H24/3352

MONASH UNIVERSITY
THESIS ACCEPTED IN SATISFACTION OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
ON..... 6 December 2002 .....
.....[redacted].....
Sec. Research Graduate School Committee

Addenda:

Research questions: (continues page 5)

Artificial neural networks have so far been little used by epidemiologists or bio-statisticians. It is not yet clear which, if any, of the frontiers of epidemiology may be advanced by their use. In a sense, the multi-layer feed-forward neural network trained by back-propagation of errors is still a 'method in search of an application'. The unifying theme of this thesis is to identify areas where 'conventional' statistical approaches have proved less than ideal (and where there are theoretical grounds to think neural networks might be useful), and attempt solutions using these new techniques.

A number of specific research questions are addressed:

1. Can feed-forward neural networks identify predictive relationships between recent weather in metropolitan Melbourne (together with other proxies for the availability of medical care and laboratory investigation) and the total number of requests for human faecal analysis made to medical pathology laboratories? If so, which summary measures of recent weather have the strongest relationship with request numbers for the coming seven days?

2. Can feed-forward neural network models using such summary weather and other measures for metropolitan Melbourne generalise their relationship with future request numbers for a given time period? If they can, are such models also able to make accurate predictions prospectively, beyond the end of the time period used for their creation? Would such predictions be useful to a disease surveillance manager concerned with the prediction and/or early detection of water-related outbreaks of diarrhoeal disease?

3. For a given time period, can feed-forward neural networks find generalised relationships between recent weekly counts of measles cases in Mozambique by province, and the weekly counts in the near future? If they can, are such models also able to make accurate predictions beyond the end of the time period used for their creation? Would such predictions be useful to a disease surveillance manager concerned with the prediction and/or early detection of measles at national or provincial level?

4. Can feed-forward neural networks accurately predict survival to two or five years for newly-diagnosed colon cancer patients aged 65 or more, using individual clinical and other information available at or shortly after the time of diagnosis? Can neural networks make more accurate predictions of survival than models made using the Cox proportional hazards regression technique? To what extent do these two modelling approaches make use of different information about each patient in making their predictions?

5. Can feed-forward neural networks be used to automate the analysis of responses to Job Specific Modules and generate assessments of occupational exposure to benzene among professional drivers? How well do the neural networks' assessments compare with those of an experienced occupational hygienist?

Further Discussion: (continues page 229)

Most of the studies presented in this thesis represent the first attempt to apply artificial neural networks to the problem, and for most there is no clearly established successful statistical technique with which to compare. This discussion focuses instead on some features and limitations of feed-forward neural networks, and their relevance to epidemiological problems and data sets.

One common observation in the studies presented here is that large amounts of training data are required, especially considering that some of the data need to be reserved at the outset for the early-stopping training technique to work. In fact what is needed is a training set that contains representatives of every different scenario or pattern the network will meet in its practical application. In practice such data sets may be impossible to obtain, either because no one has seen the need for such collections in the past, or because the

underlying dynamic 'driving' the time series is itself changing with time. The former situation is remediable with time – in ten years, for example, a much more useful data set will be available for training networks to predict gastrointestinal disease in Melbourne. However, for the problem of a change in the underlying dynamic of the time series one must be more pessimistic. For example, there would be no advantage in having even the past 100 years of measles data for Mozambique because the dynamics of measles over the past twenty years have been unique in Mozambique's history – with war, altered population concentrations and movement, and measles vaccination of varying coverage and distribution. Perhaps a model including such factors would make more accurate predictions, but such information is simply impossible to acquire.

The need for suitable amounts of representative data is most critical where the main goal of the network is to accurately categorise rare events rather than common ones. In such cases the network tends to 'concentrate' on the events which are commonest in the training data, achieving good overall errors at the expense of inaccurate outputs for the rarer events. Nevertheless the fact that networks can 'learn' the patterns in a given data set is encouraging, suggesting that the problem lies in the data rather than the method itself. For African measles, one encouraging possibility would be networks trained to predict provincial measles counts, using information from all the countries in the region. In the post-apartheid era it would be feasible to create a system for all of southern Africa, training networks with inputs from eleven or twelve countries by province. At this level even predictions for whole countries might be useful.

In other applications, such as the assessment of benzene exposure, there are other strategies that can help. In this study the network's classifications most often differ from the hygienist's for patterns that are represented rarely or not at all in the training data. Such networks are still useful if cut-off points are set so that the network has close to 100% predictive power (whether negative or positive). The network can thus be used to rule cases in or out, and only the opposite type need to be looked at more closely. This would translate into substantial savings in time and money for the occupational health researcher.

Another important observation from the studies of time series prediction is that choosing the right scale is vital, in relation to both time unit and geographical area. Unfortunately there is an inevitable trade off between predictive accuracy and the usefulness of the predictions. With data aggregated to the national level, and/or using monthly or quarterly totals, patterns are simpler and accurate prediction is not difficult. However, such models allow only very general predictions to be made. It adds little to what the local surveillance manager already knows to predict that there will be an outbreak somewhere in the country in a given month or quarter. To be useful the network must at least tell which provinces will be most affected. At the other end of the scale, using weekly data for single districts, the 'signal' is swamped by the inevitable random variation, and the network cannot converge on a general solution.

The use of an asymmetric moving average seems promising. Collapsing data down to weekly or monthly totals smooths out much of the day-to-day random variation, but has the major disadvantage of not allowing any analysis for the latest week or month until complete data for that last period are available. The MMWR 'Figure 1' graph suffers from this – the technique is very robust, but it only tells us if the past month exceeded historical limits, and has nothing to say about the last week, or the last few days. Symmetrical moving averages also suffer from these 'end effects'. For example a symmetrical seven-day moving average can only be calculated up to the fourth-last day of the time series. The data for those last three days are the most important for early outbreak detection, but the network can make no use of them. With an asymmetrical moving average the original time series is replaced by one in which every time point's value is actually the mean number of cases for the n time periods ending at that point. This way much of the random variation is smoothed out, but it is still possible to use all the data right up to the latest observations. Using a seven-day asymmetric moving average for daily data essentially removes the effect of weekends as well.

A universal observation from these studies is that careful selection and pre-processing of inputs is very important. If there is no general relationship between the chosen inputs and outputs, the network will either have to 'memorise' the training data, or else not converge on any solution at all. Although there is no mathematical model built into the structure of a neural network, the selection and processing of the inputs by the experimenter inevitably gives the network a strong 'hint' as to the kind of function relating them to the outputs. In the time series models presented in this these, for example, the selection of multiple, highly auto-correlated recent events forces an autoregressive character on to the network model. There really is no such thing as 'model-free modelling' in such practical applications. A network with 100 million training examples

can learn a very difficult pseudo-random number sequence, given only the most recent part of the sequence – but such large sets of training data are almost never seen in real-life epidemiological problems. The network practitioner must always give the network a 'hint' about the likely underlying model, to allow the network to concentrate on the underlying dynamic of the series rather than its minutiae. There were no clear time trends in the data sets used in these studies, but if there had been then de-trending would have been useful for the same reasons.

Two more points that became clear as these studies progressed were the vital importance of a technique to avoid over-training, and the need for a true 'prospective' test of the trained network under conditions as close as possible to those in which it might be used in a practical application. The networks used here were all large and complex, and there was a very real risk of over-training. The early stopping technique used made a very real difference to the performance of the trained networks on previously unseen data. Similarly striking was the difference in the time series studies between the network's performance on a test set spread throughout the training period, and one chosen from beyond the end of the training data. There does not seem to be any analogous process to early stopping in conventional statistical techniques, and the very concept seems to be new to many statisticians. (It is certainly misunderstood by some reviewers for scientific journals).

In considering the cancer prognosis models, it is clear that conventional statistical methods are very strong, and that 'universal' models and staging systems have perhaps gone as far as they ever will. One important lesson from this study, and one of the reasons for using neural networks, is the observation that localised data sets have features allowing more accurate predictions than 'universal' models such as the current bin models. The role of neural networks is to look for non-linear relationships. Where these exist neural networks have a natural advantage. Computers are now ubiquitous and becoming ever faster and more powerful, so it is now possible to contemplate making a new model every week, using the very latest data from a local registry. These models may not apply beyond that registry's geographical catchment area – but is that really a problem, given that the owners of all the other registries can make their own local models in the same way?

Difficulties remain with the interpretation of the weights within the network, and thus the contribution of individual inputs. But perhaps this is a false hope anyway, given that a non-linear relationship might mean that an apparently unimportant input for most cases suddenly becomes the differentiating one for difficult cases. The strength of neural networks lies in their making accurate predictions or classifications – not in assessing the means by which the classifications or predictions are made. There is a philosophical concern that clinicians may not (some would say *should* not) believe in neural networks' predictions because they act as 'black boxes'. But the proof of a pudding is still in the eating, and clinicians are likely to accept models that have been thoroughly tested and are proven in practice to be accurate, rather than worrying too much about exactly how they come to their conclusions.

Another 'philosophical' concern is sometimes raised, in studies attempting to predict disease incidence. If a network model accurately predicted an outbreak and steps were taken to avoid it, would that invalidate future predictions from that model? Would such a model immediately become redundant, the moment it gave its first accurate prediction? Perhaps it would, but there are two reasons to believe this would not undermine the usefulness of the overall approach. Firstly, with modern computers it would be a simple process to re-train any network model every week, once its optimum basic structure has been determined. Secondly, in incorporating new data, it would be possible to include a new input, to indicate whether or not preventive actions had been undertaken. Such networks might eventually become more powerful, not less so.

Artificial neural networks are new to epidemiology. The studies presented here represent early and somewhat tentative steps. In some areas they show that there are seemingly insurmountable difficulties, and neural networks represent an enticing but unproductive dead-end. For most, however, these studies show that neural networks hold considerable promise, and that new data collections (perhaps allowing a different selection of inputs) might lead to useful practical applications.

Applications of Artificial Neural Networks in Epidemiology: Prediction and Classification

James Francis Patrick Black

M.B.,B.S.(Hons), D.T.M.&H., M.Comm.H.

A thesis submitted for the degree of Doctor of Philosophy
Department of Epidemiology and Preventive Medicine
Monash University
Victoria, Australia

February 2002

Table of Contents

Abstract

This thesis explores the potential uses of artificial neural networks for several different types of problems in some of the key areas of modern epidemiological research. Four of the five applications represent the first attempts to use artificial neural networks in this way.

Recent weather as a predictor of gastroenteritis in Melbourne

Neural network models were trained on different types and combinations of inputs related to social events, recent requests for faecal analysis, recent weather, and day of the year. All inputs were smoothed as 7-day asymmetric moving averages. The outputs were similarly smoothed leads of the faecal analysis requests series, and the aim was to predict analysis request numbers up to seven days ahead.

Each new input added to the models' ability to generalise the relationships, evidenced by improving R^2 values for the validation data despite 'early stopping' of network training. Adding the day of year input, representing overall seasonal effects, was only as effective as adding one of the recent weather inputs. Models including all the weather inputs accounted for most of the variation in faecal request numbers up to seven days ahead.

Recent weather is one of the most important predictors of requests for faecal analysis in metropolitan Melbourne up to at least seven days into the future.

Forecasting gastroenteritis in Melbourne

Even cities with state-of-the-art water treatment facilities may still suffer large outbreaks of water-borne gastroenteritis. This study examined the use of artificial neural networks to model time series of requests for faecal analysis in the city of Melbourne, Australia.

Retrospective daily counts of requests for faecal microscopy and culture were obtained from Australia's government health insurance scheme. Weather data, public and school holidays were included to create eight multivariate time series with data from July 1, 1996 to June 30, 2000. The series were smoothed using asymmetric seven-day moving averages, and used to train feedforward neural networks by backpropagation of errors. All networks were trained to predict smoothed request numbers one to seven days ahead. Model generalisation and forecasting ability were tested separately.

In general, the more inputs included the better the model fit on the training and validation data sets. However, the larger the models the less well they predicted request numbers in the prospective 180-day test set. Smaller models using limited weather inputs gave quite accurate predictions on the prospective test sets.

At this city-wide scale artificial neural networks can generalise the relationships between past events and future daily numbers of requests for faecal analysis. The best of the models produced in this study would be very useful in the early detection of water-borne disease outbreaks.

Measles in Mozambique

Measles is a common childhood illness in Mozambique. This study assessed the potential of artificial neural networks in forecasting weekly cases, which would allow more timely and effective control measures.

