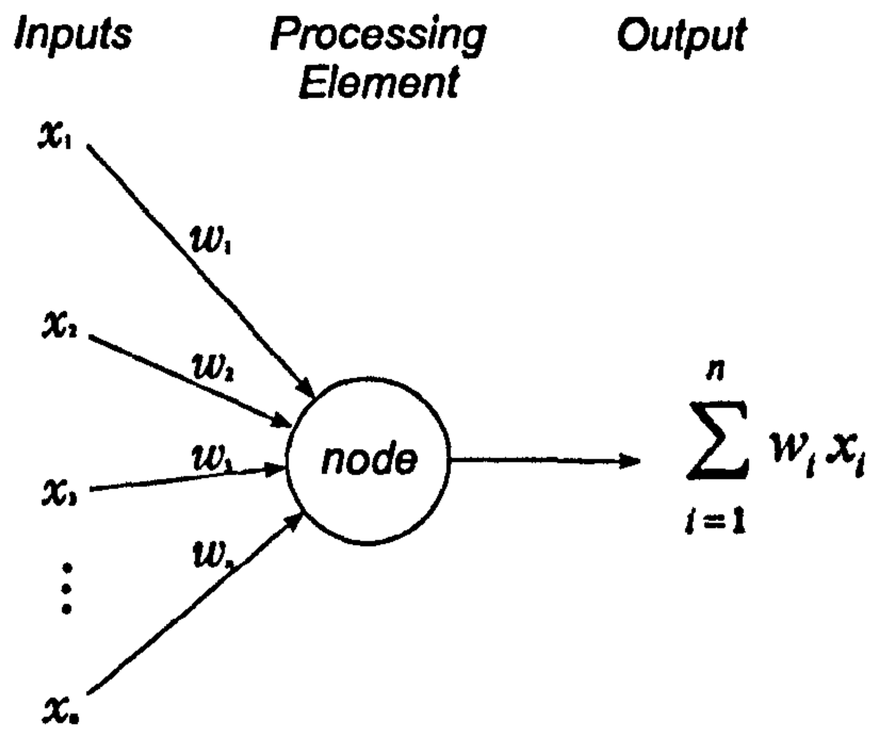


Figure 44 Processing element: The basic element of a neural network.

Figure 44 shows a schema for a linear associator (Rojas, 1996), the processing element computes the weighted input, and outputs the result. The linear associator is used as the founding component in self-organising maps (SOMs) an *unsupervised* learning technique.

An example of prediction using neural networks is the hybrid system used by Kolehmainen *et al.* (2000) to forecast urban air quality for the following day using airborne pollutant, meteorological and timing variables. This approach uses the self-organising map (SOM) algorithm (Kohonen,1998), Sammon's mapping (Sammon, 1969) and *fuzzy distance* metrics to cluster the data and then the use of Multi-Layer Perceptron (MLP) models to calculate the levels of individual pollutants from the clustered data. Predicted levels for each of the pollutants is derived by summing a combination of weighted MLP models appropriate to the situation. The methodology behind the work is particularly worth noting, the SOM is an unsupervised neural learning algorithm that finds prototype vectors which are able to represent the input data set. Further mapping then takes place using Sammon's Mapping that is used to map *n*-dimensional data into two dimensions for a graphical representation of the system.

Unsupervised learning makes initial data categorisations without intervention or guidance from an external influence, the technique is also known as *self-organised* learning and often uses a competitive learning rule. Supervised learning is usually guided by a user. Supervised techniques use a method of supervision to train a network – mapping inputs to the correct outputs of a neural network.

In neural networks competitive learning requires neurons (processing nodes of a neural network) to compete amongst each other to best identify the range of input features. Categories of data are found, which leads to a network that eventually fires a different output neuron for each recognised category of input pattern.

Unsupervised systems gradually detect characteristics, learnt by matching familiarity criteria. Euclidean distance (Manly, 2000; Haykin, 1999) is often used to evaluate vectors describing an input pattern against patterns seen by the system in the past. Naturally, rare inputs have less effect on network learning than those that occur frequently. This aspect influences the recognition process by suggesting that if all data is presented to the neural network, the network will find an optimised representation of the input data. The EMS uses the neural network to identify outliers from the original environmental data set. The use of delay characteristics focuses the analysis away from the raw data sets, where outliers are few, to one where outliers are the focus.

The euclidean distance between a pair of *m*-by-1 vectors $\mathbf{x}_i$ and $\mathbf{x}_j$,where each vector equals $[x_{i1}, x_{i2}, \ldots, x_{im}]$ is defined by:

$$d(x_i, x_j) = \|x_i - x_j\| \qquad \textit{Eq. 5.5}$$

$$= \left[\sum_{k=1}^{m} (x_{ik} - x_{jk})^2\right]^{1/2}$$

A technique called *winner takes all learning* exists where the node identifying the input pattern the best is adapted to more closely resemble the input data (vector). The winning (node) is usually chosen using a method like the euclidean distance metric, between the input and the *winner.*

The winner moves a fraction of the distance between it and the input vector, usually defined by a *learning constant* α. The winner moves towards the input vector, whilst the remaining neurons in the network are left unaffected. To prevent endless learning within the network, learning is slowed by reducing α monotonically. This brings the network to a gradual rest, with the neurons representing the input data presented to the network.

A modification to this learning rule is one where the weights of both winning and losing neurons are adjusted in proportion to their level of response to input data. This technique is useful when more subtle learning is required in the case when clusters are hard to distinguish (Zurada, 1992).

A problem can occur when using a winner-take-all learning network. Neural nodes can get attracted to isolated regions within the data set, also known as *local minima* (Ripley, 1996). Local minima are located when the error function (often the euclidean distance between the model neuron and the input is used), identifies what appears to be an optimisation after a few iterations. Further improvements can be made to the winner-take-all learning method by allowing a number of nodes to be defined as winners and moving these in proportion to closer resemble the input, also called *multiple-winner unsupervised learning*. This reduces the likelihood that false optimisations of the underlying data are found. The self-organising map (SOM) possesses these characteristics.

5.5.2 The Self Organising Map (SOM)

Self-organising maps (SOMs) were introduced by Kohonen in 1982 (Kohonen, 1982). The SOM technique requires no prior knowledge of data groupings or clustering. The technique has been investigated and discussed by Simula *et al.* (1999), Kangas (1996) and Vesanto (2000) among others. Kohonen *et al.* (1996) describe the use of SOMs in various engineering applications. The basic SOM developed by Kohonen (1982, 1987) has been continually adapted (Kohonen, 1998; Kohonen, 2006). Tamminen *et al.* (2000) consider a method for health monitoring where a SOM is utilised to combine dynamically classified *health levels*. The monitoring period covered eight weeks of physical measurements and diaries recorded in a home environment, by four test subjects. Use of a SOM found that there were some structures as well as differences between the weekdays and weekend, and that physical activities had a much stronger effect on health levels, than mental stress states which showed no clear clustering.

Oyanya *et al.* (2005) conclude after their study using Graphical Information Systems (GIS) and SOMs in a hybrid approach, that analysing spatially-oriented biomedical data and SOM as an analytical technique provides a useful exploration tool to support the formulation of new study hypotheses regarding the spatial distribution of a particular

disease. The study focused specifically on the use of adult asthma data, and gives an example of how SOMs can be applied in the problem domain.

Step two (*Predictor Identification*) focuses on the processes of extracting appropriate data for analysis, and the automatic identification of validated predictors of asthma attacks. It has been established through research (Ripley, 1996; Ritter, 1992) that the use of Self -organising Maps (SOMs) is a recognised technique in the field of pattern recognition.

The SOM technique satisfies the requirements of the EMS in several ways. It has the inherent ability to adapt to and recognise reoccurring patterns. Further, it can be used to fulfil the requirements of *Step 3* (Section 3.5.3) *Monitoring for the Environmental Predictor*, by monitoring closely related identified predictors.

Literature review has found papers related to pattern recognition in medical diagnostics (Dokur *et al.*, 1997; Schizas *et al.*, 1992), more specifically Taibi *et al.* (1992) use SOMs. Spencer *et al.* (1997) show the use of self-organising discovery in intensive care, while Pradhan *et al.* (1996) discuss the use of neural networks to detect seizure activity from EEG data, and Kolehmainen *et al.* (2000) use SOMs to predict air quality.

Data is often non-linear, noisy, and contains contradictory values making it harder to model and predict. The Self-organising Map (SOM) is a neural network approach where the network adapts to recognise common input signals. The adaptation occurs as a result of the neural network gradually increasing the selectivity of each individual neuron during the course of the learning process (Ritter, 1992). The size of the neural neighbourhoods reduce automatically as the process continues. This ultimately ensures that the neighbourhoods do not overlap. However, as Rojas (1996) points out, "Such an overlap could only be suppressed by acquiring more information". The effect of this is to validate features within the input space that are most prominent without identifying features that are described as *local minima.*

The incoming input signal can be transformed into a mapping that generalises the input signal into a number of previously unknown classifications. However, Ritter *et al.* (1992) state that the frequency with which a signal occurs is not always an indication of its importance, and that the SOM can increase its attentiveness to certain signals (using *a priori* knowledge) by adjusting the size of the learning step according to a previously

defined rule. The EMS increases the attentiveness of the neural network by pre-processing environmental data into delay characteristics that focus analysis by the SOM algorithm onto outliers within the environmental data set; the delay characteristics having been identified using FDA reference datums.

Table 9 Advantages of neural networks

Advantages	***Disadvantages***
Handle noisy or missing data.	No statistical definition of the system.
Work with a large number of variables.	Often slower than a straight forward statistical approach on batch data.
Can be used to investigate non-linearities.	The SOM reacts to the last piece of information.
Provides general solutions with good accuracy.	Does not record a history of adaptation, so is unable to give a statistical indication of reliability.
Able to adapt over time (in real-time) to new patterns.	
Will respond to exceptional input values by setting up a separate neuron.	

The SOM works by ordering the input data into an *n*-dimensional mapping that takes the form of the input data. For example, taking a time series sample for one of the air pollution measurements can be defined as a vector in the form,

$$V_x [i_1 \ldots i_n]$$

Using values of nitrogen dioxide at 6 hour intervals for 2 days would look like,

$$V_x [\,49,\ 28,\ 35,\ 24,\ 37,\ 29,\ 35,\ 39\,]$$

where each element of data is linked with a date, and taken at a regular interval. Other attributes such as location, data type, and time information are used to choose and filter the data sets.

Each constituent neuron belonging to the neural network models the input data, making a generalisation of it through an internal vector known as a *weight*. The neural weights inside the SOM are chosen so they have the same dimensions as the input vectors (giving each neuron an *n*-dimensional weight). Initialisation of the neural weights can be achieved

through evaluation of the input data distribution, but can also be completely random. Particular accuracy is not required; the primary restriction on weight choice, is that no two (n-dimensional) weights can contain the same ordered values. The input vectors (also known as the *input space*) are then presented to the neurons of the SOM network. The SOM process iteratively moves the neural weights to generalise the vectors of the input space.

The SOM uses weights that are adjusted to become more like the incoming patterns (Picton, 2000; Kohonen, 1987). The neural weights are adjusted in accordance with the rules outlined in the steps below, and after training, the full neural network becomes representative of the total distribution of input data.

The learning algorithm for the SOM network is described by Rojas (1996) as:

Start: The n-dimensional weight vectors $w_1, w_2, \ldots, w_m$ of the m computing units are selected at random. An initial neighbourhood radius r, an initial learning constant η and a neighbourhood function ϕ (the EMS uses a gaussian function) are selected.

Step 1: Select an input vector ξ.

Step 2: The neuron k with the maximum excitation is selected (that is, for which the distance between w_i and ξ is minimal).

Step 3: The weight vectors are updated according to the neighbourhood function, and update rule; $w_i \leftarrow w_i + \eta\phi(i,k)(\xi - w_i)$, for $i = 1, \ldots, m$.

Step 4: Stop if the maximum number of iterations has been reached; otherwise modify η and ϕ as necessary and continue with *Step 1*.

The modifications of the weight vectors (in *Step 3*) attract them in the direction of the input ξ. The advantage gained through the use of self-organising maps is one of classification. SOMs arrange the neural network into an ordered representation of the data. Neurons representing similar patterns will be positioned closer together within the map. This technique allows the generalisation of areas of the map, for example if it were identified that a particular neuron represented a pattern that led to an asthma attack, the surrounding neurons would also likely represent similar patterns leading to an attack (to varying degrees). The visualisation of this is an additional process to the core SOM algorithm and is not considered by this thesis. However inference can be obtained through the use of euclidean metrics.

An important property of the SOM is that it tends to represent patterns that are more commonly presented to it. This means that more neurons can be automatically assigned (and thus be more sensitive) to the types of patterns that are likely to be clustered closely together, and actively represent patterns leading to asthma exacerbations.

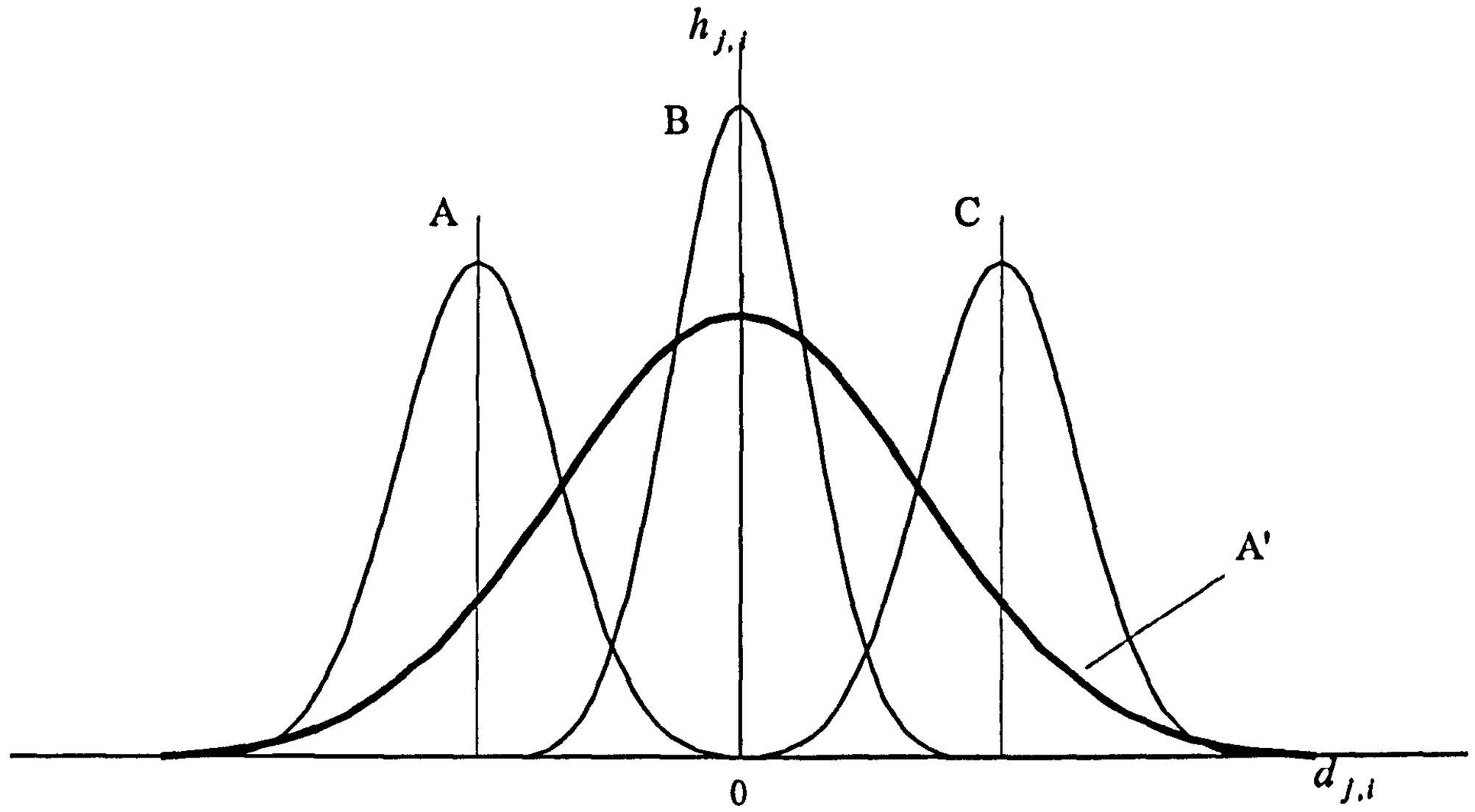


Figure 45 Neighbourhoods of the underlying neurons, where $d_{j,i}$ is the distance between the input and the neuron, and $h_{j,i}$ is the activation of the neuron. See Appendix N for further details.

Figure 45 represents three neuron neighbourhoods, of three separately identified patterns (A, B, and C). If the neural network uses a learning method that grows the number of neurons in the network, then the widest (gaussian A') neighbourhood will be the original, and in theory represents all possible values that the neural network could recognise. The three smaller curves, two of which would have been added later in the identification process, represent neurons that have been trained to recognise certain inputs over time. The neighbourhood function of the neuron should not be confused with the distribution of the input data that each neuron has reacted and identified. The neighbourhood function is used simply by the SOM algorithm as a means of deciding the activation level of each neuron in the network. As the neuron becomes trained (over time) on a particular input pattern, the neuron becomes less responsive to data outside the *range* of the neuron's neighbourhood by shrinking its width (σ).

5.5.3 Deficiencies of the Self Organising Map

During the research it was recognised that there were some deficiencies in the standard SOM algorithm. Oyanya *et al.* (2005) also noted that the SOM algorithm has a number of efficiency and convergence issues that need to be addressed. These issues are related to:

1. Speed and quality of clustering,
2. Control over the number of output neurons,
3. Updating procedure for the output neurons, and
4. Learning rate in the SOM model.

Oyanya *et al.* (2005) proposed to make some mathematical improvements to the SOM algorithm, using the above list as a basis, but do not define how the improvements would be made. A general criticism from Sarle (1994) is that neural networks are a learning technique of statistical estimation often using algorithms that are slower and less effective than algorithms used in statistical software. Haykin (1999) also makes reference to deficiencies in the SOM algorithm, specifically relating to neural network over-fitting (or over representing) data, and the lack of a neural memory; making statistical analysis of the data presented to the system difficult. Haykin (1999) suggests that the SOM algorithm fails to provide a faithful representation of the probability distribution that underlies the input data. This failure is due to the algorithm's inability to record the changes that are made to each neuron. The lack of neural memory means that the EMS would be unable to provide an indication of the probability that the pattern may exist. A further criticism of the SOM algorithm is that although the neighbourhood of a particular neuron is defined by a gaussian distribution, that distribution is controlled essentially by the number of iterations that have been performed, missing out on the potential for modelling the underlying data with a probability distribution. The neighbourhood defined by the neuron does not reflect the distribution contained in the underlying data. The fact that the distribution of the underlying data is not considered means that it is more difficult to establish a statistical basis for the results produced by the SOM algorithm. Robb (2001) found that for most datasets, neurons tend to oscillate around the input space and do not converge to a single equilibrium position. However, it was found that by using an error term the learning algorithm could be terminated when minimisation had occurred.

5.5.4 Use of the Self Organising Map by the EMS

Two important characteristics belonging to the Self-organising map algorithm, and used by the EMS are:

1. the technique's ability to converge onto an optimum solution, for a given set of input data – recognising environmental predictors of asthma exacerbation, and
2. its real-time ability to recognise when an environmental predictor has re-occurred, so that an alert can be triggered.

The technique also has several characteristics, that are not explored within this thesis. These are the use of information held by the network regarding environmental characteristics closely related to identified optimum patterns identified by the SOM algorithm. These closely related patterns are often partially activated when neighbouring neurons are activated as the best match to input data. These closely related patterns may hold further information concerning the patient's asthma exacerbants.

Visual representation of the self-organising map is not used by this thesis, owing to the techniques involved falling into the category of data dissemination. Visual representation of the SOM requires human intervention and interpretation, which does not lend itself to automation. However, it is recognised that as research tool, visualisation of the SOM would be beneficial to the EMS. Methods of vector projection and SOM visualisation are discussed by Himberg (1998) and Vesanto (1999). They both provide useful discussion over a number of techniques, namely Sammon's projection (Sammon, 1969), the U-Matrix (Ultsch and Siemon, 1990), and to a lesser extent, Curvilinear component analysis (Demartines and Hérault, 1997). Ontrup and Ritter (2001) suggest visualisation of the SOM using a hyperbolic space, and adapt the standard SOM algorithm to produce a projection of the input space using hyperbolic distance metrics, reflecting hierarchical structures within the data. All these methods improve the algorithms usefulness, however, only provide a visual representation of the identified patterns. The patterns still require interpretation, however powerful the visualisation techniques may be.

There has been some debate as to whether or not it is necessary to normalise input data to the SOM. This question is investigated in Chapter 6 (Section 6.9), and finds that normalisation of the input vectors assists in the identification of extreme values that are found within the data set.

5.6 Frequency, Boundary and Cluster Analysis (FBCA)

Whilst a neural network approach provides a technique for locating the centre of clusters and generalising the underlying data distribution, the method is difficult to control. The nature of the neural network creates two issues:

1. Setting the initial number of neurons if the network is activated with a fixed number.
2. Setting a *stop* condition if network growth is automated, starting with one or more neurons.

In order to address these deficiencies a second method of analysis was devised from original thought and building blocks, notably FDA formed during earlier research. This section outlines a set of techniques that have been designed to function as an EMS module. The module combines three processes that have been developed as an extension to the SOM methodology. The three techniques presented here are:

1. Frequency Analysis
2. Boundary Analysis, and
3. Cluster Analysis

These techniques are combined into a component providing a means to extract natural data clusters with a popularity indication. For the remainder of this thesis the technique has been termed Frequency, Boundary and Cluster Analysis (*FBCA)* by the author.

Figure 46 shown below, illustrates the subsections of the pattern identification module including FBCA, and other components such as FDA (*Feature Detection Analysis*), described in Section 5.2. and the *Hypothesis Builder*, described in Section 5.4.

The data storage, and dissemination modules, which are not developed further by this thesis, are shown to provide context. The implementation of the search mechanism to identify patient matched, environmental data is complex and computationally demanding. Therefore research has focused on the ability to incorporate this component into the system's architecture rather than implementing it. The *Air Quality Locator* component (shown in Figure 46) represents this functionality.

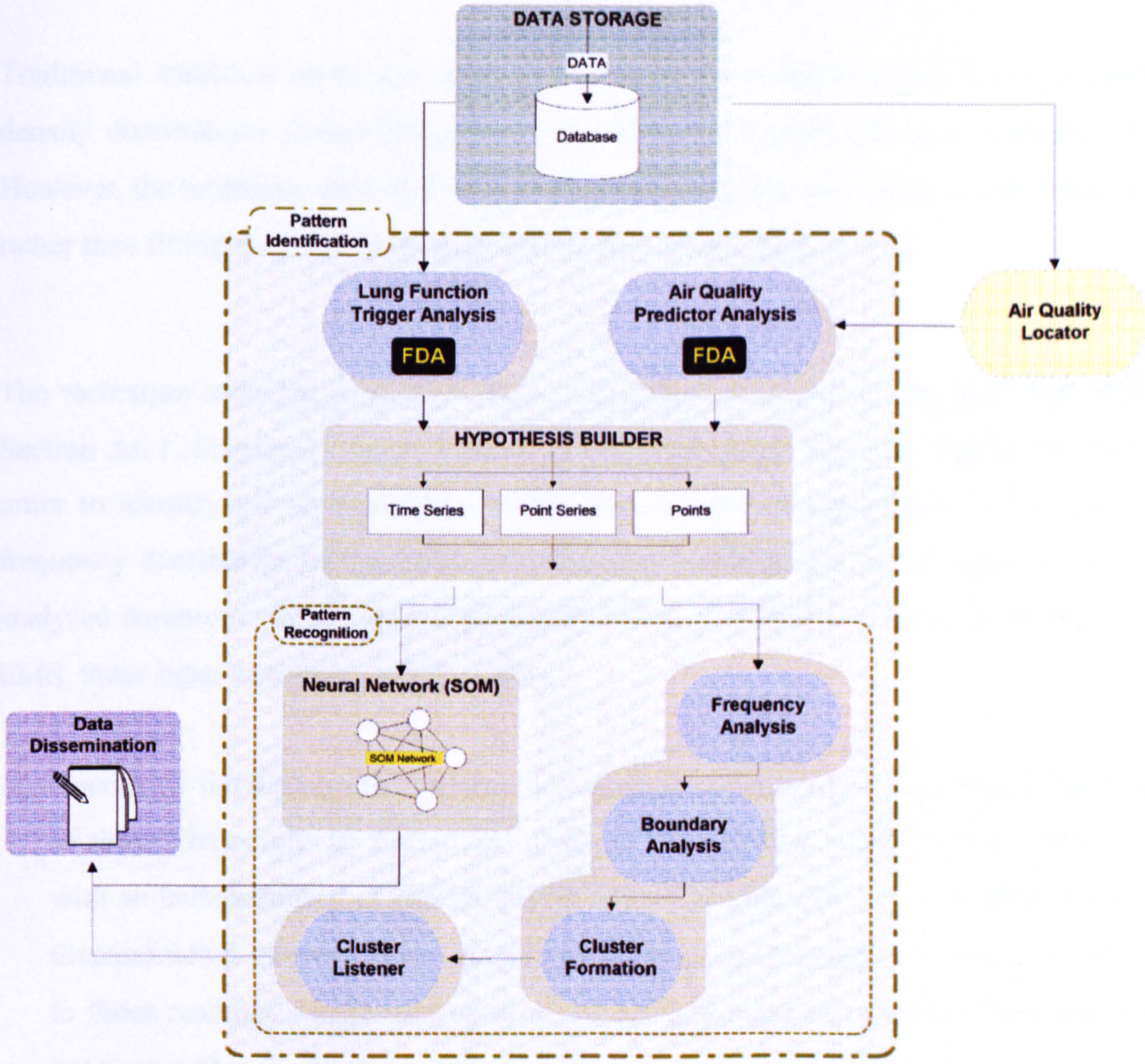


Figure 46 Pattern Identification Module Architecture (also showing data storage and dissemination).

Frequency, Boundary, and Cluster Analysis (FBCA) is designed to use the distribution of the underlying data to identify significant data classifications. Common value traits can be grouped together in order to identify both abnormal and likely event triggers. The analysis leads to an indication of the number of clusters required within the neural network. The difference between the two methods is the process by which commonly occurring patterns are identified. SOMs *'home in'* on troublesome patterns over a period of learning, whilst FBCA sets the potential cluster boundaries (and the number of clusters) by examining the statistical distribution of batch data and then proceeds to identify which clusters are *particularly active*. The usefulness of operating two supporting methods together in the same system was highlighted during evaluation where test results were corroborated by both processes. The implementation of both methods into a hybrid system increases reliability. Pattern recognition by its very nature is about removing as much uncertainty from the result as possible.

Traditional statistical techniques attempt to model the underlying data using probability density distributions (Salgado-Ugarte *et al.*, 2000), or Kernel estimates (Molina, 1994). However, the technique used by FBCA *samples* the raw data, and locates range boundaries, rather than fitting the underlying data with a distribution (Section 5.6.2).

The technique achieves *frequency analysis* deploying the FDA component, described in Section 5.6.1. Frequency analysis involves the dissection of the data signal into *bins*, in order to identify the *normal* range, outliers and general characteristics of the data. The frequency distribution of the input variables are plotted graphically in order to split the analysed parameters into logical ranges (referred to as buckets or bins). In relation to the EMS, these input variables could be either:

a) A series of fixed frequency environmental readings collected over a significant period of time. The results of *Frequency Analysis* on this data would provide clinical staff with an understanding of the full distribution of a particular input variable and enable discrimination between those environmental readings considered normal, as opposed to those readings which only occurred in exceptional circumstances. However this is not used within the EMS as analysis is focused on *delay characteristics*.

b) A batch of delay characteristics, where a *delay characteristic* is defined as a combination of two readings: i) an environmental reading (more specifically, a reference datum), combined with ii) a time *Lag*, defining the time delay between the time the asthma exacerbation took place, and the time at which the prior environmental *predictor* occurred.

The important point to note here is that analysis undertaken by the EMS at this stage of the process, is focused purely on the parameters of the delay characteristic. The original environmental and respiratory data sets are irrelevant to the analysis.

5.6.1 Frequency Analysis

Once the delay characteristics have been identified the next step is to plot the frequency distribution of each parameter (*PM2.5 Date, Value,* and *Lag*). In order to plot a frequency distribution the range of values for each parameter must be split into *buckets*.

The process for obtaining a satisfactory distribution during the frequency analysis is reliant on determining an appropriate width of bucket according to the nature of the underlying data. For example, a value (of SO_2) that fluctuates between 10 and 60*ppb* could be divided into 5 bands, each with a width of 10*ppb*, in order to capture an approximate distribution of the data. The width of buckets is often dependent on the number of data readings.

Once the frequency analysis is completed the frequency of values contained in each consecutive bucket are plotted (Figure 47) giving a frequency analysis of the underlying parameter's data, for analysis using FDA.

5.6.2 Boundary Analysis

The process of *Boundary Analysis* breaks the distribution of each parameter into identifiable sections. For example, a parameter with a bi-modal distribution will be split into its two constituent parts (or a tri-modal distribution into three parts). This process is achieved by applying the Feature Detection Analysis technique to the frequency distribution of the variable, identifying peaks and troughs and *fixing* cluster boundaries (the troughs between each peak) as shown in Figure 47 by the black vertical lines.

Analysis Options

Parameter	Sensitivity	Bucket Size	Analyse
PM2.5 - Value	70	2	☑
PM2.5 - Lag	70	3600000 (1 Hour)	☑
PM2.5 - Date	70	86400000 (1 Day)	☐

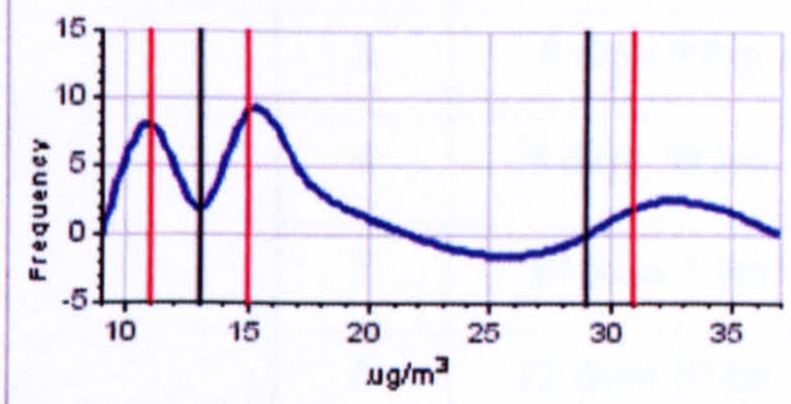
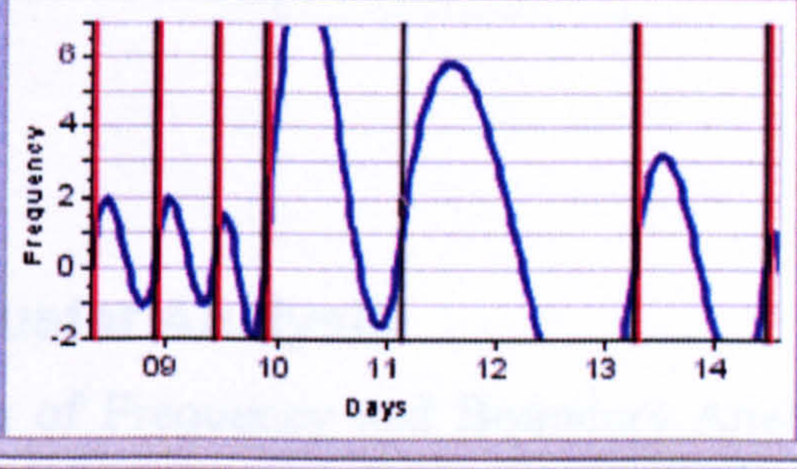
ID	Analysis	Raw Data
PM2.5-Value		{7.0=0,10.0=8,13.0=7,16.0=9,19.0=2,28.0=0,31.0=2,37.0=0}
PM2.5 - Lag		{627840000=0,631440000=1,678240000=0,681840000=1...

Figure 47 Frequency and Boundary Analysis are shown with the bucket widths given at the top of the figure. Parameters "PM2.5 - Value" and "PM2.5 - Lag" are analysed. Raw data for the "PM2.5 - Lag" data set are shown in milliseconds.

It can be seen that the distribution of the selected *$PM_{2.5}$ Values* (in Figure 47) have a tri-modal distribution. The result of Boundary Analysis (using FDA) has been to split the distribution up into the three separate ranges, by identifying limits, from 9.0 to 13$\mu g/m^3$, 13 to 29$\mu g/m^3$ and from 29 to 37$\mu g/m^3$.

It should be noted that for visualisation reasons the scale for *Lag* has been shifted forward by one day. The graph uses a date axis, and therefore starts at 01. Hence the trough shown at 08 days 20 hrs on the graph actual occurs at 07 days 20 hrs. The distribution of the *$PM_{2.5}$ Lag* shows ranges, described in Table 10.

Table 10 PM2.5 Lag range limits

Range	***From***	***To***
1	7 days 6 hrs	7 days 20 hrs
2	7 days 20 hrs	8 days 9 hrs
3	8 days 9 hrs	8 days 20 hrs
4	8 days 20 hrs	10 days 2 hrs
5	10 days 2 hrs	12 days 10 hrs
6	12 days 10 hrs	13 days 13 hrs

5.6.3 Cluster Analysis

The action of Frequency and Boundary Analysis has been to identify three $PM_{2.5}$ *Value* ranges and six $PM_{2.5}$ *Lag* ranges. This identifies eighteen (3 value x 6 lag) cluster combinations.

The clusters are defined in terms of n-dimensions, where n depends on the number of parameters analysed (in this case *two*). The clusters are formed after defining all possible permutations of identified boundaries (Figure 47 above), Table 11 sets out all eighteen possible combinations.

Table 11 Cluster matrix table, showing the range for each cluster.

	PM2.5 – Value ***($\mu g/m^3$)***	***PM2.5 – Lag (Days:Hrs)***
Cluster 1	(9.0 – 13.0)	(7 Days 6 Hrs – 7 Days 20 Hrs)
Cluster 2	(13.0 – 29.0)	(7 Days 6 Hrs – 7 Days 20 Hrs)
Cluster 3	(29.0 – 37.0)	(7 Days 6 Hrs – 7 Days 20 Hrs)
Cluster 4	(9.0 – 13.0)	(7 Days 20 Hrs – 8 Days 9 Hrs)
Cluster 5	(13.0 – 29.0)	(7 Days 20 Hrs – 8 Days 9 Hrs)
Cluster 6	(29.0 – 37.0)	(7 Days 20 Hrs – 8 Days 9 Hrs)
Cluster 7	(9.0 – 13.0)	(8 Days 9 Hrs – 8 Days 20 Hrs)
Cluster 8	(13.0 – 29.0)	(8 Days 9 Hrs – 8 Days 20 Hrs)
Cluster 9	(29.0 – 37.0)	(8 Days 9 Hrs – 8 Days 20 Hrs)
Cluster 10	(9.0 – 13.0)	(8 Days 20 Hrs – 10 Days 2 Hrs)
Cluster 11	(13.0 – 29.0)	(8 Days 20 Hrs – 10 Days 2 Hrs)
Cluster 12	(29.0 – 37.0)	(8 Days 20 Hrs – 10 Days 2 Hrs)
Cluster 13	(9.0 – 13.0)	(10 Days 2 Hrs – 12 Days 10 Hrs)
Cluster 14	(13.0 – 29.0)	(10 Days 2 Hrs – 12 Days 10 Hrs)
Cluster 15	(29.0 – 37.0)	(10 Days 2 Hrs – 12 Days 10 Hrs)
Cluster 16	(9.0 – 13.0)	(12 Days 10 Hrs – 13 Days 13 Hrs)
Cluster 17	(13.0 – 29.0)	(12 Days 10 Hrs – 13 Days 13 Hrs)
Cluster 18	(29.0 – 37.0)	(12 Days 10 Hrs – 13 Days 13 Hrs)

Each cluster combination is then activated and ready to identify any input that satisfies the range represented by it. When a cluster identifies input, it is recorded as a hit. This is shown in Figure 48.

The analysis borrows a technique from the domain of neural networks, where each cluster *listens* to the input data stream and records any input that is relevant to the cluster. If the input vector matches the internal representation of the cluster (is within each dimension's bounds) a *hit* is recorded.

Figure 48 shows an example of the clustering technique. In practice the technique would be expected to receive many more input vectors, which would also be required for validation of the cluster. The number of hits shown in Figure 48 would not be sufficient to validate any of the clusters. The number of *hits* indicating that a particular cluster is valid, would need to be significant (a minimum of at least 20 *hits*). Calculations based on the Chi-Square Distribution (See *Appendix S*) indicate that a minimum sample size of 20 data elements would be required to be 99% confident that an observed bi-modal distribution was statistically significant. Although not included in the prototype, a future development could be the use of distribution testing techniques within the FBCA module, to ensure that clusters were defined on a fully statistical basis.

Input Vectors

☑ Monitor Hits | Present Vectors to Clusters | Clear Table | Reset Hits | View Selected Cluster

PM2.5 - Value	PM2.5 - Lag	Hits
(9.0 - 13.0)	(7 Days 6 Hrs - 7 Days 20 Hrs)	
(13.0 - 29.0)	(7 Days 6 Hrs - 7 Days 20 Hrs)	1
(29.0 - 37.0)	(7 Days 6 Hrs - 7 Days 20 Hrs)	
(9.0 - 13.0)	(7 Days 20 Hrs - 8 Days 9 Hrs)	
(13.0 - 29.0)	(7 Days 20 Hrs - 8 Days 9 Hrs)	1
(29.0 - 37.0)	(7 Days 20 Hrs - 8 Days 9 Hrs)	
(9.0 - 13.0)	(8 Days 9 Hrs - 8 Days 20 Hrs)	
(13.0 - 29.0)	(8 Days 9 Hrs - 8 Days 20 Hrs)	1
(29.0 - 37.0)	(8 Days 9 Hrs - 8 Days 20 Hrs)	
(9.0 - 13.0)	(8 Days 20 Hrs - 10 Days 2 Hrs)	2
(13.0 - 29.0)	(8 Days 20 Hrs - 10 Days 2 Hrs)	1
(29.0 - 37.0)	(8 Days 20 Hrs - 10 Days 2 Hrs)	
(9.0 - 13.0)	(10 Days 2 Hrs - 12 Days 10 Hrs)	1
(13.0 - 29.0)	(10 Days 2 Hrs - 12 Days 10 Hrs)	1
(29.0 - 37.0)	(10 Days 2 Hrs - 12 Days 10 Hrs)	
(9.0 - 13.0)	(12 Days 10 Hrs - 13 Days 13 Hrs)	
(13.0 - 29.0)	(12 Days 10 Hrs - 13 Days 13 Hrs)	1
(29.0 - 37.0)	(12 Days 10 Hrs - 13 Days 13 Hrs)	

Figure 48 Available clusters and their recognition of input vectors, represented by "hits".

Figure 49 shows the contents of *Cluster 10* with two hits. *Cluster 10* is the most popular trait identified from this very small data set.

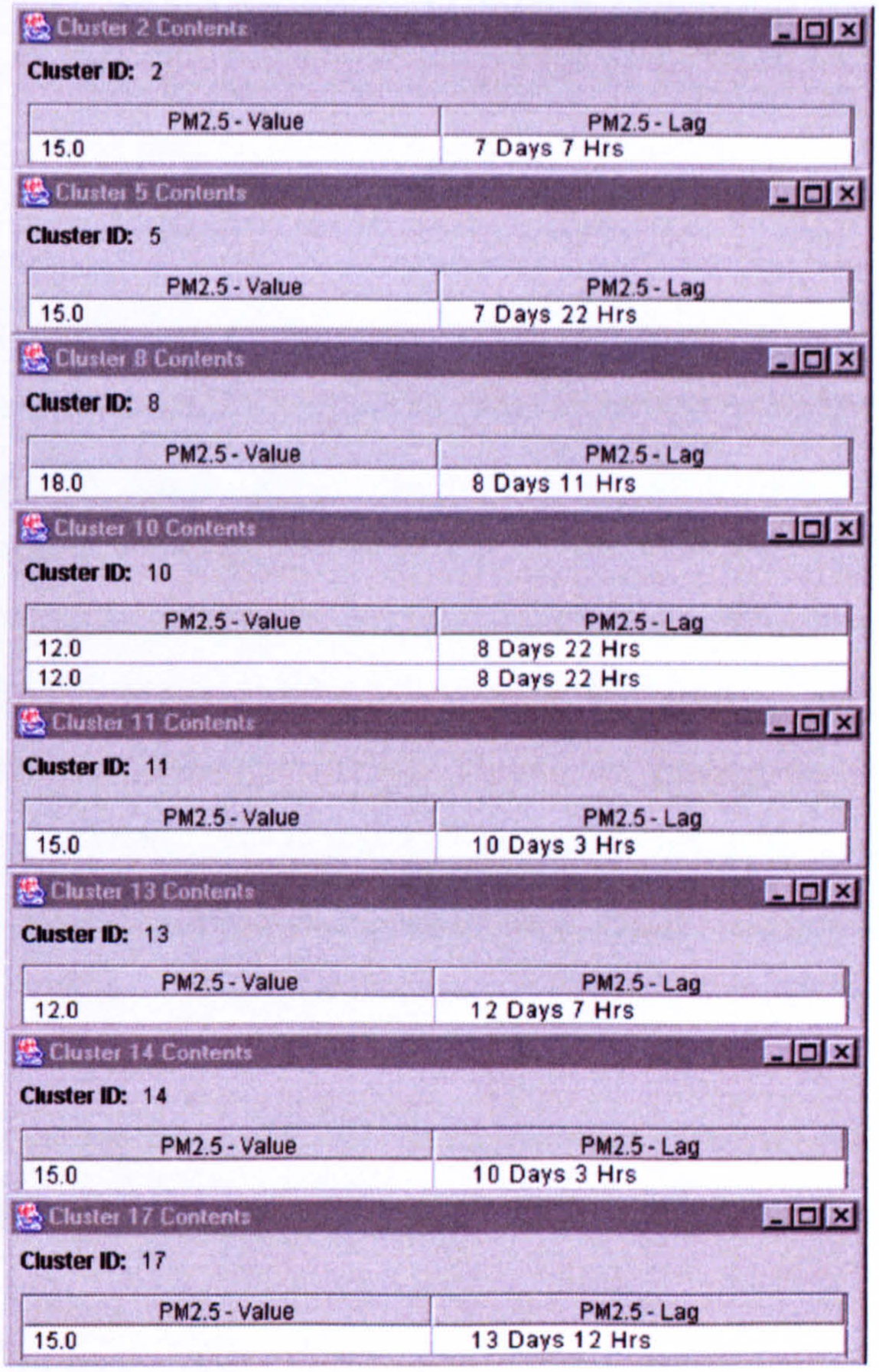


Figure 49 Cluster contents.

5.6.4 Advantages of the analysis

Breaking the original data set into a number of bands, allows the distribution of values to be analysed. The width of the bands must be suitable for the type of data being studied, for example the value of particulate matter could fluctuate between 15 and 35$\mu g/m^3$ therefore by choosing an arbitrary value of 10 *buckets*, the bucket size (for the bands) would be a value of 2$\mu g/m^3$. A bucket of this size would break the fluctuating range down into 10 bands providing data for a detailed graph showing the spread of data contained in the data set.

An advantage resulting from this analysis is that the cluster boundaries are known as soon as the FDA has been completed. Another advantage is that the width of the *buckets* used

for separating the data set into bands can be altered to reflect the data being analysed, likewise the sensitivity level of the FDA that is used to detect the boundaries can be altered.

Parameters can be discounted from further analysis if they demonstrate a uni-modal distribution. The presence of a single cluster is an indication that the parameter is unlikely to be of use in identifying a causal pattern, as there is no added benefit in analysing the parameter alongside others. A parameter type with a single cluster would appear (as a single cluster dimension) in every final cluster permutation and for this reason means that there would be no distinction between the value of that environmental parameter and cause of lung function decline. For example, a single frequency cluster for particulate matter ranging from a value of 7 to 77$\mu g/m^3$, would be particularly unhelpful as the range covers three air quality classification bands (*low*, *moderate* and *high*), and therefore needs to be split into further bands to draw a distinction between the affects of each air quality level.

However, the existence of a uni-modally distributed parameter can not be totally ruled out as being unhelpful as the patient's lung function may not decline without some presence of the parameter.

5.7 Summary

This chapter presented sections of the system architecture's data handling layer, and described how a flexible system can be built to facilitate the automatic recognition of patient-specific environmental exacerbants of asthma. The architectural design, included subsystems, their relationships, and a set of processes that are fundamental in identifying key predictive variables.

The Environmental Monitoring System (EMS) was shown to be adaptive to incorporating new analytical methods as they become available, particularly by the *hypothesis builder*. Kern *et al.* (1998) state that architecture must be flexible to allow for the use of changing technology in an architectural implementation. The concept behind the architecture purposefully allows the system to be implemented in this way, using appropriate available technologies at the time of implementation.

The architecture has been designed using a component framework, where functions are able to work together to identify when an asthma patient is likely to experience an asthma exacerbation. Due to the flexible architecture the EMS can be extended to analyse many different parameters. Figure 50 shows the workflow architecture of the EMS, giving an overview of the complete system.

Feature Detection Analysis recognises features that are naturally at the extremes of the data set (peaks and troughs). The introduction of delay characteristics (Section 3.3) focuses analyses by further analytical components to these outliers. Delay characteristics preserve the relationship between the respiratory and environmental data sets, reducing the quantity of data to a single delay and value (for each parameter). This improves scalability of the system, and simplifies analysis to the delay between the environmental predictor and time of respiratory decline, which is the information required to form an alert.

The Hypothesis Builder combines reference datums from both environmental and respiratory data sets. One Hypothesis Builder is required per monitored patient, meaning that in a large system implementation, an entire server could be dedicated to a patient if absolutely necessary.

The use of a self-organising map algorithm, is shown to be useful in identifying reoccurring patterns (Chapter 6), it is shown to be capable of: monitoring varied data types, coping with system and measurement noise, adaptation to new data trends, and classification and identification of new patterns within the data.

However additional testing and refinements are required within a fully developed clinical system to ascertain if the technique will adapt to newly identified trends on a continual basis. Integration of FBCA as an analytical component shows that additional methods of analysis can be used by the EMS, and analytical results shared with other components within the system via the *workflow manager* (Chapter 4). Analytical techniques developed during the thesis are interchangeable with new or additional techniques as they become available.

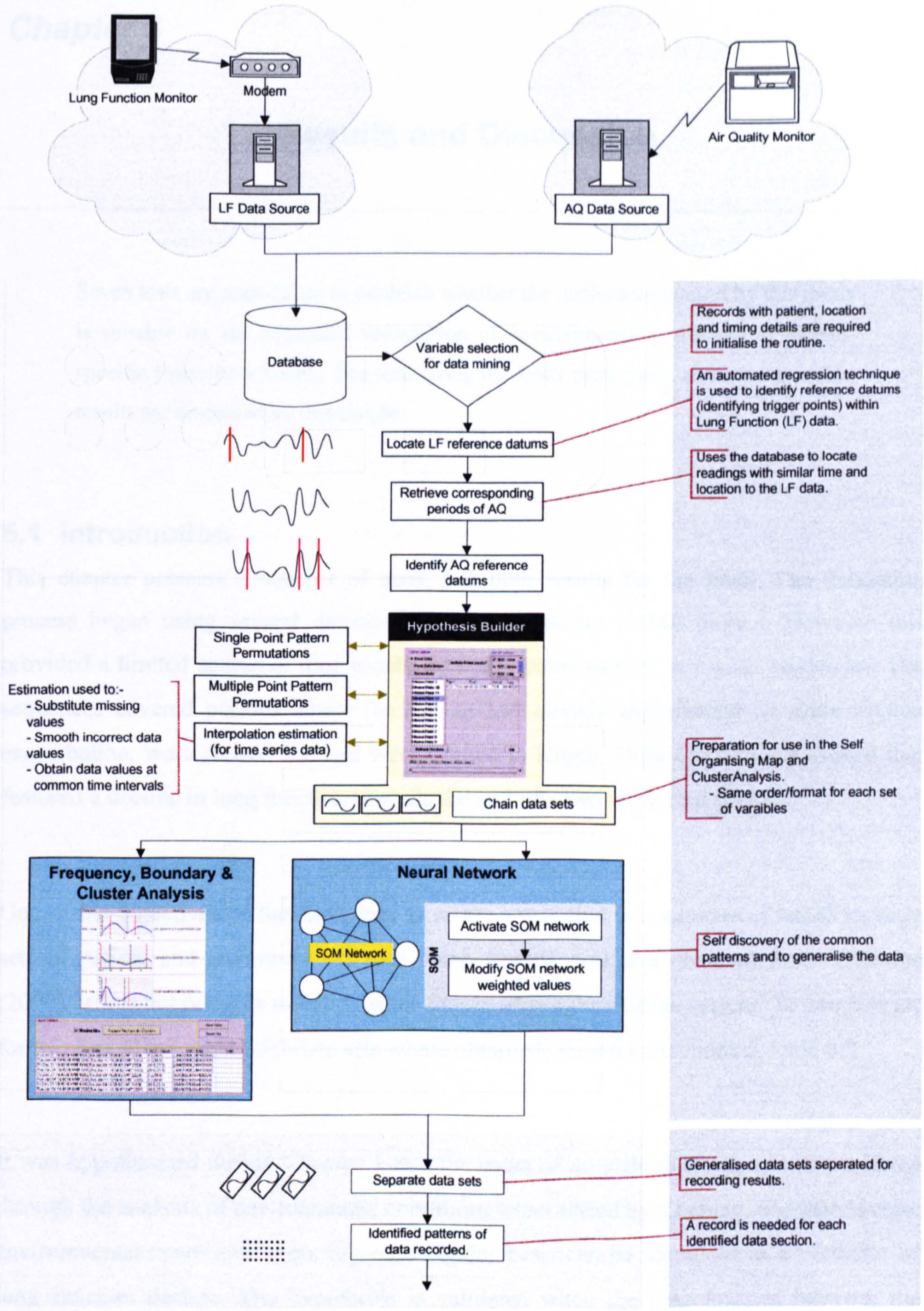


Figure 50 Workflow architecture of the system modules.

Chapter 6

Results and Discussion

Seven tests are undertaken to establish whether the methodology used by this thesis is suitable for the automatic recognition of environmental predictors of patient-specific respiratory health. The tests using the EMS prototypes, and corresponding results are discussed by this chapter.

6.1 Introduction

This chapter presents a number of tests, and their results for the EMS. The validation process began using several datasets from the Medicate (2000) project. However this provided a limited source of *nine* usable data sets, consisting of two week sequences. The sequences covered periods where the patient had already experienced an acute asthma exacerbation, were recovering, and were limited in length. Only one data set existed that featured a decline in lung function towards the end of a two week trial period.

One of the objectives of the EMS was to create a tool that was capable of handling large sets of patient and environmental data. With limited data sets obtained from Medicate (2000), it was not possible to demonstrate the scalable nature of the system. To compensate for this limitation, additional data sets where obtained; shown in Sections 6.7 and 6.8.

It was hypothesised during Chapter 1 that the onset of an asthma attack can be predicted through the analysis of environmental conditions experienced by a patient, and that adverse environmental conditions occurring on a regular basis can be identified as a predictor of lung function decline. The hypothesis is validated when the *time interval* between the environmental predictor and the asthma exacerbation (marked by a reference datum), and also known as the *Lag,* is consistently detected.

Figure 51 shows an example of the data collected during the Medicate trial, alongside a

particulate matter data set from a monitoring station in Haringey, North London. The particulate matter data set has been analysed using the FDA component, which is also marked on the chart. The red vertical lines represent the reference datums identified by the Feature Detection Analysis on the PM_{10} data.