Models were trained using a ten-year data set of measles reports from the surveillance system. Forecasting ability was tested for two kinds of hold-out test sets: a 15% set from the same time window, and a 15% set from beyond the end of the training window.

The models fit the smoothed training data well. Good generalisation was achieved for the same time window on which the models were trained ($R^2 > 0.9$ up to 8 weeks ahead), but true prospective forecasting was poor.

At an appropriate geographic scale, and with suitable pre-processing, neural networks can accurately relate future measles reports to past reports within the same time window. However, these relationships do not necessarily hold into the future within the same time series. The models created would not be useful in practical applications.

Extracting prognostic information from cancer registry and health care utilisation data

How much prognostic information for survival does a particular cancer registry contain, and what is the best way to extract that information? This study assessed the potential of prognostic models based on a localised database including cancer registry and health care utilisation (rather than more universal but less accurate 'bin models' such as the Tumour-Node-Metastases (TNM) system).

Neural networks were used to test for non-linear relationships, to determine whether the more transparent Cox proportional hazards regression technique extracts all available information from the data set.

The study used data from the U.S. National Cancer Institute's Surveillance Epidemiology and End Results-Medicare (SEER-Medicare) data set on carcinoma of the colon, including all individuals aged 65 or more diagnosed with node-positive colon cancer between 1992 and 1996, in an area encompassing approximately 14% of the U.S. population. All 4463 cases had potentially curative resection of the tumour and were followed either to death or two years; of these, 2615 patients were followed to five years. Separate models were made to predict survival to two and five years. Model inputs were parameters available shortly after diagnosis, including use of adjuvant chemotherapy with 5-fluorouracil (5FU)-containing regimens. Cox models were estimated using maximum likelihood in SAS version 8.2. Model parameters estimated with the training sets were used in the test sets to provide a survival curve for each individual, and predicted probabilities of surviving extracted. Feed-forward neural

networks were trained by back-propagation of errors using NeuroShell2, with one hidden layer and a single output (the network's estimate of survival probability).

The two techniques gave similar results. For survival at two and five years respectively the mean percentage correct was 76.2% and 69.3% for the neural networks and 76.8% and 70.1% for the Cox models. Areas under receiver operating characteristic (ROC) curves gave similar results for the two methods, with areas around 73% and overlapping confidence intervals. Both the neural network and Cox models were well calibrated. There was close agreement between the different types of models (whether correct or not) for individual patients.

The SEER-Medicare database contains considerable prognostic information for survival outcomes. The similar results for the Cox modelling and neural networks suggest that there are no important non-linearities in these data, and the Cox models capture as much prognostic information as exists. The model predictions are well calibrated, so they are of potential use to health facilities in comparing their outcomes. However, incorporation of more clinical and biochemical details in the registry might remove a large proportion of the uncertainty from the prognostic estimates.

Artificial neural networks and Job Specific Modules to assess occupational exposure

Job Specific Modules (JSMs) were used to collect information for expert retrospective exposure assessment in a community based Non-Hodgkins Lymphoma study in New South Wales, Australia. Using exposure assessment by a hygienist, artificial neural networks were developed to predict overall and intermittent benzene exposure from the Driver module. Even with a small data set (189 drivers) neural networks could assess

benzene exposure with an average of 90% accuracy. By appropriate choice of cutoff (decision threshold), the neural networks could reliably reduce the expert's workload by around 60% by identifying negative JSMs. The use of artificial neural networks shows promise in future applications to occupational assessment by job specific modules and expert assessment.

Conclusions

In summary, neural networks are promising for the forecasting of disease surveillance data – provided the underlying forces driving the variation in those data remain reasonably constant over time. They are useful as an adjunct to more conventional statistical tools in exploring the prognostic information in a disease-outcomes data set. They are likely to be useful to expert hygienists in their assessment of occupational exposure to potentially toxic chemicals.

Statement of Authorship

I hereby certify that the material in this thesis has not been submitted for the award of any other degree or diploma in any university or institution. To the best of my knowledge, this thesis contains no material previously published or written by another person, except when due reference is made in the text of the thesis.

James Francis Patrick Black

February 2002

Dedication

This thesis is dedicated to:

My partner, Jacqueline Mansourian, for 24 years of support and encouragement, and for occasionally bringing me back down to earth and reminding me what is really important in life,

and

Our children, Nayri (6) and Shen (4), because, as the Mozambicans say, *"Todo o futuro está na criança"* ("All the future is in the child".)

Acknowledgments

Many people have contributed directly or indirectly to this work, and I am grateful to them all for their assistance. In particular I would like to acknowledge the assistance of:

- My principal supervisors, initially Kit Fairley, and later Flavia Cicuttini, of the Monash University Department of Epidemiology and Preventive Medicine;
- My associate supervisor and neural network 'guru', Kate Smith, of the School of Business Systems, Faculty of Information Technology, Monash University;
- My co-authors on the various papers, both inside the Department of Epidemiology and Preventive Medicine and outside it. I especially thank Dr Vijaya Sundararajan, who created the Cox proportional hazards regression models in Chapter Seven, and Dr Geza Benke, who provided the expert assessments of benzene exposure which were the training data for Chapter Eight;
- Mr Martyn Kirk, initially of the Victorian State Government Department of Human Services, and later of the OzFoodNet, Dr John Carnie and others in the Department of Human Services;
- My Mozambican colleagues, involved in the creation and maintenance of the measles data sets used in Chapter Six, and especially Dr Avertino Barreto, head of the Epidemiology and Endemic Diseases department in the Mozambican Ministry of Health, and Mr José João Matavele, in charge of the disease surveillance databases;
- The Victorian Bureau of Meteorology, who provided the weather data used in Chapters Four and Five;

- The Health Insurance Commission of the Australian Government, who provided the data on daily requests for faecal analysis used in Chapters Four and Five;
- The National Health and Medical Research Council (NHMRC), who provided me with a scholarship;
- The Cooperative Research Centre for Water Quality and Treatment (CRCWQT), who provided my research grant;
- The creators of the linked SEER-Medicare database used in Chapter Seven. I especially acknowledge the efforts of the Applied Research Branch, Division of Cancer Prevention and Population Science, NCI; the Office of Information Services, and the Office of Strategic Planning, HCFA; Information Management Services (IMS), Inc; and the Surveillance, Epidemiology, and End Results (SEER) Program tumour registries in the creation of the SEER-Medicare database;
- My colleagues in the Department of Epidemiology and Preventive Medicine, who provided intellectual stimulation and support. Especially Hugo Stephenson (who first introduced me to artificial neural networks), Alex Padiglione (who started the 'Early detection of outbreaks of water-related gastroenteritis' project in the Department), Martha Sinclair, Margaret Hellard, Brent Robertson, Karen Smith, Pam Lightbody, Richard Hayes, Colin Fee and of course Bebe Loff;
- Last but not least, the staff of the infectious diseases unit at The Alfred Hospital in Melbourne (especially Dr Edwina Wright and Dr Orla Morrisey) for curing me of my bout of facial cellulitis, pneumonia and septicaemia in June 2000, and thus keeping me alive to complete the work!

Publications, Abstracts and Reports

The section of the literature review titled 'Machine-assisted detection of outbreaks: current approaches' is based on my contribution: 'Chapter 6 – Early detection of water related disease outbreaks', (by J Black and CK Fairley[3]), to the forthcoming book *Drinking Water and Infectious Disease: Establishing the Links*. Hunter PR, Waite M, and Ronchi, E (eds) CRC Press (to be published in July 2002). The relevant chapter was in turn based on a presentation at the OECD meeting on water and disease in Basingstoke, England, in July 2000.

Chapter Five is based on a paper called 'Artificial neural networks and gastroenteritis surveillance', by JFP Black[1], KA Smith[2], M Kirk[4] and CK Fairley[3], which has been submitted to the peer-reviewed journal *Epidemiology*.

Preliminary versions of **Chapters 4, 5 and 6** were summarised in a seminar presentation to the disease surveillance section of the World Health Organization at their headquarters in Geneva on July 2, 2001.

Chapter Six is based on a paper called 'Modelling measles in Mozambique with artificial neural networks', by JFP Black[1], F Cicuttini[1], CK Fairley[2] and KA Smith[3], which has been submitted to the peer-reviewed journal *Epidemiology and Infection*. The revised version after reviewers' comments is under consideration.

Chapter Seven is based on a paper called 'Extracting prognostic information from cancer registry and health care utilization data: Artificial neural networks and Cox proportional hazards models', by JFP Black[1], KA Smith[2], AI Neugut[5,6,7], VR Grann[5,6,7], F Cicuttini[1] and V Sundararajan[1], submitted to the peer-reviewed *International Journal of Cancer*.

Chapter Eight is based on a paper called 'Artificial neural networks and Job Specific Modules to assess occupational exposure', by J Black[1], G Benke[1], K Smith[2], and L Fritschi[8], submitted to the peer-reviewed journal *The Annals of Occupational Hygiene*. The MapMovie program presented in **Appendix One** was the subject of my oral presentation at the Master of Applied Epidemiology/Communicable Disease Network of Australia and New Zealand conference in Canberra in April 2001.

Author affiliations

[1] CRC for Water Quality and Treatment at the Department of Epidemiology and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Australia.

[2] School of Business Systems, Faculty of Information Technology, Monash University, Australia.

[3] Department of Public Health, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Australia.

[4] Department of Human Services, Government of Victoria, Australia.

[5] Department of Epidemiology, Joseph L. Mailman School of Public Health, Columbia University, New York, NY, USA.

[6] Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY USA.

[7] Herbert Irving Comprehensive Cancer Center, College of Physicians and Surgeons, Columbia University, New York, NY, USA.

[8] Department of Public Health, University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia.

Awards and Grants

This study was made possible by a research grant from the Cooperative Research Centre for Water Quality and Treatment (CRCWQT).

I was supported during the period of candidature by a National Health and Medical Research Council Scholarship, with further assistance from The Alfred Research Trusts.

List of Tables

List of Figures

List of Abbreviations

5-FU	5-fluorouracil
AJCC	American Joint Committee on Cancer
ANZAC	Australian and New Zealand Army Corps (ANZAC Day is an Australian public holiday)
AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ASCOT	A Severity Characterization Of Trauma - sometimes American College of Surgeons' Committee On Trauma (trauma scoring system)
CDC	Centers for Disease Control and Prevention (USA)
CDSC	Communicable Disease Surveillance Centre (UK)
CI	(95%) Confidence Interval
CRCWQT	Cooperative Research Centre for Water Quality and Treatment
CUSUM	Cumulative Sum (chart)
ENSO	El Niño Southern Oscillation
FINJEM	Finnish job exposure matrix
HIC	(Australian) Health Insurance Commission
JEM	Job exposure matrix
JSM	Job-specific module
MA	Moving Average
MLP	Multilayer perceptron (neural network architecture)
MMWR	Morbidity and Mortality Weekly Report
NASDAQ	National Association of Securities Dealers Automated Quotation (Stock Market)

NCDB	(United States) National Cancer Data Base
NHL	Non-Hodgkins Lymphoma
NHMRC	Australian National Health and Medical Research Council
NSW	New South Wales (Australian State)
PHLS	Public Health Laboratory Service (UK)
PSA	Prostate-Specific Antigen
ROC	Receiver operating characteristic
SEER	Surveillance, Epidemiology and End-Results (United States cancer registry)
TNM	Cancer staging system (Tumour, Lymph Node, Metastasis)
TRISS	Trauma and Injury Severity Score (trauma scoring system)
UICC	International Union Against Cancer
XOR	Exclusive OR (logical operator)

Abbreviations used in describing models in Chapters Four and Five:

C1, C2, C3	1, 2 or 3 faecal cultures done during a single episode of illness
M1, M2, M3	1, 2 or 3 faecal microscopies done during a single episode of illness
SchoolInf	Influence of school holidays (the number of days in a given 7-day moving average that were school holidays)
PublicInf	Influence of public holidays (the number of days in a given 7-day moving average that were public holidays)
SinDay	The sine of the Day-of-Year angle (used with CosDay to present Day-of-Year to a network)
CosDay	The cosine of the Day-of-Year angle (used with SinDay to present Day-of-Year to a network)

MinTemp	Minimum temperature during the 24 hour period
MaxTemp	Maximum temperature during the 24 hour period
Rainfall	Total precipitation during the 24 hour period
WindX	X-axis component of the polar coordinates representing wind speed and direction at 3 p.m.
WindY	Y-axis component of the polar coordinates representing wind speed and direction at 3 p.m.

Chapter One: Introduction

'In intensity of feeling, and not in statistics, lies the power to move the world. But by statistics must the power be guided if it would move the world aright.'

Charles Booth. (1889) *Life and labour of the people in London.* Macmillan, London.
(Quoted in Alderson, M. (1988) *Mortality, morbidity and health statistics.* Macmillan, Basingstoke.)

This thesis investigates potential applications of artificial neural networks for a range of problems spanning some of the major themes in modern epidemiology: the surveillance and control of measles in an impoverished African country, surveillance and control of gastrointestinal disease in a large Australian city, the prognosis of colorectal cancer in the United States, and the investigation of possible relationships between workplace exposure to chemicals and cancer in Australian workers.

Notwithstanding their origins in neurobiology and their links to artificial intelligence, it is in their guise as statistical tools that artificial neural networks are of greatest potential interest to epidemiologists. Artificial neural networks are a new but rapidly maturing technology, and the application of neural network analysis to epidemiological problems is still in its infancy.

The aims and contribution of this thesis

Chapter Two provides an overview of the origins of artificial neural networks and the mathematical principles underpinning their creation and assessment. It continues with a brief review of existing applications outside the medical field, then a more detailed consideration of the ways they have been used in the fields outlined above (plus some of the statistical alternatives in current use).

Chapter Three reviews current approaches to the machine-assisted detection of outbreaks in disease surveillance data sets.

Chapter Four considers what inputs are needed to forecast gastroenteritis numbers on a day-to-day basis in metropolitan Melbourne (making use of proxy indicators – the numbers of various kinds of requests for faecal analysis). Each of the weather inputs – daily rainfall, maximum and minimum daily temperature, and the speed and direction of the wind at 3 p.m. – is shown to have a separate and identifiable influence on the numbers of faecal analysis requests up to seven days ahead. Appropriate pre-processing of the data and useful geographical and temporal scales are determined. In the process questions are raised about the practical difference between true seasonal influences on gastrointestinal disease and those more correctly attributable to recent weather patterns, whatever the season. This is the first application of neural networks to this problem.

Chapter Five uses the identified inputs (data about social events, recent analysis requests, and the weather in the preceding fifteen days), in an attempt to forecast requests for faecal analysis in Melbourne up to seven days ahead. Although models

using all the identified weather inputs do not give good predictions on a prospective test set, smaller models are created that are accurate enough on the prospective tests to be useful in a practical application. This is the first application of neural networks to this problem.

Chapter Six considers the application of neural networks to the machine-assisted detection and eight-week-ahead forecasting of outbreaks of measles from routine surveillance data from Mozambique. Appropriate pre-processing of the data and suitable geographic and temporal scales are determined. Although models are made that accurately relate recent measles case numbers to future cases in a given time window, the same models do not give accurate predictions when applied to a prospective test set. This is the first application of neural networks to this problem.

Chapter Seven turns to the question of two-year and five-year survival after diagnosis and treatment of colorectal cancer, using data from a linked cancer registry and health care utilisation database from the United States. The study asks whether accurate prognostic models could be made using such localised data sets, to supplement the existing 'bin model' approaches, which are more universally applicable but less accurate. Artificial neural networks are compared with the Cox proportional hazards regression approach, to test whether the more 'transparent' Cox method captures all the prognostic information in the data set. The study concludes firstly that there is substantial prognostic information in the SEER-Medicare data set for colon cancer, and secondly that there are no important non-linearities favouring the neural network model. Although neural networks have been used in this way before, this is their first application to colon cancer and the SEER-Medicare data set.

Chapter Eight moves from cancer prognosis to cancer causation. It considers ways that neural networks might be used to make the assignment of chemical exposure in large case-control studies more consistent and less time-consuming. The neural networks are used to model the process by which an expert assessor used a standardised job-history questionnaire to assess the likelihood of drivers' occupational exposure to benzene in a study of Non-Hodgkins Lymphoma. Overall (and despite a relatively small sample) the network models are about 90% accurate. Suitable cut-off points (decision thresholds) are found that would allow the networks to confidently identify about 60% of the questionnaires as negative (and thus potentially not requiring assessment by the expert). This is the first application of neural networks to occupational exposure assessment.

Chapter Nine summarises the main findings of this research, and the future research directions suggested by the results. It summarises the implications of this research for epidemiologists.

Appendix One presents two new computer tools, MapMovie and StepGraph, which assist the analysis and presentation of disease surveillance data and neural network time series predictions. They are both included, together with data from Chapters Four, Five and Six, on a CD inside the back cover of this thesis.

Novel statistical methods and epidemiology

Charles Booth was a social investigator, commentator and activist in Victorian England. In 1893 he served on the Royal Commission on the Aged Poor, and he was the instigator and a major contributor to the inquiry into poverty in London which ran from 1886 to 1903 (London School of Economics & Political Science, 2002). He would have recognised most of the above themes from his own work in the homes, streets and factories of 19th century London. One hundred years later there are some, as there were then, who prefer to divorce science from its social and political implications (e.g. Rothman et al., 1998). Some of us, like Booth in his day, beg to differ: epidemiology's major role – perhaps indeed its principal *raison d'être* – is in influencing public policy on health (Loff and Black, 1998).

On one point we all agree: epidemiology must be as rigorous a science as possible, and must investigate every potentially relevant new methodological development in order to maintain its rigour and its relevance. Artificial neural networks represent an encouraging new technology that deserves a place in the epidemiologist's analytical toolkit.

Chapter Two: Artificial Neural Networks

I have also tried to keep my language as simple as possible. Not only because Zen teaches and advocates the greatest economy of expression, but because I have found that what I cannot say quite simply and without recourse to mystic jargon has not become sufficiently clear and concrete even to myself.

Eugen Herrigel (1953) *Zen in the art of archery.* Penguin, London.

A brief history and mathematical basis

The initial development of artificial neural networks was a fruitful interaction between computer science and neurophysiology, as developments in each field provided the other with useful insights. However, as artificial neural networks have become useful statistical tools in their own right, they have paradoxically become less realistic as models of neural function.

Neurophysiology

It has long been known that the nervous systems of humans and animals are composed of individual functional units, corresponding to the anatomical units called nerve cells (or *neurons*). A standard physiology textbook description (Guyton and Hall, 1996, Berne and Levy, 1998) goes something like the following: Each neuron has a cell body, or *soma*, plus one major projection of cytoplasm and membrane, the *axon*, and many smaller membrane projections called *dendrites*. The axons connect each individual neuron to one or more others via *synapses*, connecting either directly

to the cell body or to its dendrites. A synapse is not a direct connection, but there is a small physical gap across which the signal passes in only one direction by the release of chemical substances called *neurotransmitters*. Some synapses are excitatory, and others inhibitory. Stimulation of a single synapse does not lead to activity in the receiving neuron's axon, but the activation of multiple synapses is 'summed' by the cell body. If the summed activation exceeds a certain threshold value the neuron 'fires': a self-propagating change in electrical potential begins in the cell body and rapidly spreads along the axon. The firing of one neuron leads to the stimulation of synapses with one or more other neurons, and so on. The human brain contains billions of neurons, receiving inputs from sensory organs such as the eyes, ears and skin, and providing outputs to the muscles and other organs. Neurons are not connected randomly, but rather in organised layers and networks.

These advances in neurophysiology explained some of the simplest observed neurological phenomena, such as the persistence of deep tendon reflexes even when the spinal cord is disrupted. But they do not by themselves explain the higher neural functions such as perception, thought, consciousness, emotion or learning. Neuropsychologists sought ways of explaining how the known physiology of nerve cells could be used to explain higher neural function, while including observations such as the persistence of memory (or its global rather than localised degradation) despite localised brain damage. As Hopfield (Hopfield, 1999) explains, the study of neural networks is also aimed at discovering how to solve problems which the human brain solves rapidly and effortlessly but which are very hard on present digital machines.

The McCulloch-Pitts neural net

The first breakthrough came when McCulloch and Pitts (McCulloch and Pitts, 1943), showed that the all-or-none (binary, 'on or off') nature of neuronal activity could be understood in terms of Boolean logic, and that suitably connected networks of neurons could solve logical or mathematical problems. They were partly influenced by Alan Turing's earlier work on programmable computers, while he had used the brain as an analogy in developing his own ideas, which influenced modern computer design and formed the basis of the field of artificial intelligence. This research stimulated workers in both neuropsychology, striving to understand how the brain functions in health and disease, and in the newly-created field of artificial intelligence, striving to replicate human perception and thought in machines.

The McCulloch-Pitts conception of the neural network allowed only for simple inhibitory and excitatory influences of equal magnitude, and all-or-nothing firing of neurons when their thresholds were exceeded. They were essentially describing fixed networks of permanently connected neurons, with each net performing some specific logical calculation. Learning was not addressed by their theory.

The perceptron

As Smith points out (Smith, 1999a), the next advance was Hebb's (Hebb, 1949) suggestion that learning might occur by the alteration of weights given by the cell body to the various inputs it received from the previous layer of neurons. The cell body would thus be calculating a *weighted* sum of its inputs. This stimulated the creation of the first 'neurocomputer' in 1954, and in 1956 the field of artificial neural networks was officially launched at the Dartmouth Summer Research Project. Shortly

after, Rosenblatt (Rosenblatt, 1958) gave the field another boost by proposing the *perceptron*. This was a network based on the McCulloch-Pitts model, but where the weights could take any value (not just 1 or –1), and where the values for the weights are determined by comparison with some real-world training data and the errors used to change the weights. The perceptron had the ability to 'learn from experience', and generated a good deal of further interest and experimentation. In particular this was a boost to the 'connectionist' school, who proposed that memory might be stored in the brain by altering the strength of connections between neurons (and not by directly mapping to neurons in the way a photographic negative or computer memory does).

With the perceptron the basic concept of the artificial neural network trained by 'supervised learning' had taken shape. The general construction of a single artificial neuron is illustrated in Figure 2-1. Each neuron performs four essential functions: it receives inputs from the previous layer and calculates their sum, weighted according to its own unique set of weights. It then assesses its own level of 'activation', and finally passes on its activation to the neurons of the next layer.

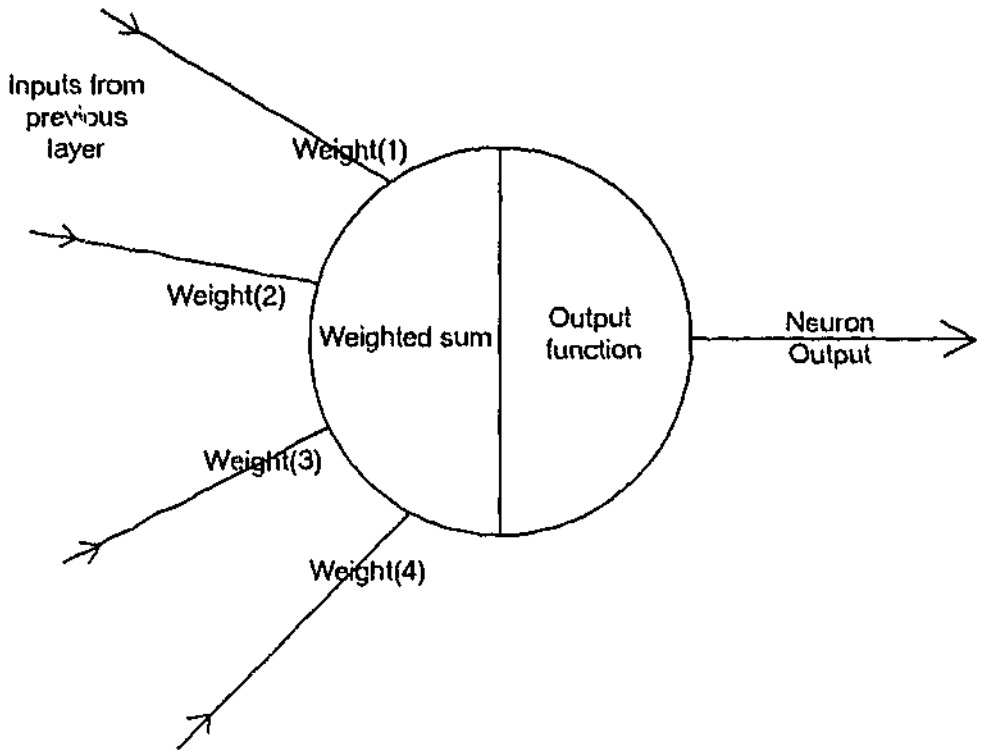


Figure 2-1: Graphical representation of a single artificial neuron.