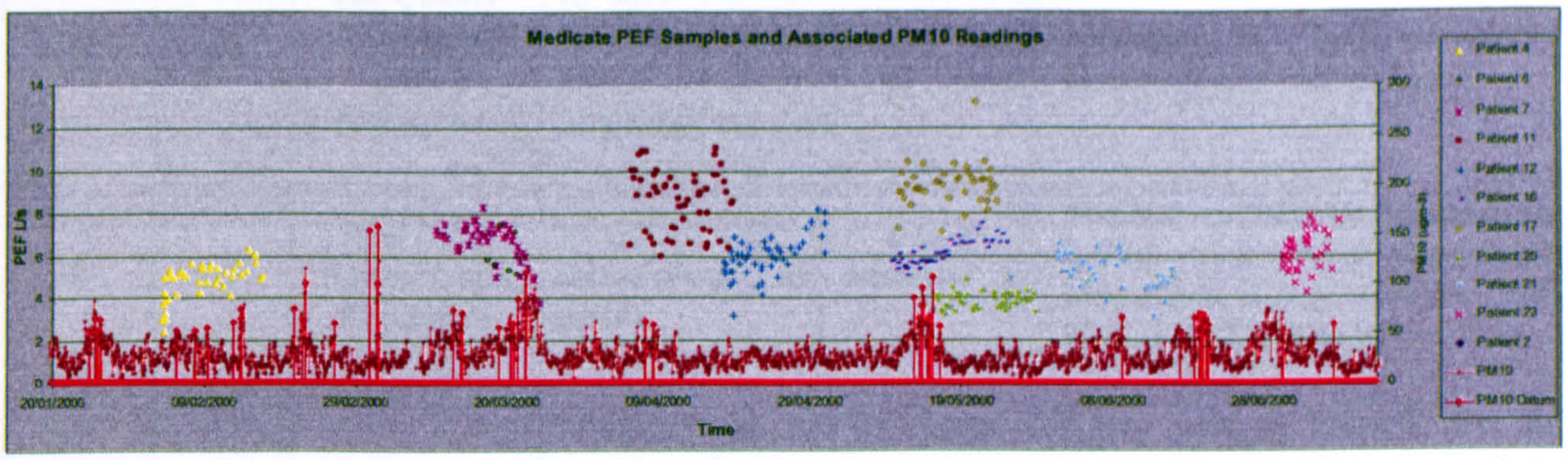


Figure 51 Lung function data sets obtained during the Medicate (2000) project and associated PM_{10} data analysed with FDA. The PM_{10} reference datums are shown by the red vertical lines.

To further validate the system's architecture, a second source of data was required that provided a data series of significant length (1 year). Twelve months of London hospital admissions data (due to asthma) were obtained from *The Information Centre* (2007), and corresponding air quality data, from four monitoring stations (Hillingdon, Brent, Marylebone, and North Kensington). The test demonstrated the use of the EMS with other types of data, and against a significant size of data set. The hospital admissions data was restricted to the regional level for location detail.

A final six month data set was obtained from an asthmatic; who took daily lung function readings, using a portable electronic lung function monitoring device. During the six month collection period the asthmatic also recorded their movements so air quality could be mapped to their specific location. This data set is shown in Figures 74 & 75. These three data sources and corresponding data sets form the main validation tools for this research. A summary of the validation process is detailed below in Table 12.

Table 12 A summary of each test presented during this chapter.

Test	*Section*	*Title*	*Purpose*	*Method*
1	6.2	Feature Detection Analysis (FDA)	To determine the accuracy of feature detection analysis when data containing an element of noise is used.	Apply FDA to a data signal with increasing levels of noise applied, to compare response.
2	6.4	Air Quality Monitoring Station Characteristics	To show the naturally occurring probability of each pollutant at a sample of London air quality monitoring stations.	Analysis of the cumulative distribution of each pollutant.
3	6.5	Real Lung Function and Air Quality	To test the EMS with data that is characteristic of expected air quality.	To use a real data set typical of air quality and a corresponding set of lung function data. Analysis with FDA, FBCA and neural analyses.
4	6.6	Multi parameter	To identify whether the system is capable of identifying known characteristics when a combination of parameters are present.	Apply the lung function data set as the second air quality parameter. Effectively making the second air quality data set highly correlated.
5	6.7	Hospital Admissions Data	To use a significant (1 year) data set to validate the system.	The EMS is designed to handle a large variety of data types, within the problem domain. Using Hospital admissions due to asthma exacerbation demonstrates the adaptability of the system to new data types.
6	6.8	Six Month Set of Patient Lung Function and Air Quality	To validate the system using a significant data set taken directly from the problem domain.	A patient specific data set of lung function and air quality was recorded over a six month period. The data sets were then analysed using the EMS and conclusions drawn.
Norma-lisation Test	6.9	Normalised, Real Lung Function & Air Quality	To test the response of the system using delay characteristics derived during test 3, once normalised	Normalised delay characteristics are analysed.

Validation of the EMS began with a test to illustrate how *Feature Detection Analysis* (FDA) is capable of identifying key features in lung function and air quality under normal conditions, and with data signals where there was noise corruption.

6.2 Validating FDA With a Signal Containing Noise

6.2.1 Creation of a Control Data Set

The control data set, which originated from a real lung function (Peak Expiratory Flow) data signal (Crabbe *et al.*, 2001) is indicated by yellow boxes in Figure 52 below. The FDA component analyses the actual data points within the time series (discussed in Chapter 5).

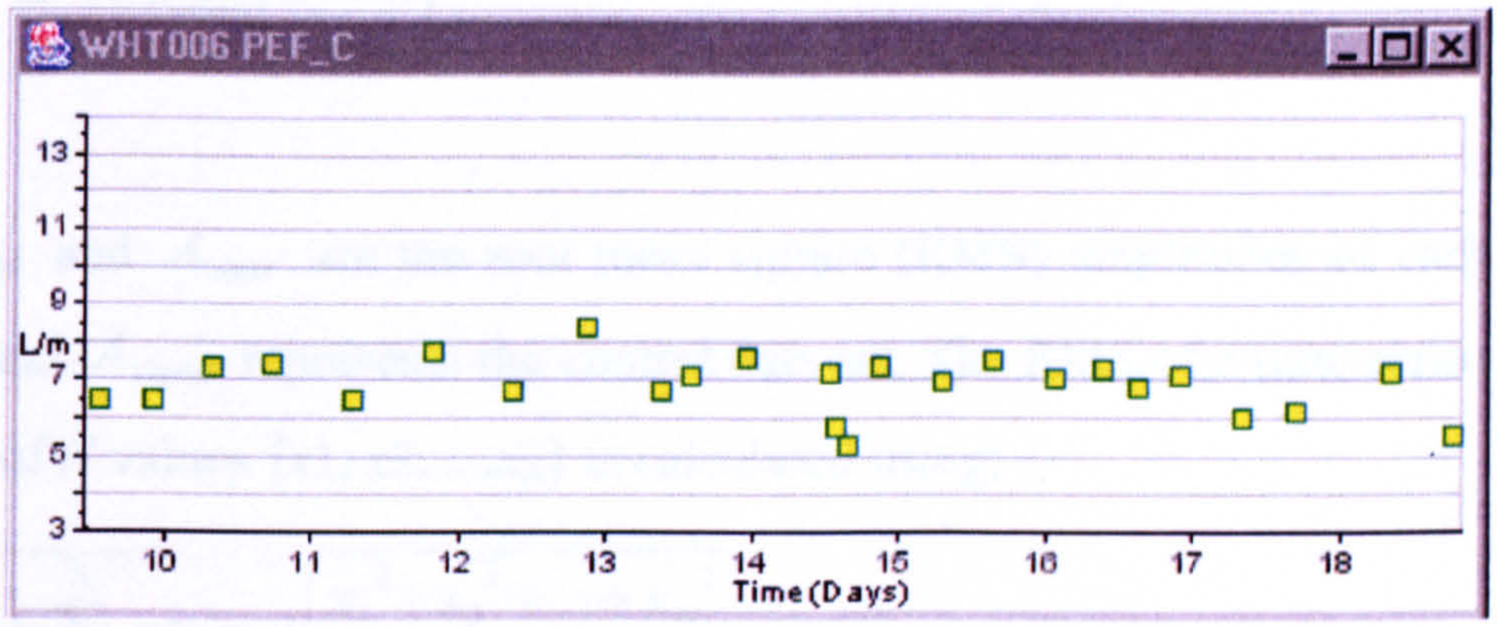


Figure 52 Visualisation by the EMS of the data set used as a control during the tests.

Four data sets containing noise and based on the real lung function time series shown in Figure 52 were created (these are shown in Figure 53). The noisy sets were created using a Gaussian random function, amplified depending on the level of noise required. A final value was then produced by adding (or subtracting, if negative) the result from the control signal. This process was applied to each data point in the control series to produce the noisy time-series signals.

6.2.2 Modelling Data Variation

Air quality data values can vary due to the type of sensors used by the London Air Quality Network (LAQN, 2008). There is an estimated uncertainty of 5 to 7 percent for NO_X at high concentrations. The LAQN, however, use a working uncertainty of ±10% for the measurements of NO_X, NO_2 and O_3 due to operational factors that can influence the value of data readings. Patient data is also subject to variation as spirometry devices designed to measure lung function can vary in accuracy depending on the design of the device. Device accuracy is generally between ±2% to ±10% (Cooper & Masden, 2000).

The first validation procedure uses one control (a sample of lung function data from the

Medicate project) and four data sets modelled from the control set, to which noise has been added to simulate variation in the data set. The signal to noise ratios (SNRs) defining each of these data sets were chosen to reflect the nature of the underlying data. Using the maximum likely inaccuracy of ±10% leads to a variance that equates to a SNR of approximately 20dB using Equation 6.1 below;

$$SNR(dB) = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \qquad \textit{Eq. 6.1}$$

where A_{signal} and A_{noise} are the root mean square (RMS) amplitudes of each respective time series and A_{signal} represents the control data set. The RMS of a time series containing a collection of N values $\{x1, x2, ...,x_N\}$ is calculated using;

$$x_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^{\,2}} = \sqrt{\frac{x_1^{\,2}+x_2^{\,2}+\ldots+x_N^{\,2}}{N}} \qquad \textit{Eq. 6.2}$$

Where N is the number of values in the time series, and x_1 to x_N are data values.

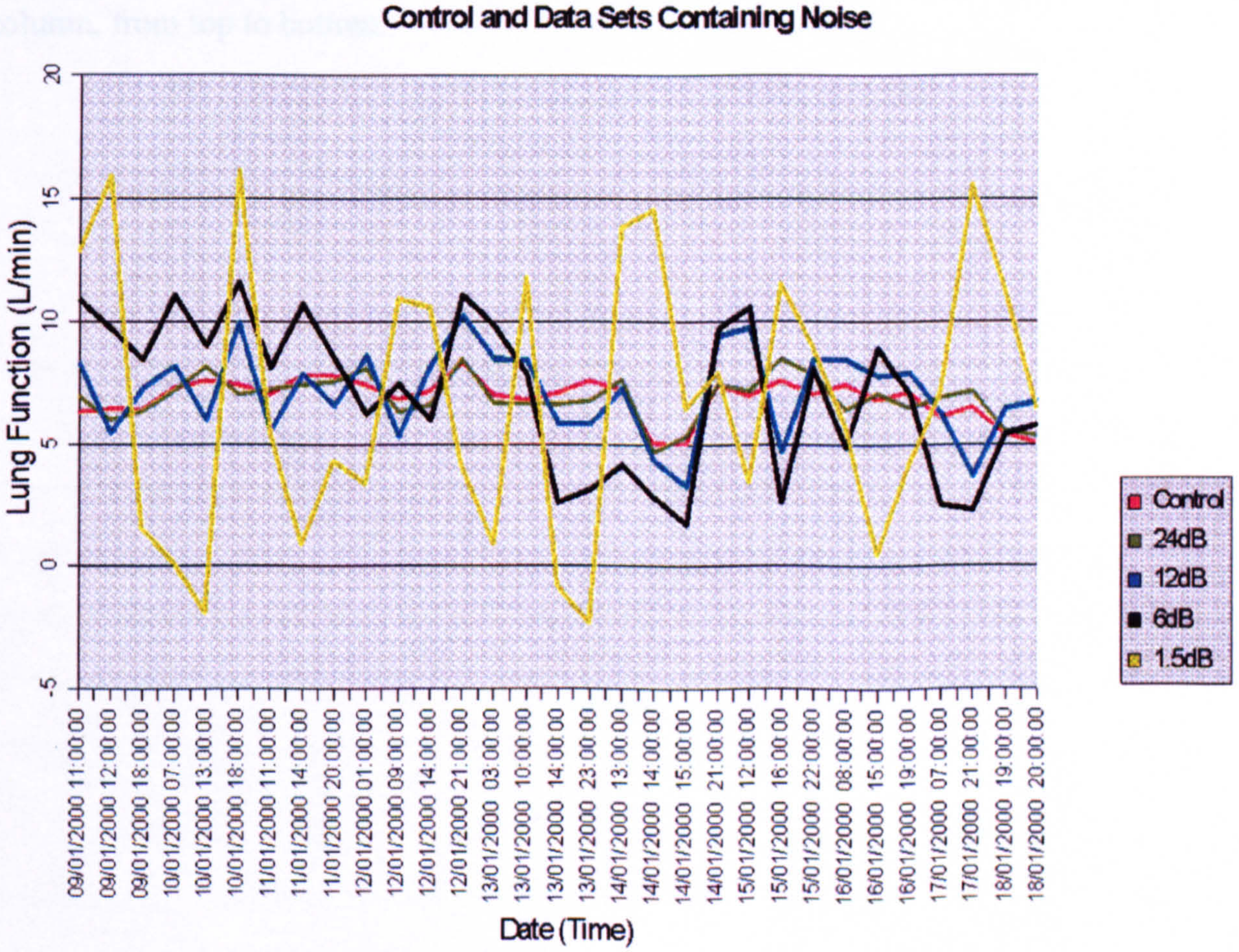


Figure 53 Control, and data sets with artificial noise.

A range of Gaussian random noise between 24dB and 1.5dB, (doubling each step) was added to the control signal (making four noisy data signals). The level of noise was chosen at 24dB in the first instance, to reflect minor inaccuracies in sensor response. Analysis of data obtained using automatic air quality monitoring sensors generally allows for ±10% error tolerance, which equates to approximately 20dB. The total noise applied to the control signal to create each of the four noisy time series signals was calculated using the signal to noise ratio (Equation 6.1).

The figures below show the results of the FDA on the control and noisy sets of lung function data created using additive noise levels of between 24dB and 1.5dB, where the noise interference on a signal is greater at a lower ratio value (e.g. 1.5dB). The identification sensitivity value used was the arbitrary value of 70%. The analysis includes all reference datums (marked in red) found in the analysis, the option to disregard values that are below a certain threshold value has not been used. The lines in green show the linear regression trend through each graph section (data peak to trough points; between red reference datum markers). The control data set has been presented twice (Figures 54 & 55) so a comparison with the sets containing noise can be made by examining the data in each column, from top to bottom.

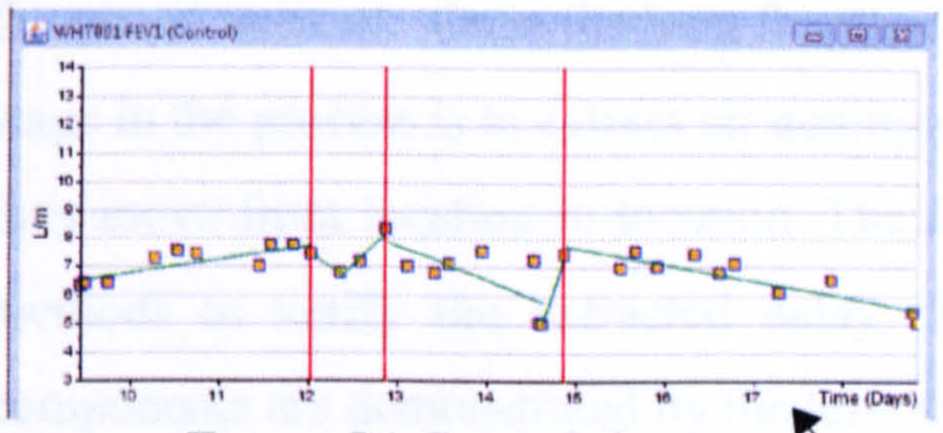


Figure 54 Control data set

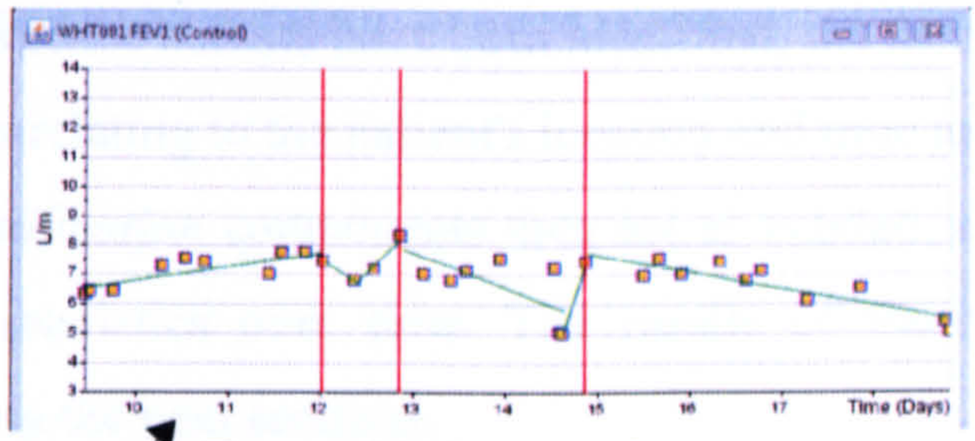


Figure 55 Control data set

same data set

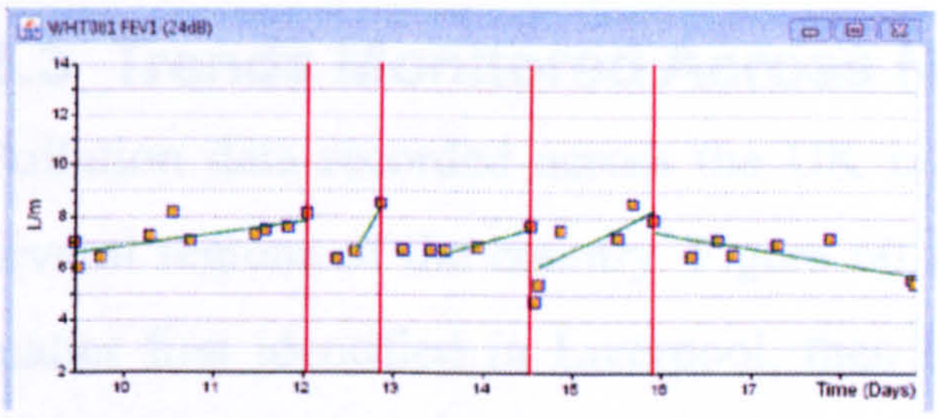


Figure 56 Control data set with additional noise at a Signal to Noise Ratio of 24dB - Set 1

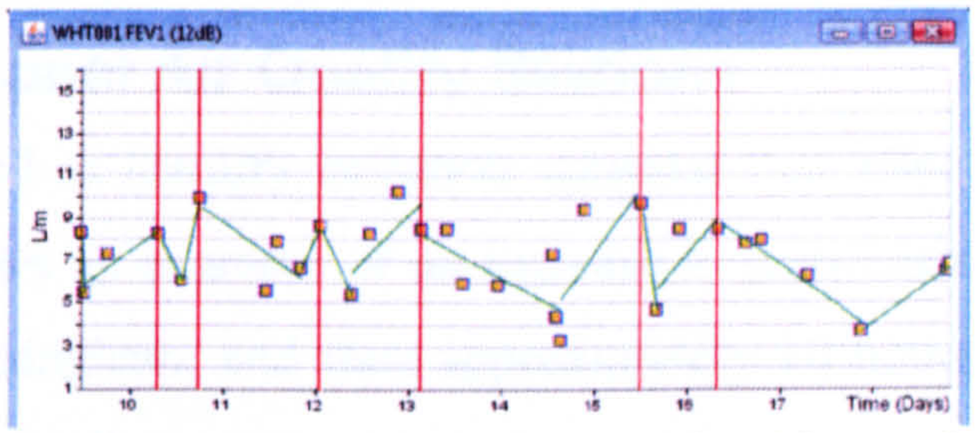


Figure 57 Control data set with additional noise at a Signal to Noise Ratio of 12dB - Set 2

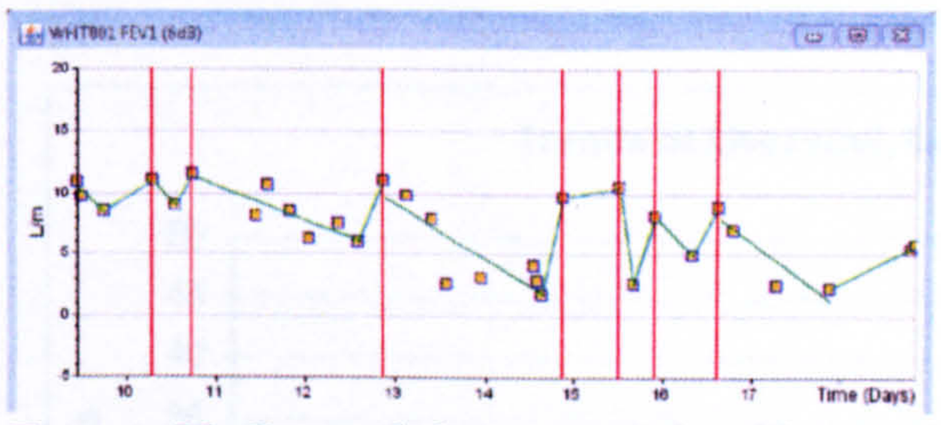


Figure 58 Control data set with additional noise at a Signal to Noise Ratio of 6dB - Set 3

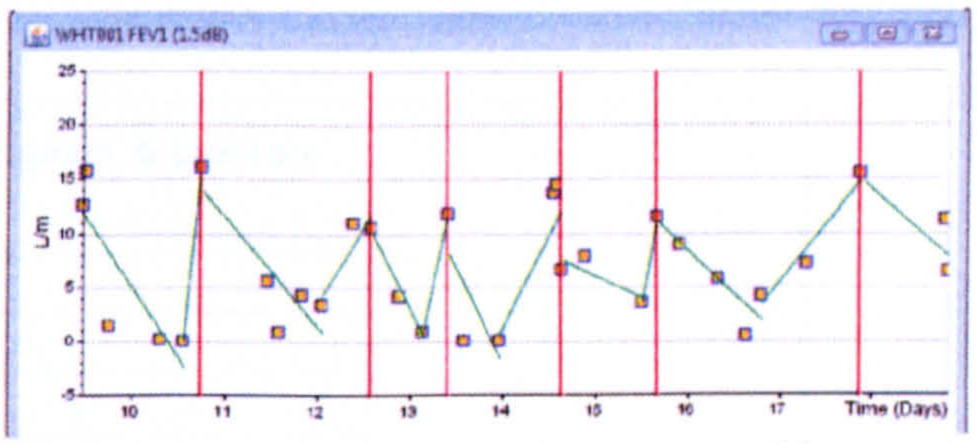


Figure 59 Control data set with additional noise at a Signal to Noise Ratio of 1.5dB - Set 4

Figures 54 to 59 show data sets with Gaussian random noise, where the standard deviation (in L/m) of each test set is n_1=0.9, n_2=1.8, n_3=3.1, and n_4=5.4, the control data set has a standard deviation of 0.8. There is 69% correlation between the control and data *Set 1*. The remaining three sets have 25%, 17% and 11% correlation to the control data set. Sets 1 and 2 both show datums around the 12th and 13th of the month, while set 3 correctly locates the identified reference datum from the control set on the 15th. Sets 1, 2 and 3 all locate the datum on the 13th (*12th 22:00*) of the month.

Since the level of noise a 'working system' would be expected to analyse (based on a ±10% reading error) is 20dB, the reduced accuracy resulting from the noisy data would not be a significant problem within the system. A reduction in accuracy is minimised by validating the delay characteristics produced from the identified reference datums over iterations of

the neural network. Once the lung function reference datums have been identified, the next stage in the process is to extract air quality data relating to the patient's location and time as they move from location to location. The *identification* components then act as validation methods to verify the extracted delay characteristics over time. The results of these components are demonstrated by the tests during the next sections.

6.3 Trends Monitored Across National Air Quality Stations

Pollution data recorded across the UK is used to identify pollution episodes that affect several regions of the country. Figure 60 shows an example of an increase in particulate matter first identified in Liverpool, then Birmingham, and then approximately 10 hours later recorded by the London (Bloomsbury) air quality monitoring station.

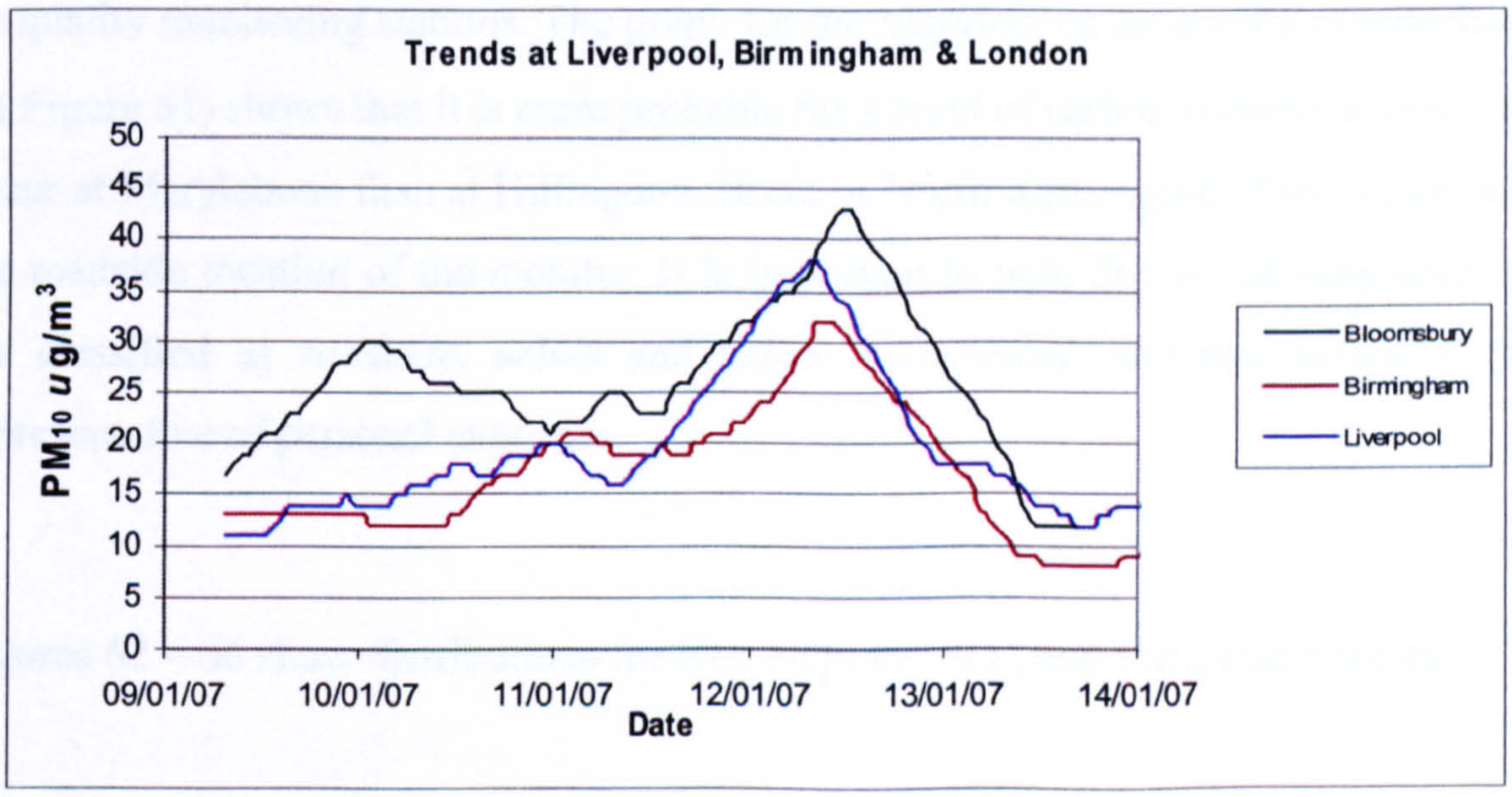


Figure 60 Multi-region pollutant episodes. A time delay of approximately 10 hours is shown between Birmingham and Bloomsbury particulate matter (PM_{10}) levels.

Identification of the occurrence of air quality episodes in this way shows that it is possible to track an episode from one air quality monitoring station to another. It should also be noted that it shows that pollution trends can be tracked over significant distances. These factors demonstrate the robustness of monitoring air quality, and support the requirements of the EMS, to be able to track patient specific air quality. In the example shown in Figure 60 it can be seen that a patient based in central London experiences a greater amount of background particulate matter compared to the readings taken in Birmingham, but

experiences comparative peak readings, although after a delay of approximately 10 hours. This is fundamental, as the time of day that a patient experiences the high levels of air pollution are varying, and therefore could stimulate different health outcomes.

6.4 Characteristics of London Air Quality Monitoring Stations

Several air quality monitoring stations from the London Air Quality Monitoring Network were analysed to find the types of pollutants recorded with greatest frequency. The analysis was undertaken to ascertain the probability of each pollutant occurring at the selected monitoring station, and demonstrate the differences between them. The analysis was achieved through the use of cumulative frequency distributions. The resulting graphs show the probability of a pollutant occurring, given a monitoring station and value of queried pollutant.

Figure 61 below shows the cumulative frequency distribution of carbon monoxide for four air quality monitoring stations. The graph for the Marylebone air quality monitoring station (in Figure 61) shows that it is more probable for a level of carbon monoxide over 1 $\mu g/m^3$ to occur at Marylebone than at Hillingdon, Brent or North Kensington. This is due, in part, to the roadside location of the monitor. It is important to note that monitoring sites in towns are classified as *roadside*, *urban* and *urban background*, and that *urban* is the most representative of personal exposure.

Figures 62 – 66 show distributions for NO, NO_2, O_3, SO_2, and PM_{10} respectively.

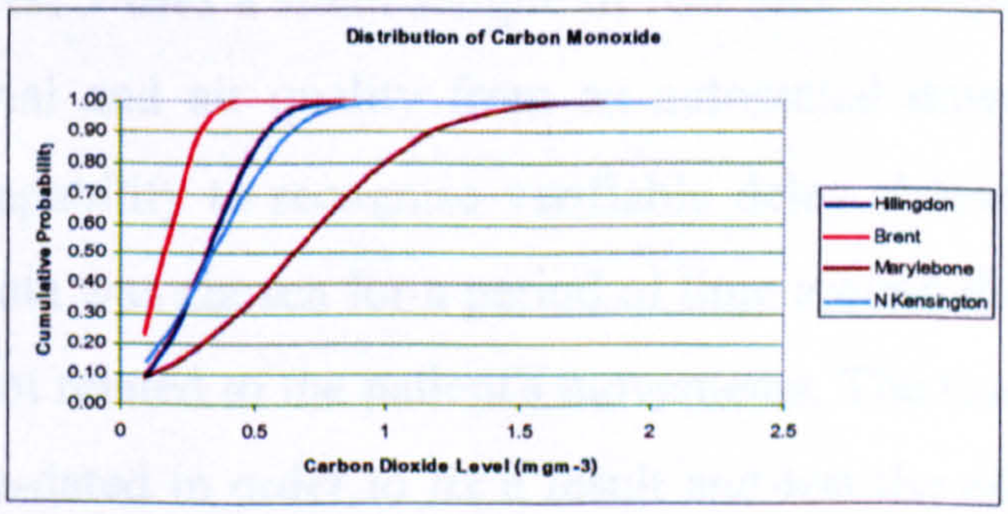


Figure 61 A sample of London air quality monitoring stations and the probability of carbon monoxide occurring at each one, against the value of pollutant.

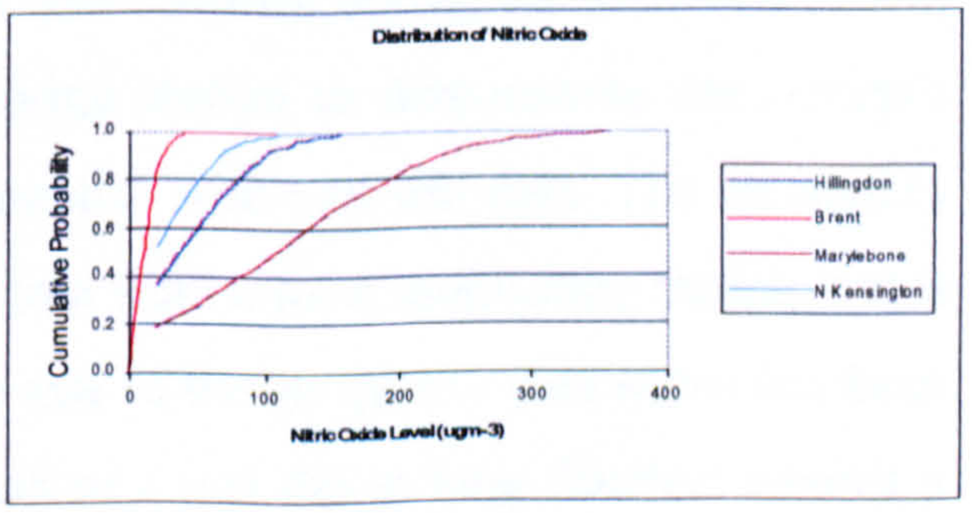


Figure 62 A sample of London air quality monitoring stations and the probability of nitric oxide occurring at each one, against the value of pollutant.

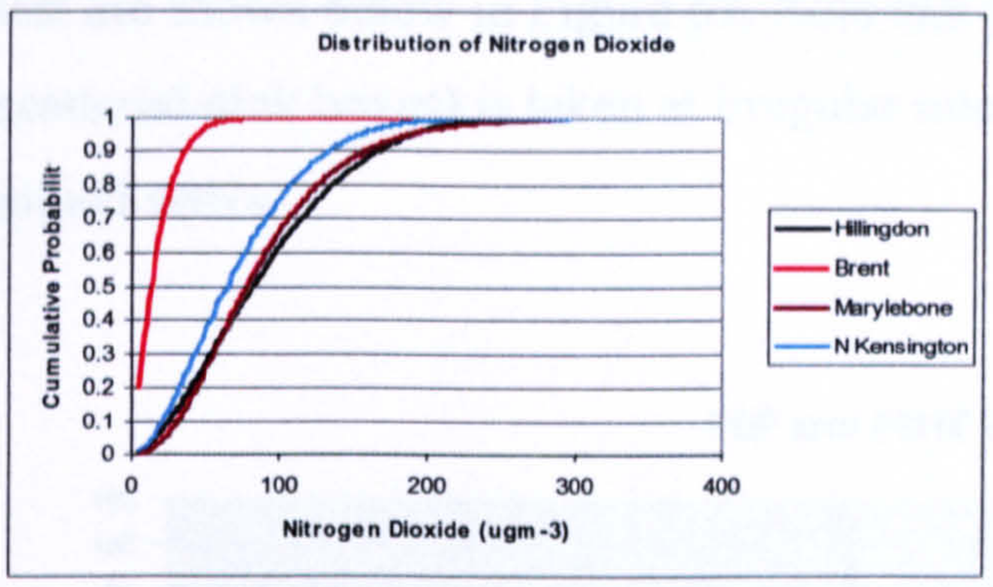


Figure 63 A sample of London air quality monitoring stations and the probability of nitrogen dioxide occurring at each one, against the value of pollutant.

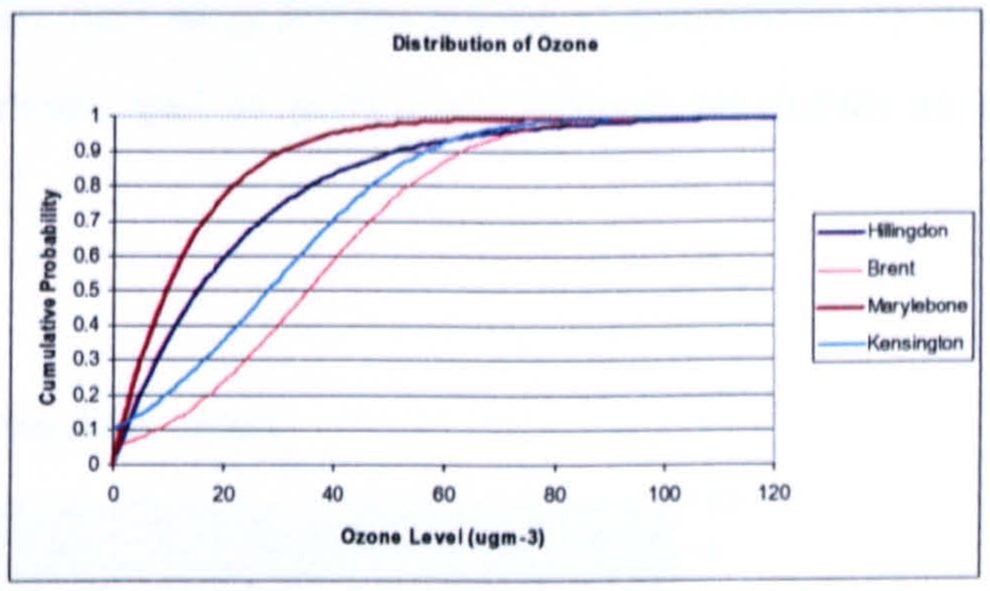


Figure 64 A sample of London air quality monitoring stations and the probability of ozone occurring at each one, against the value of pollutant.

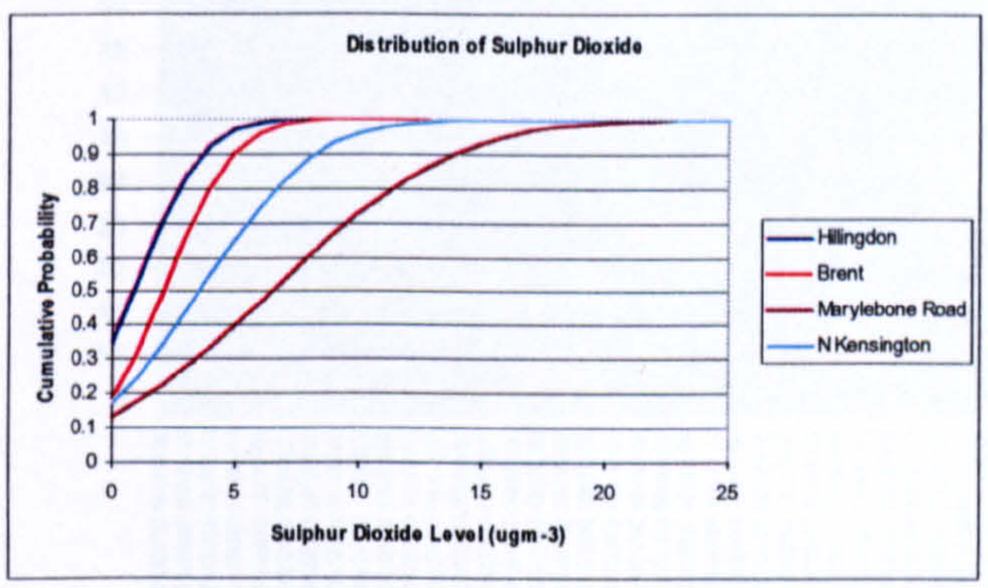


Figure 65 A sample of London air quality monitoring stations and the probability of sulphur dioxide occurring at each one, against the value of pollutant.

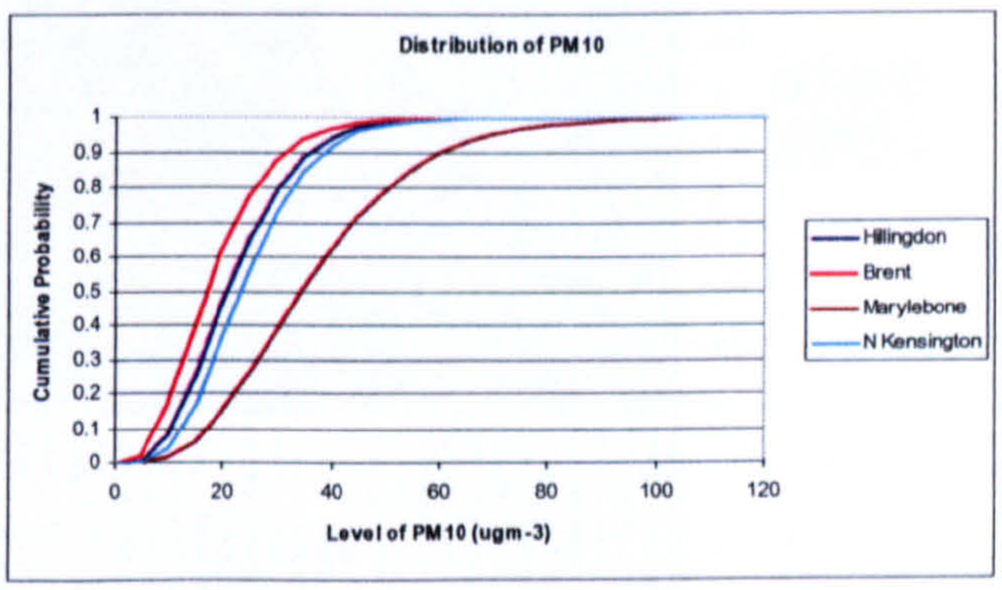


Figure 66 A sample of London air quality monitoring stations and the probability of particulate matter (PM_{10}) occurring at each one, against the value of pollutant.

6.5 Test 3 – Real Lung Function and Air Quality

Test 3 uses a small sample of real lung function data obtained during the Medicate (2000) trial and air quality from an automated monitoring station to demonstrate the system's capability to recognise verifiable delay characteristics with real life data. The air quality data was chosen for a period of time around the trial from a local monitoring station, but is not related to the patient's movements. The time axis of the air quality data series has been re-dated in order to *fix* a result and test the affect of a real dip in lung function against a real air quality peak and demonstrates how the EMS recognises this.

The graphs showing the air quality (PM_{10}) and the lung function (PEF) data used for the

test are shown below in Figure 67. Note that the real lung function data (represented by the scattered pink boxes) is taken at irregular intervals, and as such are unable to be drawn as a joined series.

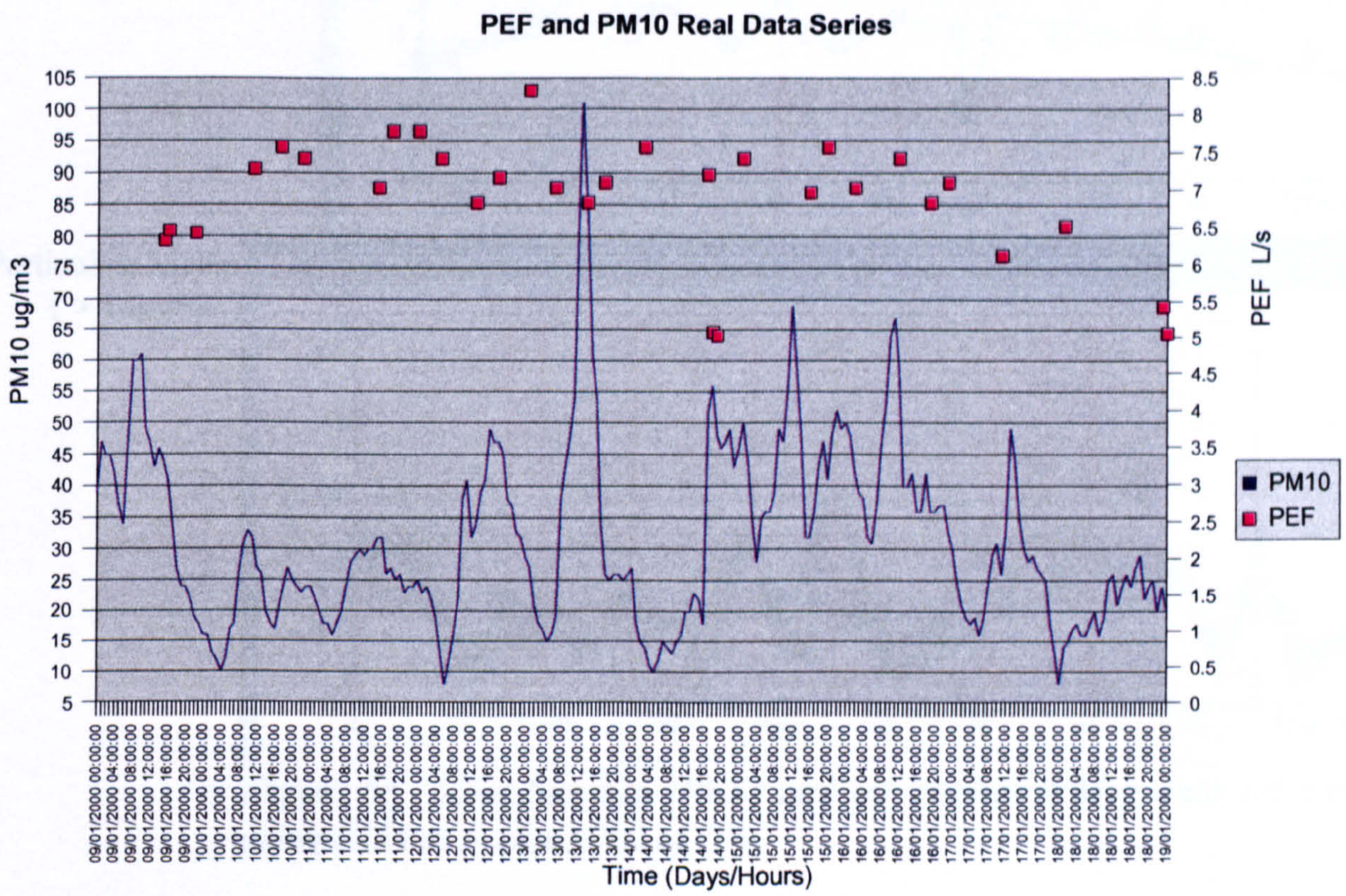


Figure 67 Showing real air quality and lung function data, matched so that a peak in air quality coincides with a decline in lung function.

The two data sets shown in Figure 67 were presented to the FDA module. The FDA results are shown for both the PEF and PM_{10} data sets in Figure 68. The analysis identified four reference datums from the peak expiratory flow data set (represented by the red lines) leading to a decline in lung function. The figure shows PM_{10} data extracted from the database matching the time and location of the PEF data. The PM_{10} time series used during the extraction of delay characteristics ends at the point of the last identified PEF datum; as the analysis is looking at the affect of historical delay characteristics rather than including an element of prediction (taking into account any reference datums that are identified after the onset of the asthma episode).

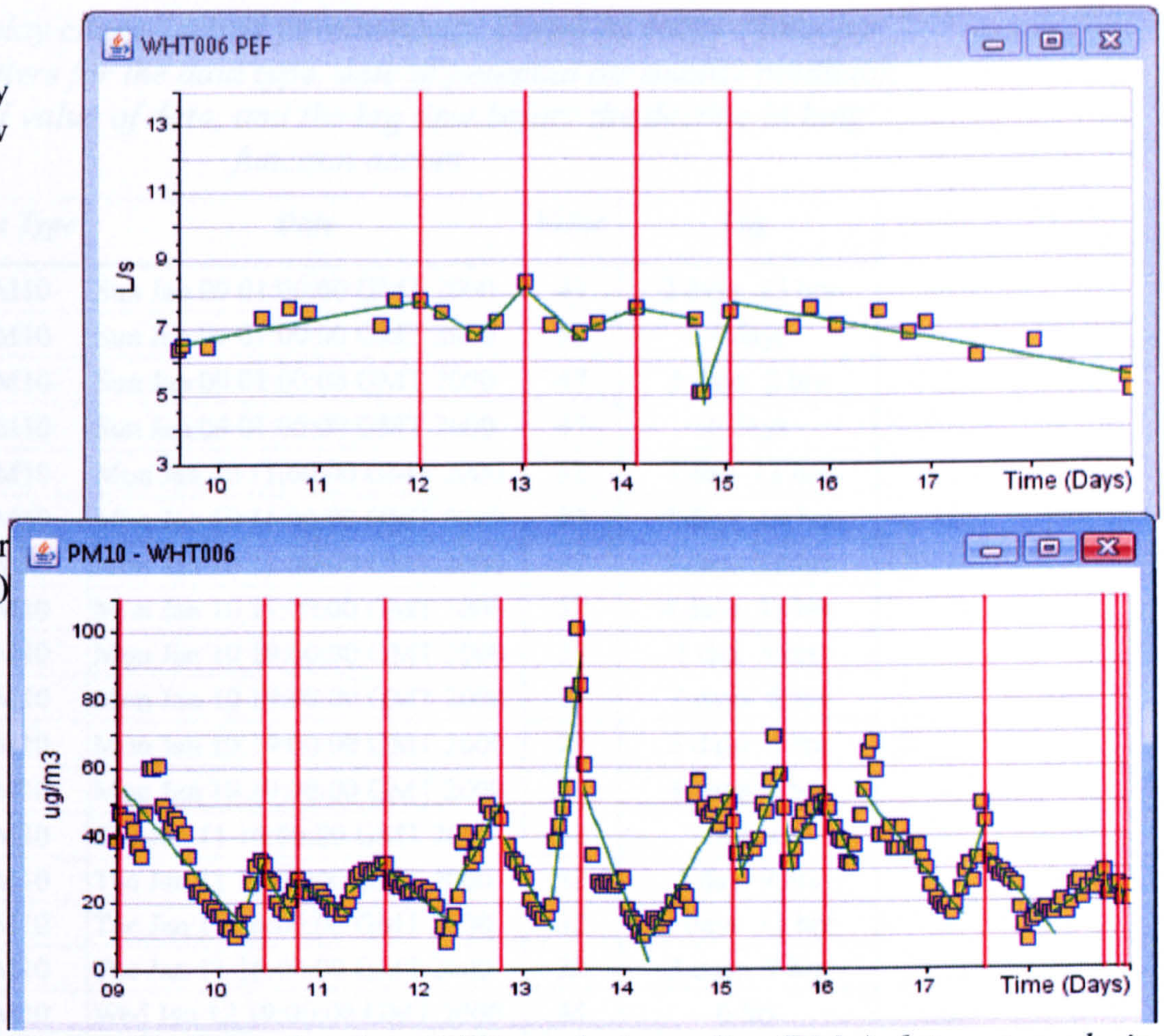


Figure 68 Lung function (PEF) and air quality (PM_{10}) data set analysis, using FDA at a sensitivity of 70%.

All possible permutations of delay characteristic drawn from the two data sets shown in Figure 68 are displayed in the table below as vectors, including *date*, *value* and *lag* time. All the vectors were presented to both the cluster analysis and neural network components for further analysis.

Table 13 Delay characteristic permutations, shown as vectors (one per row). Parameters for the data type, date of potential air quality predictor, the physical value of data, and the lag time before the decline in lung function occurs

Vector#	*Data Type*	*Date*	*Value*	*Lag*
1	PM10	Sun Jan 09 01:00:00 GMT 2000	47	2 days 23 hrs
2	PM10	Sun Jan 09 01:00:00 GMT 2000	47	4 days
3	PM10	Sun Jan 09 01:00:00 GMT 2000	47	5 days 2 hrs
4	PM10	Sun Jan 09 01:00:00 GMT 2000	47	6 days
5	PM10	Mon Jan 10 11:00:00 GMT 2000	32	1 day 13 hrs
6	PM10	Mon Jan 10 11:00:00 GMT 2000	32	2 days 14 hrs
7	PM10	Mon Jan 10 11:00:00 GMT 2000	32	3 days 16 hrs
8	PM10	Mon Jan 10 11:00:00 GMT 2000	32	4 days 14 hrs
9	PM10	Mon Jan 10 19:00:00 GMT 2000	27	1 day 5 hrs
10	PM10	Mon Jan 10 19:00:00 GMT 2000	27	2 days 6 hrs
11	PM10	Mon Jan 10 19:00:00 GMT 2000	27	3 days 8 hrs
12	PM10	Mon Jan 10 19:00:00 GMT 2000	27	4 days 6 hrs
13	PM10	Tue Jan 11 16:00:00 GMT 2000	32	8 hrs
14	PM10	Tue Jan 11 16:00:00 GMT 2000	32	1 day 9 hrs
15	PM10	Tue Jan 11 16:00:00 GMT 2000	32	2 days 11 hrs
16	PM10	Tue Jan 11 16:00:00 GMT 2000	32	3 days 9 hrs
17	PM10	Wed Jan 12 19:00:00 GMT 2000	45	6 hrs
18	PM10	Wed Jan 12 19:00:00 GMT 2000	45	1 day 8 hrs
19	PM10	Wed Jan 12 19:00:00 GMT 2000	45	2 days 6 hrs
20	PM10	Thu Jan 13 14:00:00 GMT 2000	84	13 hrs
21	PM10	Thu Jan 13 14:00:00 GMT 2000	84	1 day 11 hrs

The test does not include the *Date* parameter as part of the FBCA and neural network analysis to show the effect of analysing the lag and value attributes only. The parameters used for the FBCA are shown in Figure 69 below. The bucket sizes used for the cluster analysis are shown as $12\mu g/m^3$ for PM_{10} parameter values, and 5hrs (18000000 milliseconds) for lag. The size of bucket is also known as the sampling period.