As shown in Figure 2-2, the values for each neuron's weights are determined from the 'training' data set.

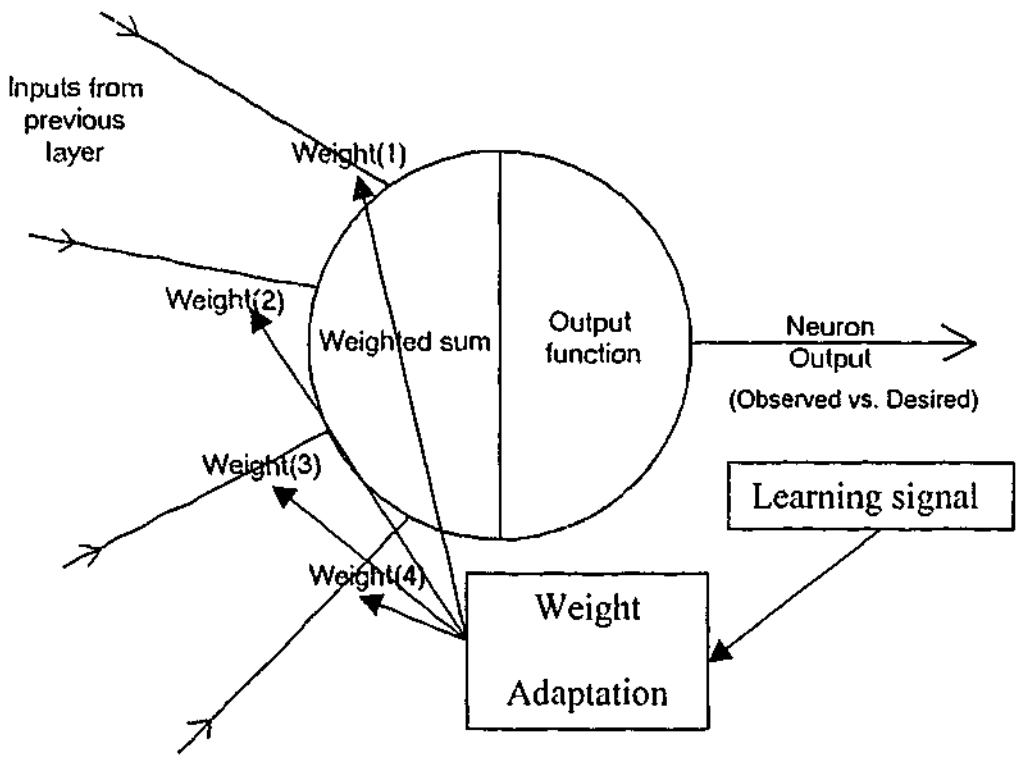


Figure 2-2: Adaptation of connection weights via a learning signal and weight adaptation rule.

Starting from different small random values for each weight, the output of the neuron is calculated and compared with the observed ('correct') value for each set of inputs in the training data. Through repeated iterations, the neuron connection weights are adjusted until the output of each neuron is close to the observed value. The perceptron can then be presented with similar but previously unseen patterns and will classify them correctly.

As Taylor (Taylor, 1997) notes, however, much of the early enthusiasm was dampened by Minsky and Papert's illustration in 1969 that the basic perceptron model is severely limited (Minsky and Papert, 1969). Most significantly, they gave a rigorous mathematical proof that Rosenblatt's perceptrons cannot solve problems (even very simple ones such as the Exclusive-OR logical problem, or XOR), where the output categories are not linearly separable. To solve such problems it is necessary to add at least one 'hidden' layer of neurons to the network. These are called 'hidden' because their outputs are passed to the final layer of neurons, and it is not possible to alter their connection weights directly from a consideration of the training data itself. Figure 2-3 shows the general structure of a fully-connected feedforward neural network with one hidden layer. The circles represent individual neurons, and the solid lines represent the connections, each with a single weight associated.

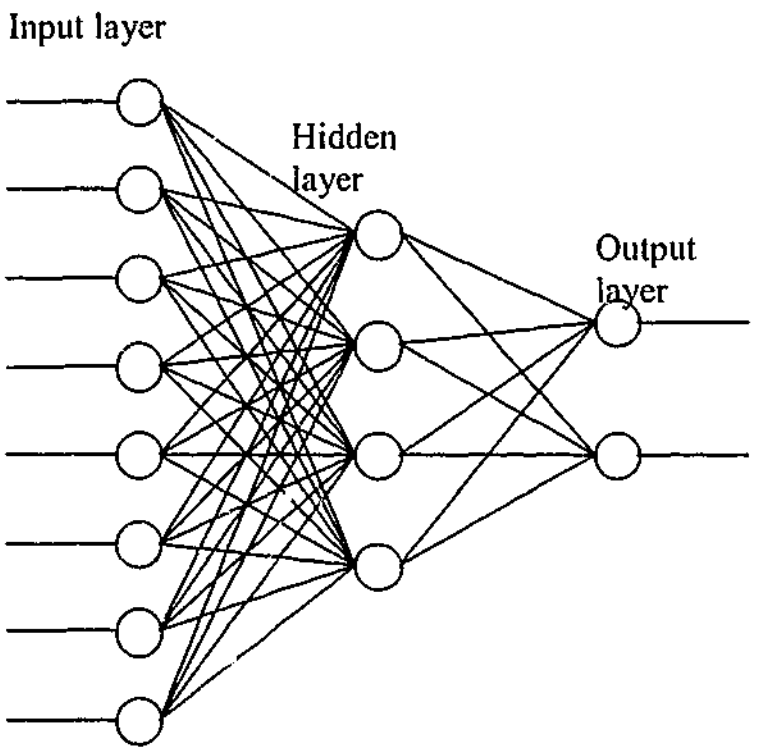


Figure 2-3: the general structure of a fully-connected feedforward neural network.

Backpropagation of errors

This requirement for hidden neurons, and the associated difficulty in adjusting their connection weights, led to a decade of stagnation in neural network research. It was only with the development of the algorithm for 'backpropagation of errors' (commonly just 'backpropagation' or even 'backprop') that artificial neural networks became truly useful for real-life problems.

The standard backpropagation algorithm was discovered and popularised by Rumelhart in 1986 (Rumelhart, 1986), although it had actually first been proposed by Paul Werbos in his unpublished 1974 Ph.D thesis, and then independently discovered by both Le Cun (Le Cun, 1985) and Parker (Parker, 1985) in 1985.

Smith (Smith, 1999a) gives a succinct description of the standard in-line backpropagation algorithm, as follows:

(x is the input pattern, w_{ji} the weight connecting the i^{th} neuron in the input layer to the j^{th} neuron in the hidden layer, v_{kj} the weight connecting the j^{th} neuron in the hidden layer to the k^{th} output neuron, d_k is the desired or target output for the k^{th} output neuron, λ is the gain, and c the learning rate.)

Step 1: Present an N dimensional input pattern x to the input layer, plus an extra input of -1 to allow for a threshold.

Step 2: Calculate the net inputs to the hidden layer neurons

$$net_j^h = \sum_{i=1}^{N+1} w_{ji} x_i$$

Step 3: Calculate the outputs of the hidden layer neurons

$$y_j = \frac{1}{1+e^{-\lambda net_j^h}}$$

Step 4: Calculate the net inputs to the output layer neurons

$$net_k^o = \sum_{j=1}^{J+1} v_{kj} y_j$$

Step 5: Calculate the outputs of the output layer neurons

$$o_k = \frac{1}{1+e^{-\lambda net_k^o}}$$

Step 6: Calculate the learning signals for the output neurons

$$r_k^o = \lambda(d_k - o_k) o_k (1 - o_k)$$

Step 7: Calculate the learning signals for the hidden neurons

$$r_j^h = \lambda(\sum_{k=1}^{K} r_k^o v_{kj}) y_j (1 - y_j)$$

Step 8: Update the weight in the output layer

$$v_{kj}(t+1) = v_{kj}(t) + cr_k^o y_j(t)$$

$$= v_{kj}(t) + c\lambda(d_k - o_k)o_k(1-o_k)y_i(t)$$

Step 9: Update the weight in the hidden layer

$$w_{ji}(t+1) = w_{ji}(t) + cr_j^h x_i(t)$$

$$= w_{ji}(t) + c\lambda(\sum_{k=1}^{K} r_k^o v_{kj})y_j(1-y_j)x_i(t)$$

Step 10: Update the error for this epoch,

$$E \leftarrow E + \sum_{k=1}^{K}(r_k')^2$$

and repeat from Step 1 with the next input pattern. At the end of each epoch, reset E=0, and repeat the entire algorithm until the error E falls below some pre-defined tolerance level (say 0.000001).

This is a description of the most common type of neural network, with every neuron in each layer fully connected to every neuron in the next layer. Steps 1 to 5 illustrate the 'Feed-forward' aspect, as the calculated output of each layer is used as the inputs for the next layer.

Steps 6 and 7 illustrate the use of the 'generalised delta rule', which is the most common way of adapting the connection weights. The learning signals for the output layer neurons are computed using the difference between the neuron's current output and the target output in the training data, and the derivative of the activation function.

For the hidden neurons the error is unknown so the errors of the final layer are propagated backwards (hence the name backprop). The only requirement is that the activation function have an easily-calculated derivative, so a number of other functions, including tanh and various logarithmic and Gaussian functions can be used instead of the more common logistic function illustrated here. (Although all the networks used in this study used logistic activation functions.)

In practical applications the end-point described in Step 10 commonly leads to over-training (see the section below titled "Generalisation, over-training and validation"), and techniques for 'early stopping' are employed to avoid over-training.

The above description includes the 'bias unit' that is included in perceptron networks (e.g. see Steps 1, 2 and 4). This is an extra 'neuron' attached to the input and hidden layers, with its output fixed at −1. This allows each neuron to effectively learn a threshold other than zero.

Different network architectures and training algorithms

The backpropagation algorithm and feedforward network architecture described above implement what is known as 'supervised learning' (or sometimes 'learning with a teacher'). The network is intended to model relationships between known inputs and known outputs. Once it has been trained, the same network can (often) give accurate estimates of unknown outputs given a new set of inputs similar to those in the training set. In this sense artificial neural networks trained by supervised learning are performing regression functions. A neural network with no hidden neurons and a linear activation function is equivalent to a linear regression model – hidden neurons

and non-linear activation functions add the potential for modelling more complex and non-linear data.

There are a number of other neural network architectures, a full description of which is beyond the scope of this review. Some, such as recurrent networks, are variants of the feedforward approach, while others, like Kohonen's self-organising maps, are analogous to cluster analysis and do not use known target outputs. These implement so-called 'unsupervised learning. (See (Masters, 1993) and (Masters, 1995) for a survey of different network types.)

There are also a number of alternatives to standard backpropagation, variously claimed to converge faster, possess simpler training parameters, or be less likely to stick in local minima (see below for an explanation of these 'minima'). Some of them, such as conjugate gradient descent, QuickProp, Rprop (Riedmiller, 1994b, Riedmiller and Braun, 1993), and TurboProp have shown considerable promise (Fischer, 1996, Riedmiller, 1994a), but none has supplanted standard backprop as the most commonly-used algorithm.

Error surfaces

A very useful concept in understanding the way neural networks function is the 'error surface', suggested by Hopfield (Hopfield, 1982). This is a graphical representation of the relationship between the values of the connection weights and the overall error for the network compared with the training data. Every connection weight between two neurons (or between the inputs and the first layer of neurons) affects the overall error, and there is thus at least one optimum value for every weight, at which the difference between the network's outputs and the target values is at a minimum. The common aim of all the different training algorithms is to simultaneously find the optimum value for every weight.

It is easy to imagine a graphical relationship between the value of a single hypothetical weight and the network error (see Figure 2-3). At extreme weight values the error tends to plateau and the network effectively ignores that weight, but there may be a quite convoluted relationship in the active range, with possibly multiple minima, some of which are better than others.

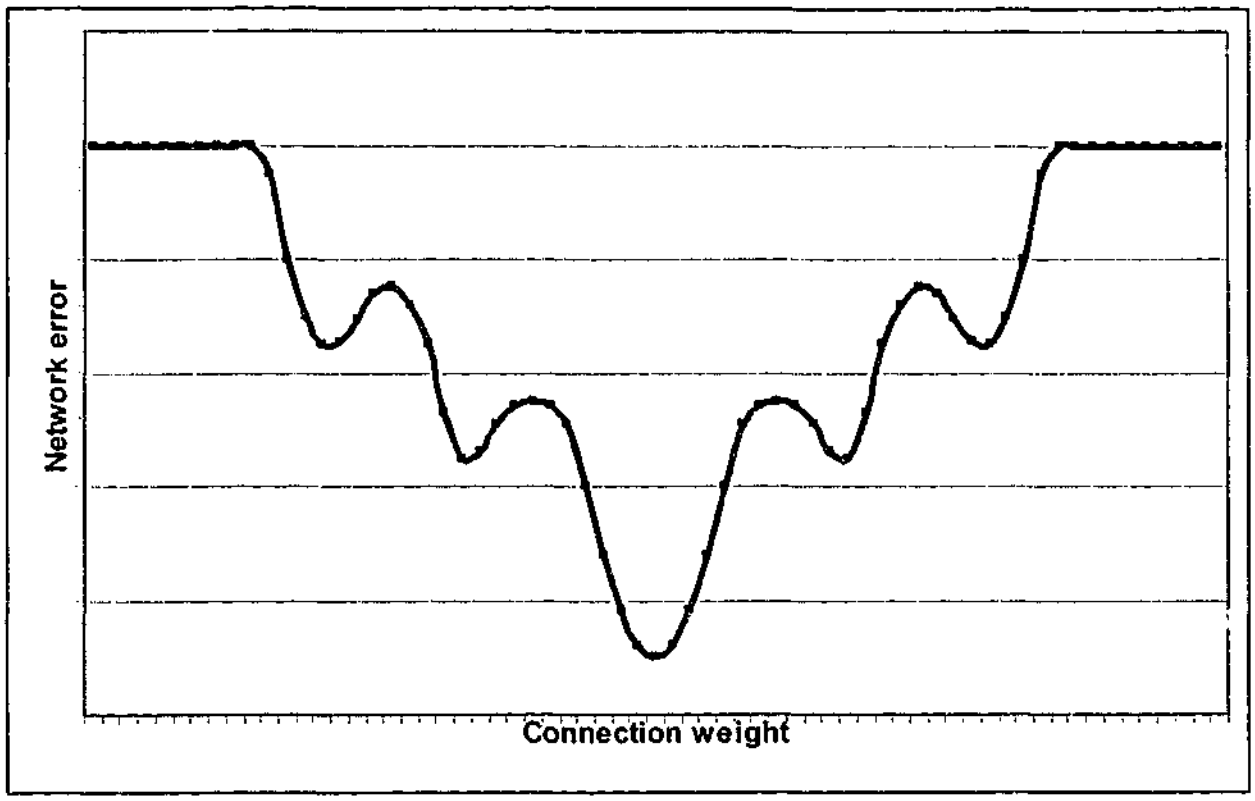


Figure 2-4: Graphical representation of the relationship between a single hypothetical network weight and the overall network error

It is not difficult to imagine a similar graphical representation, this time for a hypothetical network with two weights and thus a three-dimensional error surface (see Figure 2-4) which defines all the possible error values for that network. It is mathematically straightforward to define (but hard for inhabitants of a three-dimensional world to imagine) a multi-dimensional error surface, with one extra dimension added for every one of the network's connection weights. A network constructed for a real-world problem may have many thousands of weights, and a correspondingly complex error surface. The neural network modeller's aim is to find the lowest point on this multi-dimensional surface. A region with plateaus in some dimensions may have deep ravines in others. There may be many local minima separated by steep 'hillocks', representing traps for training algorithms. Finally, there is no guarantee that the best (or 'global' minimum) point will turn out to be very low at all. (Or in other words that there really is a strong relationship between the inputs

and the outputs.) The optimisation process has been likened to sending a skier to find the lowest valley in a mountain range. Or (somewhat whimsically, and inverting the error surface) to parachuting a blindfolded kangaroo over the Himalayas and asking it to find the top of Mount Everest (Sarle, 1994a).

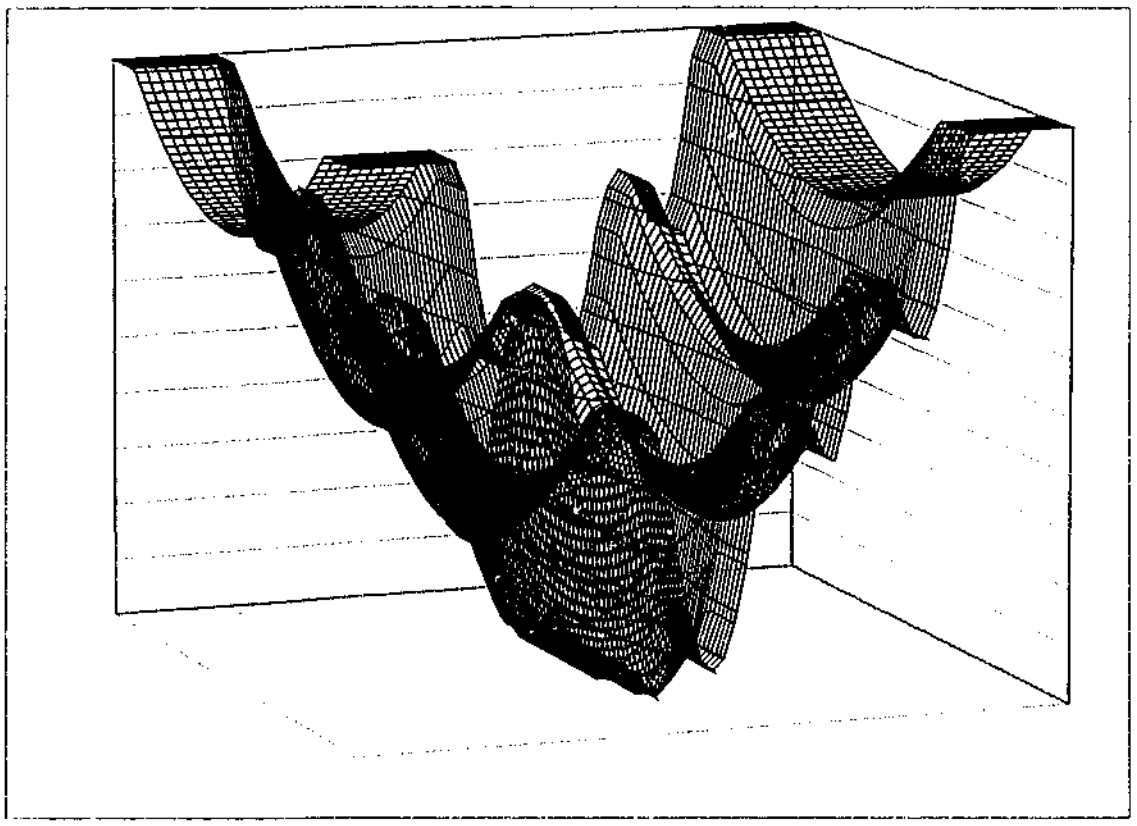


Figure 2-5: Three-dimensional error surface for a hypothetical network with only two connection weights.

Error surfaces can be defined in exactly the same way for conventional statistical algorithms such as linear, logistic or proportional regression. The objective is the same for all – to find the set of parameter values (for neural networks read 'weights', for parametric techniques 'coefficients') that give the minimum possible error. The essential difference between neural network training algorithms and the numerical optimisation methods used for parametric techniques is that neural network algorithms make no assumptions at all about the shape of the error surface. They simply start at a random point on the surface and use hints from the local terrain to

decide which may to move. Neural network training algorithms explore their error surfaces step by step. Parametric techniques assume they know the shape of the error surface in each dimension – commonly assuming it to be quadratic – so that given a known starting place they can leap very quickly to the optimum point. These are excellent and highly efficient algorithms, and are often quite robust against deviations from the starting assumptions (such as normally distributed independent variables without interactions). But where the starting assumptions are seriously violated (i.e. where the relationships are non-linear and complex) parametric techniques cannot be applied or fail. On the other hand a neural network, inching its way around a highly-convoluted error surface may be able to reach a very useful solution.

Using the concept of the error surface, the description of the backpropagation algorithm can now be completed. The derivative of the activation function tells in which direction the weights must move to improve the model fit, but it does not define the optimum size of the step. Too small a step and the network may take far too long inching across the error surface in flat areas, or become stuck in a tiny local minimum half way down the slope. Too large a step and the algorithm may bounce about at the bottom of a good minimum and fail to converge. The 'learning rate' parameter defined by the modeller, together with the size of the error, determines how big a step is taken each time. The learning rate is commonly around 0.1, but there is no way of calculating the optimum for a new problem, and it may need to be adapted for specific problems.

Another way of avoiding small local minima on the error surface is the use of a 'momentum' term. In this case the weights are changed according to a combination of

the new learning signal and the weight change at the last iteration. The momentum term can also be adapted to suit the problem, and is also commonly around 0.1 to begin.

It has been pointed out by many authors (e.g. (Sarle, 1994b, Masters, 1993)) that for the standard 'on-line' learning algorithm described above there is a separate error surface for every different input pattern. A batch-learning modification is advocated by some, in which an average learning signal is calculated only once in every epoch (i.e. once for every presentation to the network of the whole training set). The weights are then updated just once for each epoch. However, for most problems the individual error surfaces tend to have a generally similar topography for the whole problem, so that in practice the on-line algorithm works at least as well, if not better.

The 'curse of dimensionality'

The previous section described the 'embedding' dimensionality of the kinds of problems neural networks are set to solve. Each connection weight between two neurons in the network contributes one more dimension to the 'input space' for the problem, and makes the error surface more complex. For many numerical optimisation algorithms the computational difficulty increases so much with each added dimension in the input space that solutions become impossible.

This notorious 'curse of dimensionality' (Bellman, 1961) has been with us as long as mathematicians have sought computational schemes for optimisation problems. Viewed in these terms, there is not a great deal of subtlety to the backpropagation algorithm, which must compute an adjustment for each weight in the network for

every training pattern many times over. This need for sheer computational power is one of the reasons the development of artificial neural networks has mirrored the development of the personal computer – applying neural networks to real-life problems is only feasible with computers having fast processors and large amounts of memory.

Neural network practitioners sometimes present network training as a form of *tabula rasa* learning, in which the network is able to draw meaning from the data without any internal structure to guide it. Geman and others (Geman et al., 1992) point out that this is unrealistic, and practical networks trade off what they call 'bias' against 'variance'. In their terms 'bias' means assumptions made by the researcher in the design of the network or, more commonly, in the way the problem is presented to the network via the arrangement and preparation of the training data. A network truly free of this 'bias' tends to have high 'variance' (i.e. slow and difficult convergence), and the only way to avoid this is through extremely large training sets. (More correctly, a high ratio of training examples to problem dimensions.) This insight has certainly become obvious to some in the field (Burke, 1994), and strongly influences researchers' choice of input representations and pre-processing.

Even more important to the neural network practitioner is the 'intrinsic' dimensionality of the data set for a particular problem. Although they are writing about different problems, Korn and others (Korn et al., 2001) could be describing typical neural network problems when they point out that it is unrealistic to assume that every input is uniform and independent of all the others. Real data typically have skewed distributions and 'subtle dependencies between attributes'. In practice this

means that the intrinsic dimensionality for the problem – basically the complexity of the patterns represented within it – can be much less than the embedding dimensionality of the network.

Generalisation, over-training and validation

It is in the nature of epidemiology that samples are drawn from larger populations in order to draw inferences about the larger population. If the observations made apply only to the sample, then they are of little use. Most epidemiological studies are intended to be generalisable to some larger (even if at times imaginary) group.

The demonstration (Hornik et al., 1989) that, given enough hidden neurons, standard multilayer feedforward networks are capable of approximating any measurable function to any arbitrary degree of accuracy, has some very important consequences for the 'generalisability' of neural network studies.

First it implies that any failure of a network to converge on a solution must be due to insufficient learning cycles, insufficient hidden neurons, or the lack of a true deterministic relationship between the training inputs and outputs.

A second consequence is that where a deterministic relationship does exist, and the network has enough hidden neurons, the final error level can often be as low as the experimenter wishes (or computer time permits). The goodness of fit of a network model to the original training data tells us very little about the generalisability of the same model to other data drawn from a similar population. Indeed, it is very easy to 'over-train' or 'over fit' a neural network model, so that the model begins to reflect

idiosyncrasies in the training data. Figure 2-5 gives a hypothetical illustration. In this greatly simplified example, the experimental aim is to produce the non-linear function shown by the solid line. This is the general form of the relationship between the variables on the x and y-axes, and it is generally assumed that the network will pass through this stage at some time in its training process. Testing the model at this point with previously unseen data will give the optimum solutions. However, if the neural network is allowed to continue training, it begins to create an even more complex model, depicted by the dashed line. This gives the minimum error on the training set (which is what the backpropagation algorithm is designed to do), but the network's accuracy for previously unseen cases tends to degrade badly.