Analysis Options

Parameter	Sensitivity	Bucket Size	Analyse
PM10 - Value	70	12	☑
PM10 - Lag	70	18000000	☑
PM10 - Date	70	86400000	☐

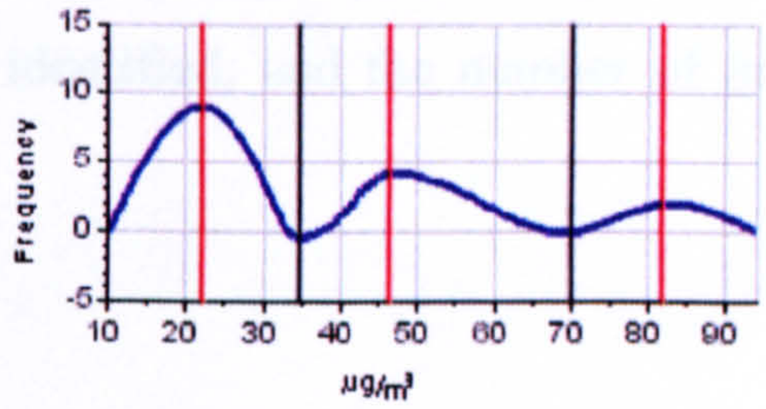
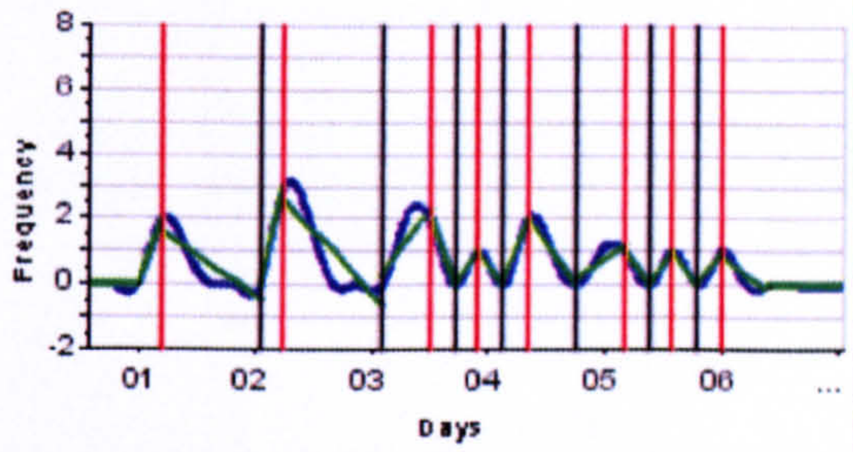
ID	Analysis	Raw Data
PM10 - Value		{0.0=0,21.0=12,33.0=0,45.0=7,57.0=0,69.0=0,81.0=2,93.(
PM10 - Lag		{-3600000=0,14400000=2,32400000=1,50400000=0,684

Figure 69 Using value and lag parameters (and excluding date). The figure above shows Boundary Analysis (using the FDA technique) locating the clusters by identifying the peaks and troughs of the Frequency Analysis. Troughs are marked with black lines, and the peaks with red.

Figure 69 shows the results of the *frequency analysis* on the test data set. Three *value* clusters and eight *lag* clusters have been identified (the boundaries of each denoted by the black lines). The graphs represent the frequency of the parameter given in the *ID* column of the table, the bucket width for the *PM10 Value* parameter (set at 12$\mu g/m^3$) can be seen in the *Raw Data* column of the figure where the PM10 value, raw data tally pairs are spaced by 12$\mu g/m^3$ from each other (pairs are separated by a comma). The *PM10 Lag* is represented in a similar way. The y-axis represents the frequency of the delay characteristic parameter (*PM10 Value* or *PM10 Lag*) that correspond, and are recognised by each bucket. The sensitivity value governs how the FDA identifies the peak and troughs of the resulting data and is arbitrarily set at 70% (discussed during Chapter 5).

The choice was made to analyse the effect of the *Value* and *Lag* characteristics and ignore the *Date* of the identified air quality points, this means that the found cluster permutations will not be date specific and can be applied to any future event, for example (as *Cluster 4* suggests in the tables below) when the level of PM_{10} is in the range 27 to 32$\mu g/m^3$, there is an increased likelihood that an asthmatic will experience a decline in lung function between *1 days 5 hrs* to *1 days 13 hrs*, and *2 days 6 hrs* to *2 days 14 hrs* time for example.

These ranges will be verified through continued validation (represented as hits) of the delay characteristics by the component. The *hits* recognised by each cluster are recorded and can be analysed further to determine a more precise measurement and statistical result.

The clusters identified, and the number of *hits* for each cluster are set out in Table 14 below.

Table 14 Value ranges covered by each cluster, and the number of delay characteristics (hits) recognised by each

Cluster#	***PM10 Value*** $\mu g/m^3$	***PM10 Lag***	***Hits***
1	0 – 33	-1 hr – 1 day 0 hrs	1
2	33 – 69	-1 hr – 1 day 0 hrs	1
3	69 – 93	-1 hr – 1 day 0 hrs	1
4	0 – 33	1 days 0 hrs – 2 days 1 hr	3
5	33 – 69	1 days 0 hrs – 2 days 1 hr	1
6	69 – 93	1 days 0 hrs – 2 days 1 hr	1
7	0 – 33	2 days 1 hr – 2 days 16 hrs	3
8	33 – 69	2 days 1 hr – 2 days 16 hrs	1
9	69 – 93	2 days 1 hr – 2 days 16 hrs	
10	0 – 33	2 days 16 hrs – 3 days 2 hrs	
11	33 – 69	2 days 16 hrs – 3 days 2 hrs	1
12	69 – 93	2 days 16 hrs – 3 days 2 hrs	
13	0 – 33	3 days 2 hrs – 3 days 17 hrs	3
14	33 – 69	3 days 2 hrs – 3 days 17 hrs	
15	69 – 93	3 days 2 hrs – 3 days 17 hrs	
16	0 – 33	3 days 17 hrs – 4 days 8 hrs	1
17	33 – 69	3 days 17 hrs – 4 days 8 hrs	1
18	69 – 93	3 days 17 hrs – 4 days 8 hrs	
19	0 – 33	4 days 8 hrs – 4 days 18 hrs	1
20	33 – 69	4 days 8 hrs – 4 days 18 hrs	
21	69 – 93	4 days 8 hrs – 4 days 18 hrs	
22	0 – 33	4 days 18 hrs – 6 days 0 hrs	
23	33 – 69	4 days 18 hrs – 6 days 0 hrs	2
24	69 – 93	4 days 18 hrs – 6 days 0 hrs	

The contents of the most verified clusters, Clusters 4, 7, 13 and 23 are detailed below. Each cluster shows the delay characteristics that have been recognised by the highlighted cluster.

Table 15 Most verified clusters

Cluster 4

	PM10 Value ($\mu g/m^3$)	*PM10 Lag*
Hit 1	27	1 days 5 hrs
Hit 2	32	1 days 9 hrs
Hit 3	32	1 days 13 hrs

Cluster 13

	PM10 Value ($\mu g/m^3$)	*PM10 Lag*
Hit 1	27	3 days 8 hrs
Hit 2	32	3 days 9 hrs
Hit 3	32	3 days 16 hrs

Cluster 7

	PM10 Value ($\mu g/m^3$)	*PM10 Lag*
Hit 1	27	2 days 6 hrs
Hit 2	32	2 days 11 hrs
Hit 3	32	2 days 14 hrs

Cluster 23

	PM10 Value ($\mu g/m^3$)	*PM10 Lag*
Hit 1	47	5 days 2 hrs
Hit 2	47	6 days

The result of the neural network analysis after 200 iterations of the input data set are summarised in the table below. The weight values show the recognised PM_{10} value and lag time.

Table 16 Summary of each neuron's weight vector. The PM_{10} value and lag time before lung function (PEF) decline is shown

Neuron ID	*PM10 ($\mu g/m^3$)*	*Lag Time (to trigger)*
N1	39.5	4 days 3 hours
N2	41.5	1 day 11 hours

The two neurons featured in Table 16 can be compared with the original input vectors presented to the neural network (Table 13), and the original data sets in Figure 67 where the results of the FDA are shown. Inspection of the graphs in Figure 67 and identification of reference datums crossing the data series at $40\mu g/m^3$ (Figure 68) or above confirm the viability of the lag time, to possible trigger identified by the neural network. A recent directive from the European Union (EU, 2008) prescribes an average reading of $50\mu g/m^3$ for particulate matter (PM_{10}) over a period of 24 hours (which is permitted 35 times per year), and an average of $40\mu g/m^3$ over a year as a guide for avoiding adverse health effects. This is in line with the identified values. The result where four neurons have been used during the analysis is given below.

Table 17 Summary of each neurons weight vector when four neurons are used in the neural network

Neuron ID	*PM10 ($\mu g/m^3$)*	*Lag Time (to trigger)*
N1	37.9	2 days 18 hours
N2	38.6	4 day 3 hours
N3	44.0	1 day 0 hours
N4	41.7	5 days 14 hours

The results of the neural analysis in Table 16 above, show a network employing two neurons. The neurons have identified the euclidean best fit for the data set, providing a representation of the data; comparing with the cluster permutations (shown in Table 14) it can be seen that the two neurons have located two areas, centred around clusters 5 and 17. Representation of the data is significantly improved with the use of four neurons (Table 17). A comparison with the results of FBCA (Table 14) shows that similar data clusters have been identified. *Neuron 1* identifies similar data sets to those which have been classified by *cluster 11*; *2* is similar to *cluster 17*; *3* similar to *cluster 2*; and *4* is similar to *cluster 23*. It should be noted however that the clusters identified by the neural network are not necessarily those clusters which identify the most hits.

FBCA has identified the clusters which received the most hits as being *clusters 4*, *7*, *13*, and *23*. While the neural network identified *cluster 4*, the other nodes identified were not those which were the most significant. This result is affected by the number of neurons

available, and the order in which data is submitted to the network.

6.6 Test 4 - Multi Parameter

The EMS has been designed to handle analysis requiring multiple parameters. For example, to identify whether or not a particular combination of particulate matter and ozone contributes to a patient's asthma episode. To simulate this requirement this *multi-parameter test* introduces a second parameter, to test the system's capability to recognise a pattern containing more than just a single time series or connection between single parameter types.

The data sets for lung function and the two air quality parameters are shown in Figure 70 below. It should be noted that the data used for a substitute second air quality parameter is actually an identical data set to the lung function time series. This has two purposes: to prove that the EMS is capable of using different data types as input to the system, and to test the ability of the system to recognise instantaneous effects on health from air quality. The test should identify a direct correlation between the air quality and the lung function data for the second parameter. The result will be shown by a lag time of *0* for the delay characteristic.

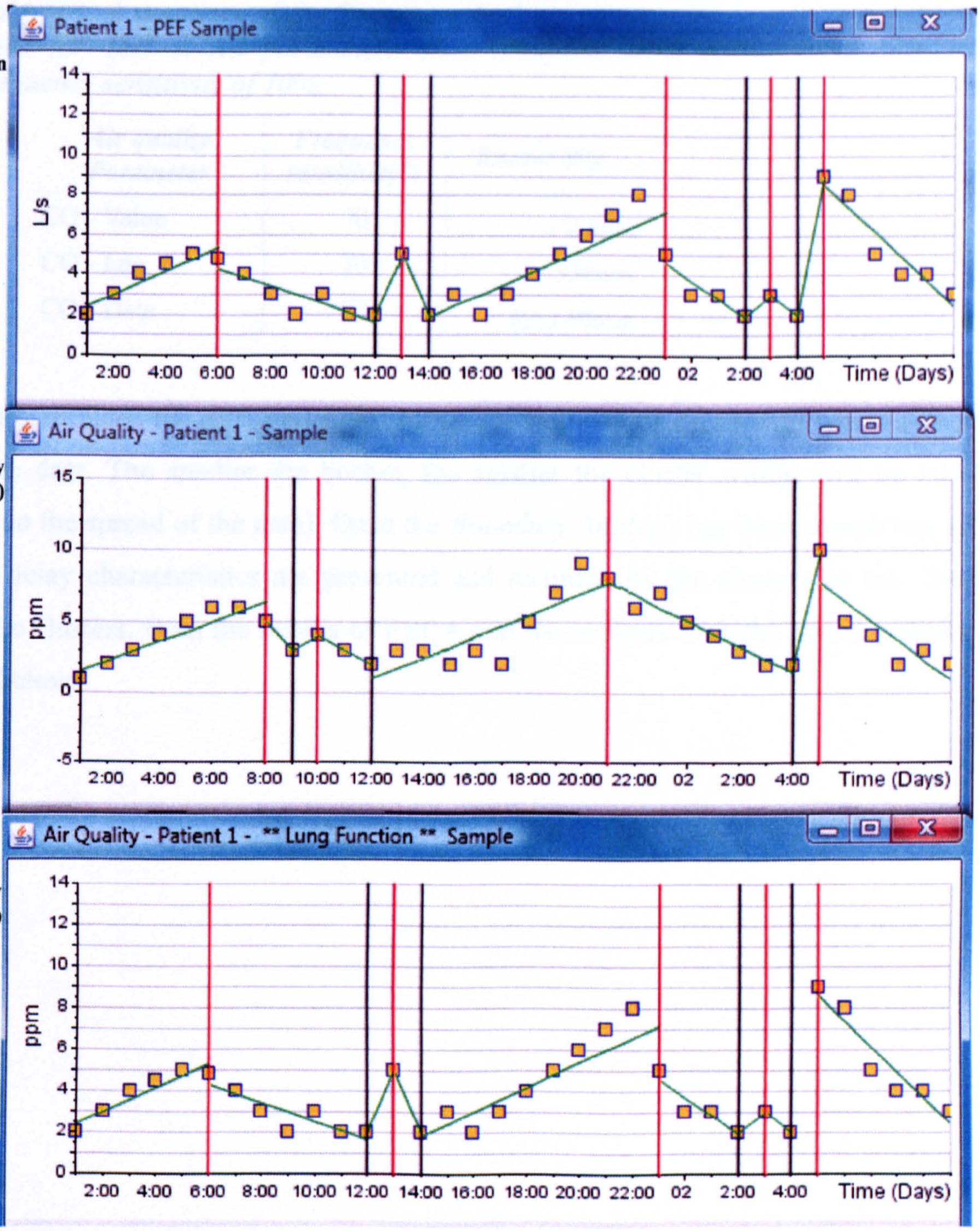


Figure 70 Lung function data and two data sets used for air quality data. The identified reference datums are shown in red.

Following the analysis shown in Figure 70 above, 131 delay characteristics were extracted between the three data sets; from the reference datums belonging to the lung function and two air quality parameters. Frequency and Boundary analysis identified 21 possible cluster permutations from analysis of the delay characteristics.

The following FBCA settings were used during the analysis:

Table 18 A summary of the Boundary Analysis parameters used for Test 4. All parameters were analysed using a frequency sensitivity of 70%.

Air quality Parameter	***Frequency sensitivity%***	***Bucket Size***
CO - Value	70	2 ppm
CO - Lag	70	33mins
CO - Date	70	8hrs 30min

The bucket sizes used for each parameter influence the level of detail with which FBCA analyses the data. The smaller the bucket, the smaller the cluster widths will be (also dependent on the spread of the data). Once the *Boundary Analysis* has been completed, all air quality delay characteristics are presented and recorded by the clusters as *hits*, thus verifying the clusters. Both the results of FBCA and the analysis with the neural network are shown below.

Table 19 Validated clusters after frequency analysis. The clusters shown here have recognised the largest number of delay characteristics, identified during the FDA as reference datums (identified in Figure 70).

Clusters							**Cluster Contents (Hits)**			
Cluster #	*Air Quality 1 Value (ppm)*	*Air Quality 2 Value (ppm)*	*Air Quality 1 Lag (hrs)*	*Air Quality 2 Lag (hrs)*	*Hits*		*Value 1 (ppm)*	*Value 2 (ppm)*	*Lag 1 (hrs)*	*Lag 2 (hrs)*
1	(1.0 - 15.0)	(1.0 - 13.0)	(4hrs - 5.25hrs)	(-1hr - 3 hrs)	4	⇒	5	3	5 hrs	0
							5	3	5 hrs	2 hrs
							5	4.8	5 hrs	0
							5	5	5 hrs	0
2	(1.0 - 15.0)	(1.0 - 13.0)	(5.25hrs - 7hrs)	(-1hr - 3 hrs)	4	⇒	8	5	6 hrs	0
							8	3	6 hrs	0
							8	3	6 hrs	2 hrs
							8	4.8	6 hrs	0
3	(1.0 - 15.0)	(1.0 - 13.0)	(7hrs - 12 hrs)	(-1hr - 3 hrs)	4	⇒	8	5	8 hrs	0
							8	3	8 hrs	0
							8	3	8 hrs	2 hrs
							8	4.8	8 hrs	0
4	(1.0 - 15.0)	(1.0 - 13.0)	(12 hrs - 14.5 hrs)	(-1hr - 3 hrs)	4	⇒	4	5	13 hrs	0
							4	5	13 hrs	0
							4	3	13 hrs	0
							4	3	13 hrs	2 hrs
5	(1.0 - 15.0)	(1.0 - 13.0)	(14.5 hrs - 16hrs)	(-1hr - 3 hrs)	4	⇒	5	5	15 hrs	0
							5	3	15 hrs	0
							5	3	15 hrs	2 hrs
							5	4.8	15 hrs	0
7	(1.0 - 15.0)	(1.0 - 13.0)	(18hrs - 20 hrs)	(-1hr - 3 hrs)	8	⇒	5	5	19 hrs	0
							5	5	19 hrs	0
							5	3	19 hrs	0
							5	3	19 hrs	2 hrs
							4	3	19 hrs	0
							4	3	19 hrs	2 hrs
							4	4.8	19 hrs	0
							4	5	19 hrs	0

Clusters containing the largest number of hits (verified as active clusters) are expanded above (on the right hand side of the table), to show the recognised delay characteristics. It can be observed that some of the *Lag 2* times are zero. This is a result of using the lung function (PEF) data as the second air quality data set (*Air Quality Parameter 2)*. The zero reading identifies when a reference datum in both (lung function and air quality) data sets coincide.

The value ranges for both parameters are the same in all six (shown) verified clusters. This

depicts a uni-modal distribution as the range spreads from the minimum to maximum values presented within the data set. Six ranges were identified (and verified) for *Parameter 1 Lag*, each range covering lag values progressively increasing in length. *Parameter 2 Lag* values however, all fell between -1 hrs and 3 hrs; which confirms correct identification of the identical data set used for the air quality (*Parameter 2*). The method also indicates success in recognising instantaneous effects.

The neural network component was activated with the same data set, using 4 neurons. The result is shown by the table below.

Table 20 Result of the neural network using 4 neurons. The results for the Value and Lag parameters are shown for both air quality types (param 1 & param 2)

Neuron ID	*Value 1 (ppm)*	*Value 2 (ppm)*	*Lag 1 (hrs)*	*Lag 2 (hrs)*
N1	7.2	4.2	7.2	0.4
N2	5.9	4.0	13.0	0.5
N3	5.9	3.9	11.1	0.5
N4	5.9	3.9	11.1	0.5

The *Lag 2* delay has been recognised by the neural network to be half an hour, which is close to an instantaneous effect. This close match is due to the neural algorithm oscillating between 0 and 2 hours. The neural network behaves in this way due to each neurons neighbourhood not being tuned to a particular range before learning, and therefore adjusting to the occurrence of new data as it is presented to the network.

6.7 Hospital Admissions due to Respiratory Episodes

Data collected during the Medicate project was typically recorded over a period of two weeks, and once a patient had been admitted to hospital. This method meant that analysis was performed on data covering periods where patients were recovering from acute asthma exacerbations, rather than before and during a decline in their respiratory health.

The EMS focuses analysis of data to outliers which means that the quantity of data available for analysis from a normal time series data set is reduced, which improves scalability, but also means that data set lengths are required to be considerably larger than if traditional statistics were used. As yet the length of time series data required to successfully identify predictors of asthma exacerbation is unknown. However it was found from the Medicate (2000) project that a data set relating patient respiratory health to the environment was required that covered a period significantly greater than two weeks.

6.7.1 Testing the Hospital Admissions Data with Correlation

Data sets relating to UK respiratory episodes are held by *The Information Centre for health and social care "The IC"*, but are restricted to hospital admissions data. Historical records for hospital respiratory episode admissions were obtained from *The Information Centre* (2007) in the form of their *Hospital Episode Statistics*. The data obtained covered a period of one year, from 1st April 2005 to 31st March 2006, and was used as a substitute to personal lung function data. This approach had two advantages in testing the EMS:

1) The data was readily available across a wide sample of the population.
2) The use of a data set outside the problem scope, demonstrated the application of the EMS architecture to a wider range of problems.

The hospital admissions and associated air pollutant data was first analysed using the traditional technique of correlation, in order to identify relationships that would formerly have been identified between poor air quality and the hospital admissions data. The technique included time lag analysis of the correlation coefficient (discussed during Section 1.1.2). The investigation used the values of daily maximum air pollutants determined across all the local air quality monitoring stations relevant to the study; a similar technique was used during the Medicate project (Crabbe *et al.*, 2004).

Figures 71 and 72 show the results of the correlation study in which patient hospital admissions due to asthma were correlated against the maximum daily pollution levels, which had occurred within the fifteen days prior to the date of admission. The analytical period of fifteen days was chosen as an arbitrary value, 50% longer than the period defined by previous research (Lebowitz, 1996) to include possible outliers.

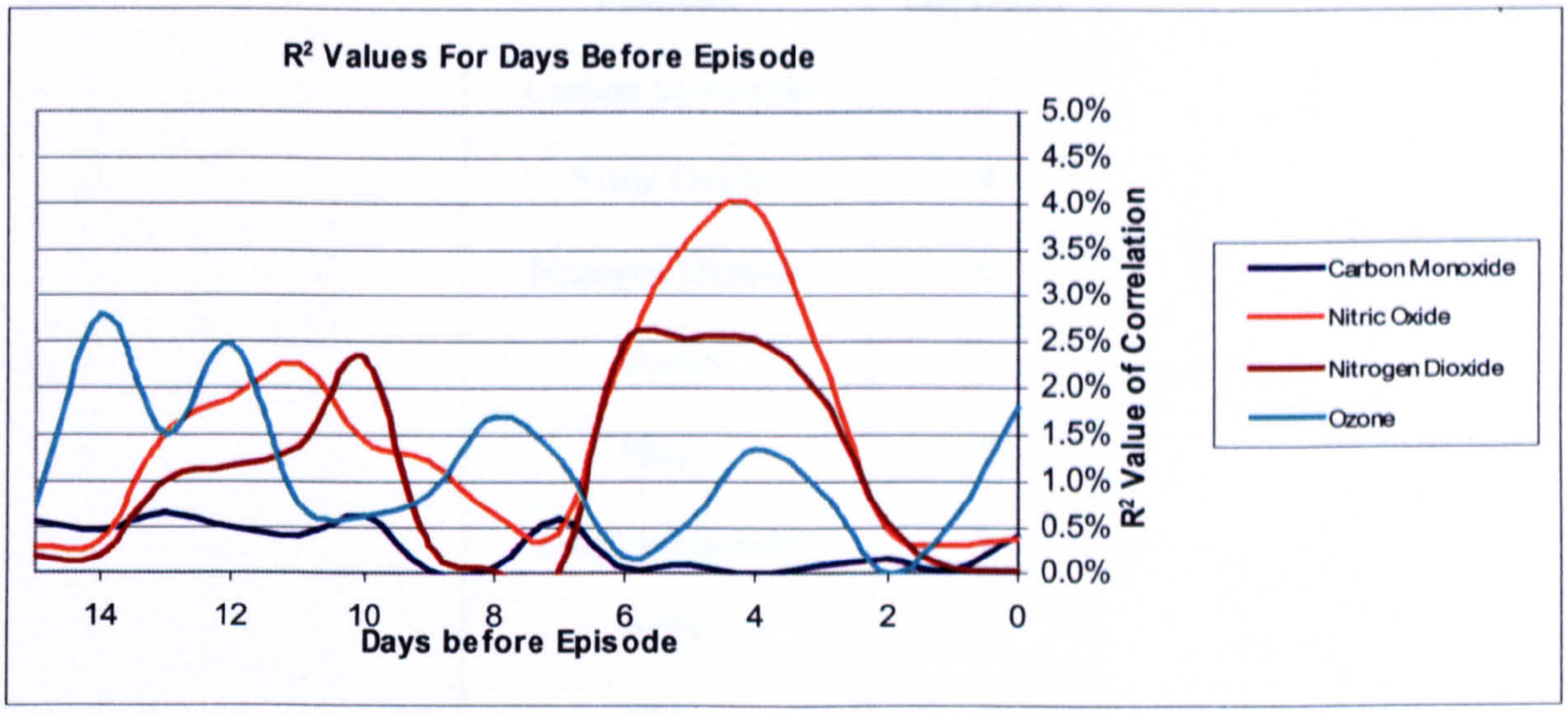


Figure 71 Correlation coefficients, depicting the correlation between maximum air pollutant readings and hospital admission episodes, for a period of 15 days before hospital admission.

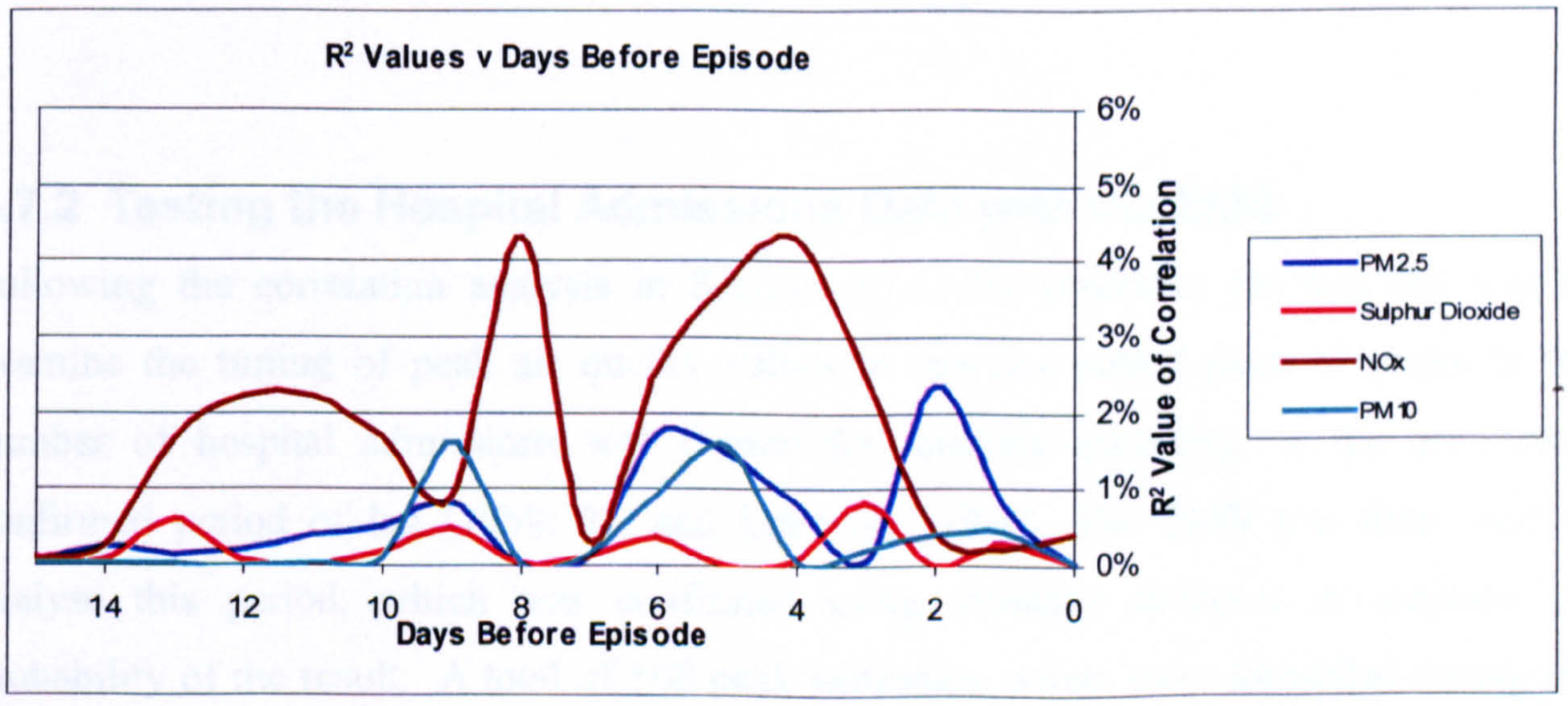


Figure 72 Correlation coefficients, depicting the correlation between maximum air pollutant readings and hospital admission episodes, for a period of 15 days before hospital admission.

The graphs shown above in Figures 71 and 72 indicate the relationships between the various pollutants and delay before admissions due to respiratory episodes. The first peak (with the lowest lag time) for each pollutant are shown in Table 21.

Table 21 Lag time before first hospital admissions

Pollutant	*Lag (days)*
Carbon Monoxide	7
Nitric Oxide	4
Nitrogen Dioxide	6
Ozone	4
$PM_{2.5}$	2
Sulphur Dioxide	3
NOx	4
PM_{10}	5

These results however were not significant enough to establish a relationship between air quality and hospital admissions due to asthma. The highest level of confidence (R^2) equalled 8.9%. However, the results did confirm previous research values for the lag effect of air quality on the asthmatic (Lebowitz, 1996).

6.7.2 Testing the Hospital Admissions Data with the EMS

Following the correlation analysis in Section 6.7.1, the objective for this test was to examine the timing of peak air quality values. A ten day period prior to peaks in the number of hospital admissions was chosen for analysis according to the previously confirmed period of lag (Table 21, and Lebowitz, 1996). The EMS was then used to analyse this period, which was confirmed using standard statistics to indicate the probability of the result. A total of 108 peak admission points were identified during the twelve month period (1st April 2005 to 31st March 2006).

Using FDA to analyse air quality data over a period of 10 days prior to the 108 admission points, (an example of which is shown in Figure 73) a number of peaks (and troughs) were identified. The identified peaks were then used by the *Hypothesis Builder* (described in Section 5.4) to derive a set of delay characteristics for further analysis by the EMS analytical components.

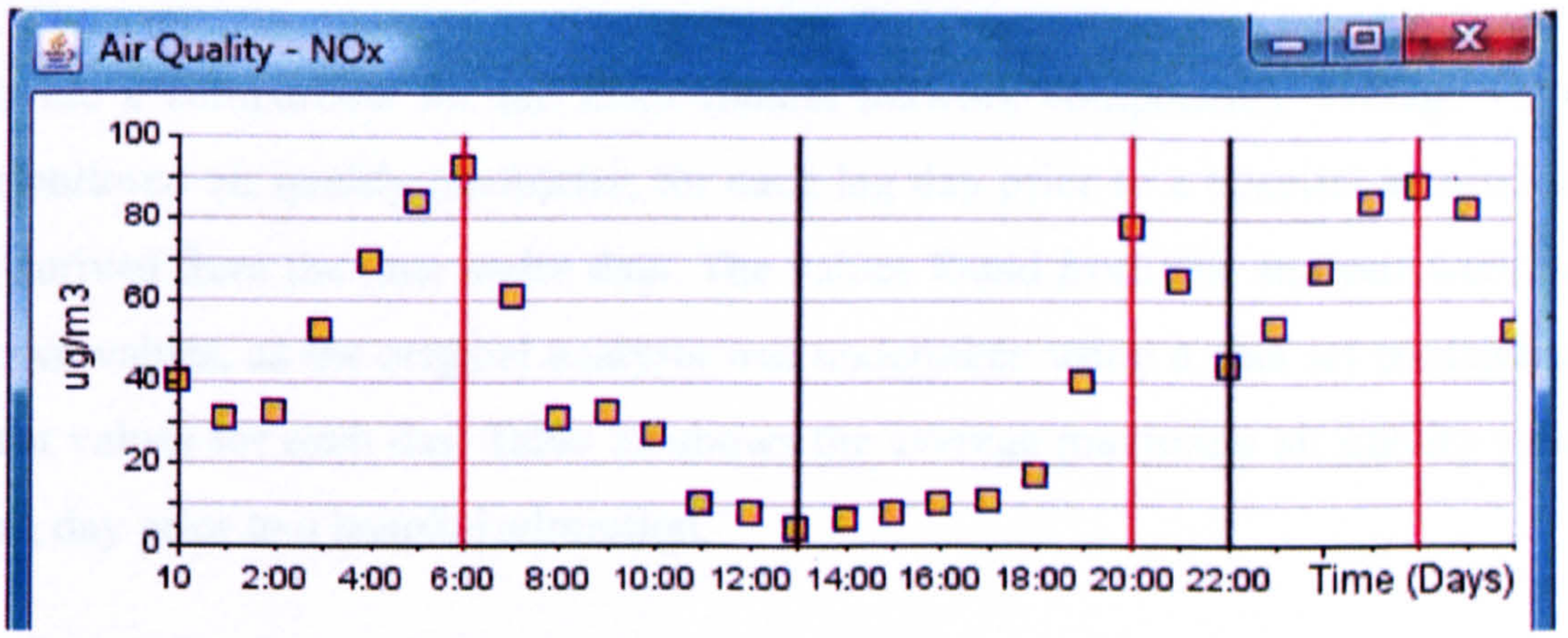


Figure 73 An example of FDA on an air quality data sample relating to a 10 day period prior to a peak in admissions. The graph shows an hourly, 24 hour segment, from the 10 day sample.

Before the delay characteristics were presented to the FBCA and neural network components a Chi-square test was undertaken using the identified peaks, and subsequent delay characteristics, to determine if their occurrence would happen by chance. During this test, it was assumed that if the distribution of the identified peaks was completely random and not associated with the number and dates of admission, there would be an equal chance of a peak occurring on any day during the ten day periods examined. That is, the expected number for each variable for each day would be 10.8 peaks (from 108 peak admission points over ten days). A data sample and associated maximum readings are illustrated below. For example, there were 15 peaks in carbon monoxide, 9 days prior to a hospital admission.

Table 22 Frequency of air quality peaks prior to a hospital admission. The maximum frequency of readings for each parameter is highlighted in yellow

Days Prior to Admission	Carbon Monoxide	Nitric Oxide	Nitrogen Dioxide	Ozone	PM2.5	Sulphur Dioxide	NOx	PM10
10	6	11	11	12	12	13	10	12
9	15	11	13	10	5	11	13	8
8	10	7	7	13	11	12	5	10
7	10	12	7	8	9	13	10	11
6	13	18	16	12	15	11	18	13
5	15	12	13	12	11	10	13	7
4	8	9	15	5	8	13	10	14
3	10	10	7	14	15	12	11	13
2	12	7	10	12	12	6	7	15
1	9	11	9	10	10	7	11	5
Total combinations	108	108	108	108	108	108	108	108

To provide a comparison for the EMS (neural network component), average values for each monitored air quality parameter, for each lag day prior to a hospital admission peak where derived from the time series data. The values found from this analysis were average maximum values, as the original analysis was undertaken using a data set containing peak pollutant values for each day. Table 23 shows the average maximum air quality values for each lag day prior to a hospital admission.

Table 23 Average maximum values for each parameter's air quality peak prior to a hospital admission. Those values that coincide with the maximum frequency of readings in Table 22 are highlighted in yellow

Days Prior to Admission	Carbon Monoxide	Nitric Oxide	Nitrogen Dioxide	Ozone	PM2.5	Sulphur Dioxide	NOx	PM10
10	3.0	491	273	93	54	41	974	155
9	2.9	488	285	100	47	43	999	149
8	3.0	417	294	101	65	48	786	153
7	2.9	510	315	103	47	45	984	146
6	2.9	434	274	95	58	30	917	173
5	3.2	436	297	99	51	37	965	127
4	3.2	573	302	109	57	47	1,164	167
3	2.7	415	261	91	49	47	865	132
2	3.5	479	322	112	63	45	1,092	167
1	2.8	551	274	112	49	41	1,067	156

108 multi-parameter vectors were presented to the neural network until the network had completed its learning period. The result of the network, which completed containing four neurons is shown in the two tables below. The first table shows the results for the *lag* component, and the second, the *value* of the parameter.

Table 24 Results from the lag analysis of air pollutant data sets related to peaks in hospital admissions (in days)

Neuron	CO	NO	NO_2	O_3	$PM_{2.5}$	SO_2	NOx	PM_{10}
N1	4.26	5.52	5.50	5.66	6.12	7.42	4.65	5.82
N2	5.04	5.66	5.51	4.95	4.24	7.23	6.27	5.69
N3	5.77	4.22	5.49	4.92	4.97	6.60	5.12	3.57
N4	5.04	5.54	5.62	5.19	3.98	4.34	4.21	3.01

Table 25 Results from the value analysis of air pollutant data sets related to peaks in hospital admissions (in $\mu g/m^3$ except CO mg/m^3)

Neuron	CO	NO	NO_2	O_3	$PM_{2.5}$	SO_2	NOx	PM_{10}
N1	3.1	527	300	105	57	44	1,045	109
N2	3.0	416	281	99	54	44	831	144
N3	3.0	496	288	99	54	42	983	150
N4	3.1	435	288	100	55	41	944	147

A comparison of neural network results in Table 24 and 25 against peak air quality episodes prior to a peak in hospital admissions (Table 23) is shown by Table 26 below. The results of the neural network's four neurons are overlaid onto the grid (of Table 23), with cells with green borders being representative of the specific lag days prior to a peak in hospital admissions where the highest frequency of poor air quality was identified. Cells with a green border and grey background, are representative of a peak in hospital admission frequency, where the neural network has failed to identify the lag characteristic. Numbers within square brackets represent the number of neurons that identified the *lag* day, while the unbracketed figure is the average value of the air quality parameter identified by the neural network at that point in time.

Table 26 Showing the results of the neural network against the maximum occurrences of input data (Table 23)

Days Prior to Admission	Carbon Monoxide	Nitric Oxide	Nitrogen Dioxide	Ozone	PM2.5	Sulphur Dioxide	NOx	PM10
10								
9								
8								
7						43 [3]		
6	3.0 [1]	459 [3]	290 [3]	105 [1]	57 [1]		831 [1]	127 [2]
5	3.1 [3]		288 [1]	99 [3]	54 [1]		1014 [2]	
4		496 [1]			55 [2]	41 [1]	944 [1]	150 [1]
3								147 [1]
2								
1								

From the table above it can be observed that the neural network has failed to identify a number of periods attributable to a peak in hospital admissions, (depicted by grey cells with green borders). This is due to the self-organising algorithm used by the network converging on to a solution fitting the multi-parameter vector, rather than each parameter individually. The network has identified several features found by earlier analysis of the data, in particular carbon monoxide at a value of $3.1 mg/m^3$ at 5 lag days, nitric oxide $459 \mu g/m^3$ and nitrogen dioxide $290 \mu g/m^3$, both at 6 lag days, and $43 \mu g/m^3$ sulphur dioxide at 7 lag days. This result would allow an alert to be generated 5 lag days before a peak in hospital admissions, if the values of each air quality parameter were higher than those being monitored by the network. Using the Chi Square test and frequency data from Table 22 this would have a 2.46% probability of happening by chance, purely taking into account the recognition of the lag component.

6.8 Analysis of a Six Month Set of Lung Function and Air Quality

To validate the EMS using a real respiratory data set, a further data set was collected. A lung function and patient specific air quality data set was collected over a continuous six month period between July 2007 and February 2008. An asthmatic patient was enrolled, and issued with an electronic lung function measuring device, a *PiKo-1* (Ferraris Respiratory Europe; Hertford, UK) for them to monitor their respiratory condition. The *PiKo-1* recorded the patient's maximum lung function reading given within a 3 minute period. The patient was asked to undertake a minimum of three respiratory manoeuvres during this time to ensure a reliable peak reading was obtained. Within the trial period, the patient experienced a decline in respiratory condition serious enough to seek medical attention. As a result of their decline in respiratory health (between the 15th September and 5th October 2007), the patient changed their medication (according to clinical advice) which is shown within the results.

The charts below (Figures 74 and 75) show the full six month data set belonging to Patient *A,* for Peak Expiratory Flow (PEF) and Forced Expiratory Volume in one second (FEV_1).

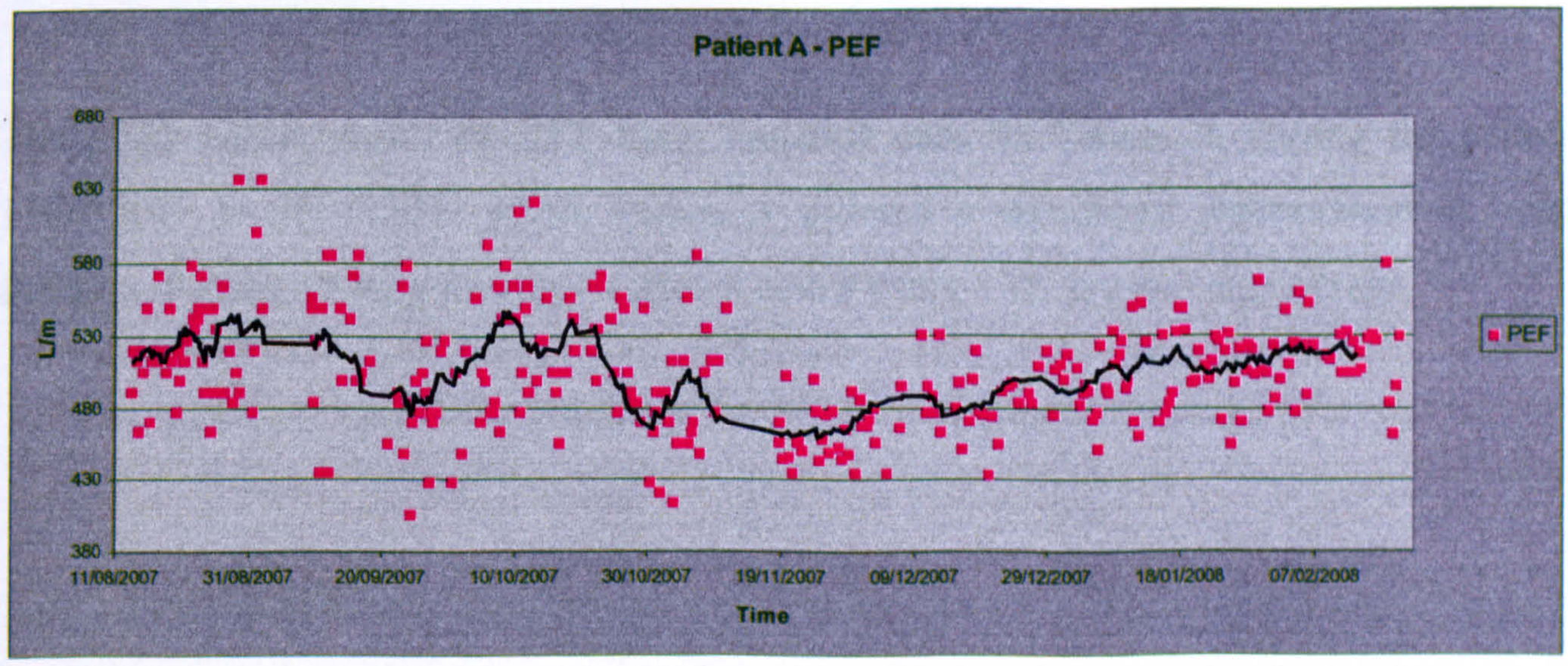


Figure 74 Six month peak expiratory flow - data sample (from Patient A), showing a 12 day moving average in black.

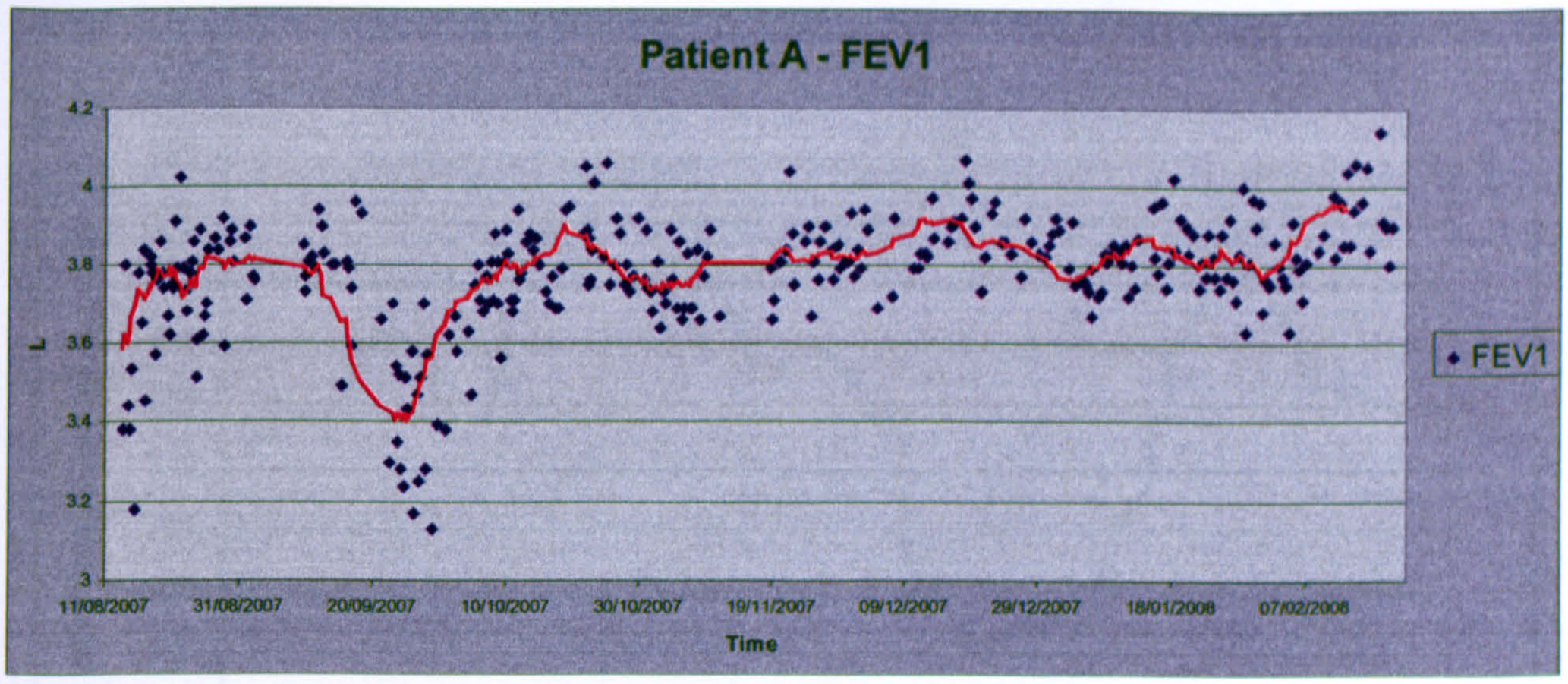


Figure 75 Six month Forced Expiratory Volume in 1 second - data sample (from Patient A), showing a 12 day moving average in red.

Figure 75 clearly shows the period of respiratory decline where clinical advice was sought, centered around 20th September. The period is marked by a drop in Forced Expiratory Volume (in one second) from 3.8 to 3.4 litres. A 12 day moving average of FEV_1 diurnal variability remained constant during the six month period, fluctuating between 1 and 6%. Diurnal variability of PEF (also using a 12 day moving average), declined gradually over the six month measurement period, where peak daily variability reduced from 15 to 8%.

The graph below shows the PEF Lung Function data for Patient *A*. During the period 15/09/2007 to 05/10/2007 where Patient *A* suffered a significant deterioration in lung function. Patient *A*'s lung function is shown below using a 12 day moving average.

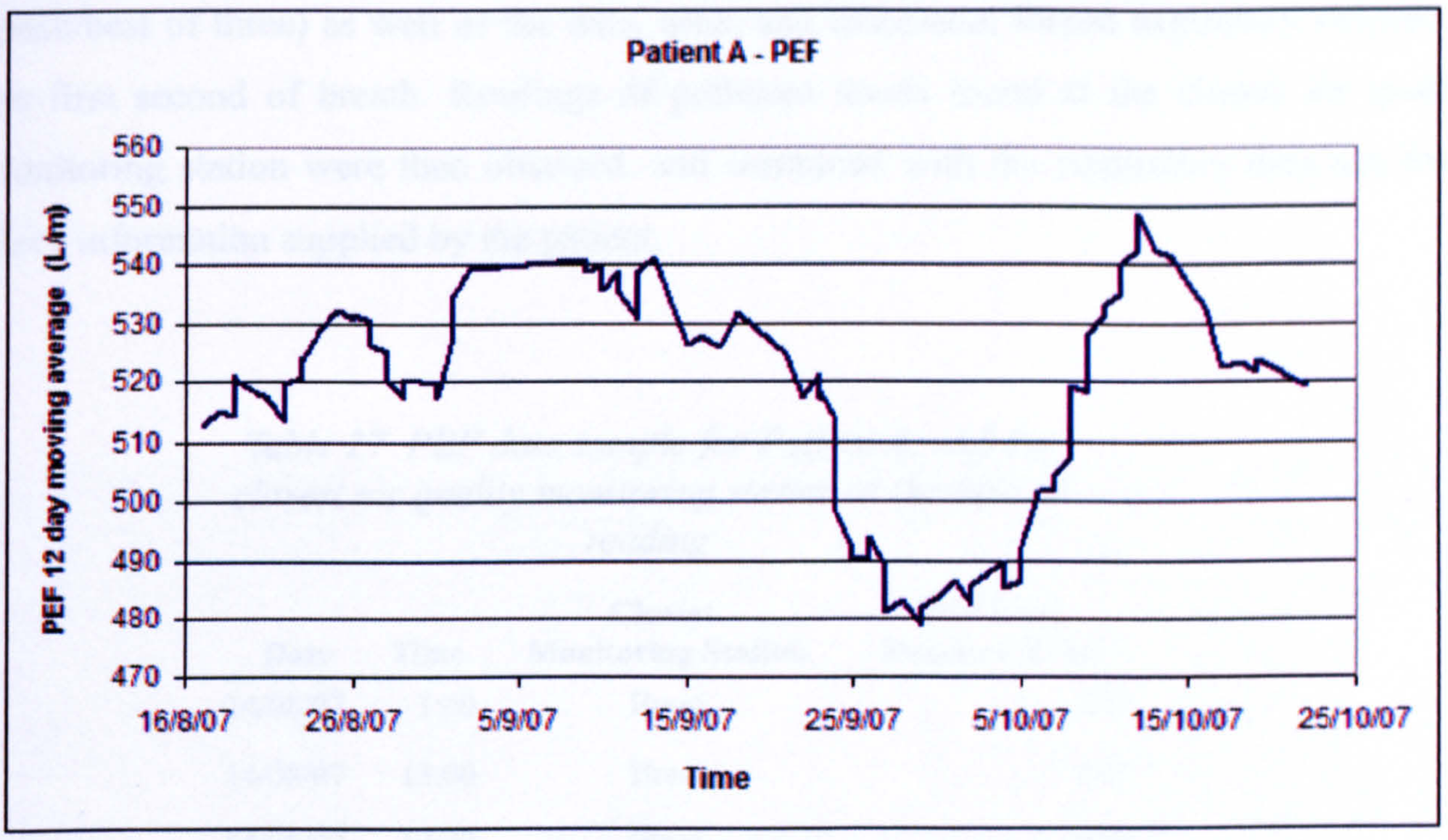


Figure 76 Sample showing a decline in respiratory condition for Patient A, between August 16th and October 25th 2007.

The raw *Peak Expiratory Flow* data was analysed with Feature Detection Analysis to obtain reference datums. The analysis is shown in Figure 77.