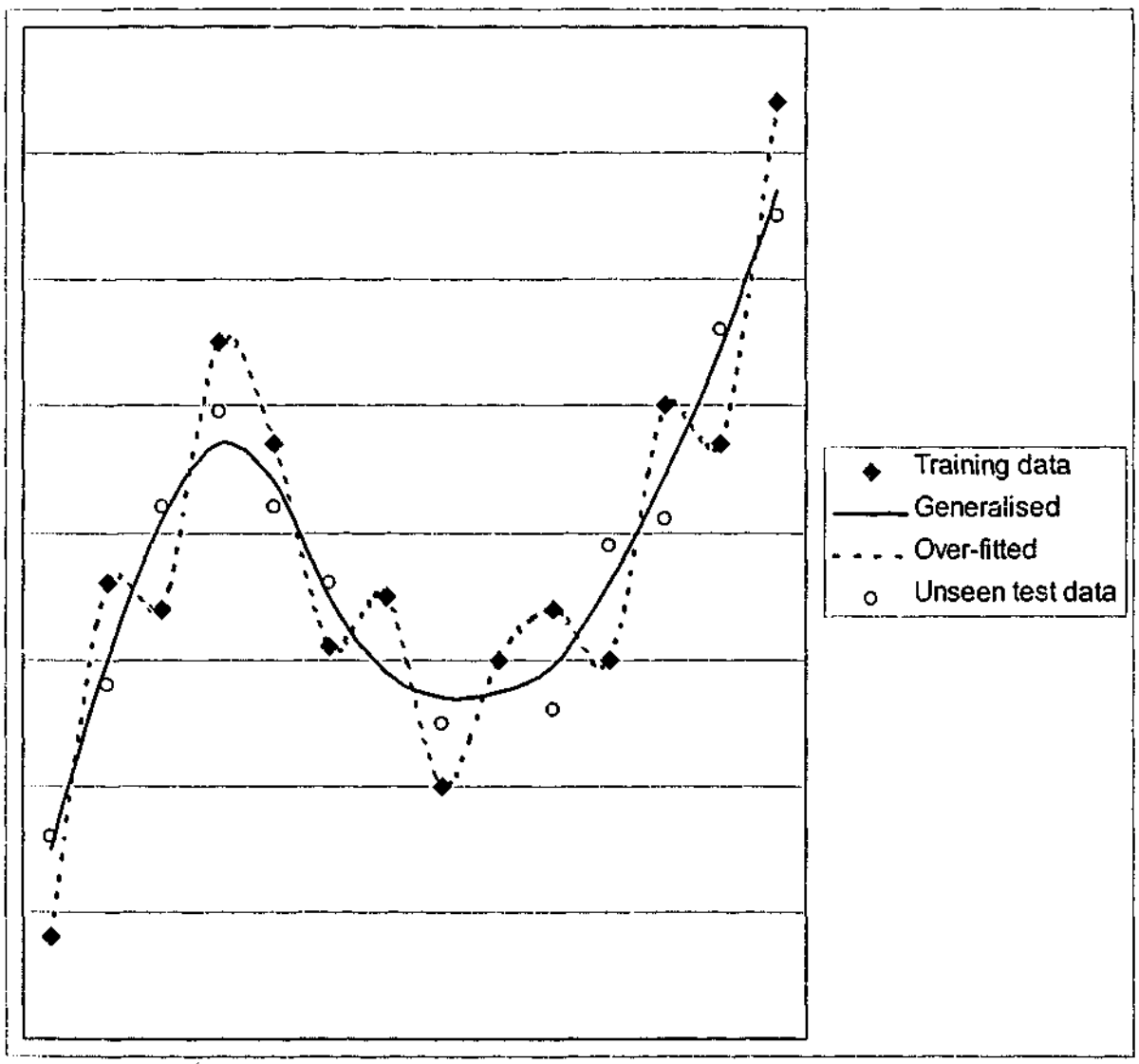


Figure 2-6: Hypothetical example of over-fitting a model.

The phenomenon of over-training ('over-fitting' in statistical terminology) is common to all inferential statistical techniques, not just neural networks. In general, the more free parameters in the model, and the greater the embedding (input) dimensionality, the worse the problem.

Limit the number of inputs

In the neural network field several approaches have been tried in efforts to avoid over-training. The first approach, common to all modelling, is to limit the number of inputs as much as possible without losing too much of the underlying information. Neural network practitioners should be wary, though, of limiting the inputs to only those variables shown by univariate linear analysis to be independent predictors of the output. Finding non-linear and complex relationships is the fundamental attraction of the neural network approach.

Limit the number of hidden layer neurons

A second common approach is to limit the number of hidden layer neurons in the network. If too few are included the network will be unable to converge on a solution. But if too many hidden layer neurons are present the network will not only be slow and computationally clumsy, but it is also more likely to begin over-training. There is no single or mathematically rigorous rule for calculating the optimum number of hidden neurons, but there are a number of 'rules of thumb' that provide a useful starting point. Smith (Smith, 1999a) mentions two: firstly the square root of the product of the numbers of inputs and outputs, and secondly, half the sum of numbers of inputs and outputs, plus the square root of the number of training patterns. (The latter is the default setting for the NeuroShell2 program used in this study (Ward

Systems Group Inc, 2000).) These are most useful as a starting point, from which a number of empirical tests can be made.

'Prune' the network during or after training

The third approach to avoiding over-training is known as 'pruning', and has been surveyed by Reed (Reed, 1993). The basic principle is to start with a larger than necessary network and remove the elements that contribute least to its decision-making. The simplest in principle (and most computationally demanding in practice) takes a deliberately over-trained network and sets the weights, one by one, to zero. Any weight in which setting to zero does not change the network's predictions is simply removed. In practice less computationally demanding techniques are required. Most of these can be considered in two broad groups. In the first type the network is first trained, and then some rule applied to decide which weights to remove (Cottrell et al., 1995). In the second type (Lozowski et al., 1996) the error function used during training is modified so that weights that contribute little are effectively removed by pushing their values close to zero. Many other different types of pruning algorithm have been proposed (Karnin, 1990), reflecting the fact that none has been convincingly shown to be effective in all situations.

Use a validation set for 'early stopping' of training

A more generally applicable approach to avoiding over-training (and the approach used throughout this thesis) is the use of a validation set to directly monitor the generalising ability of the network during training (Astion et al., 1993). Before training begins, a representative sample of the available data is chosen and separated from the training set. Training begins in the usual way, but at regular frequent

intervals (in this study every 200 presentations of a training pattern) training stops, and the network is used in its current state to calculate outputs for the reserved validation set. Almost invariably the mean squared error for the validation set is slightly higher than for the training set, but they tend to reduce in parallel as training progresses. At the point where over-training begins the two curves diverge, and the error for the training set continues to decrease while that for the validations set either remains the same or increases (the hallmark of over-training). The software keeps a copy of the best set of weights so far, so that once it becomes clear that no further improvement is going to happen the network can be re-built in its most generalised form. Figure 2-6 gives a hypothetical illustration (the actual curves are usually not so smooth).

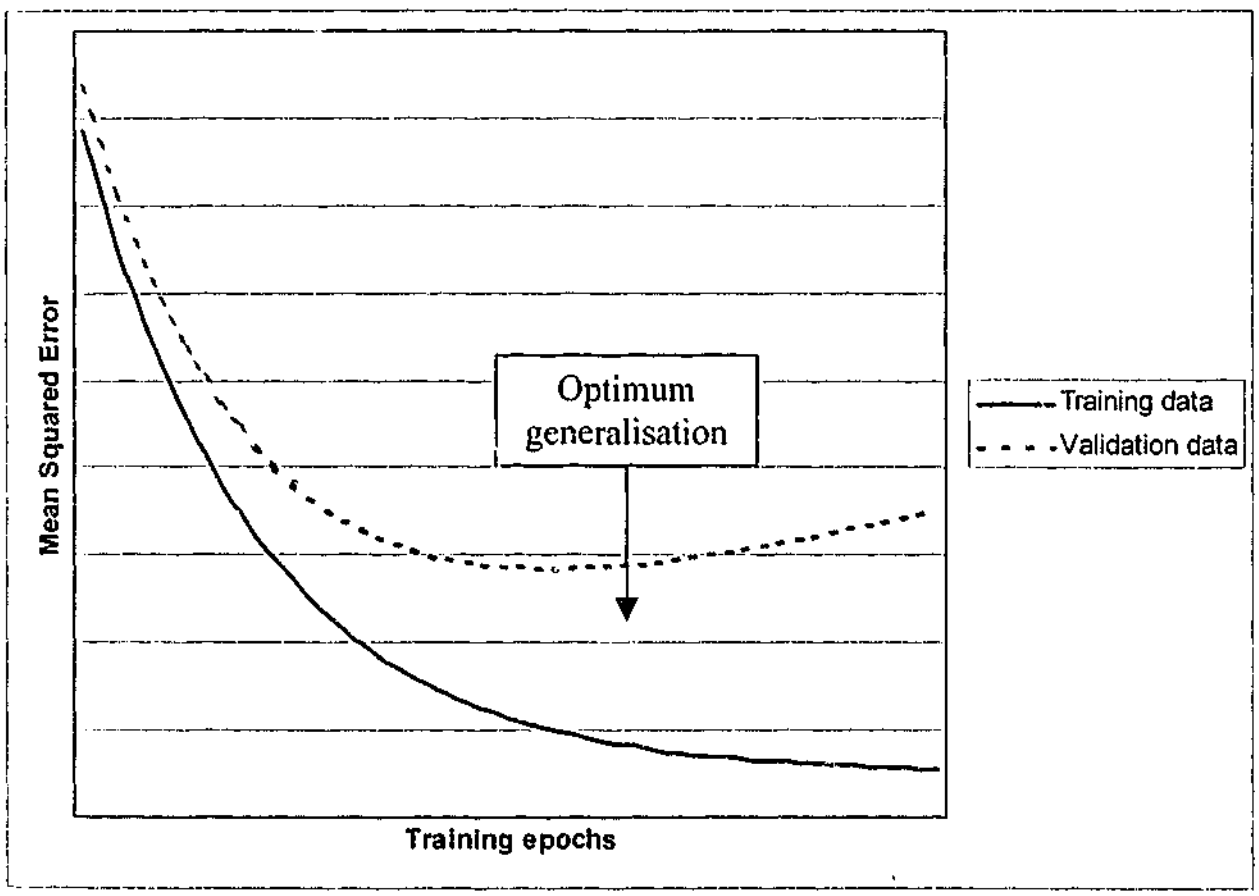


Figure 2-7: Hypothetical example of how over-training can be avoided by use of a validation set.

Using a validation set like this is known as 'early stopping' (of training), and is generally a very good way of minimising the risk of over-training. Its biggest drawback, however, is that some of the original data must be 'sacrificed' to the validation set. If the original data set is not large (in relation to the intrinsic dimensionality of the problem) this can have severe repercussions. If a particular pattern is not well represented in the training data set, but is common in the validation set, the training process can be skewed. A worse consequence is that the network is thus deprived of the opportunity to learn that pattern, and will thus not recognise it in future applications. For very small data sets, techniques have been developed (Finne et al., 2000) for 'leave-one-out' validation. In this technique only one pattern is removed each time, and a new model trained with all the other data. By training a new model for each training pattern it is possible to estimate the general accuracy of that architecture for that data set.

The number of hidden layer neurons is much less critical when early stopping is used to avoid overtraining. It may even be better to use a larger number of neurons in this case.

Test on new (unseen) data, and beware of extrapolation

It is important to note that a trained neural network model can only give meaningful results on new data points that represent an interpolation of the training data. For any non-linear model it is dangerous to extrapolate beyond the range of data used in creating the original model, as there is no guarantee that the relationships that hold for one range of values will hold for values outside that range.

Whatever the method(s) used to minimise over-training (and wherever the objective is to produce a functional model for real-life use), the performance of the network should still be tested against a completely new data set. This may ideally be data from a different source, or it may be a further segment of the original data that is reserved before training begins. The latter approach is generally used throughout this thesis, and these reserved data are called the test set.