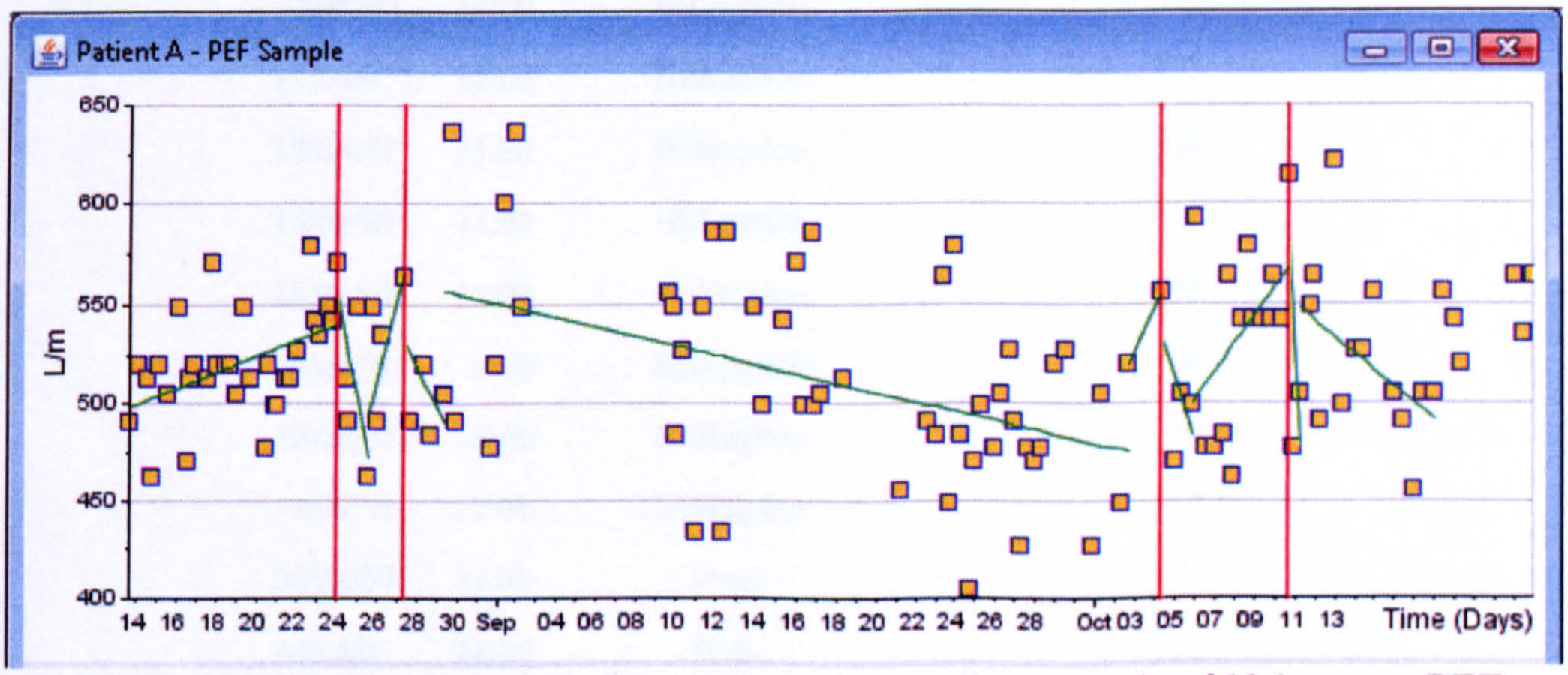


Figure 77 Identified reference datums from the raw data sample of 126 patient PEF readings. The four reference datums are marked with red vertical lines, while the trend of the data is shown in green.

A sample of raw PEF data from *Patient A* is shown alongside locations of monitored environmental pollutants in Table 27. Peak expiratory flow readings were recorded by the patient using a portable electronic monitoring device, which recorded the reading

(peak/best of three) as well as the date, time, and associated forced expiratory volume in the first second of breath. Readings of pollutant levels found at the closest air quality monitoring station were then obtained, and combined with the respiratory data set, from diary information supplied by the patient.

Table 27 PEF data sample for Patient A, and the closest air quality monitoring station at the time of reading

Date	Time	Closest Monitoring Station	PEF Lung Function (L/m)
14/08/07	1:00	Brent	520
14/08/07	13:00	Brent	513
14/08/07	19:00	Brent	462
15/08/07	1:00	Hillingdon	520
15/08/07	12:00	Hillingdon	505
16/08/07	1:00	Brent	549
16/08/07	12:00	Brent	470
16/08/07	16:00	Brent	513
16/08/07	20:00	Brent	520
17/08/07	14:00	Hillingdon	513
17/08/07	19:00	Hillingdon	571
17/08/07	23:00	Hillingdon	520
18/08/07	11:00	Hillingdon	520
18/08/07	15:00	Hillingdon	520
19/08/07	1:00	Hillingdon	505
19/08/07	8:00	Hillingdon	549
19/08/07	17:00	Hillingdon	513
20/08/07	11:00	Brent	477
20/08/07	14:00	Brent	520
20/08/07	20:00	Brent	499
21/08/07	1:00	Brent	499
21/08/07	11:00	Brent	513
21/08/07	17:00	Brent	513
22/08/07	1:00	Hillingdon	527

Air quality reference datums were identified from air quality monitoring stations which lay closest to the location of the patient at the time of a respiratory reading. This data was used as the basis for calculating appropriate *delay characteristics* from the environmental data. Figure 78 shows air quality readings that were representative of Patient *A*'s location.

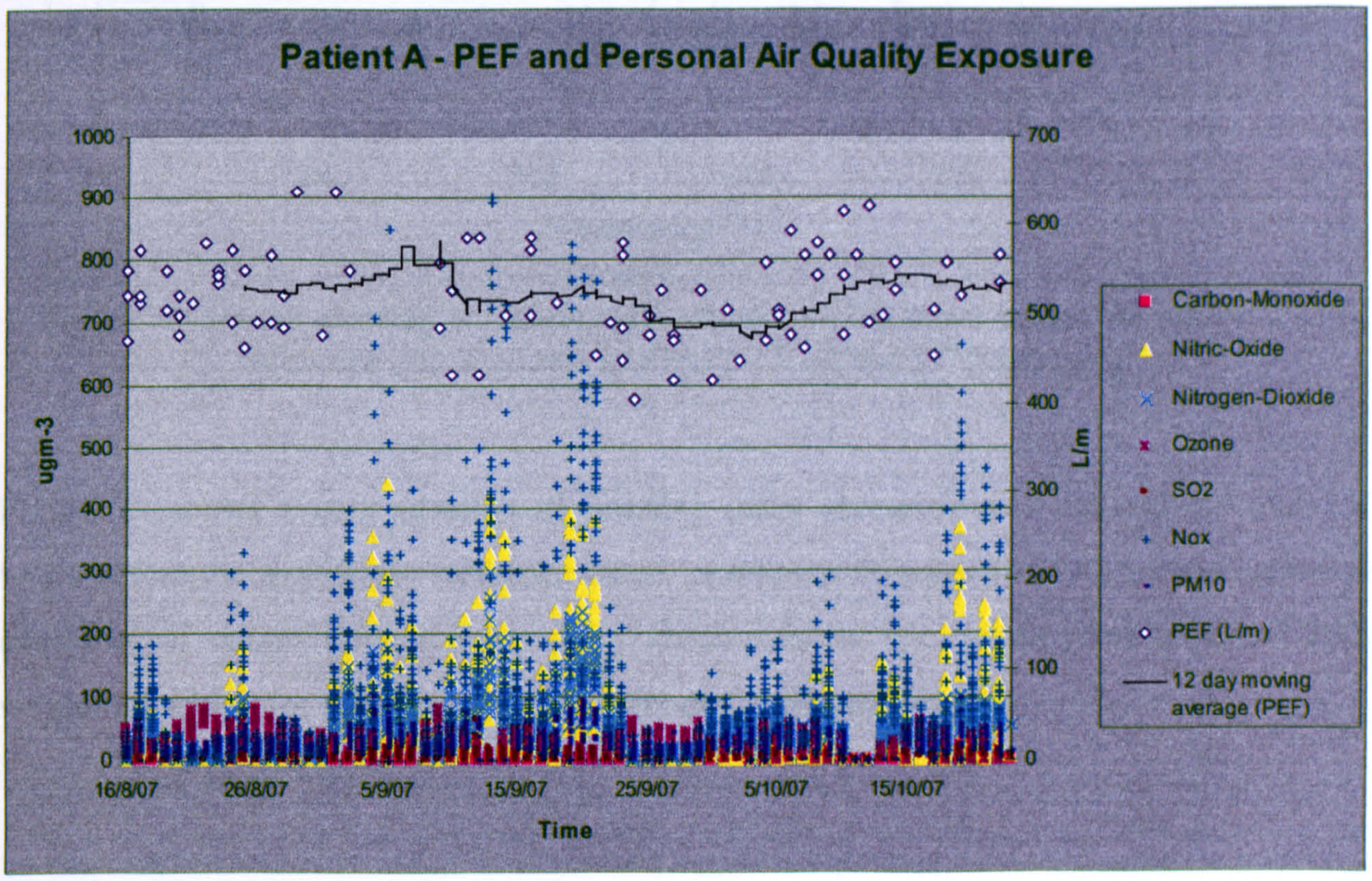


Figure 78 Personal air quality and peak expiratory flow readings for Patient A, over a period of asthma exacerbation.

Figures 79 to 81 illustrate a set of delay characteristics obtained from the ten day period prior to the PEF reference datum identified on the 27th August (Figure 77). A graph is shown for three measures of air quality. The scale on the *x*-axis of the graphs have been reversed, so the scale reads from left to right The plot begins at -240 hours (ten days before the time of the lung function reference datum) and runs to the time of the datum (time zero). The graphs show all possible delay characteristics. The EMS is selective over which of these delay characteristics are identified, and taken forward for further analysis. The EMS selects reference datums from the air quality data sets and uses these to identify the relevant delay characteristics to analyse.

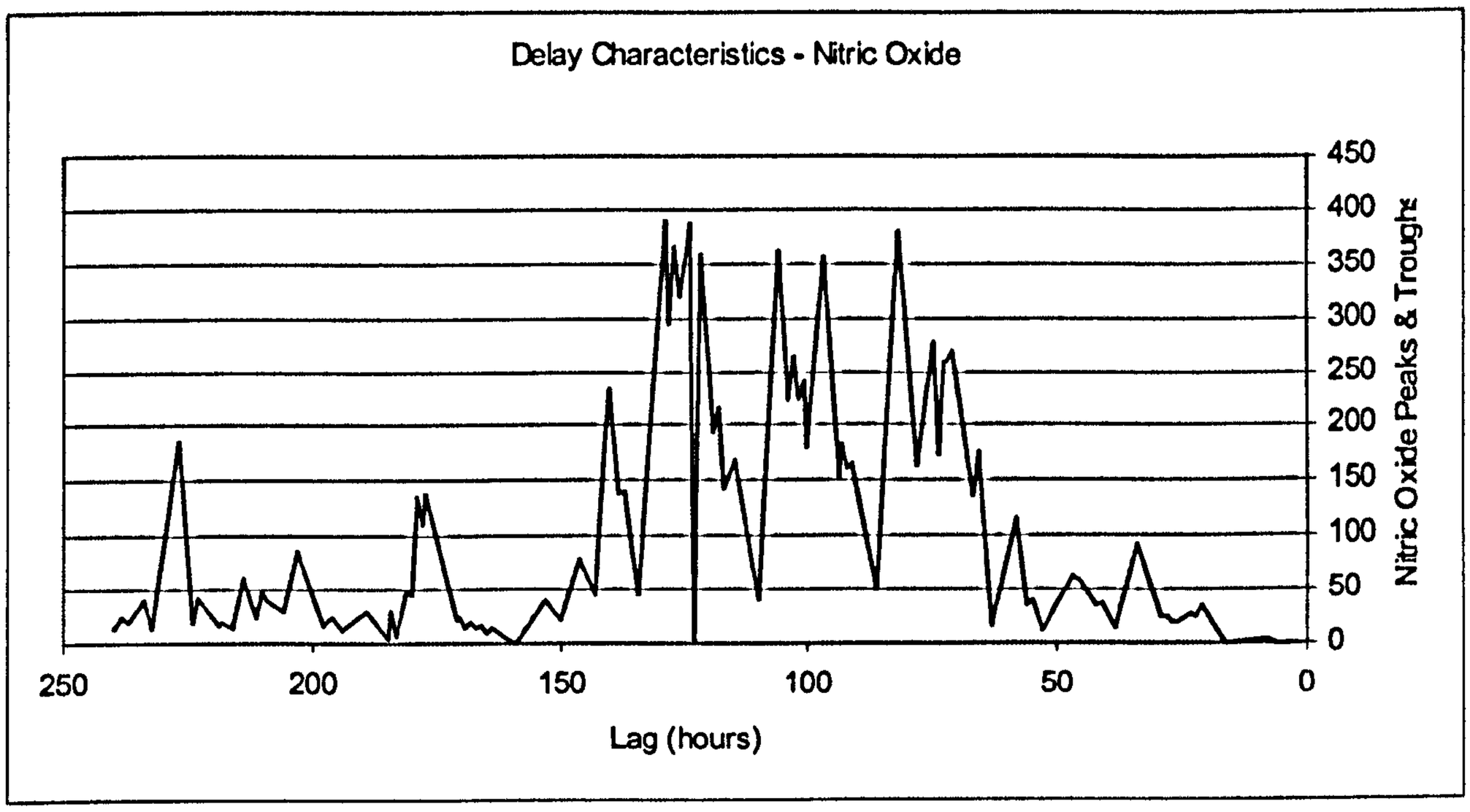


Figure 79 Graph showing the nitric oxide at varying lag times before the second PEF reference datum shown in Figure 77, on the 27th September.

Figure 79 shows a concentration of possible delay characteristics between 50 and 150 hours before the respiratory reference datum at 0 hours. Figures 80 & 81 are also given as examples of ranges that probable delay characteristics for carbon monoxide and NO_X would fall within for the respective pollutant and associated patient.

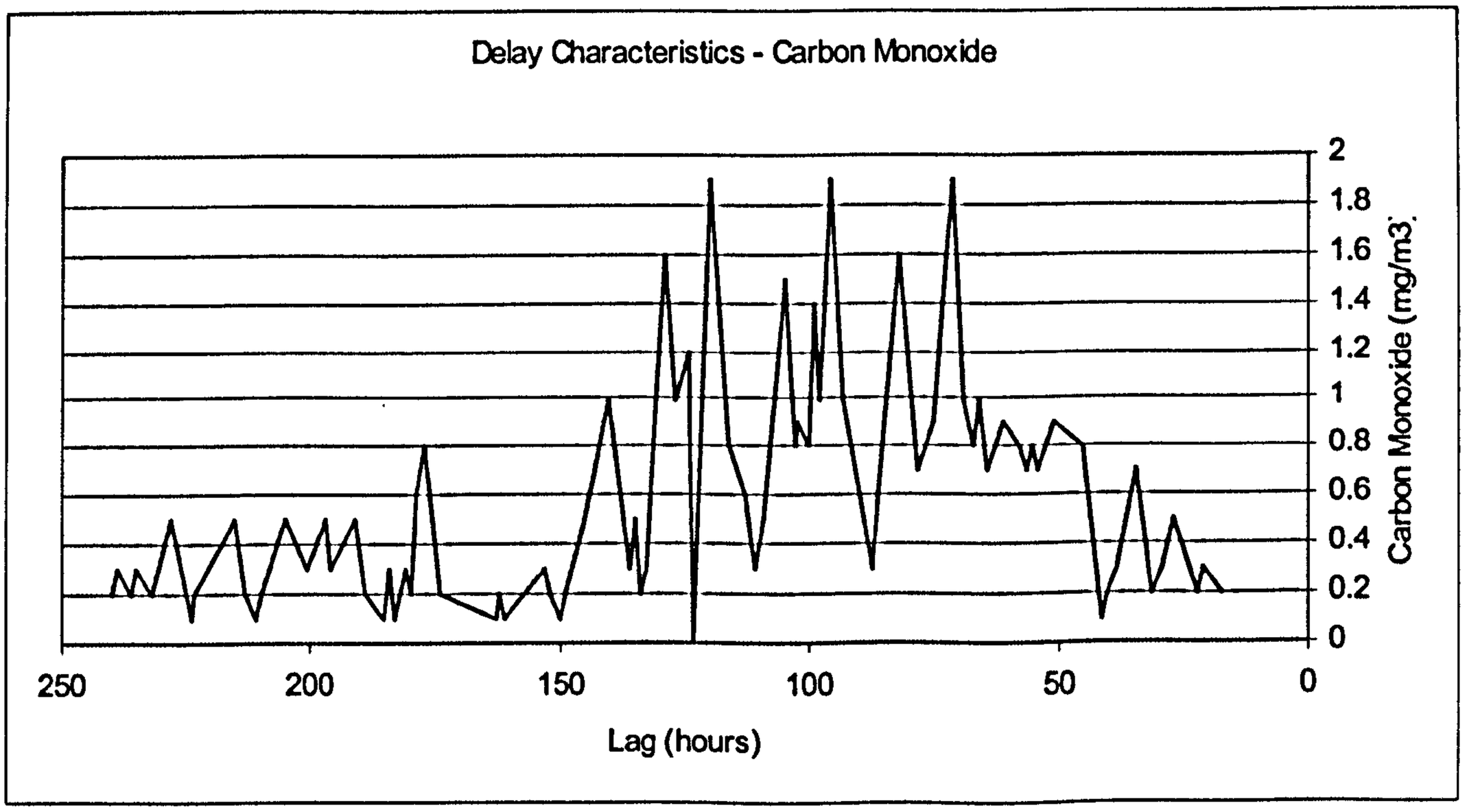


Figure 80 Graph showing the carbon monoxide at varying lag times before the second PEF reference datum shown in Figure 77, on the 27th September.

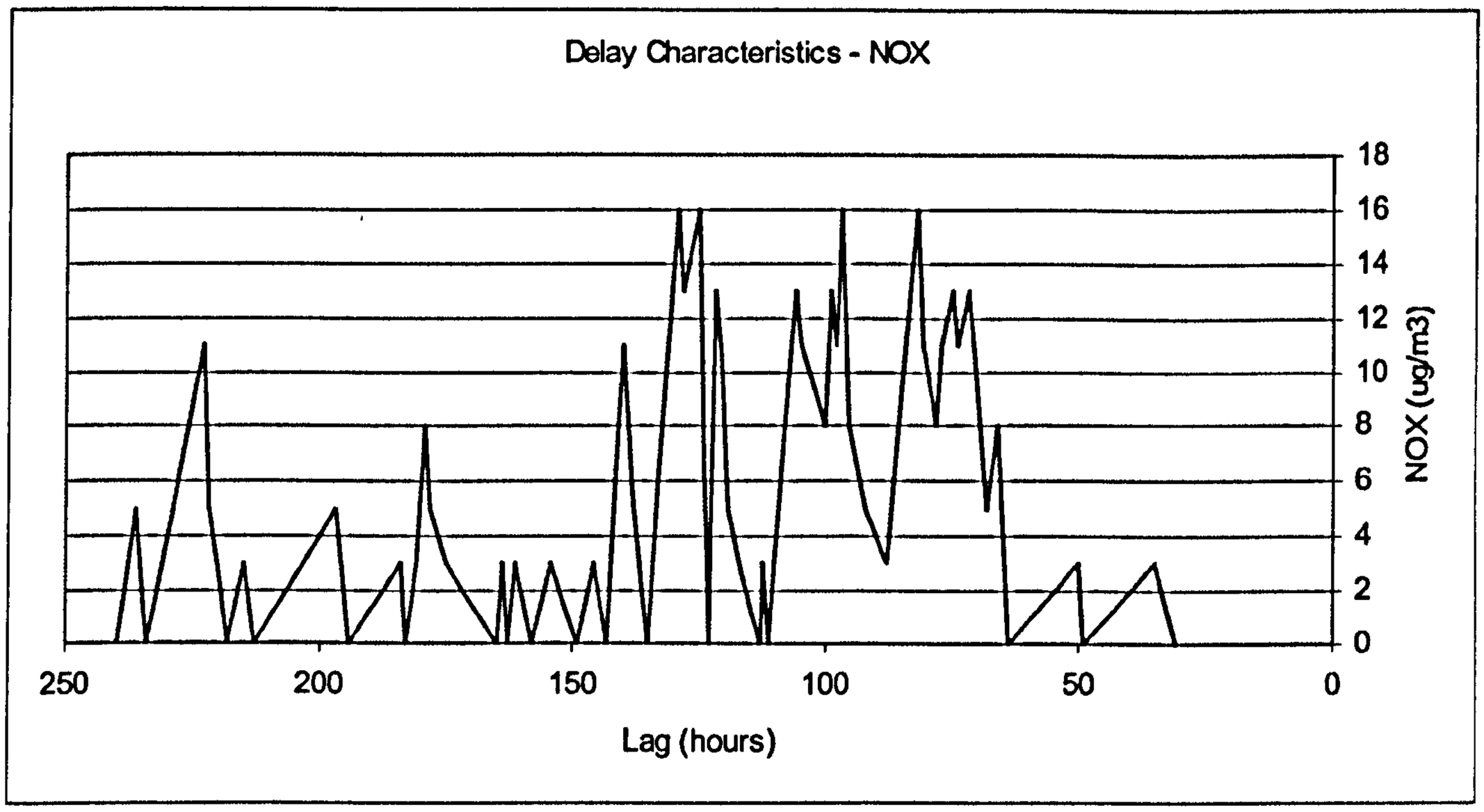


Figure 81 Graph showing NO_X at varying lag times before the second PEF reference datum shown in Figure 77, on the 27th September.

Figure 82 shows PEF lung function readings for *Patient A*, starting at 01:00 hours on 13th August 2007. A pattern of normal respiratory health follows up until the 9th September. This sequence is shown to the left of the graph, in dark blue.

The patient reported experiencing breathing difficulties on 10th September 2007, on 12th September and again on 19th September. Medical advice was sought on the 20 September 2007 after which the patient's medication was increased. This sequence is shown in red.

Then a period, between 21st September 2007 and 2nd October 2007 followed, in which the asthma condition was stabilised and medication dosage adjusted. This section of the sequence is shown in brown. The final part of the sequence, from 4th October to 22nd October 2007 shows a period returning to more normal health.

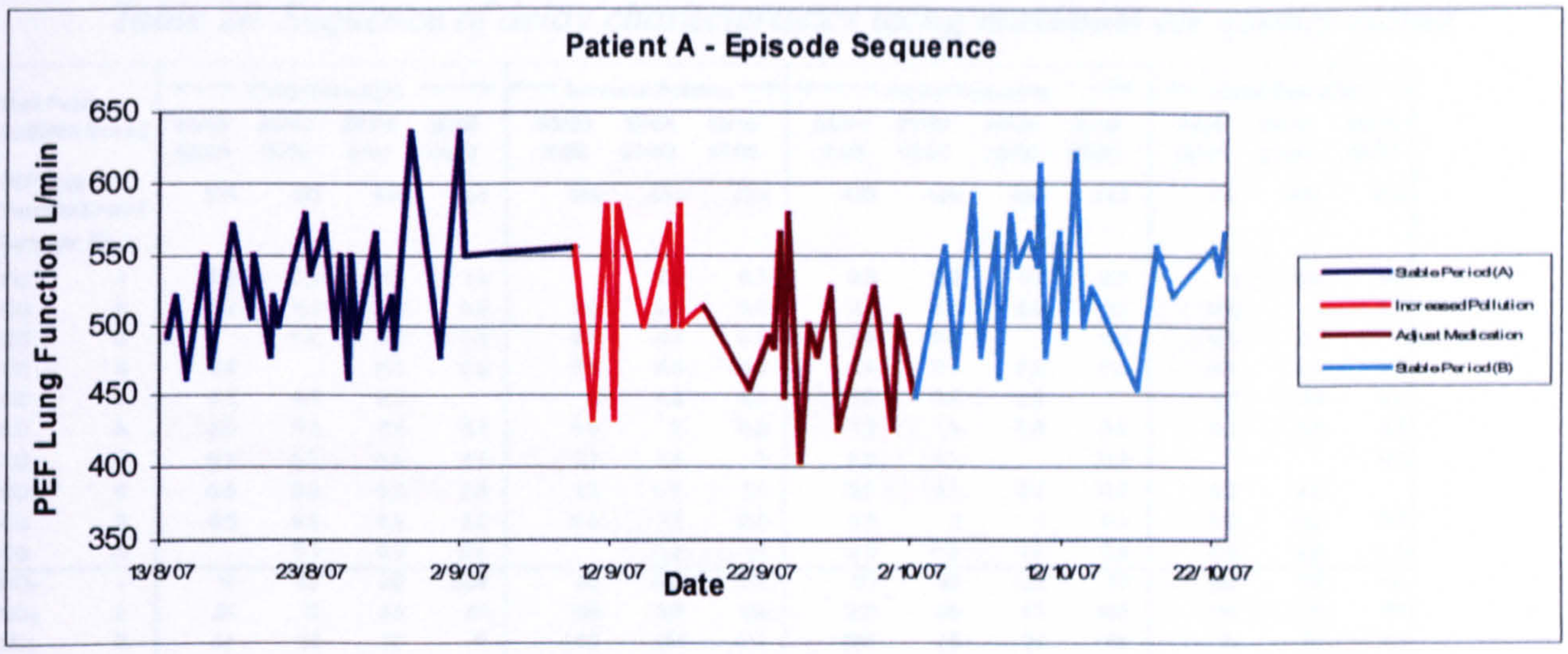


Figure 82 Sequence of data showing an asthmatic episode for Patient A.

The tabulated data shown below relates to particular data points from each part of the sequence illustrated in the figure above. The columns show the level of air quality at the time and location of each patient lung function (PEF) reading. Each pollutant is set out in a section of the table.

Rows within each section of the table refer to the number of days prior to a maximum air quality value. For example, the air quality readings in the first column occurred leading up until 02:00 hours on 23 August 2007. The maximum level of carbon monoxide occurring during the ten day period prior to 02:00 hours on 23rd August 2007, occurred six days earlier and registered $0.6 mg/m^3$. The peak value for NO_X was $183 \mu g/m^3$ and this occurred on the fifth day prior to 23rd August 2007. Nitrogen dioxide peaked at $76 \mu g/m^3$ on the sixth day before 23rd August 2007. Nitric oxide at $83 \mu g/m^3$ on the sixth day before, ozone at $82 \mu g/m^3$ one day before, PM_{10} at $38 \mu g/m^3$ eight days before, and sulphur dioxide at $5 \mu g/m^3$ three days before. All these values are well within critical thresholds prescribed by the EU (2008). Values for PM_{10} and ozone are closest to reaching their respective threshold limits at 76% and 68%. It is important to note two factors when comparing these values. This research aims to:

1) Identify predictors, and not causes of respiratory decline, and

2) Develop an analytical process that is patient specific.

Both these factors reduce the importance of threshold values designed for monitoring general populations. The important factor to consider is, if pollutant levels are capable of acting as predictors of respiratory decline or not, and triggering an alert.

Table 28 Sequence of delay characteristics using maximum air quality values

Data Point (dd/MM hh:mm)		← Stable Period (A) →				← Increased Pollution →			← Adjust Medication →				← Stable Period (B) →		
		23/08 02:00	24/08 01:00	24/08 15:00	25/08 01:00	09/09 17:00	10/09 23:00	12/09 07:00	24/09 17:00	27/09 07:00	30/09 22:00	02/10 07:00	04/10 08:00	04/10 23:00	06/10 01:00
PEF Lung Function (L/min)		535	571	491	549	556	433	433	405	426	426	448	556	470	592
Pollutant	Days														
CO	1	0.5	0.2	0.2	1.6		0.9	0.9	0.3	0.2	0.5	0.6	1	0.8	0.7
CO	2	0.2	0.5	0.2	0.2	0.7	0.6	0.9	0.7	0.2	0.2	0.5	0.6	1	0.8
CO	3		0.2	0.5	0.5	0.9	0.7	0.6	1.9	0.2		0.5	0.6	0.5	1
CO	4	0.5		0.2	0.2	0.6	0.9	0.3	1.6	0.5	0.2	0.2	0.6	0.6	0.5
CO	5	0.5	0.5	0.3		2	0.8	0.7	1.9	0.9	0.2		0.5	0.5	0.6
CO	6	0.6	0.5	0.5	0.5	0.6	2	0.9	1.9	1.9	0.2	0.2	0.2	0.2	0.5
CO	7	0.2	0.6	0.6	0.5	1.7	1.6	2	0.3	1.9		0.2			0.5
CO	8	0.6	0.2	0.3	0.6	1.3	0.6	1.6	0.8	1.9	0.7	0.2	0.2	0.2	
CO	9	0.5	0.6	0.6	0.2	0.8	1.7	0.6	0.5	1	1	0.5	0.2	0.2	0.2
CO	10		0.5	0.5	0.6		0.8	1.7	0.5	0.8	1.9	0.9	0.2	0.2	0.2
NO_x	1	17	40	38	238	42	413	479	117	25	138	97	178	157	185
NO_x	2	27	17	40	40	191	155	191	210	48	53	138	118	178	157
NO_x	3	44	27	27	17	432	191	413	596	38	25	103	71	78	178
NO_x	4	97	44	27	27	223	432	42	766	117	25	25	138	97	78
NO_x	5	183	97	44	44	850	325	229	772	210	25	25	103	138	97
NO_x	6	181	183	143	97	166	850	432	823	607	48	25	25	53	138
NO_x	7	46	181	183	183	376	709	424	181	772	38	48	25	25	103
NO_x	8	157	46	181	181	397	151	850	308	823	210	38	25	25	25
NO_x	9	31	157	147	46	292	397	166	191	512	384	117	48	25	25
NO_x	10	50	31	157	157	40	292	397	348	308	766	210	32	48	25
NO_2	1	13	31	31	78	25	124	136	61	19	71	59	73	76	84
NO_2	2	23	13	31	31	84	80	80	69	36	27	71	55	73	76
NO_2	3	36	23	23	13	107	78	124	122	32	21	40	53	55	73
NO_2	4	32	36	19	23	73	107	25	201	61	21	21	71	59	55
NO_2	5	69	32	36	36	178	96	84	233	73	19	21	40	71	59
NO_2	6	76	69	67	32	67	178	107	225	201	36	19	21	27	71
NO_2	7	40	76	76	69	97	174	122	78	233	32	36	21	21	40
NO_2	8	67	40	55	76	149	67	178	103	225	69	32	19	21	21
NO_2	9	27	67	67	40	103	149	67	97	149	122	61	36	19	21
NO_2	10	31	27	61	67	27	103	149	80	50	201	73	21	36	19
NO	1	5	6	8	149	13	189	224	36	4	44	25	69	63	76
NO	2	8	5	6	6	74	49	73	93	9	5	44	44	69	63
NO	3	9	8	5	5	211	74	189	268	4	4	41	13	15	69
NO	4	43	9	8	8	99	211	11	380	36	5	5	44	25	15
NO	5	75	43	9	9	439	149	74	363	93	4	3	41	44	25
NO	6	83	75	50	43	66	439	211	390	276	9	5	5	10	44
NO	7	14	83	75	75	165	354	196	78	380	4	9	4	4	41
NO	8	66	14	83	83	168	53	439	139	390	93	4	5	5	4
NO	9	9	66	53	14	123	168	66	85	235	176	36	8	4	5
NO	10	14	9	66	66	11	123	168	186	139	380	93	9	9	4
O_3	1	82	66	56	60	82	52	48	64	52	32	34	22	58	40
O_3	2	80	82	66	66	58	82	52	42	46	60	32	24	24	58
O_3	3	58	80	82	82	38	58	82	38	64	46	60	34	34	24
O_3	4	36	58	58	80	30	38	58	18	50	48	42	32	30	32
O_3	5	30	36	56	58	20	30	38	22	38	52	48	60	32	34
O_3	6	60	30	36	36	42	20	30	24	14	46	52	42	60	32
O_3	7	52	60	30	30	48	42	20	36	18	64	46	48	46	60
O_3	8	44	52	60	60	18	48	42	42	22	42	64	52	48	46
O_3	9	50	44	52	52	38	18	42	60	36	38	50	46	52	48
O_3	10	62	50	50	44	42	38	48	48	34	18	38	64	46	52
PM_{10}	1	30	34	25	40	27	40	73	27	23	31	42	46	35	52
PM_{10}	2	26	30	34	34	43	43	40	36	25	25	31	60	60	35
PM_{10}	3	22	26	30	30	57	43	43	68	25	26	25	42	47	60
PM_{10}	4	22	22	18	26	34	57	27	79	27	42	26	31	42	47
PM_{10}	5	30	22	22	22	99	53	43	95	36	23	42	25	31	42
PM_{10}	6	30	30	26	22	38	87	57	81	79	25	23	26	25	31
PM_{10}	7	21	30	30	30	56	99	60	33	95	25	25	42	26	25
PM_{10}	8	38	21	22	30	49	38	99	43	81	36	25	23	42	26
PM_{10}	9	13	38	38	21	35	56	38	47	69	68	27	18	23	42
PM_{10}	10	26	13	38	38	22	38	56	40	43	79	36	25	25	23
SO_2	1	3			8	3	11	11		5	29	3	8	8	13
SO_2	2		3			3	3	8	3	5		29	3	8	8
SO_2	3	5		3	3	19	3	11	8	3	3		3	3	8
SO_2	4		5	3		11	16		16		3	3	29	3	3
SO_2	5	3		5	5	21	19	3	16	3	5	3		29	3
SO_2	6	3	3			8	21	19	16	13	3	5	3		29
SO_2	7	8	3	3	3	8	16	13	3	16		5	3	3	
SO_2	8	5	8	8	3	8	5	21	8	16	3	3	5	3	3
SO_2	9	5	5	5	8	5	8	8	5	11	8		5	5	3
SO_2	10	5	5	5	5	3	5	8	11	3	16	3	3	3	5

Table 28 shows an eight week sample sequence of data (taken from the six month data set). During *Stable Period (A)* it can be seen that the levels of pollution are significantly lower than in the period of *Increased Pollution*. For example, the *Stable Period* peak value for

carbon monoxide is 0.6mg/m^3 while in the period of *Increased Pollution* the peak value has increased to 1.7mg/m^3, and later, during the period of *Adjustment* has risen to 1.9mg/m^3, before returning to a maximum peak of 0.8mg/m^3 during *Stable Period (B)*. Similar patterns are found for nitrogen oxide as nitrogen dioxide (NO_X), where in *Stable Period (A)* the maximum level is 183$\mu g/m^3$, rising to 850$\mu g/m^3$ during the *Increased Pollution* period, dropping to 823$\mu g/m^3$ in the *Adjustment* period, before reducing to 178$\mu g/m^3$ in *Stable Period (B)*. The results are summarised below by Table 29.

A number of elements in Table 29 state two figures. The un-bracketed numbers refer to peak values within the sample shown in Table 28, while numbers shown in brackets refer to peak values recorded in the remainder of the data set.

Table 29 PEF values within Patient A's six month personal air quality data set.

Lung Function (PEF) L/min	**Stable Period (A)**	**Increased Pollution**	**Adjustment of Medication**	**Stable Period (B)**
Maximum	636	585	578	621
Minimum	462	433	405	448
Average	527	523	486	532
Standard Deviation	47	57	50	50

Air Pollutant	**Maximum Values**			
Carbon Monoxide (mg/m^3)	0.6	1.7	1.9	0.8
Nitrogen Oxide as Nitrogen Dioxide ($\mu g/m^3$)	183 (298)	850	823	178 (185)
Nitrogen Dioxide ($\mu g/m^3$)	76 (78)	178	233	76 (84)
Nitric Oxide ($\mu g/m^3$)	83 (149)	439	390	69 (76)
Ozone ($\mu g/m^3$)	82 (140)	82	64	60 (66)
PM10 ($\mu g/m^3$)	38	99	95	60
Sulphur Dioxide ($\mu g/m^3$)	5	21	16 (29)	29

Table 29 shows (with the exception of ozone levels) that levels of pollution are significantly higher during the period of *Increased Pollution* and the period of *Adjustment of Medication* than in *Stable Period (A)* and *Stable Period (B)*.

The complete six month PEF data set belonging to Patient *A* (sample shown in Figure 77), and associated personal air quality were analysed using the FDA component of the EMS. Reference datums were obtained and analysed with the pattern identification components. The air quality reference datums were filtered using a threshold value during FDA, the thresholds used are shown in Table 30.

Table 30 Threshold values used for filtering air quality trend lines that did not appear above the threshold value.

	CO	***NO_X***	***NO_2***	***NO***	***O_3***	***PM_{10}***	***SO_2***
Threshold	$1.5 mgm^{-3}$	$400 \mu gm^{-3}$	$150 \mu gm^{-3}$	$250 \mu gm^{-3}$	$70 \mu gm^{-3}$	$61 \mu gm^{-3}$	$16 \mu gm^{-3}$

The EU directive on ambient air quality (EU, 2008) indicates a number of air quality levels that have been identified to have an adverse effect on health. Although these threshold values were not set in accordance with the directive, NO_2 and PM_{10} levels given in Table 30 are indicative of the directive, where NO_2 has been given a $200 \mu g/m^3$, and PM_{10}, $50 \mu g/m^3$ for limit values. Threshold values were set after Feature Detection Analysis of the six month data sample, in order to filter datums with insignificant air quality values, and to keep values that might act as predictors of a decline in lung function.

The results of the neural network, with four neurons are presented below. Each parameter shows the lag hours to the start of a potential asthma exacerbation, and the value of the air quality parameter at the time of the lag reading. The delay characteristics are ordered by the number of days prior to the asthma exacerbation from 10 days, down to 1 day.

Table 31 Delay characteristic attributes, identified by the neural network component

Lag (days)	CO		NO_X		NO_2		NO		O_3		PM_{10}		SO_2	
	Lag (hrs)	Value (mgm^{-3})	Lag (hrs)	Value (μgm^{-3})	Lag (hrs)	Value (μgm^{-3})	Lag (hrs)	Value (μgm^{-3})	Lag (hrs)	Value (μgm^{-3})	Lag (hrs)	Value (μgm^{-3})	Lag (hrs)	Value (μgm^{-3})
10													229	22
9					219	181	222	256	221	115	215	82		
8	197	1.7	195	515			185	310					198	23
7					170	200	175	257	166	86	169	82	161	23
6			138	607										
			146	406										
5					127	193			123	91	114	82		
4														
3	80	1.8	69	406	66	194			67	87			70	23
2	42	1.8					58	283			56	81		
1	16	1.8												

The delay characteristics found by the neural network in Table 31 can be compared to the maximum values of air quality given in Table 28 & 29 and the probability of air quality values in Figures 61 to 66. The important factor to note is whether or not the neural network is capable of creating an alert to warn of impending lung function decline. Figure 78 shows an increase in air pollution before a decline in lung function, however the largest increase occurs after the identified reference datums on the 24th and 27th August. This would suggest that FDA should identify the trough points of trend reversal (which it has the capability to do), so that the increase in air quality is analysed. Figure 78 also shows a period of increased air quality that occurs outside the 1 – 10 days delay range, between the 1st and 25th September. When the results above are compared with the values in Table 29 (period of increased pollution), and Table 28 (Increased Pollution, column 12/09), it can be seen that there is a very good match for all variables at a lag of 8 days (where neurons have identified a lag of 7 and 9 days, the network will identify this as a match, although at a lower activation level), except ozone which matches at 3 lag days. If further research found that the level of ozone was not important to the triggering of an alert, ozone could be removed from the analysis, and the remaining parameters would raise an alert 8 days before a decline in lung function. This alert would successfully predict the onset of Patient *A*'s decline in lung function, which occurred between 15th September and 5th October 2007.

6.9 Normalisation Test

The purpose of this test is to demonstrate how the neural network performs under the two conditions of using normalised and un-normalised input data. It is expected that the neural network, which adapts to form a fit of the underlying data will perform well with un-normalised data as has been demonstrated during the previous tests, but will have a number of restrictions. This test will identify these limitations. Using the delay characteristic permutations (shown in Table 32 below), the neural network was activated, and received the set of permutations 100 times.

Table 32 Delay characteristic permutations, shown as vectors (one per row). Parameters for the data type, date of potential air quality cause, the physical value of data, and the lag time before the event (decline in lung function) occurs.

Vector#	*Data Type*	*Date*	*Value* ($\mu g/m^3$)	*Lag*
1	PM10_ts	Fri Jan 07 17:00:00 GMT 2000	34	4 days 3 hrs
2	PM10_ts	Fri Jan 07 17:00:00 GMT 2000	34	5 days 4 hrs
3	PM10_ts	Fri Jan 07 17:00:00 GMT 2000	34	6 days 6 hrs
4	PM10_ts	Fri Jan 07 17:00:00 GMT 2000	34	7 days 4 hrs
5	PM10_ts	Sat Jan 08 11:00:00 GMT 2000	32	3 days 9 hrs
6	PM10_ts	Sat Jan 08 11:00:00 GMT 2000	32	4 days 10 hrs
7	PM10_ts	Sat Jan 08 11:00:00 GMT 2000	32	5 days 12 hrs
8	PM10_ts	Sat Jan 08 11:00:00 GMT 2000	32	6 days 10 hrs
9	PM10_ts	Sat Jan 08 19:00:00 GMT 2000	27	3 days 1 hr
10	PM10_ts	Sat Jan 08 19:00:00 GMT 2000	27	4 days 2 hrs
11	PM10_ts	Sat Jan 08 19:00:00 GMT 2000	27	5 days 4 hrs
12	PM10_ts	Sat Jan 08 19:00:00 GMT 2000	27	6 days 2 hrs
13	PM10_ts	Sun Jan 09 16:00:00 GMT 2000	32	2 days 4 hrs
14	PM10_ts	Sun Jan 09 16:00:00 GMT 2000	32	3 days 5 hrs
15	PM10_ts	Sun Jan 09 16:00:00 GMT 2000	32	4 days 7 hrs
16	PM10_ts	Sun Jan 09 16:00:00 GMT 2000	32	5 days 5 hrs
17	PM10_ts	Mon Jan 10 19:00:00 GMT 2000	45	1 day 1 hr
18	PM10_ts	Mon Jan 10 19:00:00 GMT 2000	45	2 days 2 hrs
19	PM10_ts	Mon Jan 10 19:00:00 GMT 2000	45	3 days 4 hrs
20	PM10_ts	Mon Jan 10 19:00:00 GMT 2000	45	4 days 2 hrs
21	PM10_ts	Tue Jan 11 14:00:00 GMT 2000	84	6 hrs
22	PM10_ts	Tue Jan 11 14:00:00 GMT 2000	84	1 day 7 hrs
23	PM10_ts	Tue Jan 11 14:00:00 GMT 2000	84	2 days 9 hrs
24	PM10_ts	Tue Jan 11 14:00:00 GMT 2000	84	3 days 7 hrs

The result using un-normalised data is displayed in Table 33 below.

Table 33 Neural network weights after training with un-normalised values, (converted to standard values).

Neuron (ID)	***Date***	***Value (PM10)***	***Lag***
A	Tue Jan 11 06:45:16 GMT 2000	60	1 day 10 hrs
B	Mon Jan 10 12:12:36 GMT 2000	42	3 days 5 hrs
C	Sat Jan 08 19:41:57 GMT 2000	33	4 days 6 hrs
D	Sun Jan 08 03:56:12 GMT 2000	34	6 days 3 hrs

The set of permutations (Table 32) were normalised and presented to the EMS in the same way as the un-normalised data, and the results of the analysis recorded. The results were then converted back to their un-normalised equivalents using the conversion ratios recorded during the normalisation process.

For historical interest the conversion ratios used to translate the normalised results were;

Table 34 Conversion ratios used to translate the normalised test results to un-normalised values.

	Parameter 1 (Date)	***Parameter 2 (Value)***	***Parameter 3 (Lag)***
Ratio Values	2.99E-006	17.54	1.67E-006
Bias Values	-9.47E+011	-27	-2.16E+007

The conversion ratios were applied to the respective parameter by first dividing by the ratio and the taking away the bias. This is summarised by the following formula, where the formula is used to convert each parameter into its un-normalised equivalent;

$$(\text{Parameter value } n + \text{Bias } n) \times \text{Ratio } n = \text{un-normalised value.} \qquad \textit{Eq. 6.3}$$

For example, using the normalised results (from this test) for neuron *A*;

Table 35 Normalised result for neuron A

	Date (Parameter 1)	***Value (Parameter 2)***	***Lag (Parameter 3)***
Neuron A	1791.61	234.47	999.78

The converted result for each parameter value of neuron *A* is shown below. The *Date* value is shown in milliseconds since the *epoch* (Jan 01 1970).

Table 36 Normalised and converted result for neuron A.

Neuron (ID)	***Date***	***Value (PM10)***	***Lag (millis)***
A	947599199000	84	162000000

This methodology was applied to each neuron parameter in the network. The values were then translated to a common form to make them readable, as in Table 37.

Table 37 Neural network weights after training with normalised values.

Neuron (ID)	***Date***	***Value (PM10) $\mu g/m^3$***	***Lag***
A	Tue Jan 11 13:59:59 GMT 2000	84	1 day 21 hrs
B	Mon Jan 10 19:00:00 GMT 2000	45	2 days 16 hrs
C	Sun Jan 09 03:20:26 GMT 2000	30	3 days 22 hrs
D	Sat Jan 08 02:25:47 GMT 2000	32	5 days 21 hrs

From a visual comparison between Tables 32 and 37 it can be seen that the normalised result (in Table 37) has adapted to the input data (displayed in Table 32). The generalisation of the value component of the delay characteristic is particularly accurate. However the un-normalised result (Tables 33) is *weighted down* towards lower values (33 to 60$\mu g/m^3$), the largest value of 84 recognised by the normalised result is reduced to 60$\mu g/m^3$.

The difference in results between the two methods is due to the distance measure (euclidean distance) used when identifying the closest neuron to a particular incoming pattern. This leads to greater sensitivity to large numbers in the equation (as opposed to smaller ones) such as the date and lag parameters that are usually in at least six digits as apposed to value parameters that are usually in no more than two digits.

It is worth emphasising here that the parameter values do not have to be in the same *range*, just roughly on the same scale. For example, the *Lag* component could be measured in hours which would be in a similar scale to the *Value* component.

The result of this test indicates that improved accuracy can be achieved by employing normalised data. However, the need for greater accuracy needs to be balanced against the additional processing resources that would be required.

6.10 Summary

The tests that are presented during this chapter demonstrated various features of the system architecture. *Feature Detection Analysis* (Test 1) demonstrated that the technique was sufficiently capable of identifying consistent reference datums (on which to base the further analytical components); with the capability to cope with noise levels up to 6dB. With this ability the system would be able to handle inaccuracies due to measurement and sensor irregularities.

Test 2 analysed the cumulative frequency distribution of monitored air quality parameters at four air quality monitoring stations across London. The analysis provided a basis for understanding how air quality varies between each region, and the likelihood of individual patient's exposure to levels of pollution. As the distributions indicate the probability of a pollutant being at a certain level, the graphs are useful in determining if the results given by the analytical components of the EMS are valid.

The sample of lung function and air quality data used in *Test 3* was successfully analysed

with Feature Detection Analysis. Visual inspection of the peak expiratory flow and particulate matter time series data (Figures 67 and 68) confirm the identified reference datums. The results from the subsequent FBCA and neural analysis supported each other, with the neural network identifying the data most verified by FBCA. The validation between the two analytical methods shows the advantage of a hybrid system, combining more than one method for pattern identification.

An increase in test complexity from *Test 3*, through to *Test 6* showed that the EMS is capable of handling real data sets. It should be acknowledged however, that although data sets were of significant length, use of the EMS in a clinical research environment would encounter significantly more data parameters. During *Test 4* (Multi parameter), the test successfully identified the instantaneous lag effect that a perfectly correlated air quality to lung function time series would create, and demonstrated the systems capability to analyse the effect of multiple parameters. Both the neural network and FBCA techniques performed appropriately, FBCA used a bucket width of approximately 30 minutes, identifying a uni-modal distribution for *Lag 2* (Table 19) of between *-1* and *3* hours, which correctly represented the underlying data's *boundaries*. The neural network recognised that the second lag parameter (*Lag 2*) should be allocated a neural weight (dimension) representing approximately half an hour, which correctly approximated the underlying data.

The ability to analyse a wider range of data types was corroborated through the use of hospital admissions data in *Test 5*. The test showed that the EMS can be applied to data sets outside the immediate problem domain. The analysis also showed that the neural network is capable of identifying key features within the data set; identifying similar levels of carbon monoxide, nitric oxide, nitrogen dioxide, and sulphur dioxide with three separate neurons (Table 26). The only air pollutants that were not identified by the analysis were ozone, and PM_{10}, in these cases (and in one case each for carbon monoxide, $PM_{2.5}$, and sulphur dioxide) the neural network failed to model the extremes of the parameter's data distribution due to the self-organising algorithm's use of un-normalised data.

Analysis of a six month lung function and air quality data set in Test 6 with the analytical components provided a set of results (from the EMS neural network component) that could

be compared with a significant period of patient asthma exacerbation (Figure 78). The comparison used two sources to confirm the results:

a) Table 28, containing maximum air quality values and associated delay, to peak expiratory flow reference datums.

b) The cumulative frequency distributions of four London air quality monitoring stations (Figures 61 to 66).

The result produced by the neural network (Table 31) after analysis of the full six month delay characteristic data set found that all seven air pollutants were identified as being present in significant amounts, seven to nine days before a period of asthma exacerbation.

Table 31 shows the results from the analysis, where all identified air quality values are within boundaries given by the cumulative frequency curves (Figures 61 to 66). However, the results are reflecting maximum values at the London, Marylebone roadside monitor which would suggest that personal exposure to a high level of the air pollutants has been considerable. The lag components are also representative of Figures 79 to 81 (delay characteristics). The neural network should identify patterns that are representative of the top few percent of the cumulative frequency curves, so patients are not falsely alerted to good air quality episodes. For example, a patient is notified once a month of an impending episode of poor air quality, therefore for 12 days out of 364, (or 3% of the readings) a warning should be created. However, daily peak readings would usually be grouped together within the day, further reducing the percentage of readings likely to contain peak readings by a factor of 4. It is therefore predicted that approximately 1% of air quality readings would be problematic. This indicator is useful in highlighting the (99 to 100%) region of the cumulative frequency distribution that results of the neural network would be expected to fall within.

Results from the neural network in Test 6 (Table 31) showed that carbon monoxide (CO) remained at approximately 1.8$\mu g/m^3$ throughout the ten days prior to an asthma exacerbation, but identified periods of lag were clustered between three and one day before the exacerbation period. Nitrogen oxide (NO) was found to have the opposite characteristic; lag periods of between seven and nine days before a period of asthma exacerbation were found, although a peak was also found at 2 days. The neural network

also identified a reduction in ozone (O_3) before exacerbation. Ozone levels were 115$\mu g/m^3$ nine days before an exacerbation, then reduced to approximately 87$\mu g/m^3$ from seven lag days onwards. This characteristic of ozone is found in the general literature (Anderson *et al.*, 2001). Both PM_{10} and sulphur dioxide (SO_2) readings were constant at 82$\mu g/m^3$ and 23$\mu g/m^3$ respectively throughout the ten day lag period.

The results from the *normalisation* test indicated that it is desirable, although not essential, to normalise the neural weights. In particular it was found that the use of normalisation improved the identification of the winning neuron. This provided a more precise representation of the underlying data. It was also found that the *date* component of the delay characteristic in analyses leads to a very specific analysis of the data set. Following this research it is suggested that the reference datum's *date* is only required when analysing particular causes or events within a population. Inclusion of the component at other times does not allow a predictive pattern to emerge.

Chapter 7

Conclusions and Further Research

This chapter highlights the contributions to knowledge made by this thesis, and draws conclusions about the work. Points for further research are then discussed.

7.1 Contributions to Knowledge

This thesis provided several contributions to knowledge. The following are offered as the main contributions:

- **A Process Architecture** that supports applications facilitating the identification of significant, and repeatable, environmental predictors of patient-specific periods of asthma exacerbation.
- **Feature Detection Analysis,** a method that identifies trend reversals within air quality and respiratory time-series data.
- The **Delay Characteristic** technique, that associates two *features* together, using the factor of time, and allowing the *features* within these associations to be validated. The delay characteristic holds sufficient information to form an alert.
- The application of a **Self-organising Map** (SOM), an unsupervised machine learning method, to validate the delay characteristics.
- **Frequency, Boundary and Cluster Analysis** (FBCA), a method that analyses the underlying frequency distributions of the delay characteristics, to overcome problems of *over-fitting* data within the SOM.

Associating a decline in lung function leading to a period of asthma exacerbation, to a change in environmental condition was a major research objective outlined during Chapter 1 (Section 1.2). A three-step process (Section 1.2.1) was proposed, and then shown in Chapter 5 (Sections 5.5, 5.7 and 5.8) to extract information relevant to producing patient-specific alerts.

Process Architecture

The **architecture** developed as part of this research has several features:

- designed to facilitate methods capable of identifying patient-specific predictors (Sections 3.5, 4.2.2 and 4.2.6, and Chapter 5);
- defines a scalable process (Chapter 3, and Section 4.2.3);
- component-oriented, this gave applications (such as the EMS) flexibility in the choice of implementation and enabled interchange of analytical components (Chapter 4);
- extensible, due to the ability to add analytical modules (Section 4.4);
- capable of handling multiple parameters (demonstrated in Section 6.6 and 6.7.2);
- assists research into the *cause and effect* relationship between environmental and respiratory data sets (Chapter 6).

EMS Validation of Architecture

The Environmental Monitoring System (EMS) was pivotal in forming the process architecture developed by this thesis. The EMS was developed as a prototype and implemented the process architecture defined by this thesis.

While clinical procedures are often focused on the trend of a group of patients, the analysis of data by the thesis prototypes was patient-specific. This enabled the response of the system to be tailored to particular individual's sensitivities. Results in Chapter 6 showed that an automated pattern identification system focused on the detection of events in the environment can successfully create information to alert patients, to aid them in the avoidance of environmentally induced asthma episodes.

The analysis of generic data types with a time and location component was achieved, and demonstrated through tests with air quality data from a number of sources along with admissions data and patient lung function. The supplementary test using hospital

admissions data (Section 6.7) showed the system's capability to handle data sets containing additional data types not related to lung function. The architecture implemented by the EMS application, allowed the introduction of additional methods of analysis. The analytical methods are shown in Chapter 5, in particular the *Point*, *Point Series* and *Series* methods.

The control relationships between a number of subsystems and a set of analytical techniques were identified (Chapters 4 and 5).

Feature Detection Analysis

Feature Detection Analysis (FDA), originated during this thesis and is offered as a technique to identify the onset of an asthma exacerbation (Section 5.2). The capabilities of the analysis lead to its further use for monitoring air quality, and within the boundary identification module of the Frequency, Boundary and Cluster Analysis (FBCA) prototype that identifies cluster boundaries.

It is the ability of the FDA component to identify key features from a data set that makes the analysis viable; ignoring sections of a trend that do not have significance, but recording those that do. The ability of the process to cope with highly variable data is particularly important. Inaccuracies in data measurement are emphasised as the patient moves from location to location. Various devices could be used and all may have slightly different calibrations. The London Air Quality Monitoring Network use a working error tolerance of ±10% (LAQN, 2008), which the EMS application's design has taken into account. FDA reduces the effect of inaccurate data readings by analysing the data trend, rather than using a maximum or daily average for the monitored parameter. The pattern identification components also validate any delay characteristics presented to them over a period of time. This additional benefit further reduces the probability of inaccuracies within the system. Feature Detection Analysis (FDA) also automates (with the option of guidance by the user) the identification of reference datums, options include the ability to recognise if a trend falls below a certain threshold, if the trend is greater than a given gradient, in addition to identifying periods where the trend reverses.

Delay Characteristic

This research proposed the analysis of *time lag* (Section 3.3) between a *reference datum* attributable to a possible environmental predictor, up until a reference datum attributable to the start of a patients asthma exacerbation; using the *delay characteristic*. The use of the delay characteristic *Date* parameter is useful in analysing a location-based environmental problem (due to the ability to make a connection between the date and the location of the patient). However, the inclusion of the *Date* component under normal analyses leads to too many clusters, and not enough validation of the time *Lag* which is the more important characteristic.

SOM Validation of Delay Characteristics

The Self-organising map (SOM) method is capable of handling many parallel variables. The SOM algorithm reduces the effect noise has on the system due to the algorithms convergence to a suitable answer over time. Overfitting of the underlying data can be problematic; it is a consequence of too many neurons being issued during the analysis, and the arbitrary way in which node boundaries are established. The SOM algorithm does not have a *stop* condition for the convergence of the neural network, other than a set number of iterations. The use of FBCA within the EMS to identify a probable number of clusters overcomes this deficiency.

Frequency, Boundary and Cluster Analysis (FBCA)

Frequency, Boundary and Cluster Analysis (FBCA) was developed as part of an application of the architecture during Section 5.6 as a means of overcoming the problems of overfitting data, associated with SOM networks (Section 5.5.3). FBCA analyses the actual distribution of parameter values to establish node boundaries, while the SOM algorithm establishes boundaries based on a function that reduces over time according to the number of data elements processed.

FBCA provides a particularly useful indicator, making the informed disregarding of parameters possible. As FBCA produces a frequency distribution for each analysed

parameter, a uni-modal distribution implies that only one range will be produced for that parameter during creation of the cluster permutations. This leads to the same range appearing in all clusters, a condition which provides no differentiation between clusters in respect of that particular parameter. This condition suggests that the parameter is less likely to be a factor in the initiation of an asthma attack. Further analysis of that particular parameter is required however, in relation to the other parameters before such conclusion could be confirmed. It is possible that pre-disposition to an asthma attack might be associated with specific parameter values within the unified distribution range. This association could only become apparent after the vector of values collected by each cluster had been examined in more detail.

Although not fully developed in this thesis, the use of *frequency analysis* as a means of defining cluster boundaries opens the prospect of combining the advantages of the *neural network*, with more traditional statistical analysis. The major advantage of the neural node is that it provides a means of classifying and immediately identifying a particular vector of environmental conditions. A disadvantage of the neural network technique is that it provides only an arbitrary classification of the data, and although certain data sets may be assigned to different network nodes, it does not necessarily indicate that the identified data sets are drawn from statistically-significant different data. It was this deficiency which prompted the development of FBCA.

Through prototyping the analytical components, in particular the neural network (SOM) and FBCA components, the usefulness of operating two supportive methods in the same system was highlighted. During evaluation, test results could be corroborated by both techniques. The implementation of both components into a hybrid system increases system reliability by confirmation and enhancement of results.

Use of correlation techniques during the course of the research were shown to be an option for identifying relationships between air quality and lung function. However, the technique was found to be susceptible to noise and a slow process due to the infinite number of correlation calculations that were required, especially when lag effects were analysed.

7.2 Recommendations for Further Research

This research has taken the first step in defining a scalable system architecture suitable for facilitating predictor identification of asthma exacerbations. The work begun by this thesis requires further research to refine the analytical techniques before being taken forward for testing in a clinical, and *real-time* environment. With the eventual view of developing a fully functioning clinical system. A number of recommendations can be made to guide future research into a suitable system.

7.2.1 Data Analytics

A number of general analytical aspects of the work require further research, these include: the effect of medication on automated pattern recognition; the automated choice of environmental factors to monitor; the identification, and use of predictors belonging to groups of patients; and improvements to Feature Detection Analysis.

Medication Details

Medication details were recorded for a minority of patients during the Medicate (2000) clinical trial and the work of Cobern *et al.* (2005). It is clear that the use of medication will have a significant influence as to how environmental conditions will affect a patient's lung function. Inclusion of this important information as a parameter, to explain changes in response, would be a useful addition to the system. Enabling, for example, the effectiveness of various medications and their dosage to be evaluated. Research should be undertaken into how best to incorporate this additional information into the EMS.

Factor Analysis

It is envisaged that the selection of environmental parameters for inclusion in analysis by an application (such as an extended version of the EMS) would be facilitated through the addition of some automated *factor analysis*. Methods of factor analysis appropriate to this research problem should be researched, and incorporated into the system for this purpose.

Group Studies

There is a strong argument that *a group* of patients may all suffer from the same type of environmental predictors. It may become apparent through further research that a number of identification components analysing patient-specific predictors may indicate similar

characteristics for groups of patients, and could therefore be used to confirm a particular trait for a group of patients. Analysis over a group of patients using the patient-specific techniques developed by this thesis is an area where further research may prove beneficial.

Feature Detection Analysis

It was found during module prototyping that FDA had a tendency to shift reference datums to the right of maximum values. To overcome any inconsistency resulting from this shift, FDA could be adapted by adding a peak locking mechanism. This would enable the system to lock onto peaks by locating the point of inflexion (by checking left/right), ensuring that the true peak (reference datum) had been correctly identified. However caution would be required in adoption of this technique so as not to influence the analysis away from the *trend* of the data.

7.2.2 Extending the Self-organising Map (SOM)

Further research should be undertaken into the issue of when and how neural network nodes split. For the SOM technique to achieve a representation of commonly occurring patterns, the network is required to grow and adapt to new data. As nodes contract their area of activation, less input data is recognised by the neuron. To compensate, the system requires a way to add new nodes to the network. The process is shown in Figure 83.

Work by Kohonen and summarised by Schalkoff (1997) showed that input data can be characterised by a single density function, and that the point density function of each neurons weight vector will approximate to that single density function. If the functions are compared using a variance ratio a probability that the neuron under question will be activated in comparison to the others in the network might be used as the basis for determining a method of system reliability.

Measurement of system reliability has three scenarios:

1. Initiating a false alarm to a patient or clinician, or alternatively
2. Failing to alert when a patient is at risk of an impending asthma attack, or
3. Alerting patient or clinician correctly.

Figure 83*a* shows the *neighbourhood* distribution of a node (*Node A*) at the beginning of the identification process. As more and more data is processed the distribution covered by *Node A* is contracted. This is shown by Figure 83*b*. The area covered by the system is significantly less than the area covered by the node at the initiation of the process. Hence a new node is required in order to ensure that the whole area is covered by the system. A new node (*Node B*) would be added to the system (Figure 83*c*). It is necessary to ensure that when an input falls outside an existing neighbourhood of the network, a new neuron is added to cover this new input otherwise it will not be included in any further analysis.

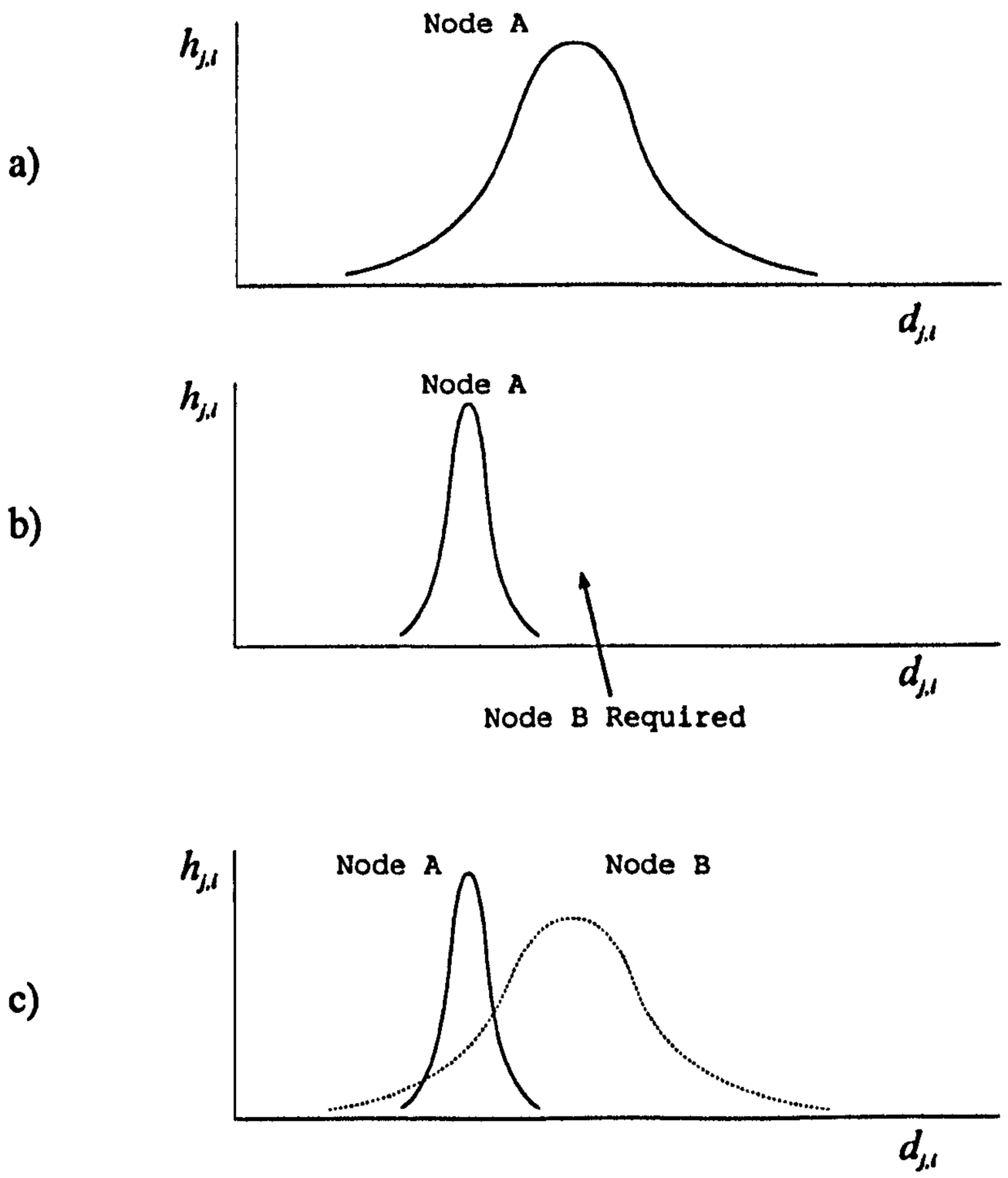


Figure 83 A second node (neuron) is added to the neural network to represent new data, not previously covered by the first neuron. The activation of the neuron is indicated by $h_{j,i}$ whilst the relative distance between neurons and the input data is represented by $d_{j,i}$.

It is important to prevent overfitting the data. The input should be represented sufficiently but with the minimum number of neurons. A minimum solution increases the model's resistance to noise. Separation between neighbouring neurons is needed in so far as each neuron is required to represent a new or significantly different pattern. The neural network should be capable of finding these patterns automatically over a period of time.

A characteristic of the SOM algorithm is that the neighbourhood function tends to lock onto a particular input pattern established by the contents of the initial processed data vectors. While this characteristic is not necessarily unwanted, it means that the algorithm does not easily adapt to data patterns which might emerge at a later time. To improve the adaptability of the SOM algorithm, Haykin (1999, *pg 481, Table P9.8*) proposed the withdrawal of a node from further learning if the node is constantly a winner. This would allow other nodes in the network to adapt to represent the data distribution.

Examination of neurons representing patterns closest to those with the most interest, possibly through the use of visualisation of the self-organising map, could be used to identify clusters of *predictors* that would not normally be identified as resulting in the same health outcome.

Through the use of the SOM technique it is possible to visualise relationships and associations between identified patterns. Similar patterns will be physically close to each other, determined by the (euclidean) distance between them. Future research could investigate the use of a priority weighting to give certain parameter dimensions more prominence, and show more reflective relations between the represented patterns.

7.2.3 Frequency Boundary and Cluster Analysis Refinement

The form of frequency analysis employed in the thesis prototypes was basic. Statistical techniques for examining frequency distributions and fixing cluster boundaries (that are of statistical significance) are well developed and could be incorporated within an application such as the EMS. Such techniques include:

a) the Anderson-Darling Test (Stephens, 1974), which could be used to test if the data is a continuous normal distribution.

b) the Chi-Square Test which could be applied to discrete distributions (Snedecor & Cochran, 1989), or

c) to test for *outliers*, that is, identifying data points which, given the statistical mean and standard deviation of the total set, are so far removed from the mean value, that it is extremely unlikely that they are members of the same data set (Grubbs, 1969).

Although these techniques are well developed, their use within the FBCA component would require further research, particularly their use within multidimensional spaces.

The ability of the *Boundary Clustering* technique to change cluster boundaries in response to the input data presented to the system, is sufficient when employed on batch data. However, to facilitate real-time analysis it will be necessary to develop a process which will systematically re-evaluate the cluster boundaries in response to changes in the input distribution, and influence the neural network component in an appropriate way.

7.2.4 Service Architecture Implementation

The use of concepts like Service Oriented Architecture (SOA) lend to the notion of extensibility and scalability that are promoted by the five features of the ISO model (ISO, 2007). SOA combines relatively large (previously un-associated) units of functionality into a *service* to achieve the execution of a specific task. The service can then be subsequently used by other systems. SOA is often used over a network to facilitate the use of computational resources.

Research into how SOA could be used to implement the EMS process architecture, and enhance the analysis of enviromedic data, through the autonomous integration of appropriate services should be considered. Integration with other healthcare providers could also be considered as a further research area.

As technology improves, the opportunity to integrate new devices, providing new monitoring functionality becomes possible. The initiative called Sensor Web Enablement (OGC, 2005) outlines a framework to help exploit the advantages of web-connected sensors and sensor systems. The aim of their work is to make sensors, instruments and imaging devices available over the Internet. The sensor's capabilities, location and interfaces can be published using XML, thereby enabling web-based (and real-time) discovery of suitable monitoring devices and collection of their information. Integration of these types of monitoring device with the EMS is encouraged, as they become available, but requires further research.

Future research could be undertaken that promotes the use of ambulatory Internet enabled, and inter-connected medical devices; facilitating around-the-clock monitoring and execution of treatment protocols. Instant alerts could be generated when medication levels were low, if environmental conditions matched a pre-determined pattern, or if adverse

events were detected or malfunctions occurred.

7.2.5 Using Grid Concepts

A grid is not subject to centralised control, and integrates and coordinates resources and users that are within different control domains. A grid addresses fundamental issues such as authentication, authorisation, resource discovery, and resource access, using standard, open, general-purpose protocols and interfaces. The goal is that the utility of the combined system is significantly greater than the sum of its parts (Foster, 2005).

The NEESGrid project (Spencer *et al.*, 2004) shows how a distributed system was implemented using Grid technology to link earthquake researchers across the United States with leading-edge computing resources and research equipment to allow complex simulations. Other Grid technology applications include: Disease and Bio-Surveillance Grid, Pervasive Mobile Environmental Sensor Grid (IET, 2006), and Personalised Healthcare. Research into how the EMS could benefit from the use of Grid methodology is required, especially to optimise scalability. Research into the use of the EMS within a geographically diverse collaboration of interacting systems could also be undertaken to discover how air quality patterns that cross large areas could be used to alert patients to impending asthma exacerbation earlier.

7.2.6 Clinical Testing

Further research and development is required if the work presented by this thesis is to be of eventual value to clinicians (and their patients). As part of this further research, clinical trials on a significant scale are vital to test the analytical processes. Clinical trials aimed at both refining the analytical processes and proving them with a wider sample of the population are necessary if confidence in the new techniques presented by this thesis is to be gained.

Eventually, research, testing, and feedback from a clinical environment should lead to a system capable of giving clinical staff and patients *real-time* information to assist in their choice of preventative action, and avoidance of asthma exacerbations.

7.3 Concluding Summary

Future systems implementing the architecture developed by this thesis, will promote improved patient care by providing advanced warning of impending periods of poor air quality. Patients will benefit from automatic prompts to use preventative medication before specific danger periods, enabling them to stabilise their condition or prevent deterioration all together. Clinical staff could also be notified that a particular patient needs extra care. The cost of patient care would be reduced by focusing treatment on individual patients at the time, and for the period they need it most. A system capable of alerting patients to environmental health effects is not new (Cobern *et al.*, 2005), but the architecture developed here, and its capability, and the depth to which the pattern recognition detail goes, along with the drawing together of the two data sets (environmental and respiratory) into the field of ***enviromedics***, and use of the *delay characteristic* to identify patient-specific trends is new and novel.

References

(AEAT, 2006) Bower J, Lampert J, Broughton G, Stedman J, Pye S, Willis P, Targa J, Kent A, Grice S, "Air Pollution in the UK: 2005", AEA Technology (2006), found at; http://www.airquality.co.uk/archive/reports/reports.php

(Aikins *et al.*, 1983) Aikins J S, Kunz J C, Shortliffe E H, Fallat R J, "PUFF: An expert system for interpretation of pulmonary function data", *Computers and Biomedical Research*, 16, 3, June 1983, 199 – 208.

(Air Monitors, 2006) Air Monitors Ltd. Unit 2, The Hawthorns, Staunton, Gloucestershire, GL19 3NY, United Kingdom; http://www.airmonitors.co.uk

(Alaert, 2002) Alaert F-A, Blobel B, Louwerse K, Barber B, "Security Standards for Health Information Systems", IOS Press 2002, ISBN: 1586030000.

(Anderson *et al.*, 2001) Anderson W, Prescott G J, Packham S, Mullins J, Brookes M, Seaton A, "Asthma admissions and thunderstorms: a study of pollen, fungal spores, rainfall, and ozone", *QJM: An International Journal of Medicine*, 2001, 94, 8, 429-433. Found also, http://qjmed.oxfordjournals.org/cgi/content/full/94/8/429

(Anderson K *et al.*, 2001) Anderson K , Qiu Y, Whittaker A R, Lucas M, "Breath sounds, asthma, and the mobile phone", *The Lancet*, 358, 9290, October 2001, 1343 – 1344.

(Anhoj and Moldrup, 2004) Anhoj J, Moldrup C, "Feasibility of collecting diary data from asthma patients through mobile phones and SMS (short message service): response rate analysis and focus group evaluation from a pilot study." *J Med Internet Res*, 2004, Dec 2, 6(4),e42.

(AQMD, 2003) The South Coast Air Quality Management District (2003), 2003 *Air Quality Management Plan (AQMP)*, http://www.aqmd.gov/aqmp/docs/2003AQMPChap2.pdf

(AsthmaUKa, 2007) AsthmaUK, "The Asthma Divide: Inequalities in emergency care for people with asthma in England", found at: http://www.asthma.org.uk/

(ATS, 2007) American Thoracic Society Quality of Life Resource, www.atsqol.org, Copyright © 2007 American Thoracic Society.

(Avgeriou and Zdun, 2005) Avgeriou P, Zdun U, "Architectural patterns revisited:a pattern language", *10th European Conference on Pattern Languages of Programs* (EuroPlop 2005), Irsee, Germany, July.

(Ayres J, 2005) Ayres J, *Understanding Asthma,* Family Doctor Publications Ltd in association with The British Medical Association, 2005, ISBN: 1898205647 also found at http://www.mypharmacy.co.uk

(Barber, 2006) Barber D, "Machine Learning, A Probabilistic Approach", http://www.idiap.ch/?barber, found at: http://web4.cs.ucl.ac.uk/staff/D.Barber/courses/mlgm_epfl_book.pdf

(Barber *et al.*, 2002) Barber A, Bayford R H, Hamilton R ; "An Informatics based approach to Respiratory Healthcare", Proceedings of *Healthcare Computing 2002;* ISBN 0953542769.

(Bertino & Martino, 1993) Bertino E, Martino L, *Object-Oriented Database Systems, Concepts and Architectures*, Addison Wesley, ISBN: 0201624397.

(Beuchat *et al.*, 2005) Beuchat A, Taub S, Saby J-D, Dierick V, Codeluppi G, Corno A F, von

Segesser L K, "Cybertools improve reaction time in open heart surgery", *European Journal of Cardio-Thoracic Surgery*, 27, 2, 2005, 266-269.

(Bewick *et al.*, 2003) Bewick V, Cheek L, Ball J; "Statistics review 7: Correlation and regression", Crit Care. 2003, 7,6, 451–459, BioMed Central Ltd, doi:10.1186/cc2401, found at http://ccforum.com/content/7/6/451

(Blanc P D *et al*, 2005) Blanc P D, Eisner M D, Katz P P, Yen I H, Archea C, Earnest G, Janson S, Masharani U B, Quinlan P J, Hammon S K, Thorne P S, Balmes J R, Trupin L, Yelin E H; "Impact of the home indoor environment on adult asthma and rhinitis"; *Journal of Occupational and Environmental Medicine*, Apr 2005, 47 (4), 362-372.

(Bray *et al.*, 2006) "Extensible Markup Language (XML) 1.0 (Fourth Edition) W3C Recommendation", Copyright © 2006 World Wide Web Consortium, (Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University), found at http://www.w3.org/TR/2006/REC-xml-20060816/

(Brunekreef and Holgate, 2002) Brunekreef B, Holgate S T, "Air pollution and health", Lancet 2002, 360, 1233 – 42.

(BTS, 1995) The British Thoracic Society et al., *The British guidelines on asthma management: 1995 review and position statement,* Supplement to February issue of Thorax 1997, 52, 51.

(BTS, 1997) The British Thoracic Society Standards of Care Committee, *Guidelines on the Management of COPD*, Supplement to December issue of Thorax 1997, 52.

(BTS, 2004) The British Thoracic Society and Scottish Intercollegiate Guidelines Network, *British Guideline on the Management of Asthma: Quick Reference Guide*, ISBN: 1899893288.

(Buschmann *et al.*, 1996) Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M; *Pattern Oriented Software Architecture, A System of Patterns*; John Wiley & Sons 1996; ISBN 0471958697.

(Carrer *et al.*, 2001) Carrer P, Maroni M, Alcini D, Cavallo D, "Allergens in indoor air: environmental assessment and health effects", *Science of the Total Environment*, 270 (2001), 33 – 42, Published by Elsevier Science B.V.

(Chauhan *et al.*, 2003) Chauhan A J, Inskip H M, Linaker C H, Smith S, Schreiber J, Johnston S L, Holgate S T, "Personal exposure to nitrogen dioxide (NO_2) and the severity of virus-induced asthma in children", *Lancet* 2003, 361, 1939 – 44.

(Cheesman and Daniels, 2001) Cheesman J, Daniels J, "UML Components – A Simple Process for Specifying Component-Based Software", Addison-Wesley, 2001, in Heisel M, Souquieres J, "Adding Features to Component-Based Systems", p137-153, *Lecture Notes in Computing Science: Objects, Agents, and Features: International Seminar*, Dagstuhl Castle, Germany, 2003, Eds Ehrich H-D, Meyer J-J, Ryan M D, Springer 2004, LNCS 2975.

(Chen *et al.*, 2007) Chen H, Gould M K, Blanc P D, Miller D P, Kamath T V, Lee J H, Sullivan S D, "Asthma control, severity, and quality of life: Quantifying the effect of uncontrolled disease", *Journal of Allergy and Clinical Immunology*, 120, 2, August 2007, 396 – 402.

(Chin-Shen *et al.*, 2007) Chin-Sheng T, Li-Te C, Hsien-Chi L, Chang-Chuan C, "Effects of personal particulate matter on peak expiratory flow rate of asthmatic children", *Science of the Total Environment* 382 (2007) 43 – 51, Published by Elsevier B.V. doi:10.1016/j.scitotenv.2007.04.016

(Choe and Yoo, 2008) Choe J, Yoo S K, "Web-based secure access from multiple patient repositories", *International Journal of Medical Informatics*, 77, 2008, 242 – 248.

(Cleland *et al.*, 2007) Cleland J, Caldow J, Ryan D, "A qualitative study of the attitudes of patients and staff to the use of mobile phone technology for recording and gathering asthma data", *Journal of Telemedicine and Telecare* 2007, 13, 85-89.

(Cobern *et al.*, 2005) Cobern W R, McSharry P E, Tarassenko L, "Telemedicine to assist patient understanding of atmospheric influence on lung function and improve real time control of mild-to-moderate asthmatics", Proceedings of *3rd European Medical and Biological Engineering Conference (EMBEC 2005), November 2005.*

(Cochrane *et al.*, 1996) Cochrane G M, Jackson W F, Rees P J, Asthma D Current Perspectives. Published: Mosby-wolfe 1996, pg 7 in National Asthma Audit 1999/2000 Summary. The National Asthma Campaign.

(Coiera, 1997) Coiera E. "Guide to medical informatics, the internet and telemedicine". London: Chapman and Hall, 1997; ISBN 0412757109.

(COMEAP, 1998) *COMEAP statement on banding of air quality*, COMEAP for the Department of Health, Copyright 1998. also found at http://www.advisorybodies.doh.gov.uk/comeap/statementsreports/

(COMEAP, 2002) *Air pollution: what it means for your health, the public information service*, Department for Environment, Food & Rural Affairs, Crown Copyright 2002.

(COMEAP, 2006) Ayres J G et.al., *Cardiovascular Disease and Air Pollution: A report by the Committee on the Medical Effects of Air Pollutants*, Department of Health, Crown Copyright 2005. also found at www.dh.gov.uk/publications

(Cooper & Masden, 2000) Cooper B G, Masden F, *Spirometry*, Eur Respir Buyers, 2000, 3. also http://www.personal.u-net.com/~ersj/

(Crabbe *et al.*, 2001) Crabbe H, Machin N, Hamilton R, Barber A, Bayford R, MEDICATE Deliverable 4.2 Final Report, *Effects of Environmental Factors on Respiratory Health (Integrating Environmental Air Quality Data)*, EC DGXIII Ten-Telecom Programme, contract: TEN 45608 (FS).

(Crabbe *et al.*, 2004) Crabbe H, Barber A, Bayford R, Hamilton R, Jarret D, Machin N, "The use of a European telemedicine system to examine the effects of pollutants and allergens on asthmatic respiratory health", *Science of the Total Environment*, 334-335, 2004, 417-426.

(Crowley, 1985) Crowley J, "Navigation for an Intelligent Mobile Robot", *IEEE Journal of Robotics and Automation*, 1985, RA-1(1), 31-41.

(Cullen, 1996) Cullen M R, "Epidemiologic Methods for the study of Occupational Asthma, Current Problems and Solutions", *Chest*, 1996, 109, 3, supplement 51S-54S.

(Dales *et al.*, 2003) Dales R E, Cakmak S, Judek S, Dann T, Coates F, Brook J R, "The role of fungal spores in thunderstorm asthma", *Chest* 2003, 123, 745-50.

(Dana, 1999) Dana P H, The Geographer's Craft Project, Department of Geography, The University of Colorado at Boulder, 1999 Peter H. Dana, http:// www.colorado.edu/geography/gcraft/notes/coordsys/coordsys.html

(Date, 1995) Date C J, *An Introduction to Database Systems*, Addison-Wesley 1995; ISBN 0201824582.

(Davison *et al.*, 1992) Davison B, Gumowitz J, Ingenito E, "Development and testing of a PC-

based system with menu-driven software for evaluating lung function in ICU patients", *Computers in Biology and Medicine*, 22, 6, November 1992, 423 – 436.

(Dawant *et al.*, 1993) Dawant B M, Uckun S, Manders E J, Lindstrom D P, "SIMON: A Distributed Computer Architecture for Intelligent Patient Monitoring", *Expert systems With Applications*, 6, 411 – 420, 1993.

(DEFRA, 2004) Nitrogen dioxide in the United Kingdom - report by the Air Quality Expert Group (AQEG) *PB9025A.*

(DeMarco, 1995) DeMarco T, *On Systems Architecture*, in Proceedings of the 1995 Monterey Workshop on Specification-Based Software Architectures, US Naval Postgraduate School, Monterey, California, September, 1995; also at http://www.systemsguild.com/GuildSite/TDM/Architecture.html

(Demartines and Hérault, 1997) Demartines P, Hérault J, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets", *IEEE Transactions on Neural Networks*, 1997.

(deSmith *et al.*, 2007) deSmith M J, Goodchild M F, Longley P A, "Geospatial Analysis – A Comprehensive Guide to Principles, Techniques and Software Tools: Second Edition", *Troubador Publishing Ltd.*, 2007, Issue version 2.14, ISBN: 13:978-1906221-522.

(DH, 2004) Department of Health, "Hospital Episode Statistics", 2004.

(DH, 2005) Department of Health, "Examples of Self Care Devices and Assistive Technologies to Support Self Care", 2005.

(Dokur *et al.*, 1997) Dokur Z, Olmez T, Yazgan E, Ersoy O K, "Detection of ecg waveforms by neural organizing networks", *Medical Engineering & Physics* 1997, 19(8), 738–41.

(Dominici *et al.*, 2003) Dominici F, McDermott A, Zeger S L, Samet J M, "National Maps of the Effects of Particulate Matter on Mortality: Exploring Geographical Variation", *Environmental Health Perspectives*, 111, 1, 39 – 43, (2003).

(Dore *et al.*, 2003) Dore C J, Goodwin J W L, Watterson J D, Murrells T P, Passant N R, Hobson M M, Haigh K E, Baggott S L, Pye S T, Coleman P J, King K R, "UK Emissions of Air Pollutants 1970 to 2001", National Environmental Technology Centre, AEA Technology , 2003. Found also at; http://www.airquality.co.uk/archive/reports/cat07/naei_report_1970-2001.pdf

(Dragonieri *et al.*, 2007) Dragonieri S, Schot R, Mertens B J A, Le Cessie S, Gauw S A, Spanevello A, Resta O, Willard N P, Vink T J, Rabe K F, Bel E H, Sterk P J, "An electronic nose in the discrimination of patients with asthma and controls (Health care education, delivery, and quality)", *Journal of Allergy and Clinical Immunology*, 120, 4, October 2007, 856 – 862, Copyright American Academy of Allergy, Asthma & Immunology, Published by Mosby Inc., doi:10.1016/j.jaci.2007.05.043.

(Dudeck, 1997) Dudeck J, "Communication Standards: Problems and Future Trends", p148 – 155, in *New Technologies in Hospital Information Systems*, Dudeck J *et al.* (Eds), IOS Press 1997, ISBN: 9051993633.

(Dysvik, 2002) Dysvik B, Jonassen S I; *J-Express v1.1, "Analysis of gene expression data"* part of a master study; University of Bergen, Department of Informatics, Norway.

(e-Government Unit, 2005) "e-Government Interoperability Framework: Version 6.1", *Cabinet Office, e-Government Unit*, Stockley House, 130 Wilton Road, London, SW1V 1LQ, Crown

copyright 2005, ISBN: 0 7115 0468 7.
Found at http://www.govtalk.gov.uk/documents/eGIF%20v6_1(1).pdf

(EAE, 2000) *Encyclopedia of the Atmospheric Environment;* Atmosphere, Climate & Environment Information Programme, Centre for Transport and the Environment, Manchester Metropolitan University; http://www.ace.mmu.ac.uk/aea

(EE&S, 2006) Environmental Equipment and Supply. 491L Blue Eagle Avenue, Harrisburg, PA 17112; http://www.envisupply.com

(Engin *et al.*, 2005) Engin M, Yamaner Y, Engin E, "A Biotelemetric System for Human ECG Measurements", *Measurement*, 38, 2005, 148-153.

(EU, 2008) "Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe", *Official Journal of the European Union*, Published June 2008.

(Ferraris, 2008) Ferraris Respiratory EuropePiKo-1 asthma monitor, UK Office: nSpire Health Ltd, Harforde Court, John Tate Road, Hertford, SG13 7NW, UK. Also found at: http://www.touchbriefings.com/pdf/886/lth041_t_ferraris.pdf

(Fielding, 2000) Fielding R T, *Architectural Styles and the Design of Network-based Software Architectures*. Doctoral dissertation, University of California, Irvine, Published 2000.

(Finkelstein *et al.*, 2000) Finkelstein J, Cabrera M R, Hripcsak G, "Internet-based home asthma telemonitoring: can patients handle the technology", *Chest*, 2000, 177(1), 148 – 55.

(Finkelstein *et al.*, 2001) Finkelstein J, O'Connor G, Friedmann R H, "Development and implementation of the home asthma telemonitoring (HAT) system to facilitate asthma self-care", *Medinfo*, 2001,10(Pt1), 810 – 4.

(Fischer *et al.*, 2004) Fischer P H, Brunekreef B, Lebret E, "Air pollution related deaths during the 2003 heat wave in the Netherlands", *Atmospheric Environment*, 38 (2004), 1083 – 1085.

(Forastiere *et al.*, 2002) Forastiere F, D'Ippolitr D, Pistelli R, "Airborne particles are associated with increased mortality and hospital admissions for heart and lung diseases", *Eur Respir Mon*, 2002, 21, 93 – 107.

(Foster, 2005) Foster I; *Globus Toolkit Version 4: Software for Service-Oriented Systems*; IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2005.

(Foster & Kesselman, 2004) Foster I, Kesselman C; *The Grid: Blueprint for a New Computing Infrastructure – Second Edition*; Morgan Kaufmann (Elsevier Inc.) 2004; ISBN 1558609334.

(Franklin, 2007) Franklin P J, "Indoor air quality and respiratory health of children (Mini-symposium: Pollutants and respiratory health in children)", *Paediatric Respiratory Reviews*, 8, 4, December 2007, 281 – 286, Copyright Elsevier Ltd., doi:10.1016/j.prrv.2007.08.007.

(Gailey and Lloyd, 1993) Gailey and Lloyd O L, "Spatial and temporal patterns of airborne metal pollution: the value of low technology sampling to an environmental epidemiology study", *The Science of The Total Environment*, 133, 3, June 1993, 201 – 219.

(Gamble *et al.*, 1987) Gamble J, Jones W, Minshall S, "Epidemiological-environmental study of diesel bus garage workers: Chronic effects of diesel exhaust on the respiratory system", *Environmental Research*, 44, 1, October 1987, 6 – 17.

(Garlan, 2003) Garlan D, "Formal Modeling and Analysis of Software Architecture: Components, Connectors, and Events", in Bernardo M, Inverardi P (Eds.), "Formal Methods for Software Architectures: SFM 2003", *Lecture notes in Computer Science* LNCS 2804, pp 1-24, 2003, Springer-verlag, Berlin, Heidelberg 2003.

(Garlan and Shaw, 1994) Garlan D, Shaw M, "An Introduction to Software Architecture", *Advnances in Software Engineering and Knowledge Engineering*, Volume 1, edited by Ambriola V and Tortora G, World Scientific Publishing Company, New Jersey, 1993.

(GINA, 2006) Global Initiative for Asthma, "Pocket Guide for Asthma Management and Prevention, A Pocket Guide for Physicians and Nurses, Revised 2006", Medical Communications Resources, Inc.,
http://www.ginasthma.com/

(Giner and Casan, 2004) Giner J, Casan P, "Spirometry at Home: Technology Within the Patient's Reach", *Archivos de Bronconeumologia*, 40, 1, January 2004, 39 – 40.

(Glykas & Chytas, 2004) Glykas M, Chytas P, "Technological innovations in asthma patient monitoring and care", *Expert Systems with Applications*, 27, 1, 2004, 121-131.

(Graham, 2004) Graham L M, "All I need is the air that I breath: Outdoor air quality and asthma (Respiratory review)", *Paediatric Respiratory Reviews: 5th International Congress on Pediatric Pulmonology*, 5, Supplement 1, January 2004, S59 – S 64, Published by Elsevier Ltd., doi:10.1016/S1526-0542(04)90012-7.

(Grubbs, 1969) Grubbs F, "Procedures for Detecting Outlying Observations in Samples", *Technometrics*, 11, 1, 1-21.

(Gupta *et al.*, 2008) Gupta R S, Zhang x, Sharp L K, Shannon J J, Weiss K B, "Geographic variability in childhood asthma prevalence in Chicago", *Journal of Allergy and Clinical Immunology*, 121, 3, March 2008, 639-645, Copyright 2008 American Academy of Allergy, Asthma & Immunology Published by Mosby, Inc. doi:10.1016/j.jaci.2007.11.036.

(Halmer *et al.*, 2002) Halmer M, Schmincke H-U, Graf H F, "The annual volcanic gas input into the atmosphere, in particular into the stratosphere: a global data set for the past 100 years", *J. Volcanol. Geotherm. Res.*, 88, 10, 969-10, 193, 1983.

(Haykin, 1999) Haykin S, *Neural Networks A Comprehensive Foundation - 2nd Edition*, Prentice Hall 1999; ISBN 0132733501.

(Hester, 1986) Hester R E, *Understanding Our Environment*, Royal Society of Chemistry 1986; ISBN 0851869076.

(Himberg, 1998), Himberg J, "Enhancing SOMbased data visualization by linking different data projections", Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 2200, FIN02015 HUT, Finland, April 1998.

(Hitchings *et al.*, 1993) D J, Dickinson S A, Miller M R, Fairfax A J) Hitchings D J, Dickinson S A, Miller M R, Fairfax A J, "Development of an accurate portable recording peak-flow meter for the diagnosis of asthma", *Journal of Biomedical Engineering*, 15, 3, May 1993, 188 – 192.

(HL7, 2007) Hinchley A, "Understanding Version 3 - A Primer on the HL7 Version 3 Interoperability Standard - Normative Edition", 4th edition (completely revised), 2007, ISBN: 3933819210.

(Ho *et al.*, 2007) Ho W-C, Hartley W R, Myers L, Lin M-H, Lin Y-S, Lien C-H, Lin R-S, "Air pollution, weather, and associated risk factors related to asthma prevalence and attack rate",

Environmental Research, 104 (2007), 402 – 409. doi.10.1016/j.envres.2007.01.007.

(Hodge & Seed, 1978) Hodge S E, Seed M L; *Statistics and Probability*; Blacki &Son Ltd / W & R Chambers Ltd 1972.

(Holgate S T, 2003) Holgate S T, Sandstrom T, Frew A J, Stenfors N, Nordenhall C, Salvi S, Blomberg A, Helleday R, Sodenberg M, Delvin R B, Wilson s J, "Health Effects of Acute Exposure to Air Pollution", Research Report, Health Effects Institute, 112, 2003.

(Hoskins *et al.*, 2000) Hoskins G, McCowan C, Neville RG *et al.*, "Risk factors and costs associated with an asthma attack", *Thorax*, 2000, 55:19-24.

(Howel *et al.*, 2001) Howel D, Darnell R, Pless-Mulloli T, "Children's Respiratory Health and Daily Particulate Levels in 10 Nonurban Communities", *Environmental Research* Section A, 87, 1 – 9, Academic Press (2001). doi:10.1006/enrs.2001.4280.

(HSC, 1998) Health Service Circular (HSC 1998/168), "Information for Health: An Information strategy for the Modern NHS", Department of Health. found also at; http://www.open.gov.uk/doh/coinh.htm

(IEEE, 2000) "IEEE Recommended practice for architectural description of software-intensive systems", E-ISBN 0-7381-2519-9, ISBN 0-7381-2518-0.

(IET, 2006) "E-Science Sensor Grid to Probe Pollution", *IEE Review and Engineering & Technology*, 136, 594, September 2006.

(*Information Centre The*, 2007), "Hospital Episode Statistics (The Information Centre for health and social care (The IC))".1 Trevelyan Square, Boar Lane, Leeds, LS1 6AE , http://www.ic.nhs.uk

(IS&S, 2006) Instrument Sales and Service Ltd. Unit 9 Cadzow Industrial Estate, Hamilton, Scotland, ML3 7QU, United Kingdom; http://www.isswww.co.uk

(ISO, 1998) "Information technology – Open Distributed Processing – Reference model: Overview", ISO/IEC 10746-1: 1998 (E), ISO/IEC Copyright Office, Case postale 56, CH-1211 Geneve 20, Switzerland. Also found at http://www.rm-odp.net

(ISO, 2007) *Systems and software engineering — Recommended practice for architectural description of software-intensive systems* , ISO/IEC 42010:2007, ISO.

(Jacobson *et al.*, 2007) Jacobson K W, Zakarian S E, Jensen J C, Moran J A, Glovsky M, "Acute Asthma with High Grass Pollen in Oregon", *Journal of Allergy and Clinical Immunology*, Volume 119, 1, Supplement 1 , January 2007, Page S187 (Program and Abstracts of Papers to be Presented During Scientific Sessions - 2007 AAAAI Annual Meeting, 2007 AAAAI Annual Meeting).

(Jaeger, 1998) Erich Jaeger GmbH, Leibnizstrasse 7, D-97204 Hoechberg, Germany, Article# 791 531, http://www.jaeger-toennies.com.

(Janetschek, 1998) Janetschek G, Bartsch G, Kavoussi L R, "Transcontinental interactive laparoscopic telesurgery between the United States and Europe", *Journal of Urology*, 160, 4, 1413.

(JCAAI, 1995) Journal of Allergy, and Clinical Immunology. Vol 96, No.5, part2. November 1995.

(Joshi *et al.*, 2005) Joshi A, Amelung P, Arora M, Finkelstein J, "Clinical impact of home automated telemanagement in asthma", *AMIA Annu Symp Proc*, 2005, 1000.

(Kangas, 1996) Kangas J, Kohonen T, "Developments and applications of the self-organizing map and related algorithms", *Mathematics and Computers in Simulation* 1996, 41, 3-12.

(Kehrl *et al.*, 1999) Kehrl H R, Peden D B, Ball B, Folinsbee L J, Horstman D, "Increased specific airway reactivity of persons with mild allergic asthma after 7.6 hours of exposure to 0.16 ppm ozone", *J Allergy Clin Immunol* 1999, 1198-204.

(Keles and Keles, 2006) Keles A, Keles A, "ESTDD:Expert system for thyroid diseases diagnosis", *Expert Systems with Applications*, 34 (2008), 242 – 246. doi:10.1016/j.eswa.2006.09.028.

(Kennedy *et al.*, 1997) Kennedy R L, Lee Y, Roy B V, Reed C D, Lippmann R P; *Solving Data Mining Problems through Pattern Recognition*; Prentice Hall 1997; ISBN 0130950831.

(Kern *et al.*, 1998) Kern H, Johnson R, Galup S D, Horgan D, with Cappel M; *Building the New Enterprise - People, Processes, and Technology*, Sun Microsystems 1998 (A Prentice Hall Title); ISBN 0130796719.

(Khoshafian, 1993) Khoshafian S, *Object-Oriented Databases*, John Wiley & Sons, ISBN: 0471570567.

(Kim J H *et al.*, 2005) Kim J H, Lim D H, Kim J K, Jeong S J, Son B K; "Effects of Particulate Matter (PM10) on The Pulmonary Function of Middle-School Children"; *J Korean Med Sci*, 2005, 20, 42-5.

(Kim J J *et al.*, 2004) Kim J J, Smorodinsky S, Lipsett M, Singer B C, Hodgson A T, Ostro B; "Traffic-related Air Pollution near Busy Roads: The East Bay Children's Respiratory Health Study"; *Am J Respir Crit Care Med*, 2004, 170, 520-526.

(Kohonen, 1982) Kohonen T, "Self-organised formation of topologically correct feature maps", *Biol. Cybern.*, 43, 59 – 69.

(Kohonen, 1987) Kohonen T, "An Introduction to Neural Computing", *Neural Networks* 1998, 1, 3-16.

(Kohonen, 1996) Kohonen T, Oja E, Simula O, Visa A, Kangas J, "Engineering applications of the selforganizing map", *Proceedings of the IEEE*, 84 (10), October 1996.

(Kohonen ,1998) Kohonen T, "The Self-Organising Map", *Neurocomputing* 1998, 1-6.

(Kohonen ,2006) Kohonen T, "Self-Organising Neural Projections", *Neural Networks*, 2006, 19, 723-733.

(Kolehmainen *et al.*, 2000) Kolehmainen M, Martikainen H, Hiltunen T, Ruuskanen J, "Forecasting air quality parameters using hybrid neural network modelling", *Environmental Monitoring and Assessment* 2000, 65, 277-286.

(Krill, 2006a) Krill P, "NetBeans IDE upgrade readied for SOA: Sun also ponders Java language, platform differentiations.", *JavaWorld*, 2006 InfoWorld Media Group, Inc; also found at, http://www.javaworld.com/javaworld/jw-10-2006/jw-1023-netbeans.html.

(Krill, 2006b) Krill P, "Telelogic eyes SOA with developer release: Tau 3.0 geared for building multiple types of enterprise applications", *JavaWorld*, 2006 InfoWorld Media Group, Inc; also found at, http://www.javaworld.com/javaworld/jw-11-2006/jw-1127-tau.html.

(Kruchten, 1995) Kruchten P B; The 4+ 1 View Model of Architecture, IEEE Software, November 1995, pp 42-50.

(Krumpe *et al.*, 1982) Krumpe P, Weigt G, Martinez N, Marcum R, Cummiskey J M,

"Computerized rapid analysis of pulmonary function test: Use of a least mean squares correlation for interpretation of data", *Computers in Biology and Medicine*, 12, 4, 1982, 295 – 307.

(Lane, 1996) Lane D J, *Asthma: the facts*, Third Edition, Oxford University Press 1996; ISBN 0192621513.

(LAQN, 2006) Air Quality Archive, AEA Technology, http://www.airquality.co.uk

(Lebowitz, 1996) Lebowitz M D, "Epidemiological studies of the respiratory effects of air pollution", *Eur Respir J*, 1996, 9, 1029-1054. doi:10.1183/09031936.96.09051029.

(Lee *et al.*, 2002) Lee J T, Kim H, Song H, Hong Y C, Cho Y S, Shin S Y, "Air pollution and asthma among children in Seoul, South Korea", *Epidemiology* 2002, 13, 481-4.

(Lee *et al.*, 2005) Lee H R, Yoo S K, Jung S M, Kwon N Y, Hong C S, "A Web-based mobile asthma management system", *J Telemed Telecare*, 2005, 11, Supplement 1, 56 – 9.

(Levy *et al.*, 2004) Levy E, Kalis M, Vo M, Lindisch D, Cleary K, "Feasibility of simultaneous respiratory function monitoring and determination of respiratory-related intrahepatic vessel excursion using the LifeShirt system" *International Congress Series* 1268 (2004) 764–769, Published by Elsevier B.V. doi:10.1016/j.ics.2004.03.338.

(Li & Gotze, 2001) Xiong Li, Hans-Jurgen Gotze, "Tutorial: Ellipsoid, geoid, gravity, geodesy, and geophysics", *Geophysics* 66, 6, 1660-1668.

(Linn *et al.*, 1999) Linn W S, Gong H, Clark K W, Anderson K R, "Day-to-Day Particulate Exposures and Health Changes in Los Angeles Area Residents with Severe Lung Disease", *J. Air & Waste Manage. Assoc.*, 1999, 49, 108 – 115.

(Luger & Stubblefield, 1998) Luger G F, Stubblefield W A; *Artificial Intelligence Structures and Strategies for Complex Problem Solving – third edition*; Addison Wesley 1998; ISBN 0805311963.

(Maclachlan *et al.*, 2007) Maclachlan J C, Jerrett M, Abernathy T, Sears M, Bunch M J, "Mapping health on the Internet: A new tool for environmental justice and public health research", Health & Place (Part Special Issue: Environmental Justice, Population Health, Critical Theory and GIS), 13, 1, March 2007, 72 – 86, Copyright Elsevier Ltd., doi:10.1016/j.healthplace.2005.09.012.

(Magnan, 2004) Magnan A, "Tools to assess (and achieve?) long-term asthma control", *Respiratory Medicine*, 98, Supplement 2, October 2004, S16 – S21.

(Maimonides, 1970) Maimonides M; from Rosner F, Munter S. (1970); "The Medical Aphorisms of Moses Maimonides"; Yeshiva University Press; from Donald J. Lane (1996); "Asthma: the facts - Third Edition"; Oxford University Press; ISBN 0192621513; p22.

(Maglaveras *et al.*, 2002) Maglaveras V, Koutkias I, Chouvarda D, Goulis G, Avramides A, Adamidis D, Louridas G, Balas E A, "Home care delivery through the mobile telecommunications platform: the Citizen Health System (CHS) perspective", *International Journal of Medical Informatics*, 68, 1-3, 2002, 99-111.

(Malveau & Mowbray, 2004) Malveau R, Mowbray T J; *Software Architect Bootcamp – second edition*; Prentice Hall Professional Technical Reference 2004; ISBN 0131412272.

(Manly, 2000) Manly B F J; *Multivariate Statistical Methods A Primer – Second Edition*; Chapman & Hall/CRC 2000; ISBN 0412603004.

(Mather *et al.*, 2004) Mather F J, White L E, Langlois C, Shorter C F, Swalm C M, Shaffer J G,

Hartley W R, "Statistical Methods for Linking Health, Exposure, and Hazards", *Environmental Health Perspectives*, 112, 14, 1440 – 1445 (October 2004).

(May *et al.*, 2000) May C, Mort M, Mair F, Ellis NT, Gask L, "Evaluation of new technologies in health-care systems: what's the context?", *Health Informatics Journal* 2000, 6, 67-70.

(McClure, 1997) McClure S, *Object Database vs. Object-Relational Databases*, IDC Bulletin #14821E August 1997, International Data Corporation (IDC), http://www.cai.com/products/jasmine/analyst/idc/14821Eat.htm

(McCoy *et al.*, 2006) McCoy K, Shade D M, Irvin C G, Mastronarde J G, Hanania N A, Castro M, Anthonisen N R, "Predicting episodes of poor asthma control in treated patients with asthma", *Journal of Allergy and Clinical Immunology*, 118, 6, December 2006, 1226 – 1233.

(Medicate, 2000) Crabbe H, Hamilton R, Machin N; "Integrating health and air quality information for use in a health telematics project"; *Air Pollution VIII 8: 753 – 763, 2000.* in, Advances in Air Pollution Series; WIT Press 2000.

(Mentor, 1999) Mentor Software Inc., 2221 East St. #203, Golden, CO 80401, Mentor Software Inc.. 1999, http://www.mentorsoftwareinc.com/CC/gistips/TIPS0699.HTM

(Michie *et al.*, 1994) Michie D, Spiegelhalter D J, Taylor C C, *Machine Learning, Neural and Statistical Classification*; MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, U.K.

(Molina, 1994) Molina R, Pérez de la Blanca N, Taylor C C, *Modern Statistical Techniques*, Chapter 4, in *Machine Learning, Neural and Statistical Classification*, Eds. Michie D, Spiegelhalter, Taylor CC, 1994, MRC Biostatistics€ Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, U.K.

(Molitor *et al.*, 2007) Molitor J, Jerrett M, Chang C-C, Molitor N-T, Gauderman J, Berhane K, McConnell R, Lurmann F, Wu J, Winer A, Thomas D, "Assessing Uncertainty in Spatial Exposure Models for Air Pollution Health Effects Assessment", *Environmental Health Perspectives*, 115, 8, 1147 – 1153, (2007).