A note on terminology

> *"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean – neither more nor less."*
>
> *"The question is," said Alice, "whether you can make words mean so many different things."*
>
> *"The question is," said Humpty Dumpty, "which is to be master – that's all."*
>
> (from *Through the Looking-Glass*, Lewis Carroll, 1872.)

One reason that artificial neural networks have not had greater recognition within the medical community is that they were not developed by statisticians, and biostatisticians commonly have little knowledge of their potential uses. This has not been helped by the tendency of neural network experts to invent new terminology for concepts that have already been explored by statisticians, making neural networks seem even more exotic and opaque than necessary.

A few statistically trained neural network enthusiasts have attempted to remedy this, with reviews that highlight the essential similarities between many neural network

techniques and statistical ones (Cheng and Titterington, 1994, Sarle, 1994b, Sarle, 1994a, Warner and Misra, 1996). They point out that, applied to some problems, neural networks are little more than inefficient alternatives to existing parametric techniques. A full coverage is beyond the scope of this review, but one important and relevant similarity is between the multilayer perceptron and both discriminant analysis and non-linear regression.

Sarle (Sarle, 1994b) points out the following differences in terminology for the same concepts (p2):

Statistical term:		**Neural network term:**
Variables	=	Features
Independent variables	=	Inputs
Predicted values	=	Outputs
Dependent variables	=	Targets (or training values)
Residuals	=	Errors
Estimation	=	Training, learning, adaptation, or self-organisation
Estimation criterion	=	Error function, cost function, or Lyapunov function
Observations	=	Patterns, or training pairs
Parameter estimates	=	(Connection)(Synaptic) Weights
Interactions	=	Higher-order neurons
Transformations	=	Functional links
Regression and discriminant analysis	=	Supervised learning or heteroassociation
Data reduction	=	Unsupervised learning, encoding or autoassociation

Cluster analysis	=	Competitive learning or adaptive vector quantisation
Interpolation and extrapolation	=	Generalisation

There is perhaps a more fundamental 'cultural' difference between neural network practitioners and statisticians, which leads them to somewhat contradictory mind-sets. Neural network modellers are concerned primarily with the accurate prediction or classification of complicated phenomena rather than on explaining how the prediction is made. Statistical models, on the other hand, are more often made in order to explore the strength of associations between predictors and outcomes – the goodness of fit is calculated to see how much variation the model 'explains'. To illustrate: Given a new model, the neural network enthusiast will immediately ask how low the error was, but the statistician will ask what model was used and what were the strongest parameters. The statistician's question can be almost meaningless when applied to a neural network model with many hidden neurons and thousands of connection weights (Breiman, 1994).

Lisboa and Wong (Lisboa and Wong, 2001) point out, too, that clinicians only really trust biostatisticians. But clinicians and epidemiologists are pragmatic people and, Dybowski and others (Dybowski et al., 1996) are probably right when they argue that neural network models could be accepted more widely, provided their predictive capacity has been rigorously evaluated in appropriate field trials.

Even within the neural network field much of the terminology has yet to stabilise. A 'fully-connected feed-forward network trained by back-propagation of errors' is the

same thing as a 'multilayer perceptron'. Many, but not all, use the term 'epoch' to mean one presentation of all the patterns in the training set to the training algorithm. Many count the inputs (although no actual processing happens at that level) as a separate layer, while others only count the hidden and output layers. Thus, what is a three-layer network to one researcher is a two-layer network to another. To avoid this confusion, it is preferable to describe a network in terms of the number of hidden layers.

One more example is particularly relevant here. In order to minimise over-training, it is common to split the available data into three sections. One section, usually called the training set, is used in the backpropagation algorithm to adapt the connection weights. The second portion of the data is used during training to check the mean squared error (but not to change the weights). When the error on this second set levels out or begins to rise, it means that over-training has begun. This portion of the data is called by different authors the validation set, the test set, or the cross-validation set. The third portion, not used at all during training, is used to test the final model. It is sometimes called the test set, but may be called the verification set, the production set, or even the validation set. Throughout this thesis the first portion is called the training set, the second portion the validation set, and the third portion the test set.

Testing and assessing neural networks

Training a network model is only one step along the road to proving its use in a real-world setting. The problem of over-training is well known, and there is no guarantee that the training data truly represent all the patterns with which the network will be presented in use. In some cases the raw output from the network is not in a suitable

form for direct use. Further validation and testing is always required, and its form depends on the kind of task the network will be expected to perform.

For time series prediction

Measuring goodness-of-fit of neural network models

The commonest training strategy is to minimise the mean squared error for the whole training set. The lowest mean squared error achieved, especially for the reserved validation or test sets, gives a good overall indication of the fit of the final model. However, as the error is calculated for all the outputs together, it is not useful for assessing individual outputs in a network with multiple outputs. Similarly, where the objective is classification the mean squared error can be deceptive. In such cases the selection of an appropriate cut-off (threshold) can give very useful outcomes even where the mean squared error is high.

The coefficient of multiple determination (R^2) is useful but its properties are different

For networks used to predict continuous variables, a useful general measure of goodness of fit for each individual output is the coefficient of multiple determination (R^2). There are several variants of R^2 with different bounds and different properties (sometimes called 'adjusted' R^2). See (Hosmer and Lemeshow, 2000), (Altman, 1991) or (Norman and Streiner, 2000). The version used throughout this thesis effectively compares the accuracy of the predictions for a single output with a trivial model that simply predicts the mean value in every case. If y denotes the observed value, $\bar{y}$ the mean of all the observed y values, and $\hat{y}$ the model prediction for a given pattern, then –

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

An R^2 value of 1.0 represents a perfect prediction, but some of the other well-known properties of R^2 when applied to parametric regression do not necessarily apply when it is used to assess artificial neural networks where training is stopped early through the use of a validation set. (See the section above titled "Use a validation set for 'early stopping' of training".) For example R^2 values below zero are common, and show that the model makes worse predictions than if it had simply predicted the mean value in every case. With early stopping through the use of a validation set, there is also no reason to assume that R^2 must always rise when new inputs (and thus new free parameters) are added to a network. This is of particular importance in Chapter Four.

Testing using contiguous or non-contiguous test data

In time series applications there are two main questions. First, how well has the network generalised the patterns in the training data? Second, if it has generalised those patterns, how well does it perform on data from a different time window? The two questions require two different approaches to the selection of the test data set.

To check for generalisation within the same time window as the training data, the test sample can be non-contiguous time points randomly selected throughout the same time series window as the training data. These are reserved before training begins (or perhaps collected prospectively after the network is trained).

The potential users of the forecasting network will be much more concerned with the second kind of test. To check for performance outside the training time window, a set of contiguous time points are selected, beyond the end of the training set. The network's performance on this kind of test set is a much more direct assessment of its likely performance in the field.

Relative costs of different types of error

Another consideration in time series forecasting is the cost of making a mistake, whether in financial or other terms. For a network predicting a financial time series, occasional failures to predict upward or downward swings may be acceptable, provided the final balance for the whole sequence is favourable. Similarly, the cost of a false alarm in predicting a measles outbreak in a Mozambican province may not be high, as the overall effect of extra vaccination, training or resources will probably still be beneficial. (Though where resources are scarce the network will soon be abandoned if it makes repeated mistakes.) On the other hand, in the case of a gastroenteritis surveillance system for an urban Australian setting, the result of just one false alarm and an unnecessary 'boil water' alert could be major litigation and considerable financial loss. Some of litigation might even be directed at the creator of the network model! Each application needs to include an assessment of the cost of both false positive and false negative predictions.

For classification

In the kinds of classification problems encountered in this study the objective is to assign each case to only one category (usually 'true' or 'false'), but the neural network output for an individual case is commonly neither exactly one nor exactly zero. It is better interpreted as the network's estimate of the probability that the case actually belongs to the positive class. The experimenter interested in a question with a dichotomous outcome – will the patient live or die, or did the worker suffer toxin exposure or not – is forced to decide how to deal with these continuous outputs. There are several options. One is to ignore outputs that are not one or zero – but that may mean discarding most of the network outputs. A second is to define a decision threshold point (more often called a 'cut-off' point in the medical literature) and assign all network outputs above or equal to that point to the positive class, and all those below it to the negative one. Receiver operating characteristic methodology provides some sophisticated tools to optimise this decision.

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves first appeared during World War II in the field of radar detection. A 'receiver' (the radar equipment and its operator), had certain 'operating characteristics' (the rate at which it correctly identified enemy aircraft, and the rate at which it correctly identified other things not to be enemy aircraft). The concept was developed in signal detection theory throughout the 1950s (Green and Swets, 1966) (where it is sometimes called 'relative operating characteristics') and came into medicine via radiology. It has since been applied to many tasks that involve classification into dichotomous classes such as healthy/diseased, abnormal/normal, will survive/will die.

The ROC curve relates false positives to false negatives, by plotting the sensitivity of the test on the y-axis and (1-specificity) on the x-axis, for many different possible decision threshold (cut-off) settings (Swets, 1988, Sackett et al., 1991). The resulting curve has some useful attributes. The area under the curve, lying between 0.5 and 1.0, can be thought of as a measure of the overall probability of a correct ranking of a positive/negative pair, without making any assumptions about the distribution of either positives or negatives in the sample (Hanley and McNeil, 1982). Another way to put it is an area of, for example, "0.8 means that 80% of the time a single random selection from the positive class will have a larger test value than a single random selection from the negative class" (DeLeo, 1993). Figure 2-8 is an example of an ROC curve.

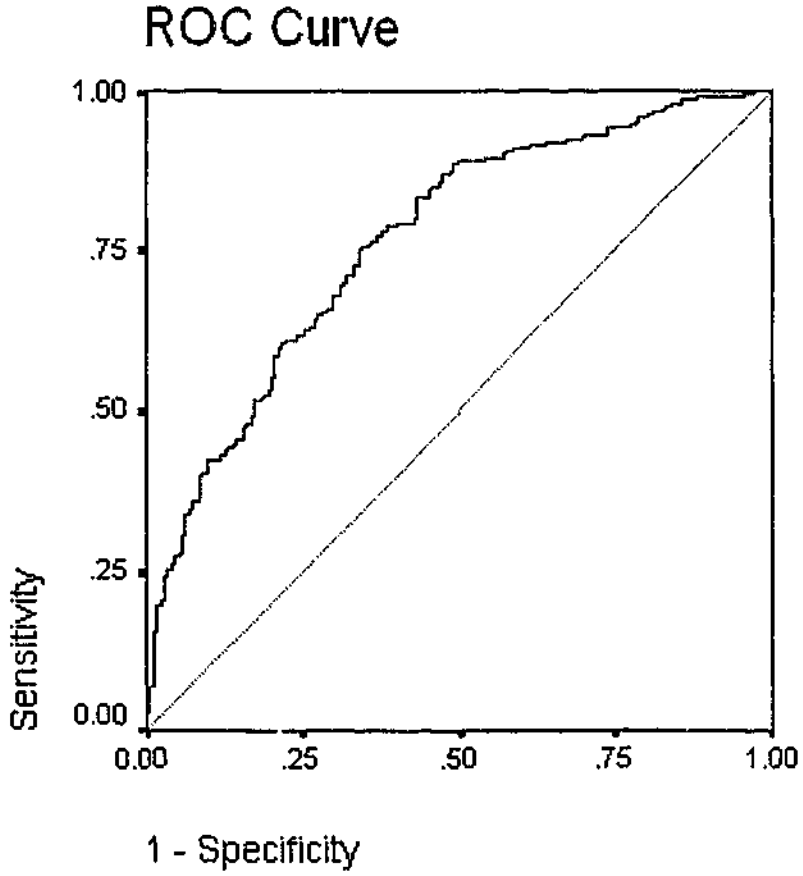


Figure 2-8: Example ROC curve (calculated from data in Chapter Seven).

The area can be estimated in various ways – the non-parametric approach is usually unbiased (DeLeo, 1993). Provided the classifications in the test data come from a trusted gold standard not related to either technique (Dayhoff and DeLeo, 2001), the area under the ROC curve is a very useful way of comparing two different classification techniques, such as neural networks and logistic regression. Because it is related to the Wilcoxon statistic (Hanley and McNeil, 1982), standard error and confidence intervals can be estimated for the area under the curve, making such comparisons much more useful.

The point on the ROC curve closest to the upper left hand corner (where sensitivity and specificity are both equal to one) represents the cut-off point that minimises the sum of false-positives and false-negatives (Sackett et al., 1991) – although this may not be the ideal cut-off in practice, for reasons outlined below.