(Moore and Peters, 2006) Moore W C, Peters S P, "Severe asthma: An overview", *Journal of Allergy and Clinical Immunology*, 117, 3, March 2006, 487 – 494.

(Moshammer & Neuberger, 2003) Moshammer H, Neuberger M, "The active surface of suspended particles as a predictor of lung function and pulmonary symptoms in Austrian school children", *Atmospheric Environment*, 37 (2003) 1737 – 1744.

(Muller & Lemke, 2000) Muller J-A, Lemke F; *Self-Organising Data Mining – Extracting Knowledge From Data*; Dresden, Berlin 2000; ISBN 3898118614.

(Muller *et al.*, 2001) Muller M L, Ganslandt T, Eich H P, Lang K, Ohmann C, Prokosch H-U, "Towards integration of clinical decision support in commercial hospital information systems using distributed, reusable software and knowledge components", *International Journal of Medical Informatics*, 64 (2001), 369 – 377.

(NAA, 2000) *National Asthma Audit 1999/2000 Summary*, The National Asthma Campaign, http://www.asthma.org.uk/infofa18.htm

(NAO, 2006) National Audit Office, "Department of Health, The National Programme for IT in the NHS", Report by the Comptroller and Auditor General, HC 1173 Session 2005-2006, 16 June 2006, © National Audit Office 2006.

(NEGTAP, 2001) National Expert Group on Transboundary Air Pollution, *Transboundary Air Pollution: Acidification, Eutrophication and Ground-Level Ozone in the UK*, DEFRA Contract EPG 1/3/153, NEGTAP 2001; ISBN 1870393619 (ch3), also http://www.nbu.ac.uk/netgap/

(NETCEN, 2000) http://www.airquality.co.uk

(NHLBI, 1997) U.S. Department of Health and Human Services - National Heart, Lung, and Blood Institute, The Lungs in Health and Disease, NIH Publication No. 97-3279, National Institutes of Health.

(NHLBI, 1997a) U.S. Department of Health and Human Services - National Heart, Lung, and Blood Institute, *Guidelines for the Diagnosis and Management of Asthma: Expert Panel Report 2*, NIH Publication No. 97-4051, National Institutes of Health.

(NHLBI, 1998) U.S. Department of Health and Human Services - National Heart, Lung, and Blood Institute, *Pocket Guide for Asthma Management and Prevention*, NIH Publication No. 96-3659B, National Institutes of Health.

(NHS, 2002) NHS Confederation, *Delivering 21st Century IT Support for the NHS - National Strategic Programme*; NHS Confederation Publication Sales (2002); http://www.nhsconfed.org

(NHS, 2003) NHS Confederation, *Briefing No. 88 (August 2003) - The national strategy for IT in the NHS*; NHS Confederation Publication Sales (2003); http://www.connectingforhealth.nhs.uk/publications/

(NIH, 1995) National Institutes for Health. Global Initiative for Asthma D Global Strategy for Asthma Management and Prevention NHBLI/WHO Workshop Report. Published: National Institues for Health, 1995; xiii. In National Asthma Audit 1999/2000 Summary. The National Asthma Campaign.

(Nitto and Rosenblum, 1999) Di Nitto E, Rosenblum D, "Exploiting ADLs to specify architectural styles induced by middleware infrastructures", In Proceedings of the 1999 International Conference on Software Engineering, Los Angeles, May 16-22, 1999, pp. 13-22.

(Nunn & Gregg, 1989) Nunn A J, Gregg I, *Br Med J*, 298, 1068-70, 1989.

(O'Connor *et al.*, 2008) O'Connor G T, Neas L, Vaughn B, Kattan M, Mitchell H, Crain E F, Evans R, Gruchalla R, Morgan W, Stout J, Adams G K, Lippmann M, "Acute respiratory health effects of air pollution on children with asthma in US inner cities", *Journal of Allergy and Clinical Immunology*, Volume 121, 5, May 2008, 1133 – 1139, Copyright American Academy of Allergy, Asthma & Immunology, Published by Mosby Inc., doi:10.1016/j.jaci.2008.02.020.

(OGC, 2005) Reichardt M et.al., *Sensor Web Enablement – An OGC White Paper*; OGC Document 05-063; Open Geospatial Consortium (OGC) Inc. July 2005; http://www.opengeospatial.org

(ONS, 1997) Office of National Statistics, Mortality Statistics: Cause 1997, Office of National Statistics, Series DH2 No.24, Annual Report of the Registrar General for Scotland, 1997. Registrar Annual Report 1997, General Registrar of Northern Ireland. In National Asthma Audit 1999/2000 Summary. The National Asthma Campaign.

(ONS, 2007) "Health Statistics - Quarterly 34: Summer 2007", *Office for National Statistics*, Crown copyright 2007, ISBN 978-0-230-52597-9, ISSN 1465-1645

(Ontrup and Ritter, 2001) Ontrup J, Ritter H, "Hyperbolic self-Organising Maps for Semantic Navigation", in proceedings of *NIPS 2001.*

(OS, 2001) "A guide to coordinate systems in Great Britain, An introduction to mapping coordinate systems and the use of GPS datasets with Ordance Survey mapping". *Ordanance Survey 2001*, http://www.ordsvy.gov.uk (D00659.doc v1.2 Aug 2001)

(OS, 2008) Ordnance Survey, *A Guide to Coordinate Systems in Great Britain: What is Position?*, Ordnance Survey © Crown copyright 2008; found at, http://www.ordnancesurvey.co.uk/oswebsite/gps/information/coordinatesystemsinfo/guidecontents/

(Oudinet *et al.*, 2006) Oudinet J-P, Méline J, Chełmicki W, Sanak M, Magdalena D-W, Besancenot J-P, Wicherek S, Julien-Laferrière B, Gilg J-P, Geroyannis H, Szczeklik A, Krzemień K, "Towards a multidisciplinary and integrated strategy in the assessment of adverse health effects related to air pollution: The case study of Cracow (Poland) and asthma", *Environmental Pollution*, 143, 2, September 2006, 278 – 284, Copyright Elsevier Ltd., doi:10.1016/j.envpol.2005.11.034.

(Oyanya *et al.*, 2005) Oyana T J, Boppidi D, Yan J, Lwebuga-Mukasa J S, "Exploration of Geographical Information Systems-based Medical Databases Using Self Organizing Maps (SOM): A Case Study of Adult Asthma", in *Proceedings of the 8th International Conference on GeoComputation*, University of Michigan, United States of America, 31 July - 3 August 2005.

(Parsaye *et al.*, 1993) Parsaye K, Chignell M, *Intelligent Database Tools & Applications*, John Wiley & Sons Inc., ISBN 0471570664.

(Pearson & Turton, 1993) Pearson J C G, Turton A, *Statistical Methods In Environmental Health*, Chapman and Hall 1993, ISBN 0412484501.

(Picton, 2000) Picton P, *Neural Networks – Second Edition*, Palgrave Publishers Ltd. 2000, ISBN: 033380287X.

(Pradhan *et al.*, 1996) Pradhan N, Sadasivan P K, Arunodaya G R, "Detection of seizure activity in EEG, by an artificial neural network: a preliminary study", *Computers and Biomedical Research* 1996, 29(4), 303–13.

(Primiano, 1998) Primiano F P, in Webster.John G, *Medical Instrumentation – Application and Design*, third edition, John Wiley & Sons, Inc., ISBN 0471153680; Chapter 9.

(Protti, 1995) Protti D J, *The synergism of health/medical informatics revisited, Methods of information in Medicine*, 34, 441-5, (1995) in Coiera.E, "Guide to medical informatics, the internet and telemedicine", Chapman and Hall Medical; ISBN 0412757109, pg. xxi.

(Pande *et al.*, 2003) Pande R U, Patel Y, Powers C J, D'ancona G, Karamanoukian H L, "The telecommunication revolution in the medical field: present applications and future perspective" *Current Surgery*, Volume 60, Issue 6, 2003, 636-640.

(Peled *et al.*, 2005) Peled R, Friger M, Bolotin A, Bibi H, Epstein L, Pipel D, Scharf S, "Fine particles and meteorological conditions are associated with lung function in children with asthma living near two power plants", *Public Health* (2005), 119, 418 – 425.

(Perry, 2002) Perry I, "Workflow by the Back Door? Using XML Systems in Health Service Processes, and Changing the System", *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, 2002 IEEE Computer Society, ISSN: 076951875.

(Pope and Kalkstein, 1996) Pope C A, Kalkstein L S, "Synoptic weather modeling and estimates of the exposure-response relationship between daily mortality and particulate air pollution", *Environ Health Perspect*, 1996, 104, 414 – 20.

(Pountain and Szyperski, 1994) Pountain D, Szyperski C, "Extensible software systems", *Byte,*

May 1994, pp. 57-62.

(Rabinovitch *et al.*, 2004) Rabinovitch N, Zhang L, Murphy J R, Vedal S, Dutton S J, Gelfand E W, "Effects of wintertime ambient air pollutants on asthma exacerbations in urban minority children with moderate to severe disease", *J Allergy Clin Immunol* 2004, 114,1131-7, 2004 American Academy of Allergy, Asthma and Immunology, doi:10.1016/j.jaci.2004.08.026

(Reddel *et al.*, 1999) Reddel H, Ware S, Marks G, Salome C, Jenkins C, Woolcock A, "Differences between asthma exacerbations and poor asthma control", *The Lancet*, 353, 9150, 30 January 1999, 364 – 369.

(Reed, 2006) Reed C E, "The natural history of asthma", *Journal of Allergy and Clinical Immunology*, 118, 3, September 2006, 543 – 548.

(Reznik *et al.*, 2005) Reznik M, Sharif I, Ozuah P, "Classifying asthma severity: prospective symptom diary or retrospective symptom recall?", *J Adolescent Health*, 36 (2005), 537 – 538, doi:10.1016/j.jadohealth.2004.05.009.

(Rialle *et al.*, 2003) Rialle V, Lamy J-B, Noury N, Bajolle L, "Telemonitoring of patients at home: a software agent approach", *Computer Methods and Programs in Biomedicine*, 72, 3, 2003, 257-268.

(Ripley, 1996) Ripley B D, *Pattern Recognition and Neural Networks*, Cambridge University Press 1996, ISBN: 0521460867.

(Ritter, 1992) Ritter H, Martinetz T, Schulten K; *Neural Computation and Self-Organising Maps*; Addison-Wesley Publishing 1992, ISBN: 0201554429.

(RKI, 2004) RKI Instruments Inc. 1855 Whipple Road, Hayward, CA 94544, USA; http://www.rkiinstruments.com

(Robb 2001) "Virtual Reality in Medicine and Biology", p9 - 13, in *Information Technologies in Medicine*, Vol 1, Akay M, Marsh A (Eds), John Wiley & Sons 2001, ISBN: 0471388637.

(Robins 1986) Robins J, "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect", *Mathematical Modelling*, 7, 9-12, 1986, 1393 – 1512.

(Rojas, 1996) Rojas R, *Neural Networks, A Systematic Introduction*, Springer-Verlag 1996, ISBN: 3540605053.

(Roos, 2003) Roos R M, *Java Data Objects*, Addison Wesley 2003, ISBN: 0321123808.

(Salgado-Ugarte *et al.*, 2000) Salgado-Ugarte I H, Shimizu M, Taniuchi T, Matsushita K, "Size Frequency Analysis by Averaged Shifted Histograms and Kernel Density Estimators", *Asian Fisheries Science*, 13, 2000, 1-12.

(Samet *et al.*, 1998) Samet J, Zeger S, Kelsall J, Xu J, Kalkstein L, "Does weather confound or modify the association of particulate air pollution with mortality? An analysis of the Philadelphia data, 1973 – 1980", *Environ Res*, 1998, 77, 9 – 19.

(Sammon, 1969) Sammon J W, "A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers*, 18, 1969, 401-409.

(Sanders & Aronsky, 2006) Sanders D L, Aronsky D, "Biomedical Informatics Applications for Asthma Care: A Systematic Review", *Journal of the American Medical Informatics Association*, 2006, 13, 418-427.

(Sarle, 1994) Sarle W S, "Neural Networks and Statistical Models", in *Proceedings of 19th Annual SAS User Group International Conference*, Dallas 1994, pp.1538-1549.

(Schalkoff, 1997) Schalkoff R J, *Artificial Neural Networks*, McGraw Hill 1997, ISBN: 007057118X.

(Schizas *et al.*, 1992) Schizas C N, Pattichis C S, Middleton L T, *A new approach to medical diagnosis*, In Ulgen Y (editor), Proceedings of the 1992 International Biomedical Engineering Days (Cat. No. 92TH04648), IEEE New York, 207–12.

(Schlesinger *et al.*, 1997) Schlesinger J M, Blumenfeld B, Broverman C, "Component Architecture in HIS: A Drug Order Entry Case Study", p54 – 60, in *New Technologies in Hospital Information Systems*, Vol 45, Dudeck J *et al.* (Eds), IOS Press 1997, ISBN: 9051993633

(Sekerel *et al.*, 2006) Sekerel B E, Gemicioglu B, Soriano J B, "Asthma insights and reality in Turkey (AIRET) study", *Respiratory Medicine*, 100, 10, October 2006, 1850 – 1854.

(Shanit *et al.*, 1996) Shanit D, Cheng A, Greenbaum R A, "Telecardiology: supporting the decision-making process in general practice", *Journal of Telemedicine and Telecare*, 2, 7-13, (1996). In Coiera E, "Guide to medical informatics, the internet and telemedicine", London, Chapman and Hall 1997, ISBN: 0412757109, pg 231.

(Sharma *et al.*, 2007) Sharma H P, Hansel N N, Matsui E, Diette G B, Eggleston P, Breysse P, "Indoor Environmental Influences on Children's Asthma", *Pediatric Clinics of North America*, 54, 1, February 2007, 103 – 120, Children's Health and the Environment: Part I, Copyright Elsevier Inc., doi:10.1016/j.pcl.2006.11.007.

(Shaw, 1995) Shaw M, "Comparing architectural design styles", *IEEE Software*, 12(6), Nov 1995, pp. 27-41.

(Shaw and Garlan, 1996) Shaw M, Garlan D, "Software Architecture: Perspectives on an emerging discipline", 1996, Prentice Hall.

(Sheldon, 2001) Sheldon T, *Encyclopedia of Networking and Telecommunications (Network Professional's Library)*, Osborne McGraw-Hill 2001, ISBN: 0072120053.

(Singh *et al.*, 2004) Singh I, Brydon S, Murray G, Ramachandran V, Violleau T, Stearns B, *Designing Web Services with the J2EETM 1.4 Platform JAX-RPC, SOAP, and XML Technologies*, Sun Microsystems Inc. Addison-Wesley 2004, ISBN: 0321205219.

(Srdanovic *et al.*, 2005) Srdanovic M, Schenk U, Schwieger M, Campagne F, "Critical evaluation of the JDO API for the persistence and portability requirements of complex biological databases", *BMC Bioinformatics*, 2005, 6, 5.

(Simula *et al.*, 1999) Simula O, Vestanto J, Alhoniemi E, Hollmén J, *Analysis and Modeling of Complex Systems Using the Self-Organizing Map* in *Neuro-Fuzzy Techniques for Intelligent Information Systems*, Springer Verlag 1999, 3-22, ISBN: 3790811874.

(Sinnott, 1984) Sinnott R W, "Virtues of the Haversine", *Sky and Telescope*, 68, 2, 159.

(Snedecor & Cochran, 1989) Snedecor G W, Cochran W G, *Statistical Methods, 8th Edition*, Iowa

State University Press.

(Soni *et al.*, 1995) Soni D, Nord R, Hofmeister C, "Software Architecture in Industrial Applications", in *Proceedings of the 17th International Conference on Software Engineering*, pp.196-207, ACM Press, April 1995.

(Souquieres and Heisel, 2004) Heisel M, Souquieres J, "Adding Features to Component-Based Systems", p137-153 in *Lecture Notes in Computing Science: Objects, Agents, and Features*: International Seminar, Dagstuhl Castle, Germany, 2003, Eds Ehrich H-D, Meyer J-J, Ryan M D, Springer 2004, LNCS 2975.

(Spencer *et al.*, 1997) Spencer R G, Lessard C S, Davila F, Etter B, "Selforganising discovery, recognition and prediction of haemodynamic patterns in the intensive care unit", *Medical & Biological Engineering & Computing*, 35(2), 117–23.

(Spencer *et al.*, 2004) Spencer Jr B, Finholt T A, Foster I, Kesselman C, Beldica C, Futrelle J, Gullapalli S, Hubbard P, Liming L, Marcusiu D, Pearlman L, Severance C, Yang G, "NEESGrid: A Distributed Collaboratory for Advanced Earthquake Engineering Experiment and Simulation", *13th World Conference on Earthquake Engineering* (2004), Paper 1674. Found at; http://www.globus.org/research/papers/13worldconferenceonEarthquakeEngineering-rad8A451.pdf

(Stedman, 2001) Stedman J R, King K, Holland M., Department for Environment, Food and Rural Affairs, The National Assembly for Wales, The Scottish Executive and the Department of the Environment in Northern Ireland, *Quantification of the health effects of air pollution in the UK for PM_{10} objective analysis*, AEAT/ENV/R/0728 Issue1, AEA Technology Environment (NETCEN).

(Stephens, 1974) Stephens M A, "EDF Statistics for Goodness of Fit and Some Comparisons", *Journal of the American Statistical Association*, 69, 730-737.

(Stocks, 1996) Stocks J; from Susan M, Hinchliff, Susan E, Montague, Jackson R, "Physiology for Nursing Practice - Second Edition", Bailliere Tindall 1996, ISBN: 0-7020-1638-1, diagrams from p533, p543.

(Stukus *et al.*, 2007) Stukus D R, Nock N, Lang D M, "Patterns of Asthma Mortality in Ohio: 1999-2003 Compared with 1990-98", *Journal of Allergy and Clinical Immunology*, 119, 1, Supplement 1, January 2007, S75, Program and Abstracts of Papers to be Presented During Scientific Sessions – 2007 AAAAI Annual Meeting, 2007 AAAAI Annual Meeting , Copyright American Academy of Allergy, Asthma & Immunology Published by Mosby Inc., doi:10.1016/j.jaci.2006.11.319.

(Sulkava, 2008) Sulkava M, "Learning from environmental data: Methods for analysis of forest nutrition time series", Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D24. Found at, http://lib.tkk.fi/Diss/2008/isbn9789512291540/isbn9789512291540.pdf

(Sun, 2003a) Sun Microsystems, Jini™ Technology Core Platform Specification Version 2.0, Sun Microsystems, Inc. 2003, http://wwws.sun.com/software/jini/specs/core2_0.pdf (p47).

(Sun, 2003b) Sun™ OpenR$_X$ Framework, A Comprehensive Platform for Open Health Industry Solutions; A Technical White Paper, September 2003, Found at: http://www.sun.com/products-n-solutions/healthcare/collateral.html/

(Taggart *et al.*, 1996) Taggart S C O, Custovic A, Francis H C, Faragher E B, Yates C J, Higgins B G, Woodcock A, "Asthmatic bronchial hyperresponsiveness varies with ambient levels of summertime air pollution", *Eur Respir J*, 1996, 9, 1146 – 1154.

(Taibi *et al.*, 1992) Taibi G, Vassallo G, Sorbello F, "Self organizing maps for medical diagnosis", In Caianiello E R (editor), *Neural Nets* Wirn Vietri 92, *Proceedings of the 4th Italian Workshop on Neural Nets*, Singapore. World Scientific.

(Tamburlini *et al.*, 2002) Tamburlini G *et al.*, eds, "Children's health and environment: a review of evidence: a joint report from the European Environment Agency and the WHO Regional Office for Europe", *Copenhagen, European Environment Agency*, 2002:48–49 (*Environmental issue report*, No. 29).

(Tamminen *et al.*, 2000) Tamminen S, Pirttikangas S, Roning J, "Self-organizing maps in adaptive health monitoring", *Neural Networks*, 2000, in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), Volume 4, 2000 Page(s):259 – 264, doi: 10.1109/IJCNN.2000.860782.

(Tang *et al.*, 2007) Tang C-S , Chang L-T, Lee H-C, Chan C-C, "Effects of personal particulate matter on peak expiratory flow rate of asthmatic children", *Science of The Total Environment*, 382, 1, August 2007, 43 – 51.

(TAQA, 2008) The Air Quality Archive, *www.airquality.co.uk*, website prepared and hosted by AEA Energy & Environment, on behalf of the UK Department for Environment, Food & Rural Affairs and the Devolved Administrations, Crown copyright 2002.

(Thacker *et al.*, 1996) Thacker S B, Stroup D F, Parrish R G, Anderson H A, "Surveillance in environmental public health: issues, systems and sources", *Am J Public Health*, 86, 633 – 638, 1996.

(TOGAF, 2007) "The Open Group Architecture Framework (TOGAF) 8.1.1 Enterprise Edition", *The Open Group* (2007), ISBN 1-9316-2462-3, Also found at http://www.opengroup.org/architecture/togaf/

(Toledano, 2004) Toledano M D D, "How to develop a software architecture: pattern languages", found at http://www.moisesdaniel.com/wri/htdsalpEn.pdf

(Topac, 2003) Topac Inc. 231 CJC Highway, Suite 103 Cohasset MA 02025 USA; www.topac.com

(Tuecke *et al.*, 2003) Tuecke S, Czajkowski K, Foster I, Frey J, Graham S, Kesselman C, Maguire T, Sandholm T, Snelling D, Vanderbilt P; *Open Grid Services Infrastructure – Version 1.0.*; Global Grid Forum, Draft draft-ggf-ogsi-gridservice-29, 2003; www.ggf.org/documents/Drafts.

(Tuomisto *et al.*, 2006) Tuomisto M T, Terho T, Korhonen I, Lappalainen R, Tuomisto T, Laippala P, Turjanmaa V, "Diurnal and weekly rhythms of health-related variables in home recordings for two months", *Physiology & Behavior*, 87,2006, 650 – 658.

(Turner *et al.*, 1999) Turner C R, Fuggetta A, Lavazza L, Wolf A, "A conceptual basis for feature engineering", *Journal of systems and Software*, 49(1), 3-15, 1999.

(Ultsch and Siemon, 1990) Ultsch A, Siemon H P, "Kohonen's self organizing feature maps for exploratory data analysis", In Proc. *INNC'90, Int. Neural Network Conf.*, pages 305–308, Dordrecht, Netherlands, 1990. Kluwer.

(Unser & Aldroubi, 1992) Unser M, Aldroubi A; *Polynomial spline signal processing algorithms.* Proc. ICASSP 92, III, 177-80, Academic Press 1992.

(van den Hazel P J, 2007) van den Hazel P J, "International strategies in children's environmental health", *Int. J. Hyg. Environ. Health* (2007), doi:10.1016/j.ijheh.2007.03.002.

(Van Hulle, 2000) Van Hulle M M, "Faithful Representations and Topographic Maps From Distortion- to Information-Based Self-Organisation", John Wiley & Sons 2000, ISBN: 0471345075.

(Vázquez *et al.*, 2006) Vázquez J C D, Martínez A C, Gómez A, Varela B A, "Intelligent Agents Technology Applied to Tasks Scheduling and Communications Management in a Critical care telemonitoring System", *Computer in Biology and Medicine*, 2006, Volume 37, Issue 6, Pages 760 – 773.

(Velsor-Friedrich *et al.*, 2005) Velsor-Friedrich B, Pigott T, Srof B, "A Practitioner-Based Asthma Intervention Program With African American Inner-City School Children", *Journal of Pediatric Health Care*, May/June 2005, doi:10.1016/j.pedhc.2004.12.002.

(Versant, 2008) "FastObjects J1" – closest product now "Versant Object Database 7"; http://www.versant.com/

(Vesanto, 1999) Vesanto J, "SOM-Based Data Visualization Methods", Laboratory of Computer and Information Science, Helsinki University of Technology, P.O.Box 5400, FIN-02015 HUT, Finland, Found at: http://www.cis.hut.fi/projects/ide/publications.

(Vesanto, 2000) Vesanto J, Alhoniemi E, "Clustering of the Self-Organizing Map", *IEEE Transactions on Neural Networks* 2000, 11, 3.

(Watson *et al.*, 2007) Watson L, Turk F, JamesP, Holgate S T, "Factors associated with mortality after an asthma admission: A national United Kingdom database analysis", *Respiratory Medicine*, 2007, 101, 1659 – 1664.

(W3C, 2004) Booth D, Haas H, McCabe F, Newcomer E, Champion M, Ferris C, Orchard D, "Web Services Architecture: W3C Working Note 11 February 2004", World Wide Web Consortium 2004, http://www.w3.org/TR/ws-arch

(W3C, 2006) Calladine J, Cowe G, Downey P, Lafon Y, "Basic XML Schema Patterns for Databinding Version 1.0: W3C Working Draft 22 November 2006", *World Wide Web Consortium 2006*, http://www.w3.org/TR/xmlschema-patterns

(Webster, 1998) Webster J G, *Medical Instrumentation – Application and Design*, third edition, John Wiley & Sons, ISBN: 0471153680.

(Watkins *et al.*, 2007) Watkins T, Dimmick F, Holland D, Gilliland A, Boothe V, Paulu C, Smith A, "Chapter 7.5 Air quality characterization for environmental public health tracking (Air Pollution Modeling and Its Application XVIII)", *Developments in Environmental Sciences*, 6, 2007, 717 – 727, Copyright 2007 Elsevier Ltd., doi:10.1016/S1474-8177(07)06075-5.

(WHO, 2007) "Asthma", Fact sheet N°307, August 2006, *World Health Organization* 2007.

(Wilhelm *et al.*, 2008) Wilhelm M, Qian L, Ritz B, "Outdoor air pollution, family and neighborhood environment, and asthma in LA FANS children", *Health and* Place, article in press, Copyright 2008 Elsevier Ltd., doi:10.1016/j.healthplace.2008.02.002.

(Witten & Frank, 2000) Witten I H, Frank E, *Data Mining – practical machine learning tools and techniques with Java implementations,* Morgan Kaufmann Publishers 2000, ISBN: 1558605525.

(Wooldridge, 2002) Wooldridge M, *An Introduction to Multi-Agent Systems*, John Wiley 2002.

(Wooten, 2001) Wootten R, "Recent Advances - Telemedicine", *BMJ* 2001, 323, 557-560.

(WS-I) Web Services Interoperability Organisation; http://www.ws-i.org

(Wyatt, 1987) Wyatt J, *The evaluation of clinical decision support systems: a discussion of the methodology used in the ACORN project*, Proceedings of the AIME 87 (Lecture notes in medical informatics) 1987, 33, 15-24.

(Zimmermann *et al.*, 2007) Zimmermann O, Grundley J, Tai S, Leymann F, "Architectural Decisions and Patterns for Transactional Workflows in SOA", *in* Kramer B, Lin K-J, Narasimhan P (Eds.), "Service-Oriented Computing - ISOC 2007", *Lecture notes in Computing Science, LNCS* 4749 pp81-93, Springer-Verlag, Berlin, Heidelberg 2007.

(Zurada, 1992) Zurada Jacek M, *Introduction to artificial neural systems*, West Publishing Company 1992, ISBN: 0314933913.

Papers Written by the Author

1. Barber A, Bayford R H, Hamilton R ; "An Informatics based approach to Respiratory Healthcare": in the proceedings of *Current perspectives in healthcare computing HC2002: Healthcare Computing 2002;* Weybridge, BCS HIC, 2002.

2. Barber A, Bayford R H, Hamilton R ; "The use of self organising maps in the field of enviromedics" ; presented as a poster and paper at *Artificial Intelligence 2003.*

3. Crabbe H, Barber A, Bayford R H, Hamilton R, Jarret D, Machin N ; "The use of a European telemedicine system to examine the effects of pollutants and allergens on asthmatic respiratory health." ; *Sci Total Environ. 2004*, 335:417-26.

Appendices

Appendix A

Medicate Project

A.1. Functional Specification

The Medicate project required an Internet based tool for viewing and analysing lung function and environmental data. The use of the Internet to transfer data had been shown to be of value in health informatics and therefore formed the basis of the approach to the work.

Software developed during the project was the first prototype version of the Environmental Monitoring System ($EMSv_1$). The $EMSv_1$ aimed to provide an environment for the analysis and management of lung function and air quality (enviromedic) data. Air quality data is obtainable from archived web pages (from the National Environmental Technology Centre, NETCEN). Lung function archives were originally stored in a legacy database (using Dbase), this was converted to a Microsoft Access database using Standard Query Language (SQL) during the Medicate (2000) project. Lung function data was then available from this database on a continual basis throughout the Medicate clinical trial.

The system architecture was composed of several distinct elements, an object database, Java classes (programs) run via a web server, and a user interface accessed via a standard web browser.

A.2. Prototype Architecture used in the Medicate Project

The overall architecture of the system, shown in Figure 84 consisted of two main constituents:

1. A *Client*, which was the user interface to the system and the main method of communication to the data.
2. A *Server*, where all the work in the manipulation of the data was achieved.

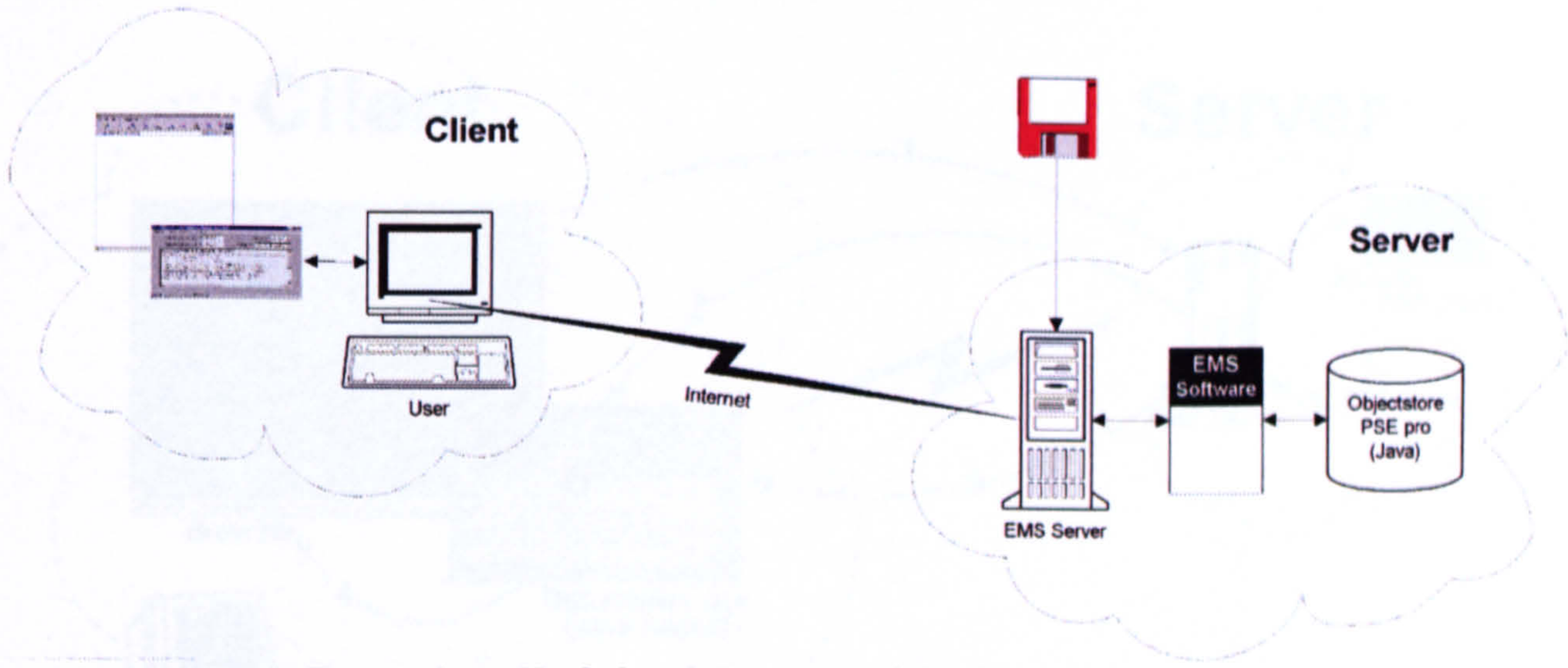


Figure 84 High-level diagram of the EMS architecture

A.3. Communication

Communication between the client and the server was achieved through the Transmission Control Protocol (TCP) and the implementation of sockets that used server ports and an Internet Protocol (IP) address. This method was chosen to comply with the web browser technology available at the time of the research. The client used Java (applet) technology to facilitate the transfer of instruction from the user to the server. Applets are small pieces of software written in accordance with the Application Programming Interface (API) specification developed by Sun Microsystems and are capable of running inside any Java enabled web browser. Applets were useful as they were capable of controlling their environment as well as existing within it, for example applet software was used to instruct the browser to reference a particular web page.

Figure 85 shows the communication process between the client and the server. The user first opened a web browser and connected to the server using an appropriate URL (Uniform Resource Locator) for example, http://ems.mdx.ac.uk. The Java applet (classes) were then transferred to the clients browser for execution (*Stage 2* in Figure 85). At this point the $EMSv_1$ Primary Graphical User Interface (GUI) was displayed to the user and could be used like a normal program. Once the user instructed the GUI to make a query on the database, an instruction was sent to the server to execute the query (*Stage 3*). The query was then made on the database and the results returned to the GUI over the Internet. At the same time a web page showing the results was created automatically and stored on the web server, the browser on the client then *pointed* to this page automatically (*Stage 4*).

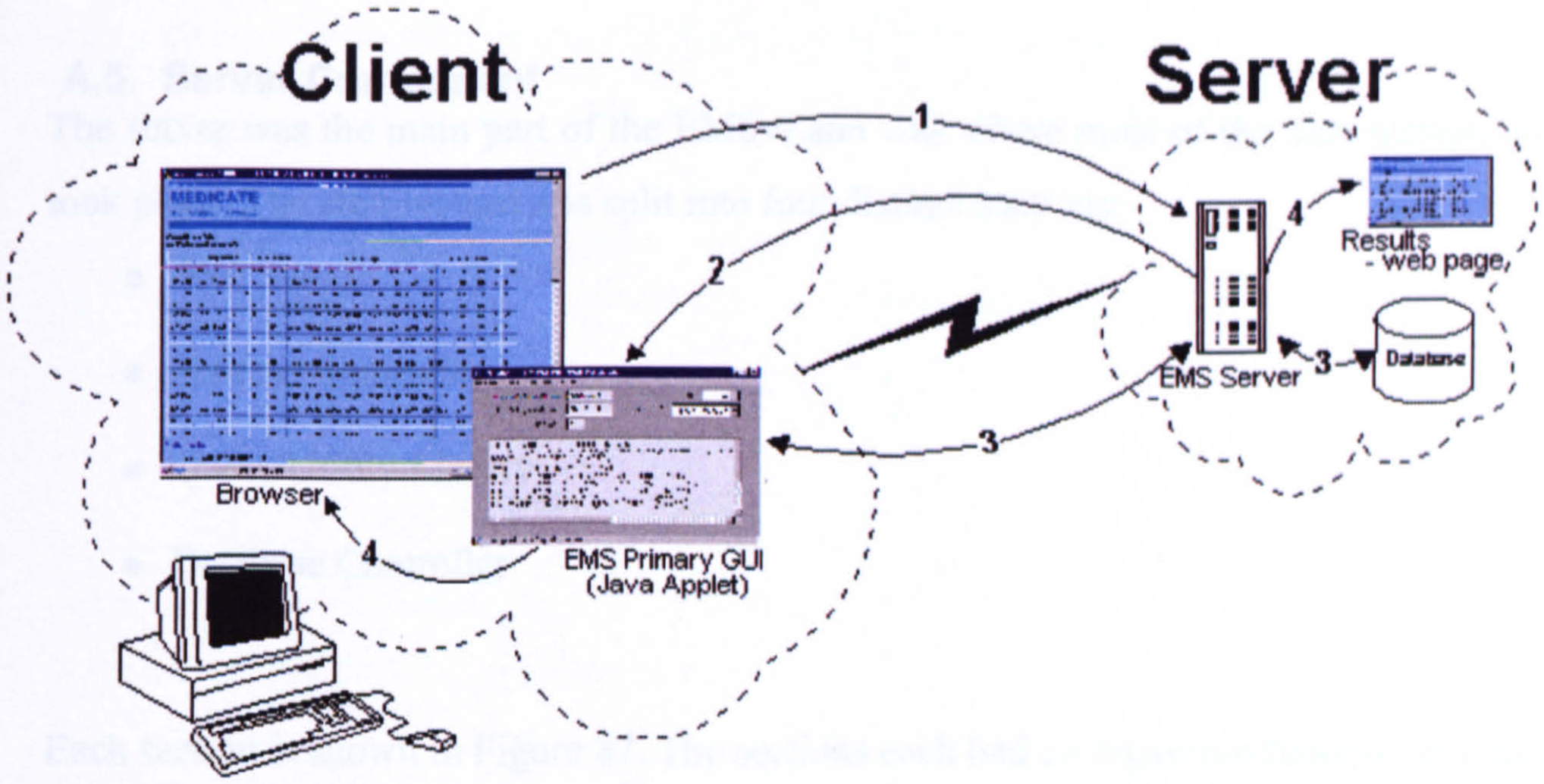


Figure 85 Communication Process

A.4. Client Component

Figure 86 below shows the primary GUI for the $EMSv_1$ system. The GUI consisted of a menu bar like many found on standard desktop computers. The menu bar gave a number of options to the user in a simple format. The interface also provided text fields and drop down menus for entering query information. Parameters available for the user to modify were: start and end dates for the query, patient identifier (presented to the user using a drop down menu which showed the patients available for a particular hospital), location of air quality readings (also presented to the user in a drop down menu), and time lag (measured in days) between lung function and air quality data sets.

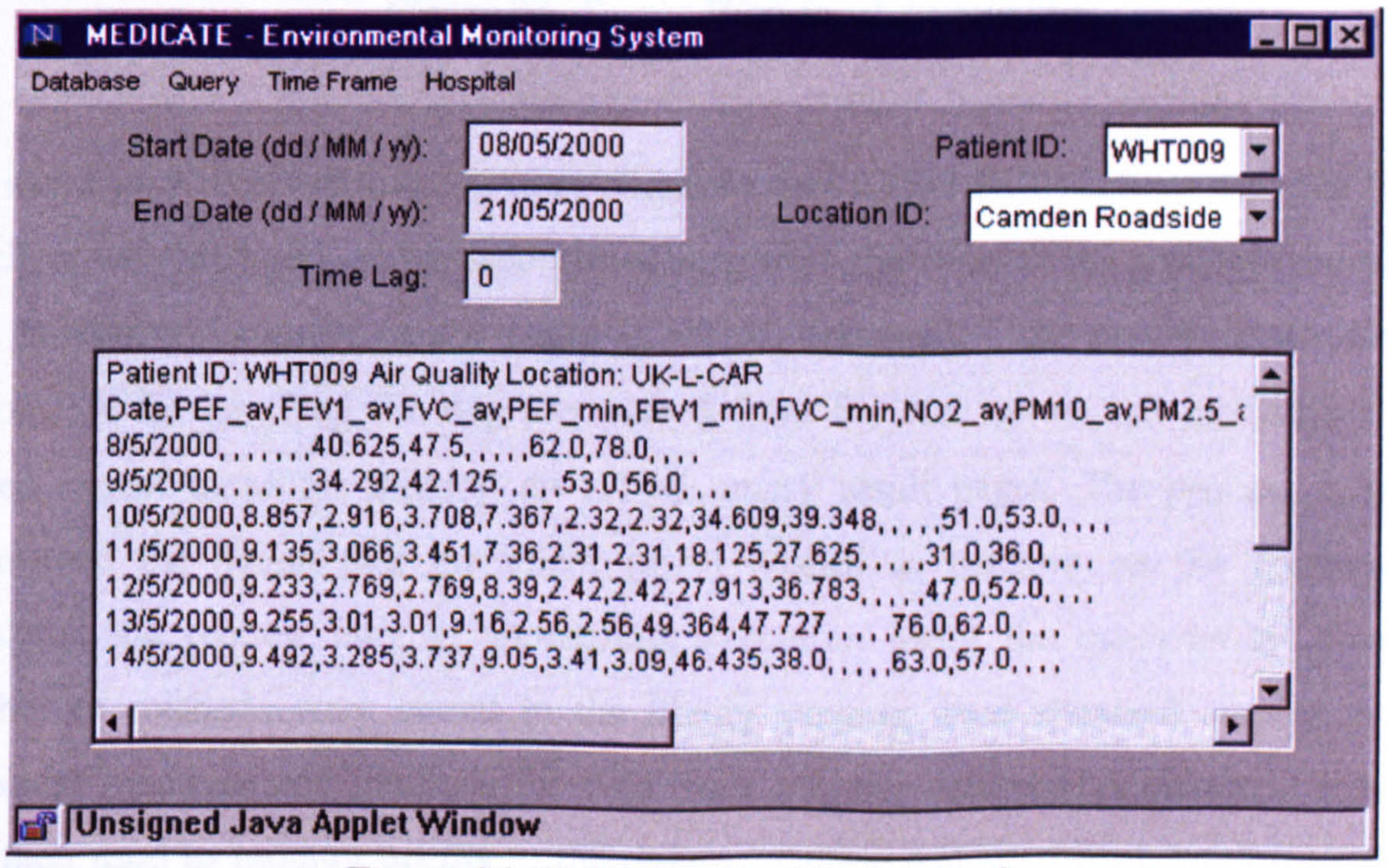


Figure 86 EMS Graphical User Interface

A.5. Server Component

The server was the main part of the $EMSv_1$ and was where most of the data manipulation took place. The architecture was split into four distinct sections:

- Web Server
- Process Controller
- Query/Utility Library
- Database Controller

Each section is shown in Figure 87. The sections each had an important function within the architecture.

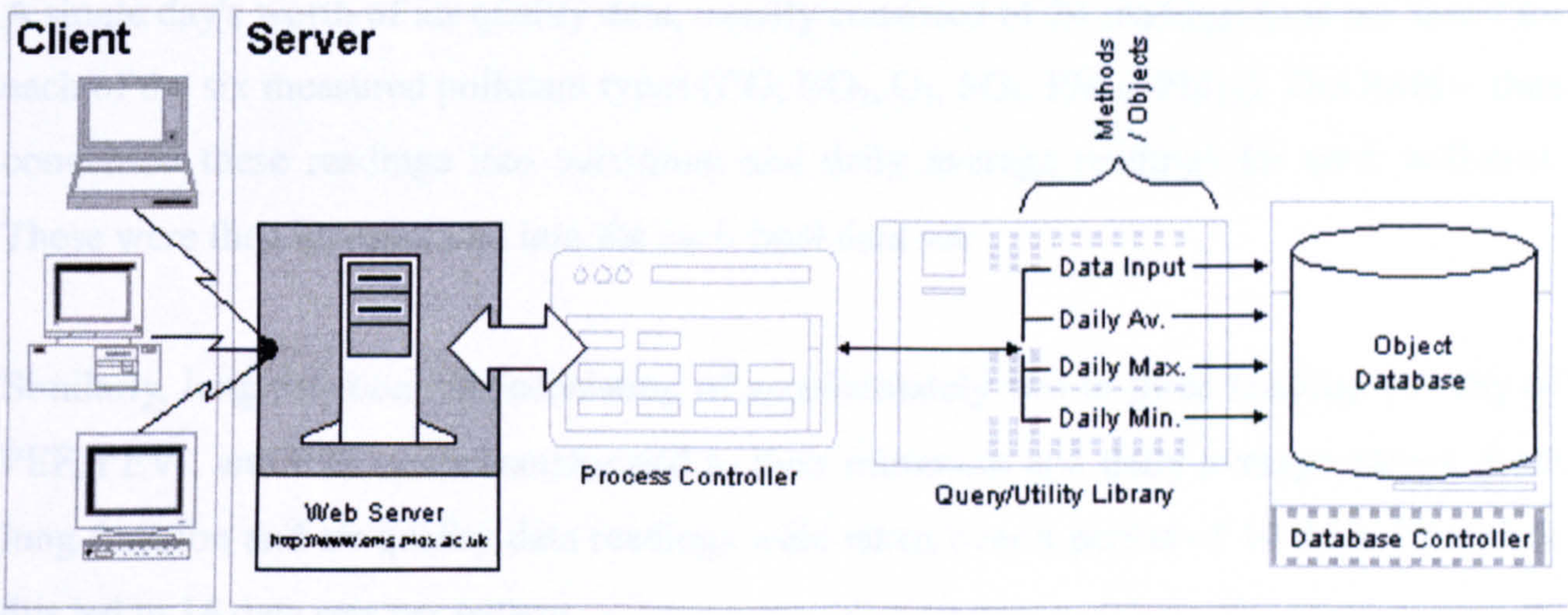


Figure 87 Server High-level Architecture

The *web server* received requests from clients to start-up the $EMSv_1$ client software (within the client web browser), and passed instructions from the client to the *process controller* as well as storing the query results pages as HTML documents. The *process controller* was responsible for handling the requests passed from the web server, all instructions were passed except those for looking up HTML query result pages. The *process controller* determined the action that the client (user) wished to perform on the database and translated the request into an appropriate execution using the q*uery/utility library.* A number of methods were stored in the library ranging from standard queries to other statistical functions and methods for data entry into the database. A *database controller* was then used to monitor access to the object database and send back any query results, complications or errors to the *process controller* for further manipulation (results page

creation etc.) and for notification to the client that the query has been completed (via the *web server*). Once this cycle has been completed and the client notified, the client automatically looked for the results page stored on the *web server.*

A.6. Testing the Medicate Prototype

To test the system a simple demonstrator was created. The demonstrator used a total of approximately 370 different data sets. Incorporating data from two major locations: London, UK and Barcelona, Spain.

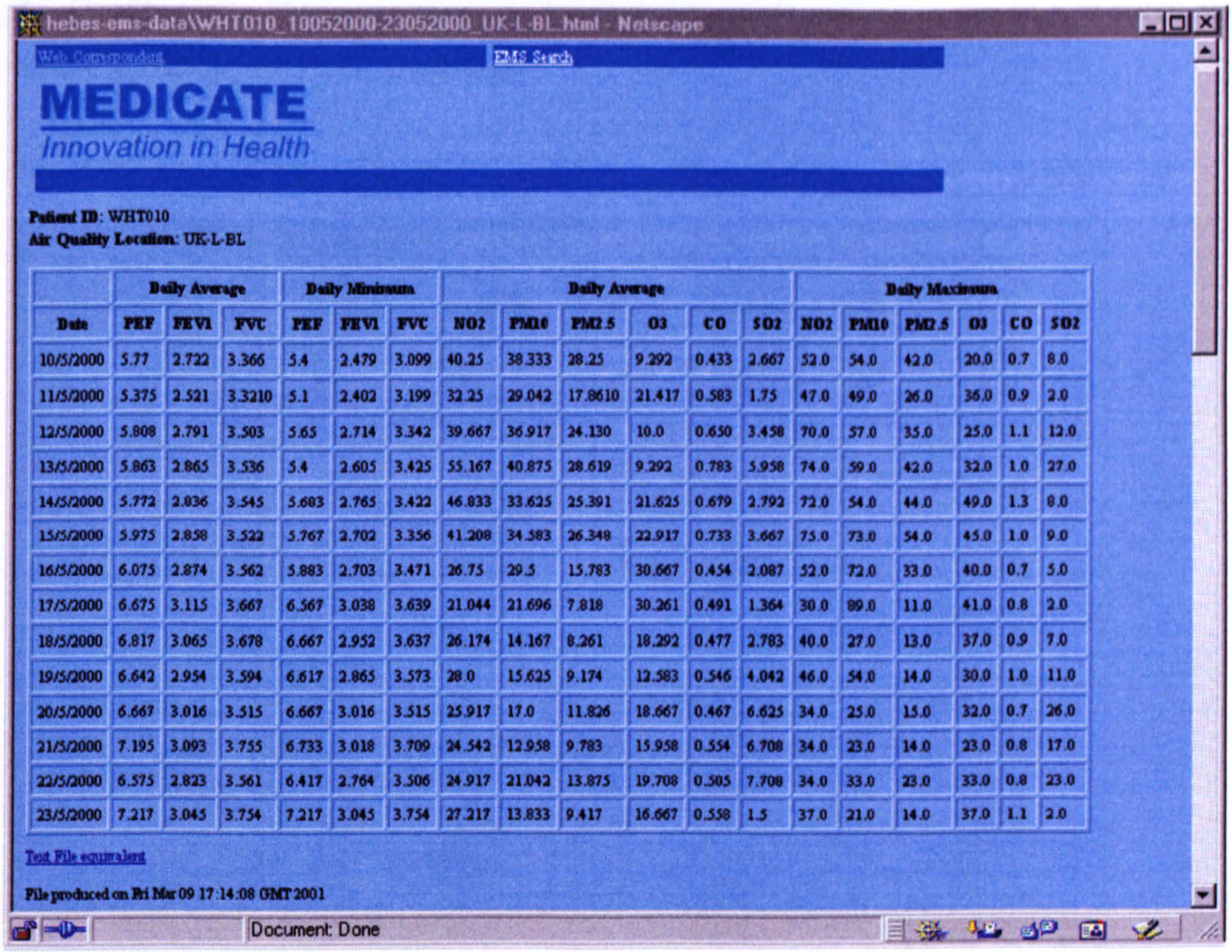
A data set was defined as, *one day's record for all data types*. Therefore if six pollutants were measured and incorporated into a dataset, the data set would increase by the six parameters (or pollutants) in size. The number of data sets would not increase. A new data set was created for each separate day.

A single day's worth of air quality data, usually consisted of 24 readings (one per hour) for each of the six measured pollutant types (CO, NO_2, O_3, SO_2, PM_{10}, $PM_{2.5}$). The $EMSv_1$ then condensed these readings into maximum and daily average readings for each pollutant. These were then incorporated into the each final data set.

Similarly, lung function data consisting of approximately two to three readings per day of PEF, FEV_1, and FVC were transformed to their minimum and daily average values. Both lung function and air quality data readings were taken over a period of 14 days. Therefore this led to 14 data sets per patient.

For UK patients a total of four air quality monitoring stations were chosen as local sites to the patients' daily activities: Bloomsbury, Haringey Park, Haringey Roadside, and Camden Roadside (TAQA, 2007). The demonstrator was tested with each of these four air quality monitoring sites, depending on the patient's locations in relation to them. A total of 16 patients returned good data readings using the electronic spirometry device, therefore a figure of approximately 224 (14 days x 16 patients) different data sets were created and analysed, each containing up to 18 (9 data types x 2 statistics) parameters. Data sets from Barcelona, Spain totalled approximately 150.

The results of one computation are shown below, where one row represents one data set.



	Daily Average			Daily Minimum			Daily Average						Daily Maximum					
Date	PEF	FEV1	FVC	PEF	FEV1	FVC	NO2	PM10	PM2.5	O3	CO	SO2	NO2	PM10	PM2.5	O3	CO	SO2
10/5/2000	5.77	2.722	3.366	5.4	2.479	3.099	40.25	38.333	28.25	9.292	0.433	2.667	52.0	54.0	42.0	20.0	0.7	8.0
11/5/2000	5.375	2.521	3.3210	5.1	2.402	3.199	32.25	29.042	17.8610	21.417	0.583	1.75	47.0	49.0	26.0	36.0	0.9	2.0
12/5/2000	5.808	2.791	3.503	5.65	2.714	3.342	39.667	36.917	24.130	10.0	0.650	3.458	70.0	57.0	35.0	25.0	1.1	12.0
13/5/2000	5.863	2.865	3.536	5.4	2.605	3.425	55.167	40.875	28.619	9.292	0.783	5.958	74.0	59.0	42.0	32.0	1.0	27.0
14/5/2000	5.772	2.836	3.545	5.683	2.765	3.422	46.833	33.625	25.391	21.625	0.679	2.792	72.0	54.0	44.0	49.0	1.3	8.0
15/5/2000	5.975	2.858	3.522	5.767	2.702	3.356	41.208	34.583	26.348	22.917	0.733	3.667	75.0	73.0	54.0	45.0	1.0	9.0
16/5/2000	6.075	2.874	3.562	5.883	2.703	3.471	26.75	29.5	15.783	30.667	0.454	2.087	52.0	72.0	33.0	40.0	0.7	5.0
17/5/2000	6.675	3.115	3.667	6.567	3.038	3.639	21.044	21.696	7.818	30.261	0.491	1.364	30.0	89.0	11.0	41.0	0.8	2.0
18/5/2000	6.817	3.065	3.678	6.667	2.952	3.637	26.174	14.167	8.261	18.292	0.477	2.783	40.0	27.0	13.0	37.0	0.9	7.0
19/5/2000	6.642	2.954	3.594	6.617	2.865	3.573	28.0	15.625	9.174	12.583	0.546	4.042	46.0	54.0	14.0	30.0	1.0	11.0
20/5/2000	6.667	3.016	3.515	6.667	3.016	3.515	25.917	17.0	11.826	18.667	0.467	6.625	34.0	25.0	15.0	32.0	0.7	26.0
21/5/2000	7.195	3.093	3.755	6.733	3.018	3.709	24.542	12.958	9.783	15.958	0.554	6.708	34.0	23.0	14.0	23.0	0.8	17.0
22/5/2000	6.575	2.823	3.561	6.417	2.764	3.506	24.917	21.042	13.875	19.708	0.505	7.708	34.0	33.0	23.0	33.0	0.8	23.0
23/5/2000	7.217	3.045	3.754	7.217	3.045	3.754	27.217	13.833	9.417	16.667	0.558	1.5	37.0	21.0	14.0	37.0	1.1	2.0

Figure 88 Results - Matrix

Figure 88 shows the top third of the results page produced from a query on the EMS database. The section shown in the figure shows readings of daily lung function (average and minimum) and daily air quality (average and maximum) in chronological order. Daily average results are obtained from the complete day's worth of figures for a particular type of data (*e.g.* PEF or O_3) while basic statistics are used to obtain the mean value. Daily maximum and minimum results for air quality and lung function respectively are also obtained from the data belonging to the particular day's readings for each data type.

Figure 89 shows standard statistics in four tables, one table each for:

- Lung function daily minimum.
- Lung function daily average.
- Air quality daily maximum.
- Air quality daily average.

where the minimum, maximum, range, mean and standard deviation along with the number of readings in the result set are displayed. The results are all derived from the individual result sets, so for example, the *maximum* reading of the *Lung Function Daily Minimum* results would be the maximum of the daily minimum results, and the *maximum* reading of

Air Quality Daily Maximum would be the true maximum reading in the 14 day result set for the particular air quality parameter.

Lung Function Daily Minimum

	Min	Max	Range	Mean	Std.Dev.	No. of Records
PEF	5.1	7.2166667	2.1166668	6.1261897	0.6432992	14
FEV1	2.402	3.045	0.6430001	2.7905715	0.20853582	14
FVC	3.099	3.754	0.655	3.4747856	0.18659468	14

Lung Function Daily Average

	Min	Max	Range	Mean	Std.Dev.	No. of Records
PEF	5.375	7.2166667	1.8416667	6.3160357	0.5801415	14
FEV1	2.52125	3.1145	0.59325004	2.898444	0.16378593	14
FVC	3.32975	3.7546666	0.4249165	3.5633204	0.12406171	14

Air Quality Daily Maximum

	Min	Max	Range	Mean	Std.Dev.	No. of Records
NO2	30.0	75.0	45.0	49.785713	16.535177	14
PM10	21.0	89.0	68.0	49.285713	20.974081	14
PM2_5	11.0	54.0	43.0	27.142857	14.378708	14
O3	20.0	49.0	29.0	34.285713	8.203457	14
CO	0.7	1.3	0.59999996	0.9142858	0.17913099	14
SO2	2.0	27.0	25.0	11.357142	8.670082	14

Air Quality Daily Average

	Min	Max	Range	Mean	Std.Dev.	No. of Records
NO2	21.043478	55.166668	34.12319	32.852486	10.082262	14
PM10	12.958333	40.875	27.916668	25.656832	10.130348	14
PM2_5	7.818182	28.619047	20.800865	16.896008	8.040039	14
O3	9.291667	30.666666	21.375	18.38173	6.861967	14
CO	0.43333337	0.7833333	0.34999993	0.56534207	0.10855685	14
SO2	1.3636364	7.7083335	6.344697	3.793443	2.1170366	14

Figure 89 Results - Statistics

The third section of the results page (shown in Figure 90) shows correlations between the result sets. The correlation matrix contains all possible combinations of results from the values contained in the four tables shown in Figure 89. The correlation uses the Pearsons correlation formula.

$$r = \frac{\sum_{n} xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum_{n} x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum_{n} y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Eq. A.1 (Pearsons Correlation Formula)

where x represents values of air quality, y values of lung function, and n the number of values. A significance rating of 0.05 was also used to determine which correlation results

were showing levels of correlation. A typical results table is shown below. Where the results highlighted in green represents correlation combinations with a degree of significance (greater than 5%), and the *Av.PM10* and *Av.PM2_5* lines show particularly high levels of inversely correlated combinations against minimum PEF, FVC and FEV_1 values and average PEF.

Correlations

	Min.PEF	Min.FVC	Min.FEV1	Av.PEF	Av.FVC	Av.FEV1
Av.NO2	-0.716	-0.568	-0.598	-0.685	-0.454	-0.476
Av.PM10	-0.900	-0.802	-0.786	-0.884	-0.685	-0.667
Av.PM2_5	-0.842	-0.787	-0.752	-0.813	-0.653	-0.649
Av.O3	0.132	0.226	0.213	0.064	0.164	0.166
Av.CO	-0.419	-0.159	-0.271	-0.394	-0.085	-0.2210
Av.SO2	0.161	0.144	0.143	0.246	0.089	0.140
Max.NO2	-0.708	-0.542	-0.566	-0.716	-0.414	-0.469
Max.PM10	-0.495	-0.364	-0.345	-0.527	-0.302	-0.212
Max.PM2_5	-0.782	-0.7010	-0.695	-0.769	-0.564	-0.588
Max.O3	-0.004	0.165	0.113	-0.118	0.127	0.066
Max.CO	-0.108	0.0510	-0.0010	-0.184	0.121	-0.094
Max.SO2	0.029	0.040	0.061	0.098	-0.026	0.097

key:

Not significant

Significant at p = 0.05

Figure 90 Results - Correlations

A.7. Conclusions

The $EMSv_1$ was used successfully to produce required data sets and additional information for further analysis by independent users. The system functionality satisfied specifications, allowing users of the system remote access (to data) across the Internet. The user interface was evaluated by its users to be sufficiently flexible for the purpose of correlating the air quality and lung function data sets.

Using the $EMSv_1$ in the analysis of environmental and lung function data was particularly useful in facilitating multiple correlations and introducing time lag effects into result sets. Significantly correlated variables (using the 95-percentile) such as those shown in Figure 90 in green were then chosen as sub sets of data for further analysis. For one particular patient (WHT010) from the Whittington Hospital London, this was related to mainly PM_{10} (daily average), $PM_{2.5}$ (daily averages and maxima) and NO_2 (daily averages and maxima)

recorded at the London Bloomsbury air quality monitoring site. Further analysis used a backward stepwise multiple linear regression procedure resulting in the following equation (Crabbe *et al.*, 2001);

$$PEFav = (-5.059\text{x}10^{-2})\,(PM_{10}av)+7.614 \qquad Eq.\ A.2$$

which suggested that PEF daily average is inversely related to PM_{10} daily average values. This result could be used in automated monitoring systems if the average PEF value was known for a particular patient. If the monitoring system detected that the average PM10 value had risen to a value where the balance of Equation A.2 (*PEFav* value) fell below the patients danger threshold it could alert both the clinician and patient. However it was also noted that as the data sample was small, this could not be drawn as a conclusive result.

The system architecture performed well, with all parts of the system functioning reliably. Time taken to retrieve results across the Internet and the database varied slightly due to variations in network connections and internal processes of the database. Initially, the average time taken to retrieve a full result set was approximately seven seconds. The measurements of time taken were achieved using the server to monitor user connections to the EMS, times were clocked for various stages in seconds. The actual database query took the majority of the time, approximately 4 seconds. This was found to be reduced when another similar query was performed. The most recently used objects were cached within the database, saving query time. Result sets which matched an existing query were served straight from the web server in a matter of seconds, the maximum request time taking around three seconds but usually reducing to approximately one second.

The accuracy of the data presented to the user was assured through continuous testing during development. Formulae were tested and checked by hand by independent users of the system. Calculations were executed using floating point numbers to an accuracy of 10^{38} then rounded up to 3 significant figures for display to the user, this ensured an accuracy that was actually greater than needed. Crabbe *et al.* (2004) provide further discussion.

Appendix B

Data from the Great London Smog (1952)

B.1. Correlation Example

Data from the Great London Smog in 1952 (shown in Figure 91) presents at least two sets of highly correlated data; the first between the number of deaths per day and the level of sulphur dioxide in the atmosphere, and the second between the level of smoke particulates and the death rate. The trends during December 1952 can be seen in the figure below.

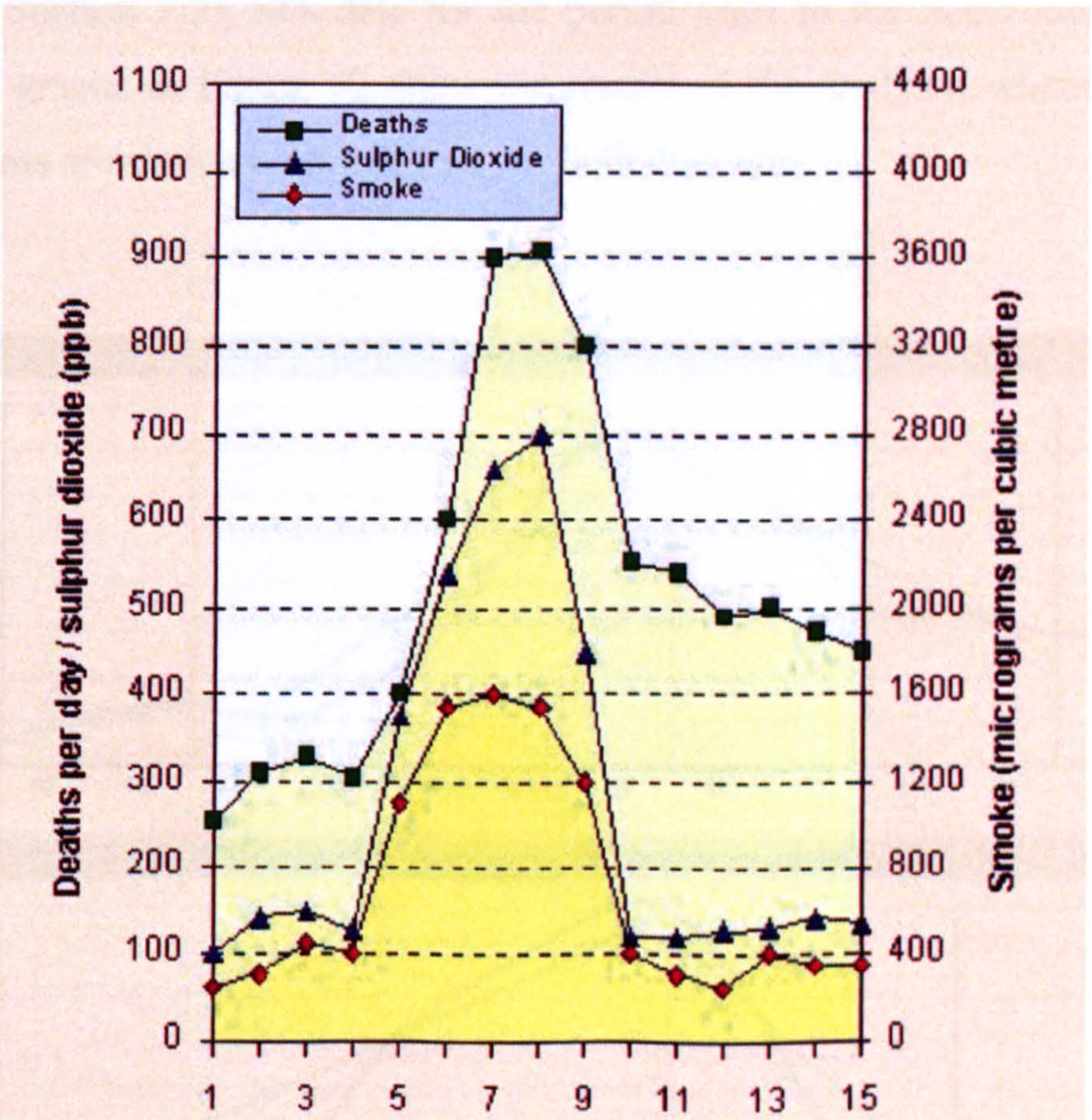


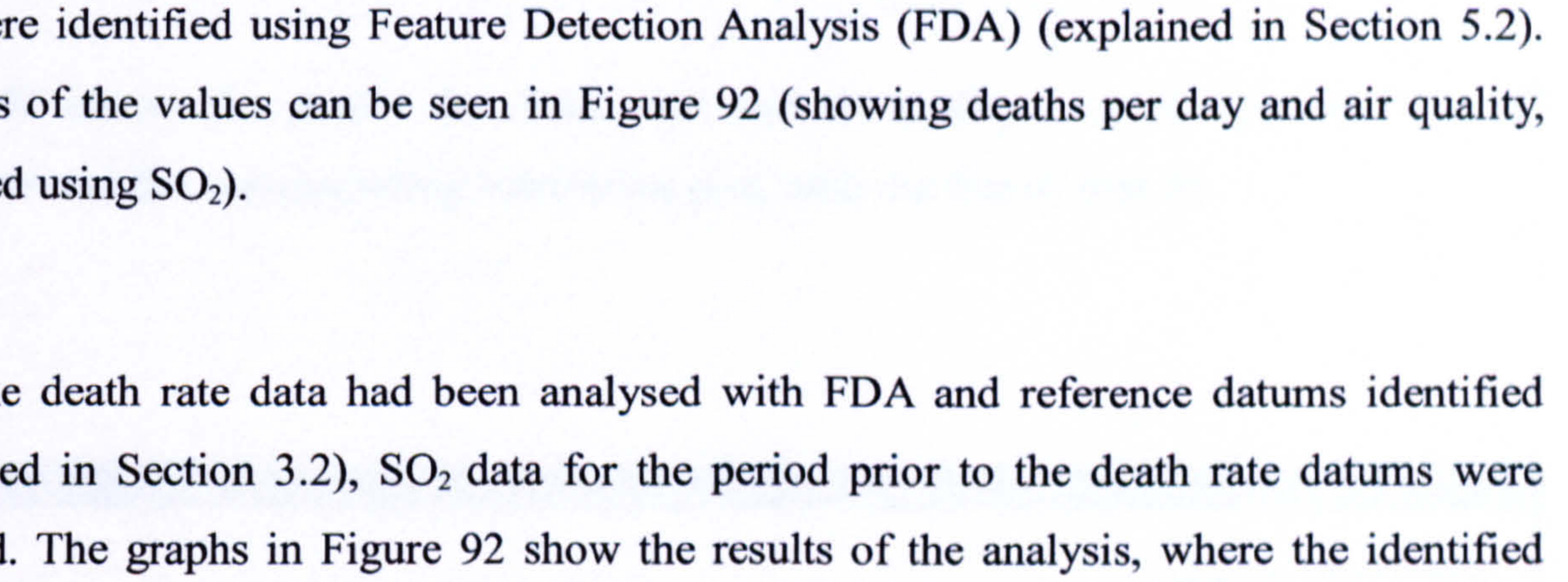
Figure 91 Deaths per day during the Great London Smog (December 1952) against sulphur dioxide and levels of smoke. © EAE 2000

A correlation component was devised for the Environmental Monitoring System (EMS) to test a set of interpolated values from a well known real-world data set. The purpose of the test was to demonstrate the difficulty in defining the period over which to correlate. The component read off corresponding *y*-axis values at regular intervals along the time series (*x*) axis.

Death rate and sulphur dioxide data sets were extracted from the graphs above, and a small London smog database created. The periods that the correlation measurements were taken over were identified using Feature Detection Analysis (FDA) (explained in Section 5.2). Analysis of the values can be seen in Figure 92 (showing deaths per day and air quality, measured using SO_2).

Once the death rate data had been analysed with FDA and reference datums identified (explained in Section 3.2), SO_2 data for the period prior to the death rate datums were analysed. The graphs in Figure 92 show the results of the analysis, where the identified reference datums are shown with red lines for both data sets.

Death Rate



Air Quality

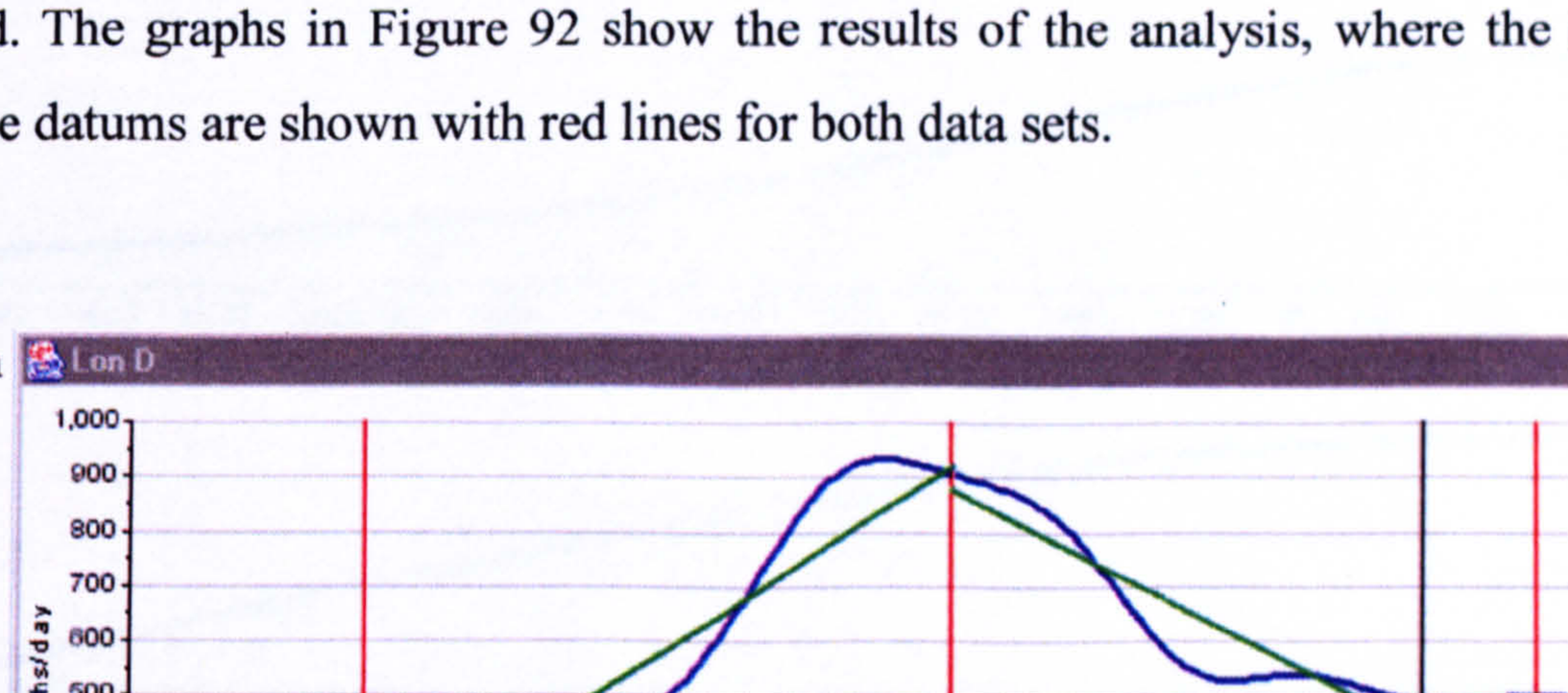


Figure 92 Graphs showing the 1952 London Smog data plotted by the EMS FDA, and identified reference datums (red markers) in each data set

The period of time between the start of the time series on the 1st December 1952 and the death rate reference datum on the 3rd December was analysed using the correlation component. By inspection of the original data (found in Figure 91) it was expected that there would be a high correlation between the data sets. Two further reference datums (from the death rate graph) between the 8th and 14th December (shown above in Figure 92) were used. To investigate the correlation between the data sets corresponding to these two reference datums, the periods from 1st December to 8th December, and from 1st December to 13th December were examined.

Figure 93 shows the graphs for death rate and air quality over the period 1st to 3rd December and the corresponding correlation plot, with the line of best fit.

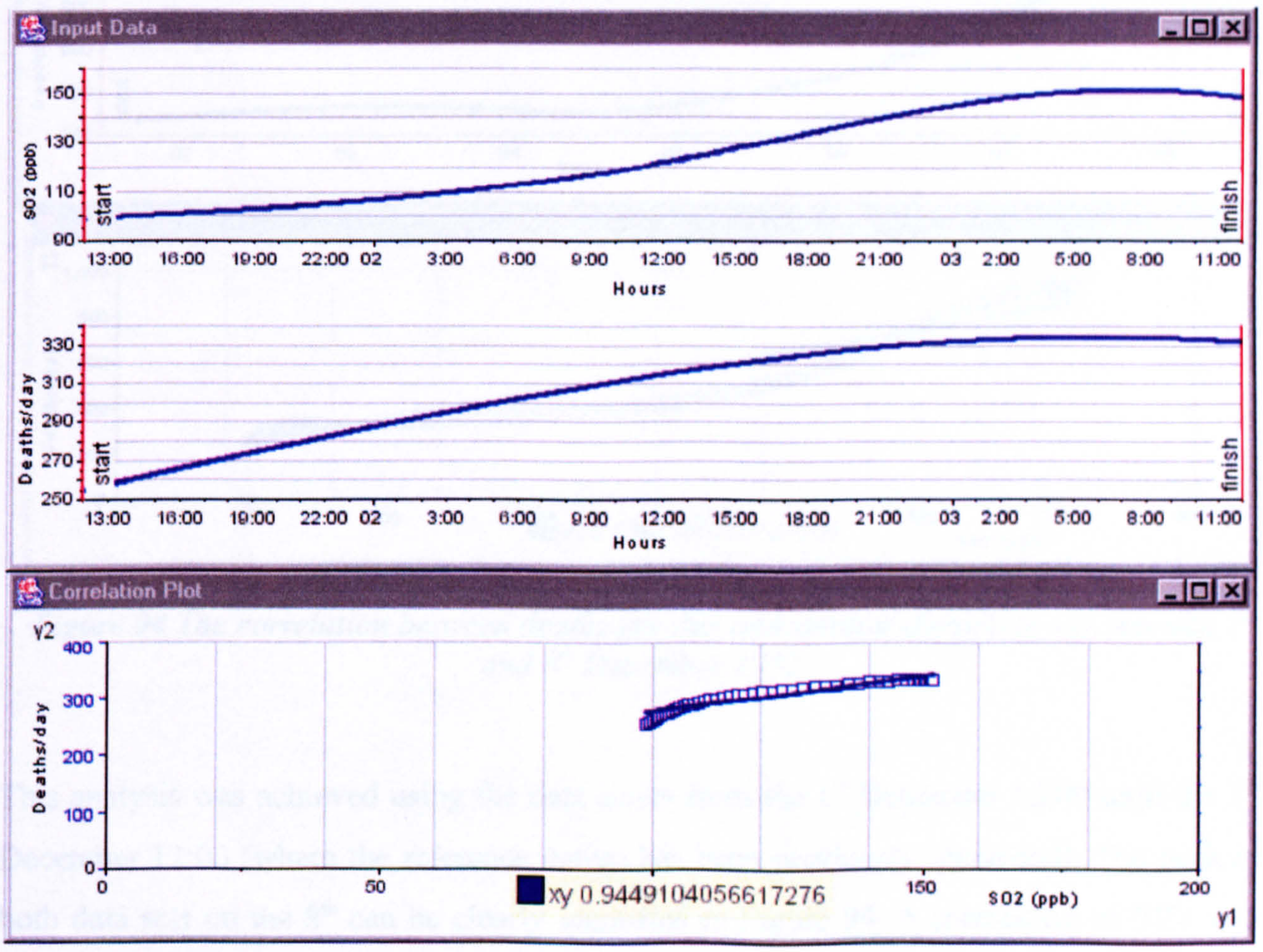


Figure 93 The correlation between deaths per day and sulphur dioxide levels between 1st and 3rd December 1952

The period of SO_2 and death rate data from the 1st December 12:00 to 3rd December 12:00 were plotted and a correlation of 89% ($0.9449^2 x100$) found between them after interpolating values at 1 hour intervals from the *Input Data* data plots in Figure 93.

B.2. Second Correlation between 1st and 8th December

The data sets used for correlating the second period (1st December to the 8th December) from the start of the time series until the 2nd death rate reference datum are shown by the graphs in Figure 94 below. The correlation coefficient identified between the two data sets is 92%.

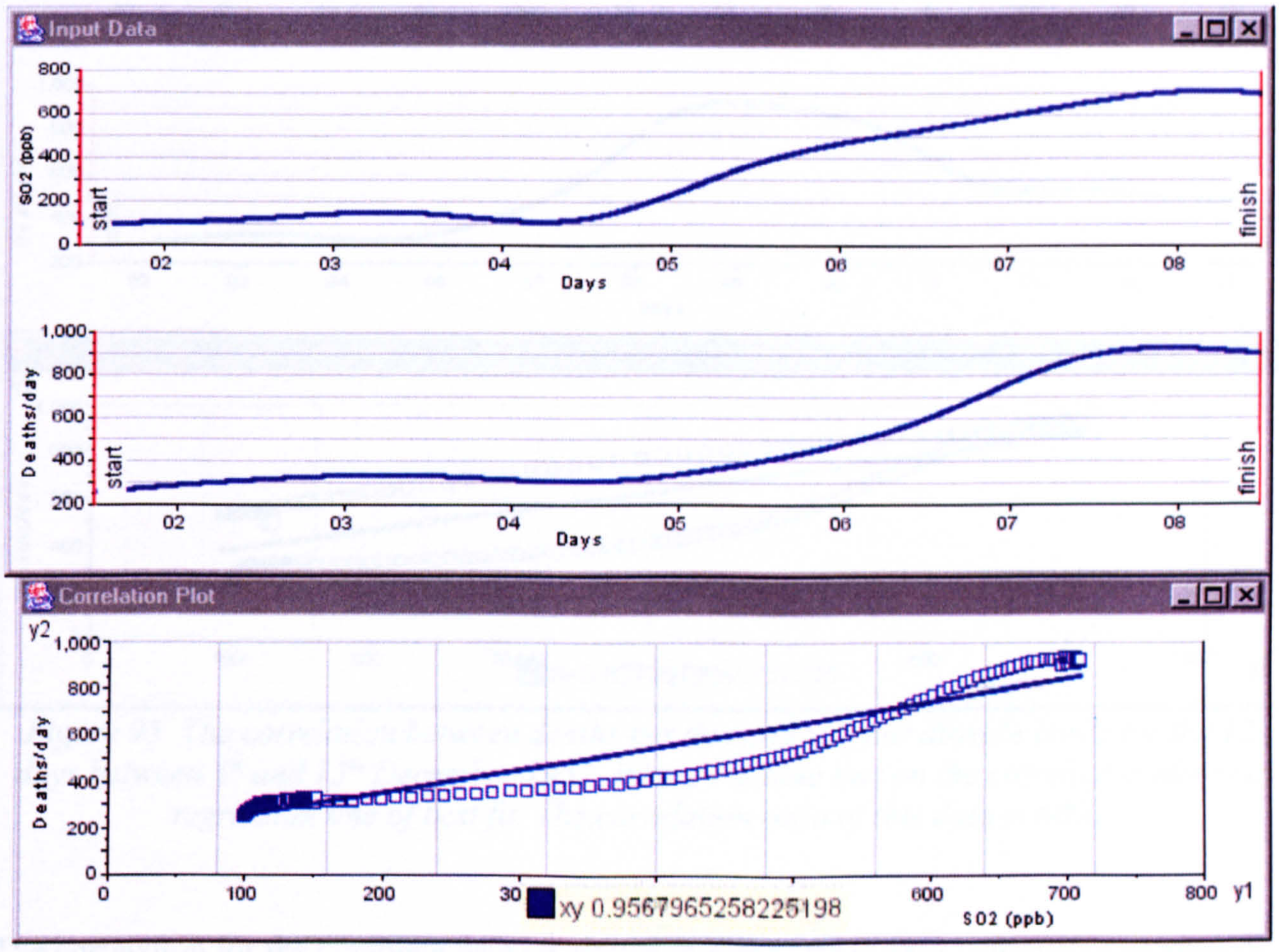


Figure 94 The correlation between deaths per day and sulphur dioxide levels between 1st and 8th December 1952

This analysis was achieved using the data series from the 1st December 12:00 until the 8th December 12:00 (where the reference datum has been previously identified). The peak in both data sets on the 8th can be clearly identified in Figure 94. A correlation of 92% was calculated.

B.3. Correlation of the full Great London Smog data set

The full data set (up until the final reference datum) for the Great London Smog (shown in Figure 92) was analysed by the correlation component. The result is presented in Figure 95 below.

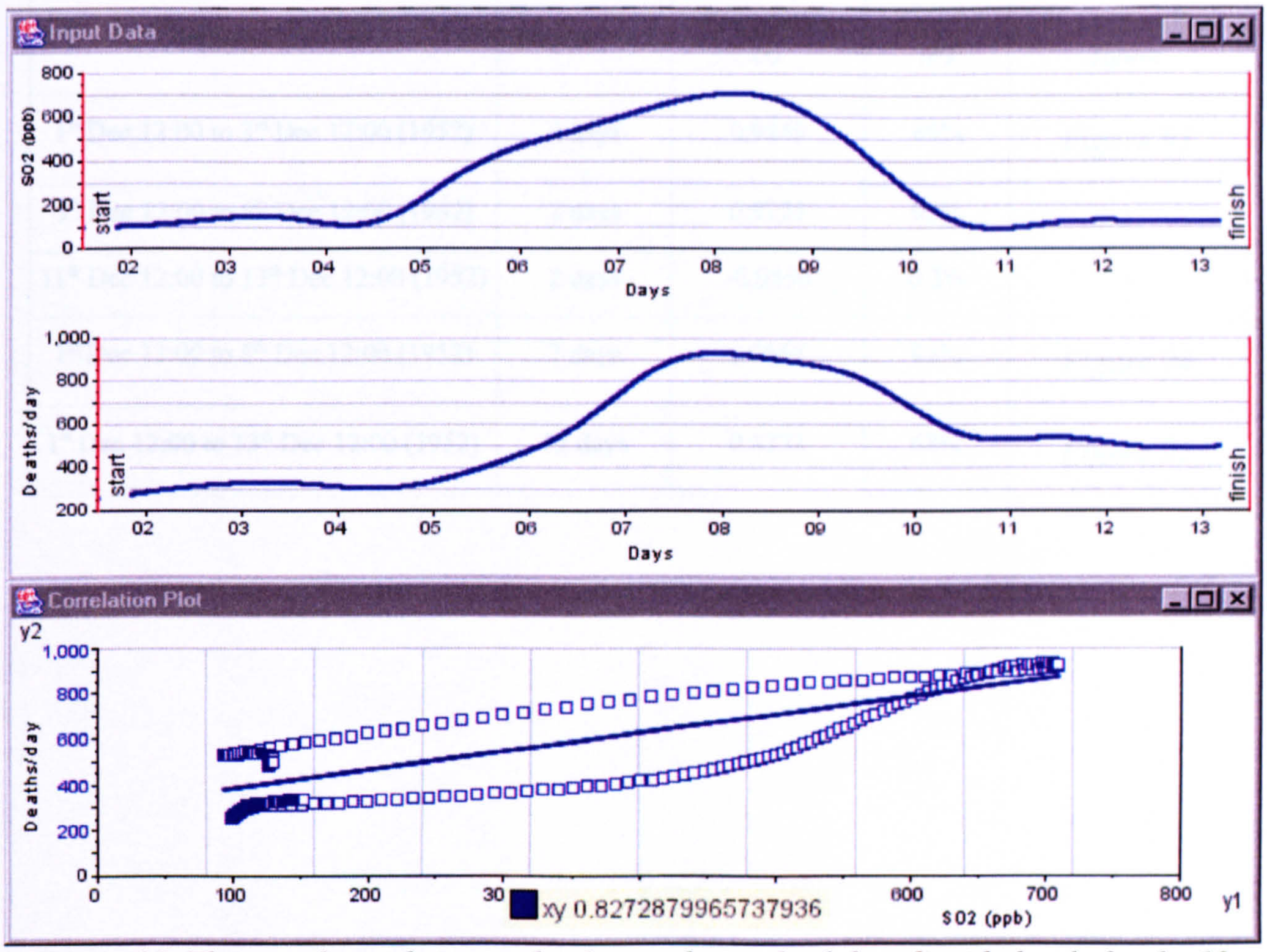


Figure 95 The correlation between deaths per day and sulphur dioxide levels for the 12 days between 1st and 13th December 1952. Where the blue line on the correlation plot is a regression line of best fit. The correlation ratio of this data is 68%.

The correlation for the complete data set shown in Figure 95 is 68%, although a correlation of 92% was identified for the period 1st to 8th December (Figure 94), shown above as the rising edge of the death rate and SO_2 curves (until the 8th December). However, the trailing edge from the 10th December 12:00 onwards shows no correlation, represented by the correlation of the 2 day period before the 13th December (result in the table below). The significance of this result is not calculated (is irrelevant), as the correlation example is demonstrating how correlation ratios are effected as the correlation period is extended.

Table 38 Summary of correlated sections, resulting from the Great London Smog data set, including 2 day periods before each identified reference datum of Figure 10 (two are not shown graphically).

Correlated Section	***Time period***	***Correlation (r)***	***Ratio% (r^2)***	***Corresponding figure***
1st Dec 12:00 to 3rd Dec 12:00 (1952)	2 days	0.9449	89%	*Figure 93*
6th Dec 12:00 to 8th Dec 12:00 (1952)	2 days	0.9725	95%	-
11th Dec 12:00 to 13th Dec 12:00 (1952)	2 days	-0.0550	0.3%	-
1st Dec 12:00 to 8th Dec 12:00 (1952)	7 days	0.9568	92%	*Figure 94*
1st Dec 12:00 to 13th Dec 12:00 (1952)	12 days	0.8273	68%	*Figure 95*

Appendix C

Architectural Patterns

There is a general overlap between the four categories (Table 39 below), with only Interactive Systems being distinct from the others. The overlap is to be expected as the architectural features of extensibility, scalability and simple management require a mixture of the patterns for an effective architectural design. For the purposes of the EMS the three architectural patterns chosen to fulfil the requirements of the architectural views are Layers, Model-View-Controller, and Reflection patterns. These three patterns have been chosen to provide separation between components and sub-components; increasing extensibility, scalability and re-use.

Table 39 Showing the use of architectural patterns in the construction of system architecture, (data provided by Buschmann et al.,1996 p26).

Category Name	*Use*	*Patterns falling into the Category*
From Mud to Structure	Breaks down the overall system into a number of cooperating subtasks, avoiding a *sea* of components.	• Layers pattern. • Pipes & Filters pattern. • Blackboard pattern.
Distributed Systems	Provides a complete infrastructure for distributed applications	• Broker pattern. • Microkernel pattern. • Pipes & Filters pattern.
Interactive Systems	Supports the structuring of software systems that feature human interaction	• Model-View-Controller pattern. • Presentation-Abstraction-Control pattern
Adaptable Systems	Supports extension of applications and evolution to new technologies.	• Reflection pattern. • Microkernel pattern.

C.1. Layers

Layers help structure applications that can be broken up into a number of subtasks at a similar level of abstraction. The layers must be strictly separated from each other. No component may spread over more than one layer. An example of a layer is *network architecture*. Each layer deals with a specific aspect of communication and uses the 'services' of the next lower layer.

C.2. Pipes and Filters

The Pipes and Filters architectural pattern gives structure to systems needing to process streams of data. The pattern divides the task of a system into several sequential processing steps. Data is passed through pipes between filters, while the filters can be used in succession to build a processing system. The method provides a simple mechanism to exchange sections of the pipeline with new filters making it a flexible approach to system design. It is not, however, regarded as good systems design when a large amount of global data has to be shared with each of the pipeline filters.

C.3. Blackboard

This pattern is useful where the system's domain is immature. It can be used to experiment with different algorithms for the same subtask, for this reason individual modules should be easily inter-changeable. The blackboard is the central data store and provides an interface that enables knowledge sources to read and write from it. Elements from a solution can appear on the blackboard, and if rejected later in the process can then be removed.

C.4. Broker

A broker pattern can be used in systems where components are decoupled from each other, usually distributed over a network. A broker component coordinates communication between software systems. The pattern uses remote service invocations to access services over a network without having to know their physical location. This allows replication and ease of service migration. By partitioning functionality into independent components the system becomes potentially distributable and scalable. A drawback to using the broker pattern is a possible reduction in efficiency, broker systems are usually slower due to the extra layers needed to maintain the decoupled aspects of the pattern. The pattern can increase the fault tolerance of a system through the ability to replicate components.

C.5. Microkernal

The Microkernal pattern is suited to systems that must be able to adapt to change. It separates a minimal functional core from extended functionality and customer specific parts. The pattern also allows extensions to be plugged into the system and coordinates them. The microkernal includes functionality that enables other components running in separate processes to communicate with each other.

C.6. Model-View-Controller (MVC)

This design pattern allows flexible use and reuse. Changes in any part of the system can be achieved without the need to change functionality in all components. The pattern divides an application (usually an interactive one with a graphical interface) into three components:

1. Model: containing the core functionality and data.
2. View: displays information to the user.
3. Controller: handles user input and receives events from the model. This can be achieved through the Publisher-Subscriber pattern (Appendix D.6).

The user interface usually consists of both the *View* and *Controller* components. Different users place conflicting requirements on the user interface and this pattern method facilitates the use of multiple user interfaces that essentially use the same data but show it and engage with it in different ways. The change-propagation mechanism is the only link between the three components.

C.7. Presentation-Abstraction-Control (PAC)

Defines a structure that is designed for a hierarchy of cooperating agents. Each agent is responsible for a particular aspect of the application's functionality, and consists of three components:

1. Presentation, provides a particular view of the corresponding semantic concept.
2. Abstraction, represents the data of the agent in a relevant way.
3. Control, provides control and communication aspects of the architecture.

This separation of components divides the presentation (human-computer interface) from the functional core, much like the MVC pattern. The first use of PAC was in the area of artificial intelligence (Crowley, 1985). Communication between the abstraction and presentation components is decoupled by the control component. In systems developed

through the use of the PAC architecture cooperating agents, all with specific tasks, work together to provide the system functionality.

C.8. Reflection

Use of the reflection pattern enables dynamic changes to the structure and behaviour of software systems. This pattern has two parts:

1. Meta-level, provides information about system properties and makes the software self-aware. This level encapsulates system internals that may change, changes in this level affect base-level behaviour.
2. Base-level, implements the application logic. The implementation of which builds on information provided by the meta-level.

The general structure of a reflective architecture is very similar to a layered system. The difference however is in the dependencies between layers. The levels used for reflection (Meta and Base) build on each other whilst in a pure layered architecture each layer only builds on those layers that are below. Reflection has benefits when modification of code is required, existing code can be modified by calling a function of the metaobject protocol. The metaobject protocol integrates the change requests and if necessary recompiles the changed parts and links them to the application while it is still executing. The architecture pattern may lead to a lower efficiency however, as the two levels consult each other, retrieve information and modify objects at run time.

Appendix D

Design Patterns

A design pattern can be described as a commonly occurring structure of components that work together to solve a general design problem, usually within a particular context. Design patterns specify functionality on a smaller scale than architectural patterns, but are at a higher level than programming language specifics. Implementation of a specific design pattern does not effect the fundamental structure of a software system, but may influence how a subsystem is designed. The table below shows eight design patterns that fall into five categories, detailed below;

Table 40 Use of design patterns in the construction of system architecture, (data provided by Buschmann et al.,1996).

Category Name	*Use*	*Patterns falling into the Category*
Structural Decomposition	Decomposes systems and complex components into cooperating parts. Independent components can be simpler to handle, are easier to understand and changes to them can be made more easily.	• Whole-Part
Organisation of Work	Component collaboration to solve complex problems.	• Master-Slave
Access Control	Guard and control access to the system.	• Proxy
Management	Handles collections of components in their entirety.	• Command Processor • View Handler
Communication	Organises communication between components.	• Publisher-Subscriber • Forwarder-Receiver • Client-Dispatcher-Server

D.1. Whole-Part

An object described as *Whole* represents a grouping together of smaller objects called *Parts* into a collaboration. The functionality of the *Whole* object is an ordered culmination of the functionality of all the smaller *Part* objects. The pattern attempts to model real world relationships. Services belonging to a *Whole* object are visible to external clients, whilst the internal workings of the *Parts* are wrapped inside the *Whole* and can not be accessed unless they are allowed by method calls from the *Whole* object.

D.2. Master-Slave

The Master-Slave pattern applies the *divide and conquer* principle, where work is partitioned into several subtasks for processing independently by the slaves. The master component splits the work and distributes it to the slave components and computes a final result from the results that the slaves return. The slaves usually perform semantically identical sub-tasks. The pattern supports fault tolerance, parallel computation and computational accuracy.

The master component provides a service that allows external entities to access the service. They are unaware that the service implemented by the master partitions the work into several equal sub-tasks, computing a final result from all the results obtained.

D.3. Proxy

A proxy is usually used to encapsulate the interface and remote address of a server. The proxy pattern is often used with a forwarder (Appendix D.7) that takes the message and transforms it into IPC (Inter-process communication) level code. The design makes the client using a component communicate with a representative rather than to the component itself. This interface can be used for many purposes: enhanced efficiency through caching mechanisms, easier access, or protection from unauthorised access. The proxy pattern assists access to services provided by other systems especially when direct access with the system could be inappropriate or subject to regular change.

The use of proxies can have an adverse affect on efficiency due to the extra layers of computation in the object communication channels. Caching only solves this inefficiency when the services or objects being cached are relatively static. If the rate at which the objects change are high then the overhead needed to invalidate old copies in the cache may

defeat any benefit from using the cache in the first place.

D.4. Command Processor

The Command Processor design can be used in applications requiring a mechanism to schedule or undo certain actions. The pattern separates the request from its execution. This separation of operations means that the execution section of the request can be stored for future use as a separate object entity, either as part of a scheduled action or as a record of what to undo if an outcome was not desired.

The ability to que actions promotes flexibility, if a different user interface was required then as long as the interface created scheduled objects that the core functionality could understand the interfaces can be changed and modified at will. The methodology could similarly be used to log actions and events within the system without affecting its operation. Another use is the automatic rollback of transactions if they do not go to plan, keeping records intact. The pattern in some circumstances may also allow commands to be executed in concurrent threads. As with all patterns that use a decoupling methodology the additional indirect cost of storage time before execution must be considered before implementing the pattern.

D.5. View Handler

A View Handler pattern can be used where a software system provides multiple views of the same document or supports working on multiple documents at the same time. The pattern overcomes the problems associated with multiple views of the same document and enabling editing in each one by providing an efficient update mechanism for propagating changes between windows containing the views. An update made to one view of the document will automatically be reflected in the other views too. It is worth noting that both the Model-View-Controller (MVC) and Presentation-Abstraction-Control (PAC) architectural patterns share the principles of the View Handler.

D.6. Publisher-Subscriber (or Observer)

The Publisher-Subscriber pattern is especially useful when one or more components of a system wish to receive notification of an event or receive a particular object from the *Publisher.* A *Publisher* notifies any number of *Subscribers* about changes to its state, using this method cooperating components are kept synchronised.

D.7. Forwarder-Receiver

Using the Forwarder-Receiver design pattern, components are decoupled from the underlying communication mechanism. On the client side a *Forwarder* receives a request from the client and handles the mapping to the communication facility used. On the server (or *Receiver*) the request is received via a general interface that has functionality for receiving and unmarshalling (extracting) the message. The design pattern addresses the issue of where to send a request by using a *name space* to identify the *Receiver* or group of Receivers that will receive the message. The message protocols are also defined.

The *Forwarder* contains all the functionality for sending messages across process boundaries, using a name repository (or registry) to identify the physical address of the recipient. The *Forwarder* connects to the remote peer (or Server) via the registry. The *Receiver* is responsible for all the functionality required to receive the message, and decouple the actual message part.

D.8. Client-Dispatcher-Server

A dispatcher allocates, opens and maintains a direct channel between a client and server. The dispatcher component acts as an intermediate layer between the client and server. This allows access to a service running on a server without knowing its physical location, and keeps the code implementing the server connection separate from the core functionality of the service. A server registers itself with the dispatcher by its name and address. Once this registration has taken place the dispatcher is now able to establish communication with the server when a client makes a service request. The connection is established once the look-up of the service (required by the client) has been successful.

Appendix E

Service Oriented Architecture

Service Oriented Architecture (SOA) provides a useful foundation on which to build an implementation of the EMS architecture. The foundations for SOA exist (*W3C*, 2004) (*W3C*, 2006). The use of web services is one way to implement a SOA. Advantages are twofold: for the service, communication between additional system components is simple, and for the client, connection is simple; the latest version of system software is delivered each time the service is used and availability is made continuous.

Foster and Kesselman (2004) promote the ideal that all entities of a service-oriented architecture are services; this implies that any user visible operation is the result of a message exchange. The standard service architecture usually consists of at least a pair of services able to communicate with each other. It is envisaged that as a long term objective, the EMS would be used in this type of component environment. Foster and Kesselman (2004) define a service as, "an entity that provides some capability to its client by exchanging messages" they go on to say that "a service is defined by identifying sequences of specific message exchanges that cause the service to perform some operation".

The prototype modules were designed during this thesis using an iterative approach. The modules communicate through the exchange of messages that could be extended to operate in a Service Oriented Architecture (SOA) environment. Zimmermann *et al.* (2007) discuss the use of Service Oriented Architecture (SOA) and web services, suggesting that it reinforces general software architecture principles such as separation of concerns and logical layering. Web services are loosely coupled, communicating directly with other web services and end users via the Internet. A web service that adheres to the Open Grid Services Infrastructure (OGSI) (Tuecke *et al.*, 2003) is called a *Grid Service*. The implementation of distributed systems working together to complete complex tasks is

made possible by the effectiveness of the network. The use of a network can often lead to substantial cost and time savings when used to assist with complex calculations.

E.1. Lookup Service

The purpose of a lookup service is to help devices, systems and other such processes to be found by each other on a single machine or over a network. Lookup services allow various combinations of functionality to be achieved by allowing a plug and play type of capability to a system, the look up service keeps a record of all the services that exist on a particular network. There are a number of examples of lookup services. JNDI (Java Naming and Directory Interface), Jini lookup service:

> "The Jini lookup service is a fundamental part of the federation infrastructure for a *djinn,* the group of devices, resources, and users that are joined by the Jini technology infrastructure. The *lookup service* provides a central registry of services available within the djinn. This lookup service is a primary means for programs to find services within the djinn, and is the foundation for providing user interfaces through which users and administrators can discover and interact with services in the djinn."
>
> (Sun, 2003a)

Novell's Directory Service (NDS) eDirectory, and Microsoft Active Directory are example of a lookup service. Most lookup services are LDAP (Lightweight Directory Access Protocol) compatible.

The primary protocol used by web services is Simple Object Access Protocol (SOAP). SOAP is based on eXtensible Markup Language (XML) and allows dissimilar applications to interact regardless of their underlying platform, programming language or internal application calls. This description also loosely defines web services.

E.2. Web Service Architecture

Through the use of web services, healthcare organisations are able to integrate IT assets and services across the network (Sun, 2003b). Web service protocols include: SOAP, XML, EDI, WSDL, UDDI.

E.3. Health Industry Model Architectures

The *Patient health network*, supports emerging wireless, wearable medical devices. This

development, which enables patient mobility, has particular application to the EMS architecture. Another development is the use of a correlation identifier to handle users, and file patient specific reports (Singh *et al.*, 2004).

Integration of the EMS with a wider healthcare system is not widely considered by this thesis, however consideration has been made with respect to the types of input and output that the system can except and create. Choe and Yoo (2008) propose the use of XML for data interchange within a secure web-based repository system for healthcare records. The proposed system uses XML messages for communication, which enables a flexible design for the exchange of healthcare data in a common format.

Messaging and Collaboration Services

Provide consistent electronic communication using common medical terminology, code sets, transaction protocols, or application messaging. The industry standard for application to application message interchange.

Health Level 7 Application Messaging

HL7 messaging addresses the interoperability requirements of the healthcare industry. It is a standard providing a means of clinical and administrative data supporting patient care to be exchanged between healthcare applications..

Electronic Data Interchange (EDI) transactions

EDI allows entities within the healthcare system to exchange medical, billing, and other information, as well as process transactions quickly and cost effectively.

Standardised medical terminology and code sets

There is not a single standard that meets the needs of healthcare providers. The most common terminology and code sets are: LOINC, SNOMED, ICD, CPT, UMLS, and DSM.

DICOM

The Digital Imaging and Communications in Medicine (DICOM) standard facilitates the interoperability of medical imaging equipment.

Appendix F

EMS Service Implementation Architecture

Figure 96 shows how the components of the EMS coordinate their work flow to achieve automatic recognition of air quality features leading to a decline in patient lung function. The flow begins with the *workflow coordinator*, where the initial parameters controlling the identification are entered by clinical staff. Once the system has been initialised, the work flow passes to the second component (*Data Analyser*) where data relevant to the patient is extracted from the database using the *database query interface*; also available as a service. The data is passed back to the analyser, where control is passed to the *FDA* component. Reference datums are returned to the *Data Analyser* and subsequently the *workflow coordinator.*

Control passes to the *Data Locator* component, where environmental data relating to the movements of the patient are extracted from the database (using the *query interface*). Depending on the analysis initially activated by the member of clinical staff, the process then either uses the FDA component to identify the environmental reference datums, or extracts time series data from the environmental data set, leading up to each identified lung function datum.

The preprocessing stage of the data analysis is complete, and the work flow transfers to the *Vector Organiser.* This component extracts the delay characteristics between the two data sets and prepares the input to the pattern recognition modules. The *workflow coordinator* stores the results from both the lung function and environmental analysis modules for this purpose. The parameters of the delay characteristic sent to the pattern recognition modules (*date* of environmental reference datum, *value*, and *lag* time to the lung function datum) can at this point be chosen (in the prototype). However this choice could be made in the initial system initialisation by clinical staff.

The remaining pattern recognition modules divide into two categories:

1. FBCA (Frequency, Boundary, and Cluster Analysis).
2. Neural Analysis (Self-organising map algorithm).

Frequency, Boundary and Cluster Analysis (FBCA) is split into two modules; the functionality of the frequency and boundary analysis is transferable to other applications as a statistical tool, and the cluster analysis module.

The event driven nature of the existing prototypes make them capable of adaptation to service driven environments. Defacto standards for the implementation of service architectures are XML and web services (e-Government Unit, 2005), they provide a *loose* coupling between multiple applications sharing the same infrastructure.

Integration with existing healthcare systems is also simplified through the use of XML. Muller *et al.* (2001) show an example of a decision support system integrated with a hospital information system using XML, and distributed object technology.

Architecture using a Service Methodology

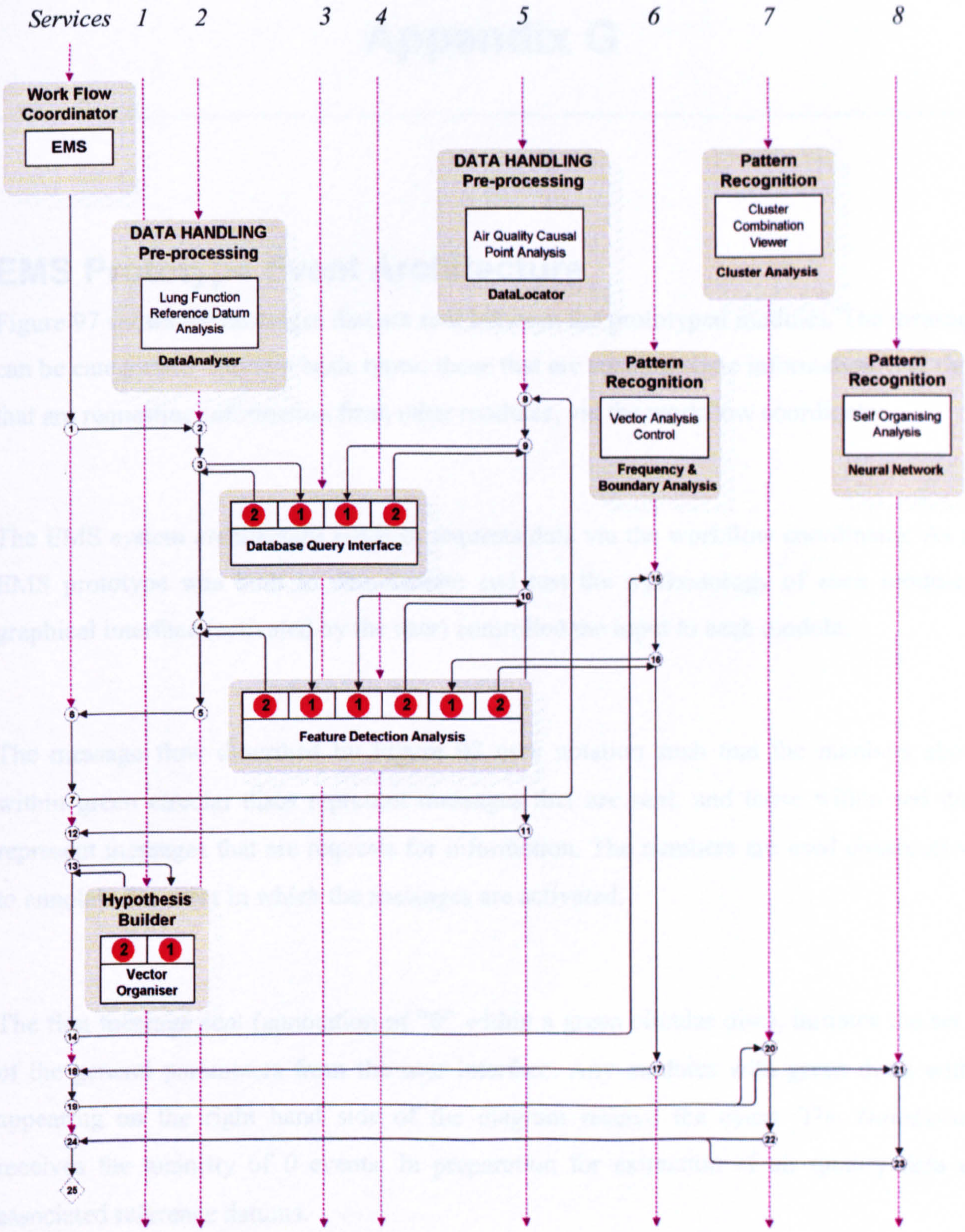


Figure 96 EMS Conceptual Service Architecture. Where the dotted purple lines represent each service, and solid black lines show the workflow.

Appendix G

EMS Prototype Event Architecture

Figure 97 shows the messages that are sent between the prototyped modules. The messages can be categorised into two basic types; those that are sending some information, and those that are requesting information from other modules, via the workflow coordinator.

The EMS system architecture sends or requests data via the workflow coordinator. As the EMS prototype was built to demonstrate and test the methodology of each module, a graphical interface (activated by the user) controlled the input to each module.

The message flow described by Figure 97 uses notation such that the numbers shown within green circular discs represent messages that are sent, and those within red discs represent messages that are requests for information. The numbers are used consecutively to annotate the order in which the messages are activated.

The first message sent (annotation of "0" within a green circular disc), initiates the set up of the general parameters from the user interface. Any modules with green discs with *0* appearing on the right hand side of the diagram receive the event. The *DataLocator* receives the majority of *0* events. In preparation for extraction of air quality data and associated reference datums.

The Data Analyser is also made ready to begin the analysis of the lung function data and identifies the appropriate reference datums which it passes back through the workflow coordinator (annotation of "1" within a green circular disc).

The *Data Locator* responds to the receipt of the reference datums by extracting the

appropriate environmental data from the database. The datums found using the FDA component are then used in one of two ways defined by the user interface; as markers to extract time series data leading to each datum, or the reference datums themselves are sent back to the workflow coordinator.

The *VectorOrganiser* (part of the hypothesis builder) then collects the data from the workflow coordinator and orders the data according to the analysis required, passing the ordered data back to the workflow coordinator (annotation of "9" and "10" within green circular discs).

At this stage, control is given over to each of the analytical components (*Neural Network* and *FBCA*) by the workflow coordinator. The Correlation component can be activated earlier during the work flow, once both corresponding air quality and lung function data sets are received by the component. (annotation of "2" and "3" within green circular discs).

The results from both the FBCA and neural analyses are held within the respective prototype components. However, in a full implementation the results would be passed back to the workflow coordinator for further use by the dissemination components.

Send **Receive**

Actual Event Architecture

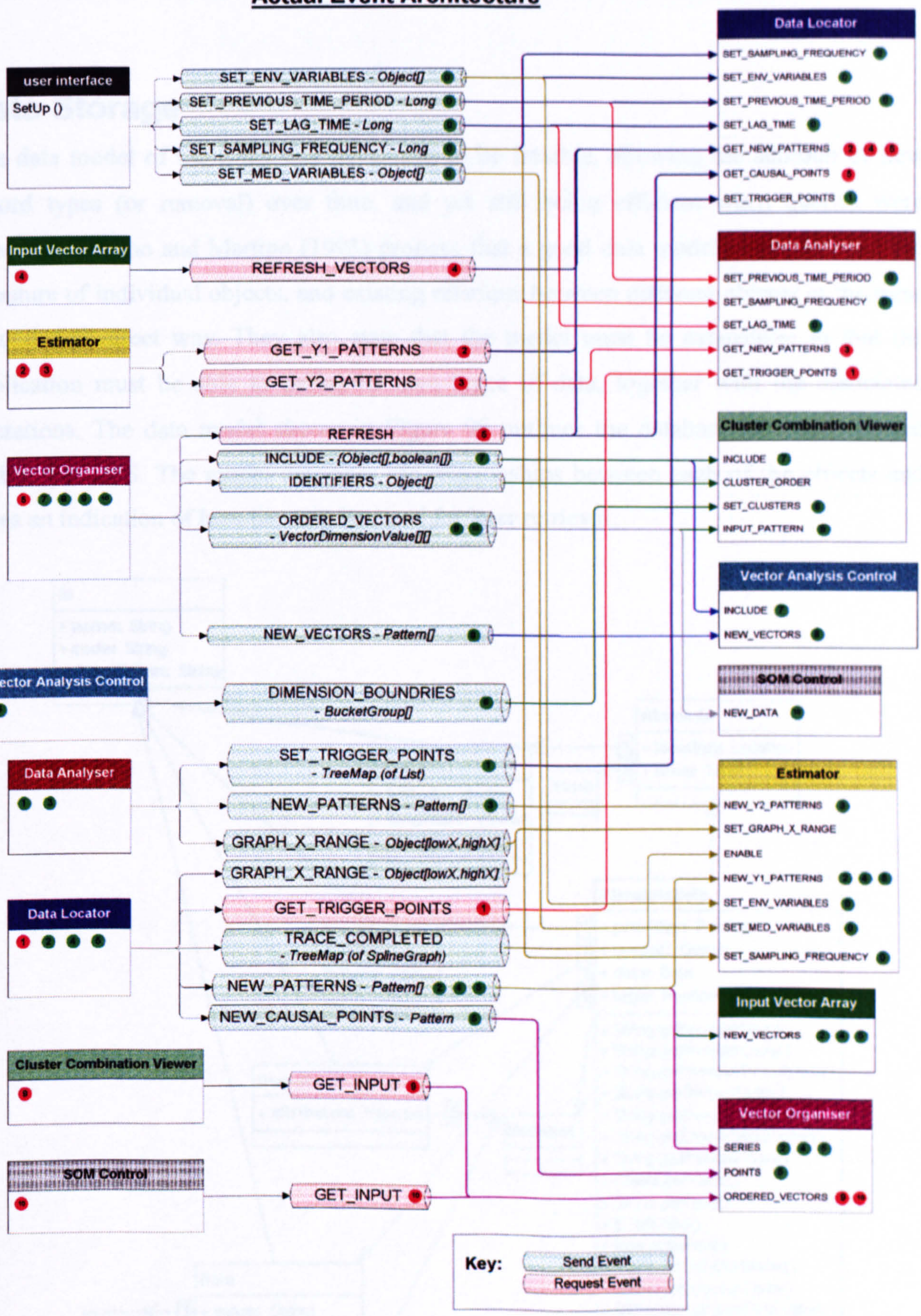


Figure 97 The actual event architecture used within the prototypes. Relates to Figure 'Prototype Modules' in Chapter 4.

Appendix H

Data Storage

The data model of the EMS was developed to be flexible, allowing the addition of new record types (or removal) over time, and yet still being efficient when queries were executed. Bertino and Martino (1993) propose that a good data model expresses both the structure of individual objects, and existing relations between different objects in the most natural and direct way. They also state that the model must be extensible; in that the application must be able to define its own types of data, together with the associated operations. The data model shown in Figure 98 outlines the database schema deployed within the EMS. The model describes the relationships between each of the objects and gives an indication of how the data is stored for later retrieval.

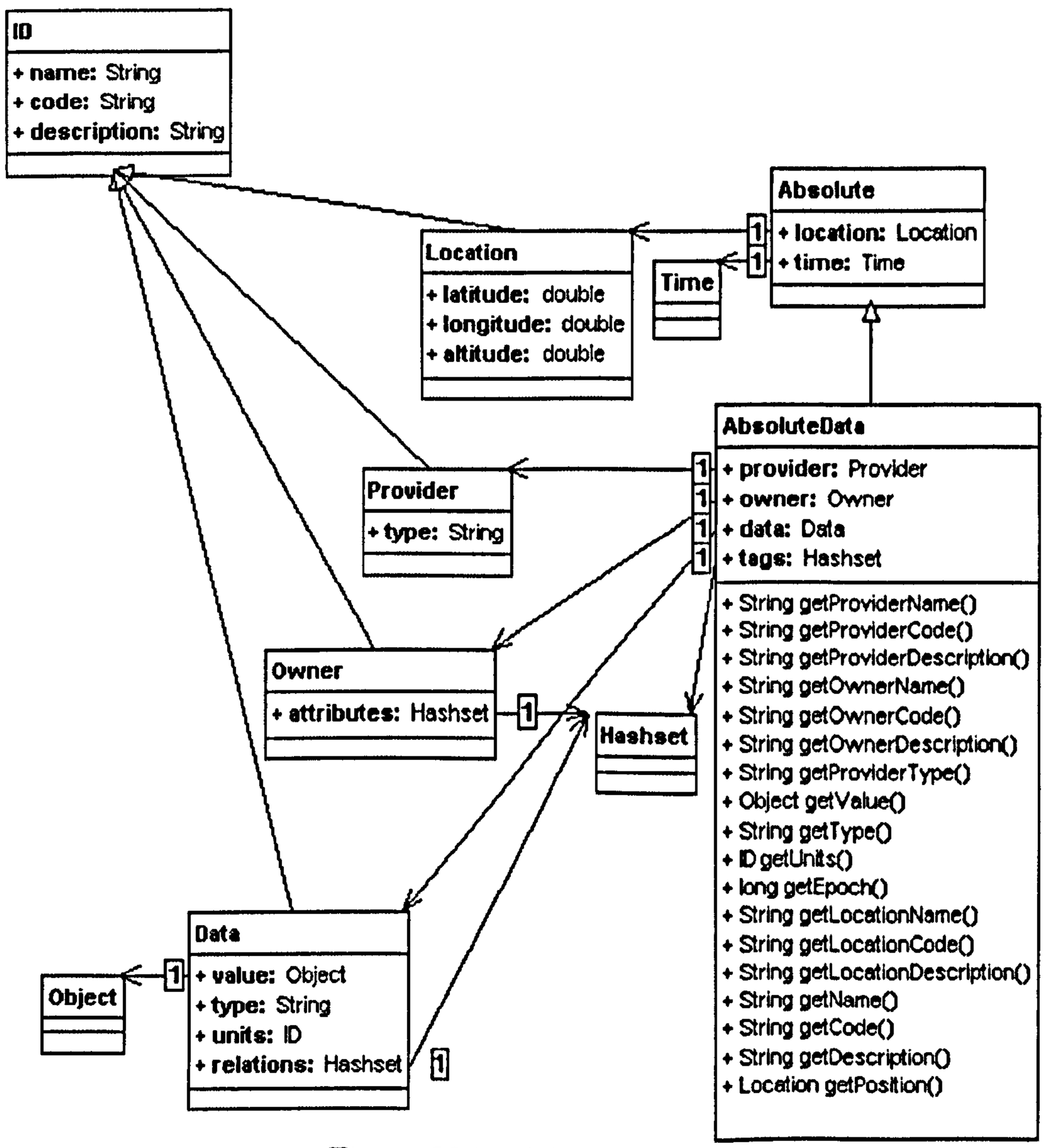


Figure 98 EMS Data Model

It is useful to note that the structure has been devised so that types of data stored and manipulated within the EMS are not dictated by the system architecture. The architecture of the data model has been devised so that any object (or type of data value) can be stored by the system; shown by the *Data* section of Figure 98, where the value can be of any *Object* type.

Appendix I

Frequency Analysis Implementation

The structure to record the cluster boundaries within the FBCA component, and monitor for matching patterns is shown in Figure 99. There are three types of bucket: *Date*, *Lag*, and *Value*. These three buckets extend the functionality contained by the *AbstractBucket* class, which implements the *Bucket* interface. The implication of this is that the three different types of bucket can be treated similarly when making comparisons, due to meeting the requirements of the Bucket interface.

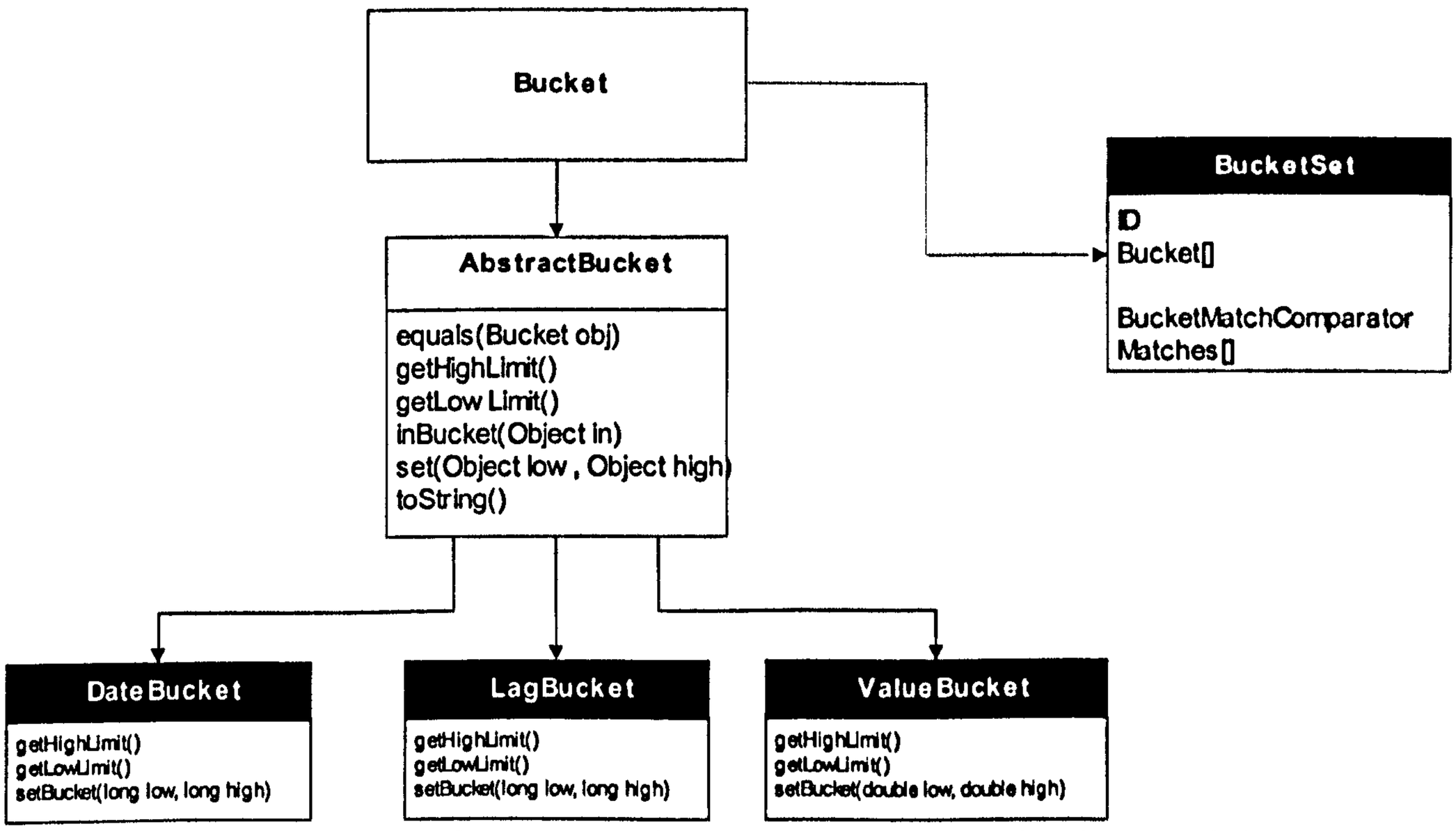


Figure 99 The EMS uses three types of bucket; Date, Lag and Value. Each bucket type implements the Bucket interface which defines core functionality that each bucket must implement. All bucket types can then be compared against the same types automatically and matches logged within the BucketSet.

The *BucketSet* class collates the vector that defines the ranges of a cluster. Each vector dimension (*Bucket*[1, ..., *n*]) holds a bucket, which in turn defines the limits of that dimension of the cluster. Once the limits of the cluster are defined in this way, the *BucketMatchComparator* monitors input data (in the form of vectors matching the bucket set). Each match is recorded so that patterns allocated to each cluster can be analysed further. Over time, bucket sets with a significantly high number of recognised input patterns become verifiable as possible predictors of asthma exacerbation.

Appendix J

Reference Datum Vector Examples

An example of the three types of ordering (discussed in Chapter 5) are shown in their vector form by the Figures (100a to 100c):

a) *Type 1* (Series, one parameter): Consists of a time series that could be taken at regular intervals if leading up to a trigger point, or irregular intervals if representing a *series of points* analysis, for one parameter.

b) *Type 2* (Series, many parameters): Similar to *Type 1*, except that the vector will generally have a greater length in order to incorporate all the parameter data.

c) *Type 3* (Many parameters, snapshot): Used for point analysis involving more than one parameter. This type could be used for a single parameter, in which case the length of the vector would be restricted to this single parameter.

Although only three types of analysis are given here, the use of a software interface within the prototype implementation facilitates the integration of other possible analysis types. Note, that the examples given here simplify the vector as they only include *Values*, rather than the possible inclusion by the prototypes of *Lag* and *Date* values. Although the date of the reading is shown in order to show a series.

It should be noted that the examples given here have been simplified. The lag elements (as mentioned in example *c* below) have been excluded, but would normally be one of the most important data parameters to analyse.

a) Shown in the grey boxes are values for the given data type (SO_2) at a regular time interval and a single patient (WHT008). This is typical of a time series data set.

Time Series one parameter

	25/04/00 09:00 WHT008 SO_2	25/04/00 15:00 WHT008 SO_2	25/04/00 21:00 WHT008 SO_2	26/04/00 03:00 WHT008 SO_2	26/04/00 09:00 WHT008 SO_2	26/04/00 15:00 WHT008 SO_2	26/04/00 21:00 WHT008 SO_2	27/04/00 03:00 WHT008 SO_2	27/04/00 09:00 WHT008 SO_2
$N_{(x)}$	5.65	6.41	6.11	6.41	6.42	6.67	7.37	8.23	6.96

Figure 100a Example of a time series vector. Time Series many parameters

b) Here the time series contains three parameters, SO_2 , PMO_{10} , and CO. Each parameter has a short time series of three values, presented at regular intervals.

	25/04/00 09:00 WHT008 SO_2	25/04/00 15:00 WHT008 SO_2	25/04/00 21:00 WHT008 SO_2	26/04/00 03:00 WHT008 PM_{10}	26/04/00 09:00 WHT008 PM_{10}	26/04/00 15:00 WHT008 PM_{10}	26/04/00 21:00 WHT008 CO	27/04/00 03:00 WHT008 CO	27/04/00 09:00 WHT008 CO
$N_{(x)}$	5.65	6.41	6.11	3.64	4.16	3.23	3.02	3.46	2.96

Figure 100b Example of a time series vector with multiple parameter types.

c) An example of a vector used in analysis for many parameters is shown opposite. The data consists of a snapshot (or a single identified point from each parameter). The date and time values shown are fairly irrelevant to the analysis. The important factor is the time lag between these possible causal points and the identified lung function trigger point. The dates are shown as part of the example as they give an indication that the values are not all taken at the same time.

Many parameters (snapshot)

	25/04/00 09:00 WHT008 CO	25/04/00 08:50 WHT008 NO_2	25/04/00 09:03 WHT008 O_3	25/04/00 09:15 WHT008 SO_2	25/04/00 08:48 WHT008 PM_{10}	25/04/00 08:59 WHT008 $PM_{2.5}$
$N_{(x)}$	0.60	21.0	10.0	2.00	25.0	14.0

Figure 100c Example of a vector with multiple parameter reference datums.

Appendix K

Classification Models

Three types of classification model are: decision-region boundaries, probability density functions, and posterior probabilities. Figure 101 (below) taken from Kennedy *et al.* (1998), shows the three models in graphical form. An explanation of each model is then given.

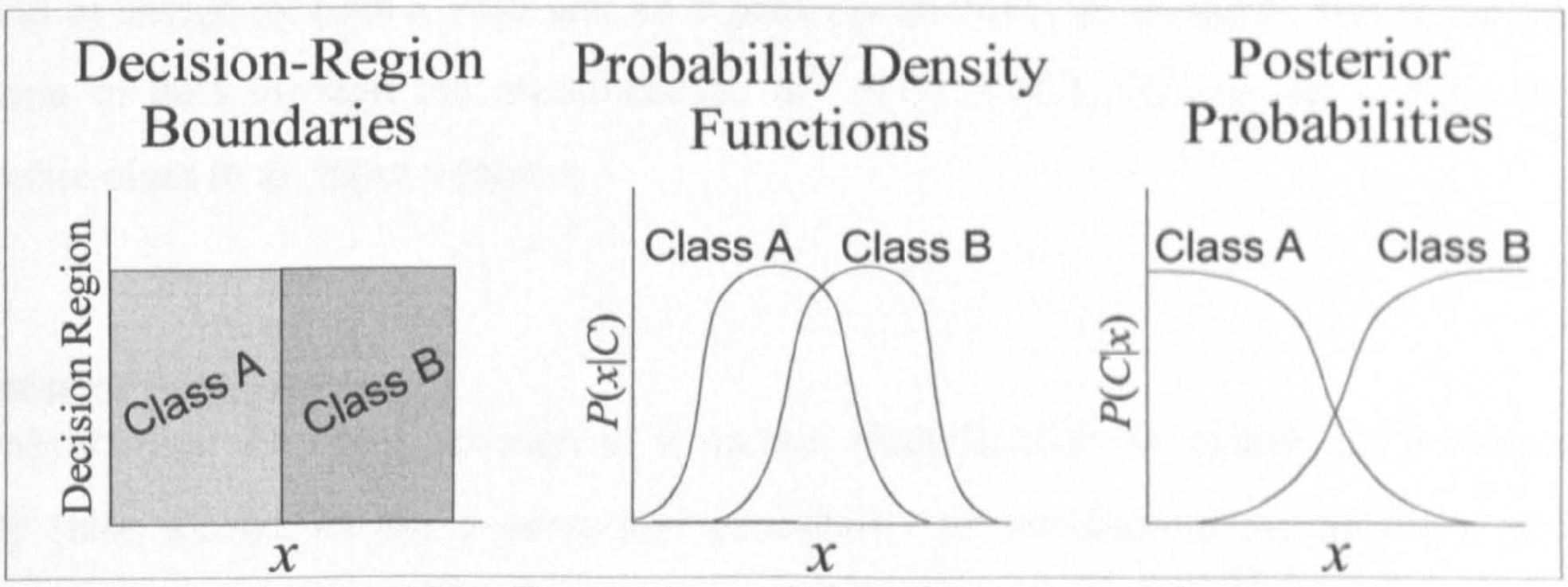


Figure 101 Comparison of classification model types.

K.1. Decision-Region Boundaries

This type of model constructs decision boundaries within the input space and is the simplest way to minimise classification error. This type of model works best when the data is clearly defined into classification type, for example everything below x is Class A, everything above x is Class B.

K.2. Probability Density Functions

The probability density function (PDF) for a normal distribution is given by the expression;

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad \textit{Eq. K.1}$$

where μ is the mean of the distribution, and σ the standard deviation (Witten & Frank, 2000).

PDF is a mapping of the input data to probability density values reflecting the statistical distribution of the data. The PDF for a class C evaluated at a point x in the input space, is denoted by $p(x|C)$.

Prior probabilities must be used alongside PDFs to enable the creation of a classification model. Prior probabilities (also known as *a priori* probabilities or *unconditioned probability*) denoted by $P(C)$ where C represents a class, should be assigned to a class before any analysis of data input for its classification. $P(C)$ is set by determining the number of data samples within a class then dividing by the number of class classifications available. Therefore, if 5 data samples where classified as similar and there existed a total of 20 different samples the probability assigned to the class C would be 5/20 (or 25%).

The method of assigning both a PDF and an *a priori* probability to an input, facilitates the classification of data through the maximisation of $P(C)\ p(x|C)$. This model assigns the most probable class to an input vector x.

K.3. Posterior Probabilities

The probability that an input belongs to a certain classification is called the posterior probability (also known as the *a posteriori* probability or *conditional probability*), and denoted by $P(C|x)$. Luger & Stubblefield (1998) give the following example to demonstrate posterior probability. The posterior probability of a person having a disease d with symptom S is;

$$P(d|S)=\frac{|d \cap S|}{|S|} \qquad \textit{Eq. K.2}$$

where the right side of the equation reads, "the number of people having both the disease d and symptom S divided by the total number of people having symptom S." This is Bayes' theorem. The posterior probability associated with an input x can be thought of as a vector with components $P(C_1|x), \ldots, P(C_m|x)$. This is an estimation problem where the total of the components should sum to 1.

Haykin (1999) suggests that the ratio of two conditional PDFs is the equivalent of a likelihood ratio. The likelihood that a pattern belongs to a certain class. The ratio is also known as *quantity* $\Lambda(x)$.

Appendix L

Modelling Approaches

L.1. Fixed Models

Fixed models are typically described by a set of mathematical equations that define a transformation between the input vector and the model's output. An accurate model requires the complete understanding of the relationships between all the variables in the input vector and the output variables. Fixed models are usually only used when the classification problem is understood. The general mapping for a single variable is defined by;

$$F(X) = X \qquad \textit{Eq. L.1}$$

Fixed models of this type often oversimplify the relationships between variables which can lead to inaccuracies.

L.2. Parametric

Parametric modelling involves two stages. The first stage is similar to fixed modelling except that a set of *free* parameters (parameters that can be given numerical values) are inserted into the fixed model. The process can be described using the following diagram (Kennedy *et al.*, 1997).

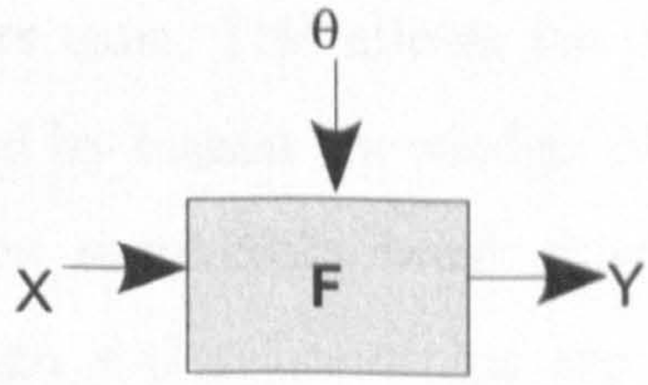


Figure 102 A parametric form

The second stage uses data to select numerical values for the free parameters. The parameters are usually chosen to minimise error on a given data set. An example of a

parametric model using linear regression (good for linear data) is given below;

$$F(X|0) = 0_0 + 0_1x_1 + \ldots + 0_nx_n \qquad \textit{Eq. L.2}$$

Where the input vector is represented by $X = [x_1, ...,x_n]$, and the parameter vector by $0 = [0_0, ...,0_n]$. The parameter vector contains n+1 parameters (using the linear regression equation). With an input space with two or more dimensions the mapping becomes a plane.

Linear regression methods can often lead to poor results when the underlying data is not linear. When the input data is in n>2 dimensions it becomes difficult to determine its characteristics.

Parametric methods make assumptions about the nature of data distribution. They often assume that the data has a Gaussian distribution. The problem is reduced to estimating the parameters of the distributions (means and variances in the case of Gaussians). Parameters are typically selected using the mean of the distribution.

Another form of parametric technique often used is Logistic Regression. This method, rather than using a linear parametric form (as used in linear regression), uses a sigmoid function instead. Sigmoids are more natural for estimating probabilities as they take on values between 0 and 1 with a smooth transition between the two extremes.

L.3. Nonparametric

Nonparametric modelling can be used where little or no prior understanding of relationships between data parameters exist. This allows the potential discovery of new relationships not previously conceived by human knowledge. Nonparametric modelling is particularly useful when visualisation capabilities break down in estimation problems where the dimension of data is high and relationships are complex. Non-parametric methods make no assumptions about the specific distributions involved, and are therefore described, perhaps more accurately, as distribution free (Michie *et al.*, 1994).

As there are always an infinite number of candidate models for a given data set, the problem is in finding the one that will generalise to other data sets the best. It is common to

resort to intuition (or heuristics) to arrive at a solution. Nonparametric methods use a particular set of rules to guide the heuristics: smoothness, complexity, out-of-sample error.

L.4. Smoothness

In classification the meaning of smoothness varies between different types of model. For decision boundaries, smoothness is a measure of how the transition between regions is smooth (or that there are few decision regions by nature). Probability Density Functions (PDFs) refer to smoothing as the nature of the mapping defining the (profile of the) function. Whilst posterior probability refers to smoothing in a similar way to estimation, that smoother mappings restrict output values and is a measure of how much the output of a mapping changes due to small changes in the input.

L.5. Complexity

The larger the number of different models that can be produced, the greater the complexity of the mapping. The more complex a parametric form, the more likely there will be a large number of models. For example (Kennedy *et al.*, 1997) use the following example;

the quadratic, $F(X|0) = 0_0 + 0_1x + 0_2x^2$ *Eq. L.3*

is more complex than the linear form

$F(X|0) = 0_0 + 0_1x$ *Eq. L.4*

as the linear mapping can be generated from the quadratic if the last term (0_2) is reduced to zero.

L.6. Out-of-Sample Error

The objective of nonparametric modelling is to arrive at the mapping that provides the best form of generalisation across a number of similar data sets. This is done via an iterative process controlled by a measurement of error. The error is derived from the estimation of an 'out-of-sample' error, which is found from the analysis of parametric fits through the data set. The out-of-sample error is allocated to each parametric fit after comparison with a test set of data. The parametric fit with the lowest out-of-sample error becomes the chosen fit for the generalised data samples. In cases where the errors are similar the lowest order of parametric model is chosen for simplicity (and greater smoothness).

Appendix M

Nonparametric Methods

Many different types of architecture are commonly used for data classification. Decision boundary methods often use learning vector quantisation (LVQ), K nearest neighbour classifiers, and decision trees. Probability Density Functions (PDFs) commonly use Gaussian mixture methods. Posterior probabilities however are the most flexible and can use any estimation architecture. The most common however are, multilayered perceptron (back propagation neural network), radial basis functions, and the Group Method of Data Handling (GMDH).

The following summary has been derived from work by Kennedy *et al.* (1998).

Table 41 Summary of non-parametric methods

Algorithm	*Parametric*	*Non-parametric*	*Classification*	*Estimation*	*Kernal Function*	*Boundaries (for classification)*	*Mapping (for estimation)*
Nearest cluster		×	×		Euclidean Norm	Piece-wise Linear	Linear combination of weighted Euclidean distances
Linear regression	×		×	×	Line	Hyperplanes	Best fit line
Logistic regression	×		×	×	Sigmoid	Hyperplanes	Weighted sum of inputs passed through a sigmoid nonlinearity
Back-propagation		×	×		Sigmoid	Hyperplanes	Weighted sum of inputs passed through a sigmoid nonlinearity
Gaussian Mixture		×	×		Gaussian	Overlapping radial (receptive) fields	Weighted sum of Gaussian outputs

Algorithm	*Parametric*	*Non-parametric*	*Classification*	*Estimation*	*Kernal Function*	*Boundaries (for classification)*	*Mapping (for estimation)*
Unimodal Gaussian	×		×		Gaussian	Overlapping radial (receptive) fields	Weighted sum of Gaussian outputs
Radial basis function		×	×	×	Gaussian	Overlapping radial (receptive) fields	Weighted sum of Gaussian outputs
K nearest neighbour		×	×	×	Euclidean Norm	Piece-wise linear	Linear combination of weighted Euclidean distances
K means		×	×	×	Euclidean Norm	Piece-wise linear	Linear combination of weighted Euclidean distances
Projection pursuit		×	×	×	Line	Hyperplanes	Linear inner product
Estimate-Maximise clustering		×	×	×	Euclidean Norm	Piece-wise linear	Linear combination of weighted Euclidean distances
MARS		×	×	×	Polynomial decision tree	Piece-wise polynomial	Piece-wise polynomial
GMDH		×	×	×	Polynomial	Piece-wise polynomial	Nonlinear spline
Parzen's window		×	×	×	Gaussian	Overlapping radial (receptive) fields	Weighted sum of Gaussian outputs
Linear decision tree		×	×	×		Hyperplanes parallel to input axis	N/A
Binary decision tree		×	×		Decision tree	Hyperplanes parallel to input axis	N/A
Hypersphere		×	×		Hypershpere	Overlapping hyperspheres	N/A
Learning vector quantisation		×	×				

Appendix N

The Self Organising Map

N.1. Competitive Process

The process which the self-organising map (SOM) uses is as follows;

An input pattern (or vector) taken from the input space (initial data set) takes the form;

$$\mathbf{x} = [x_1, x_2, x_3, \ldots x_m] \qquad \textit{Eq. N.1}$$

where m denotes the dimension of the input space.

The weight vector for each neuron in the network has the same dimension as the input space. Therefore a weight vector for neuron j would take on the form;

$$\mathbf{w}_j = [w_{j1}, w_{j2}, w_{j3}, \ldots w_{jm}] \qquad j = 1, 2, 3 \ldots, l \qquad \textit{Eq. N.2}$$

where l is the total number of neurons in the network.

To find the best matched neuron for the input vector the weight vector of the neurons can be used. One method is to compare all the inner products ($\mathbf{w}_j \mathbf{x}$) of the neurons and select the largest. The second method is to find the weight vector with the minimum Euclidean distance to the input vector.

The winning neuron, or the neuron that best matches the input is identified by $i(x)$ the index of the neuron which is chosen by applying the following condition

$$i(x) = \arg\min_j \|x - w_j\|, \qquad j = 1, 2, \ldots, l \qquad \textit{Eq. N.3}$$

The competitive process seeks to find the index $i(x)$ of the winning neuron in preparation for interaction with neighbouring neurons. The neighbourhood is defined topologically, that is to say that a neighbouring neuron is not influenced by the stretching or bending of the output space (or mapping). The interaction follows the methodology thought to be achieved in the human brain

where the firing of a neuron tends also to excite those that are in the immediate neighbourhood.

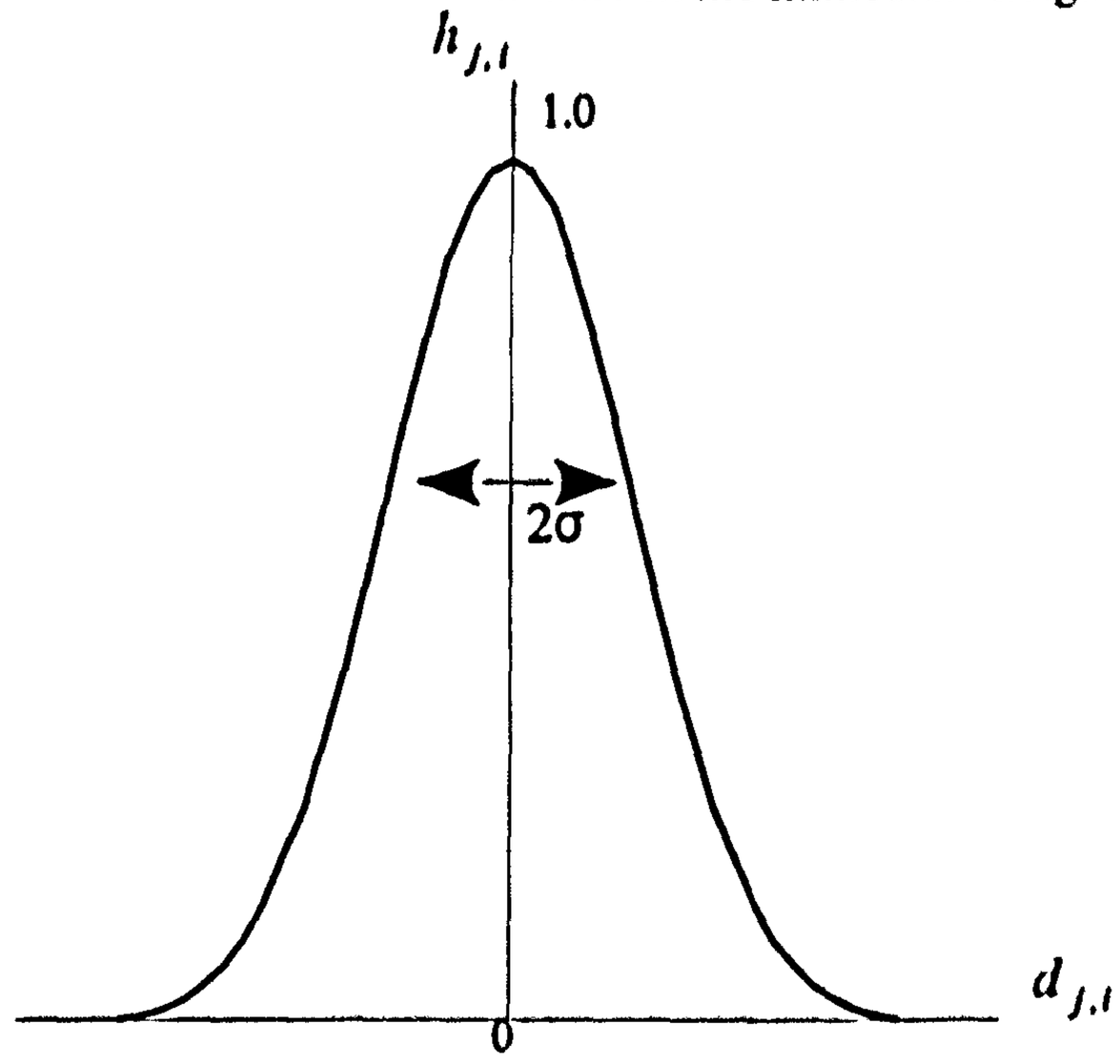


Figure 103 A Gaussian neighbourhood function.

Two requirements to enable the firing of local neurons are:

1. The topological neighbourhood $(h_{j,i})$ should be symmetrical about the winning neuron.
2. The extent to which neurons further away from the winning neuron should be excited should decrease monotonically with increasing distance, decaying to $d_{j,i} \rightarrow \infty$. This is a necessary condition for convergence.

A typical choice of neighbourhood $(h_{j,i})$ function that satisfies the two requirements is the Gaussian function, shown in Figure 103 above and given by;

$$h_{j,i(x)} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right) \qquad \textit{Eq. N.4}$$

So that neighbouring neurons cooperate, it is necessary that the distance $d_{j,i}$ between winning neuron i and excited neuron j is defined by the topological output space rather than with a distance measure defined by the original input space. The neighbourhood function $h_{j,i(x)}$ is also known as the *Probability Density Function* (Haykin, 1999).

Appendix O

Distance Formulae

Table 42 Options for the choice of distance formulae.

Distance Measure	*Formula*	*Accuracy*	*Comments*
Pythagorean Theorem	$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$	Varies by earths latitude but between ±9 and 30 meters.	Good if distance between locations <20km. *Note 1.*
Haversine Formula	$d=2\mathrm{R}\cdot atan2(\sqrt{a},\sqrt{1-a})$	≈±1m per 10km of distance.	Accurate over all distances. *Note 2.*
Polar Co-ordinate, Flat-Earth Formula	$d=R\cdot\sqrt{a^2+b^2-2\cdot a\cdot b\cdot\cos(lon2-lon1)}$	≈±1m per 1km of distance.	Computationally more expensive than Pythagorean theorem but more accurate. *Note 3.*
Law of Cosines for Spherical Trigonometry	$d=R\cdot\arccos(a+b)$	A large number of significant numbers are required during calculation to cope with the $\cos^{-1}$.	Not good for small distances. *Note 4.*

Note 1. The use of Pythagorean theorem will also incur a computational cost of transforming the co-ordinates from the spherical measurement of latitude and longitude to Cartesian co-ordinates.

Note 2. The Haversine Formula (Sinnott, 1984) where,

$$a=\sin((lat2-lat1)/2)^2+\cos(lat1)\cdot\cos(lat2)\cdot(\sin(lon2-lon1/2))^2$$

The arguments of trigonometric functions are expressed in radians. Longitude and Latitude measurements need to be converted from decimal (degrees, minutes and seconds) to radians (times degrees by $\pi/180$).

Note 3. The *a* and *b* for the Polar Co-ordinate formula are;

$$a = \pi/2 - \mathrm{lat1}$$

$$b = \pi/2 - \mathrm{lat2}$$

Note 4. The *a* and *b* for Law of Cosines for Spherical Trigonometry are;

$$a=\sin(lat1)\cdot\sin(lat2)$$

$$b=\cos(lat1)\cdot\cos(lat2)\cdot\cos(lon2-lon1)$$

Appendix P

Third Party Java API Used During Prototype Development

The architecture of the system has been designed as generically as possible to allow flexible implementation. Particular mention should be made of the selection of Java API that have been used during the construction of the prototypes. The Java APIs that have been used are:

Table 43 Java API used in the development of the EMS.

Java API	*Description*	*Use/Notes*
J2SE (Jdk2.2.1)	The Java 2 Platform, Standard Edition is at the core of Java technology.	Defines the manner by which an applet or application can make requests to and use the functionality available in the compiled Java class libraries.
JAXB - Java Architecture for XML Binding	Automates the mapping between XML documents and Java objects.	Used during preliminary database construction as a means to transfer data from XML documents.
PSEpro for Java API	An object database.	This was used during the preliminary stages of the project, before the JDO specification became available. Developed by ObjectStore.
Java Data Objects (JDO)	A standard interface-based model for data persistence. JDO technology can be used to directly store Java domain instances into the persistent store (database).	Uses XML descriptors to enhance the persistent classes for storage. Applications written with the JDO API are portable, and independent of the underlying database.
FastObjects (implementation of the JDO specification)	A true object database.	FastObjects' power comes from the ability to represent complex object networks in the database on a one-to-one basis.
Monarch Charts	A collection of pure Java charting components that allows customised, cross platform, data visualization applications.	Assists the presentation of data using graphical means.
Jama (jmat)	A Java Matrix API	Used in multidimensional scaling techniques which were experimented with as part of the project for the display of information.
JINI	Network technology created by Sun Microsystems.	For building distributed systems that are highly adaptive to change. Used to create systems that are scalable, evolvable, and flexible.
JExpress.	A tool that was used to verify that some of the concepts were feasible.	Written in Java and is capable of analysing data sets with Principle Component Analysis, Hierarchical Clustering, K-means cluster, SOMs and visualisation techniques such as Sammons Mapping. Freeware, developed by Molmine (www.molmine.com).

Appendix Q

Database index descriptor XML file

An example of a database index descriptor file is given below. Indexes are specific to the particular database implementation used, and as such should follow the specification given with each database. The file given below is for an object database implemented with FastObjects.

```
<field name="time">
    <collection element-type="mdx.ems.odb.Time" />
</field>
<field name="location">
    <collection element-type="mdx.ems.odb.Location" />
</field>
<field name="data">
    <collection element-type="mdx.ems.odb.Data" />
</field>
<field name="provider">
    <collection element-type="mdx.ems.odb.Provider" />
</field>
<field name="owner">
    <collection element-type="mdx.ems.odb.Owner" />
</field>

<extension vendor-name="FastObjects" key="index" value="dataIndex">
    <extension vendor-name="FastObjects" key="member" value="data">
                    <extension    vendor-name="FastObjects"    key="lexicalOrder"
value="true"/>
    </extension>
</extension>
<extension vendor-name="FastObjects" key="index" value="ownerIndex">
    <extension vendor-name="FastObjects" key="member" value="owner">
        <extension vendor-name="FastObjects" key="lexicalOrder" value="true"/>
        <extension vendor-name="FastObjects" key="significance" value="3"/>
    </extension>
</extension>
<extension vendor-name="FastObjects" key="index" value="providerIndex">
    <extension vendor-name="FastObjects" key="member" value="provider">
        <extension vendor-name="FastObjects" key="lexicalOrder" value="true"/>
    </extension>
</extension>
<extension vendor-name="FastObjects" key="index" value="timeIndex">
    <extension vendor-name="FastObjects" key="member" value="time"/>
</extension>
<extension vendor-name="FastObjects" key="index" value="locationIndex">
    <extension vendor-name="FastObjects" key="member" value="location">
        <extension vendor-name="FastObjects" key="lexicalOrder" value="true"/>
    </extension>
</extension>
<extension vendor-name="FastObjects" key="index" value="absoluteDataIndex">
    <extension vendor-name="FastObjects" key="member" value="time"/>
    <extension vendor-name="FastObjects" key="member" value="location"/>
    <extension vendor-name="FastObjects" key="member" value="data"/>
    <extension vendor-name="FastObjects" key="member" value="owner"/>
    <extension vendor-name="FastObjects" key="member" value="provider"/>
</extension>
```

Appendix R

Manual Data Entry

A graphical interface was created to facilitate the manual entry of data values into the database, and to view contents held within it. The figure below shows the interface and gives an overview of the data parameters held for each piece of location and time dependant data.

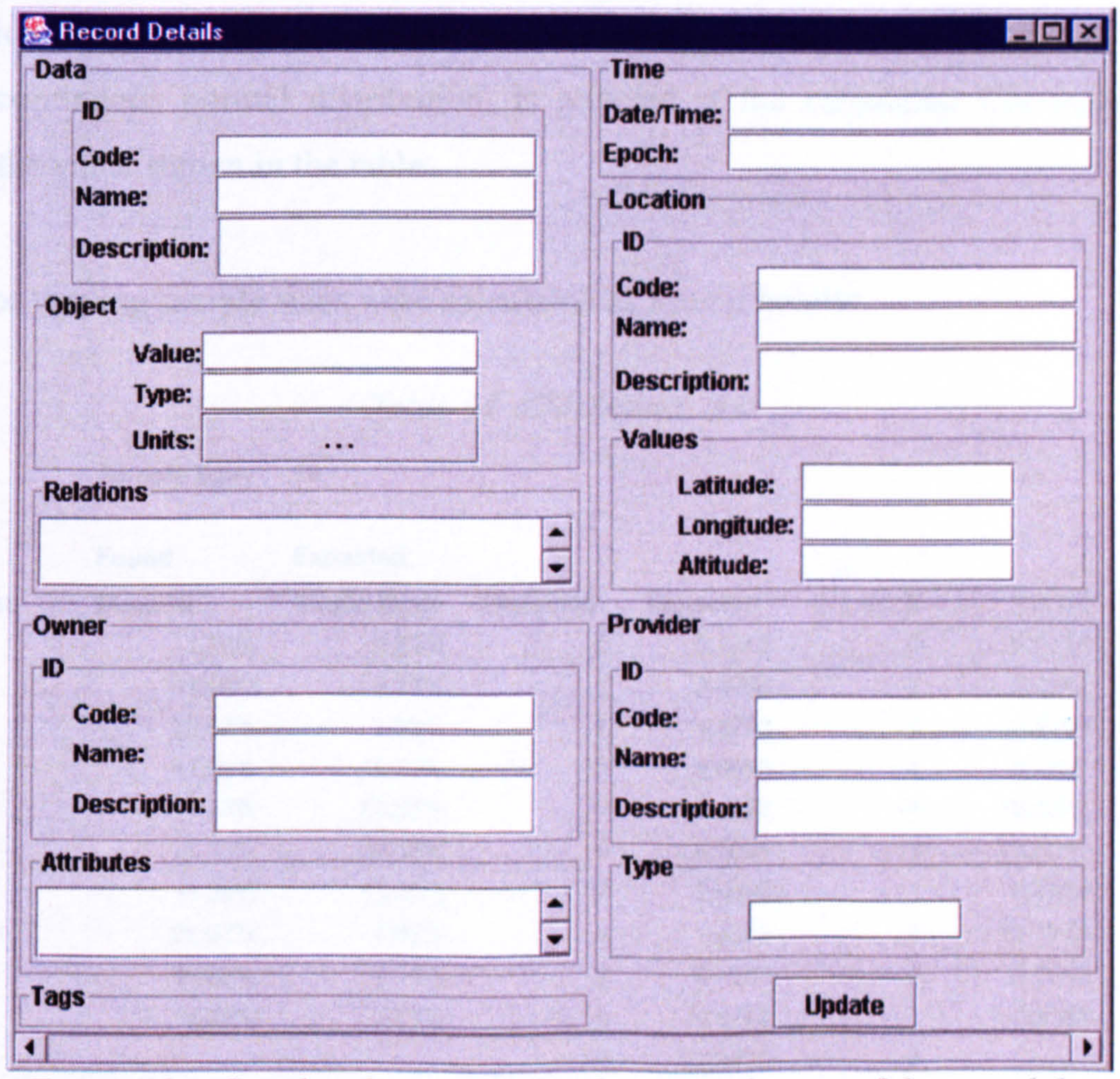


Figure 104 Graphical interface for the manual entry of data, and the viewing of database contents.

Appendix S

Quantity of data to justify splitting clusters

An estimate of the minimum number of *hits* required to statistically justify the splitting of a distribution into separate parts, can be made using the Chi-Square Distribution Test.

In this test the data is examined by bucket, and the number of *hits* found in the sample, compared with the number of *hits* expected. A measurement statistic, based on the formula shown below, is then calculated and compared with values shown in a Chi-Square distribution table. The hypothesis that the data under examination could have been drawn from a continuous normal distribution, is rejected if the calculated Chi-Square value, exceeds the value shown in the table.

Values for varying sample sizes were calculated as shown below:

Table 44 Chi-Square test

Sample Size: ***18***

Bucket	Found Bimodal	Expected Single-Mode	Observed	Expected	O - E	$(O-E)^2$	$(O-E)^2/E$
1	1.73%	0.69%	0	0.1242	0	0.0154	0.1242
2	11.92%	2.77%	2	0.4986	2	2.2542	4.5211
3	22.57%	7.92%	5	1.4256	4	12.7763	8.9621
4	11.92%	15.92%	2	2.8656	-1	0.7493	0.2615
5	1.73%	22.57%	0	4.0626	-4	16.5047	4.0626
6	1.73%	22.57%	0	4.0626	-4	16.5047	4.0626
7	11.92%	15.92%	2	2.8656	-1	0.7493	0.2615
8	22.57%	7.92%	5	1.4256	4	12.7763	8.9621
9	11.92%	2.77%	2	0.4986	2	2.2542	4.5211
10	1.73%	0.69%	0	0.1242	0	0.0154	0.1242
			18	17.9532	0		35.8628

The example above shows how 18 observations have been classified into each *bucket*. For example 2 *hits* have been sorted into bucket 2, and 5 into bucket 3. The number of *hits* in each bucket are then compared with the number of hits that would be expected, if the distribution was in fact a continuous normal distribution. The example shows that 0.1242

observations would be expected in buckets 1 and 10. While 4.0626 observations would be expected in buckets 5 and 6.

The statistic for each bucket is calculated, by squaring, difference between the actual number of *hits* and the expected number of *hits*, and dividing by the expected number of hits; the statistic for each bucket is shown in the last column.

The total statistic, in this case 35.86 is compared with the value in the Chi-Square distribution table (Table 45), to test whether or not the apparent differences between the two sets of data are in fact significantly different.

Table 45 Summary of results for different sizes of sample

		FROM X^2 TABLE						
	Calculated Sample	5%P	2.5%P	1.0% P	0.5% P	0.1% P	.05% P	
Sample Size	X^2	X^2	X^2	X^2	X^2	X^2	X^2	*Note*
10	15.2	16.919						a)
20	35.9	30.144	34.17	36.191				b)
30	50.98	43.773	46.979	50.892				c)
40	64.16	55.758	59.342	63.691	66.766			d)
50	77.99	67.505	71.42	76.154	79.49			e)
80	131.755	101.879	106.629	112.329	116.321	124.839	128.261	f)

Note

a) Sample X^2 below 5% level, not enough evidence to reject the hypothesis that the observed data is not a single distribution

b) X^2 of Sample higher than 5% level, the observed Bi-Modal distribution is different from single distribution, only 5% chance of being wrong

c) Only about 1% chance of being wrong, if distribution is split

d) Only about 1% chance of being wrong, if distribution is split

e) Only about 1% chance of being wrong, if distribution is split

f) At a sample of 80 only 0.05% chance of being wrong in separating into two distributions

To all those who have lost something great:

look forward to the future with hope.

Index

F

G

H

I

J

K

L

T

U

V

W

X