Prognosis

Wyatt and Altman (Wyatt and Altman, 1995) and Wyatt and Spiegelhalter (Wyatt and Spiegelhalter, 1990) have summarised the key issues that challenge the creators of any new prognostic model.

The first is clinical credibility. All the relevant clinical inputs should be included, and no known predictor excluded from the model. The chosen inputs should be easily collected, with high reliability, in time to guide decisions. Modellers should avoid arbitrary thresholds for continuous variables (an acknowledged shortcoming of 'bin' models such as the 'tumour-lymph node-metastasis', or TNM, cancer staging system). They point out that, whatever the statistical method, it must be correctly employed

without transgressing its assumptions. More controversially, they suggest that the decision-making structure and rules in the final model should be apparent, and its decisions should 'make sense' to the clinicians who will use them. They mention the legal concerns considered elsewhere in this review, and single out neural networks as unsuitable 'black boxes' in this regard. They also argue that it should be simple for the clinician to calculate the model's prediction for a given patient, without recourse to a computer – although the advances in computing in even the few years since then have surely made this a less important consideration.

Their second key issue is the accuracy of the model. Before it can be put to general use any new model must be tested with a large test set (with five to ten test cases per item of clinical data used), collected prospectively according to a pre-defined protocol. They do not mention the fact that the prospective collection of testing data is costly, and is unlikely to occur until a new model has been developed and tested on a smaller set. They correctly insist on a second kind of accuracy, called 'calibration'. If the model predicts an outcome with a probability, then the probability should be accurate (when applied to a large group of test cases).

Third, Wyatt and Altman point out the need for evidence of generality. They rightly point out that consistent and widely accepted definitions must be used, even for apparently obvious inputs like age. They argue that the model should be tested in a different time and place with new data. This principle can, however, be overstated – for example a model intended to relate to a particular subset of the population need not be proven accurate on other subsets on which it will never be used.

Finally, they correctly argue that a model only deserves clinicians' confidence if there is empirical evidence from unbiased clinical trials that the model actually improves patient care.

All these issues must be considered in assessing any new neural network prognostic model, although some can only be achieved after much preliminary work has been done, and the necessity of others can be questioned.

Exposure assessment

A further option is possible in some types of problem, such as the exposure assessment problem explored in Chapter Eight. It may be useful to seek out one or more decision thresholds that are highly accurate at either excluding or confirming class membership (i.e. provide 100% predictive value, whether positive or negative), even though these may not be the levels that give the best overall accuracy. In the benzene exposure problem, for example, if the assessor can be completely confident that a worker assigned a negative rating by the network truly did not have exposure, then s/he need not examine that questionnaire. Although it would still be necessary for the expert to examine all the other questionnaires, this might still mean an important saving in time and money. In some cases it may be possible to assign two threshold values: one which classifies cases as positive with a high degree of certainty, and the other which reliably classifies negatives. Then only the cases falling between the two cut-off points need be examined.

Sensitivity analysis and assessing individual inputs

Some authors (perhaps influenced by commercial neural network programs) present an analysis of the relative importance of different inputs to the decisions of their neural network models. This is usually done by comparing the means of the weights connecting each input to the neurons of the hidden layer. This is not to be recommended, however. Where it works it is analogous to an estimate of the strength of the linear relationship between that input and the output. As such it is better estimated using a parametric technique. Where the relationships are complex it can be very misleading, as a high weight at the input may be offset by one or more lower weights at the hidden or output layers.

Other authors present a 'sensitivity analysis' based on a more complex technique (Masters, 1993). A set of contrived inputs is created from a sample of training patterns, in which only the values of a single input are allowed to vary. The variation in output of the trained network in relation to changes in the chosen input is taken as evidence of the contribution of that the particular input to the network's decision-making. Once again this may be valid if the relationships are linear, but there are much more efficient linear techniques for this assessment. Where the relationships are non-linear and/or there are important interactions between variables, this technique can become quite misleading.

For these reasons no 'reading' of the weights or assessment of individual inputs will be presented in this thesis.

Existing applications of artificial neural networks

Once it had been shown that an artificial neural network could approximate any nonlinear function, it was clear that neural networks have the potential for solving major problems in a wide range of application domains (Hornik et al., 1989). Although brain researchers continue to be interested in neural networks, the major impetus for further research became their application as a new class of statistical tool. Since about 1990 the number of applications of neural networks in fields not related to brain research or artificial intelligence has increased yearly, into the hundreds, if not thousands (Wong et al., 1997).

Unfortunately the many variations in problem type, network architecture, training algorithm, training parameters, number of training patterns, form of testing and so on preclude any type of meta-analysis of the many published accounts of practical applications of neural networks. Even when organised by the application area, a review is thus reduced to descriptions of many individual studies. This rather *ad hoc* approach among the neural network community is one of the reasons for apparent inconsistencies in assessment of the potential use of neural networks.

Non-medical applications

Artificial neural networks have been applied in a very broad range of applications outside the health field. This section gives a broad overview of the areas and types of these applications. Within each application area studies are considered in order of publication date.

Speech recognition

One of the earliest applications of neural networks was to machine recognition of speech (Shamma, 1987). In 1990 Morgan and Bourlard (Morgan and Bourlard, 1990) published moderately successful attempts to recognise a 1000-word German vocabulary by combining hidden Markov models with multi-layer perceptrons. By 1999 Ellis and Morgan (Ellis and Morgan, 1999) had created networks to recognise typical large-vocabulary speech (from various American broadcast media), and they had moderate success (around 30% word error rates at best) using very large networks with thousands of hidden neurons, a purpose-built multi-processor computer, and 10 to 74 hours of training material. Venayagamoorthy and others (Venayagamoorthy et al., 1998) were concerned to recognise the speaker rather than the words *per se*, for applications in security and crime detection. They achieved moderate success at recognising a small number of speakers saying standard words.

Character and handwriting recognition

Hopfield described the use of neural networks in optical character recognition (Hopfield, 1999). Pacut and Czajka (Pacut and Czajka, 2001) used neural networks to recognise hand-written signatures, using static features (the appearance of the written signature) and dynamic ones (pen angles and pen tip pressures over time). They achieved false acceptance rates of 0% and false rejection rates of 11.11%.

The physical sciences, industry and military

Artificial neural networks have been widely used in the physical sciences and industry.

There are a number of real-world combinatorial optimisation problems that are similar to the 'travelling salesman problem' (in which the salesman must find the shortest route between a number of cities, visiting each city exactly once). Many researchers have successfully used artificial neural networks (often hybridising them with other techniques) to solve these problems (e.g. Smith et al., 1996, Joppe et al., 1990, Kwok and Smith, 2000). Smith (Smith, 1999b) has reviewed this work.

Marzban used neural network to predict the formation of tornados from Doppler-radar information about developing storms (Marzban and Stumpf, 1998).

In the military they have been used in automatic target recognition, control of flying aircraft, and engine combustion optimisation (Dayhoff and DeLeo, 2001). Dayhoff also referred to their use for fault detection in complex engineering systems.

In the food industry Mittal used neural networks to predict temperature, moisture and fat contents in meatballs during deep-fat frying, so as to allow continuous control during automated frying processes (Mittal and Zhang, 2001).

In the water industry, Fletcher and others (Fletcher et al., 2001) gave 3 examples, including water colour prediction using visible water colour, turbidity and conductivity. The neural network was more accurate than the chemist's bench measurements of samples. They used autoassociative networks to detect sensor failure (where one of a number of sensors – of conductivity, pH, temperature, dissolved oxygen, turbidity, total organic carbon, UV absorption and colour – started to give results inconsistent with the others), and to reconstruct the likely estimate from the

defective sensor while it was being repaired. (Autoassociative networks have the same number and type of outputs as inputs, and are used to extract consistent 'signals' from a noisy series.) Neelakantan and others (Neelakantan et al., 2001) used easily-measured parameters in surface water in the Delaware River to predict the concentrations of *Cryptosporidium* and *Giardia* cysts.

Biological sciences

In the biological sciences, Chan and Prager (Chan and Prager, 1991) used neural networks to accurately predict the variation in annual populations of the Canadian lynx.

Gonzalez and Arnaldo (Gonzalez and Arnaldo, 1993) used a radio-frequency capacitance field transducer to automatically record the movements of laboratory rats before and after doses of apomorphine. They were able to classify the different movements (such as paw-licking, sniffing and facial grooming) consistently using a feedforward network with one hidden layer.

Wu (Wu, 1997) reviewed the development of useful neural network tools for a number of molecular sequence analysis problems, ranging from gene identification to protein structure prediction and protein family classification (e.g. Wu et al., 1996).

Chang and others (Chang et al., 1997) described a combined neural and proportional-integral-derivative control system for control of knee joint position via electrical stimulation of the quadriceps muscle – of potential use to paraplegics.

Business and finance

Neural networks have been widely and successfully applied in the business and financial world. Wong and others (Wong et al., 1997) reviewed the use of neural networks in business up to 1995. They found 213 published reports, mostly by academics, covering 134 fields from audit opinion prediction to wire bonding process. More than half used commercially available neural network programs such as NeuroShell. Smith and Gupta note their use in marketing, retail, banking, and insurance (Smith and Gupta, 2000), and Charles (Charles, 1998) describes their use in Medicaid fraud detection in Texas. They have also been used in exchange rate prediction (Wu, 1995) and evaluating loan applications (Hopfield, 1999).

Time series forecasting

There are good theoretical reasons and some empirical experience to suggest that neural networks are useful for the forecasting of time series into the near future. Time series are of particular interest to epidemiologists involved in surveillance and control of infectious diseases, and Chapters Four, Five and Six of this thesis investigate the use of neural networks in the forecasting of disease incidence.

Baba and Kozaki (Baba and Kozaki, 1992) used an artificial neural network to predict rises and falls in the stock price of a single Japanese company. Their networks gave good out-of-sample predictions, provided the general trend in the test set was similar to the trend in the training data. Relatively little has been published regarding the use of neural networks in day-trading of stocks and shares (perhaps for reasons of commercial secrecy), but the Internet abounds with claims that day-traders make millions using them.

Chakraborty and others (Chakraborty et al., 1992) used neural networks to forecast the market price of flour in three American cities. They found by experiment that simple three-layer perceptron networks gave very good results – including better results than recurrent networks.

Caire and others (Caire et al., 1992), and later Brierley and Batty (Brierley and Batty, 1997, Brierley and Batty, 1998), developed neural network models for predicting electricity load requirements for a commercial electricity grid.

Zhang and Thearling (Zhang and Thearling, 1994) used a massively parallel 'supercomputer' to forecast a derived time series designed to have high-dimensional dynamics. (It was based on the equations of motion for a damped, driven particle in a complex potential field, and 100,000 data points were generated.) Neural network models were at least as effective as a k-nearest neighbours technique.

Lezos and others (Lezos et al., 1999) emphasised the importance of finding the optimum input dimensionality (the number of past time points used as inputs) and noted that this depends on the particular problem.

Dostál (Dostál, 1999) compared neural networks with Auto-Regressive Integrated Moving Average (ARIMA) models in forecasting numbers of airline passengers, sunspots, heat consumption in the city of Brno (in the Czech Republic), and the value of shares in the Komercní Bank. He concluded that, although their performance was

generally similar, the neural networks had an advantage in forecasting non-linear series, but were not as good for series with aspects of deterministic periodicity.

De Olivera and others (de Olivera et al., 2000) used neural networks to forecast chaotic series created using the Lorenz system. They got good results, though they noted the importance of the numbers of hidden layers and hidden neurons. They suggested m:2m:m:1, where m is the embedding dimension of the attractor of the dynamical system in consideration.

In their review Zhang and others (Zhang et al., 1998) noted that many forecasting problems have been solved with neural networks, mostly using straightforward multilayer perceptron networks with logistic activation functions for the neurons and trained by simple backpropagation minimising the mean squared error. Some hybrid models have also shown promise. At least a training and test set are used, often with a third portion used for validation during training to avoid over-training. The input vector for a time series application almost always consists of a moving window of fixed length along the series. Although the number of inputs is the most critical variable, there is no proven method for determining the optimum. Direct forecasting of multiple forward steps is possible with neural networks, and they state that it is probably better than single-step forecasting.

Building on earlier reviews (Hill et al., 1994), Remus and O'Connor (Remus and O'Connor, 2001) also reviewed this field, including the results of a large 'competition' between different modelling techniques for a standard set of time series. They came to several conclusions and made some recommendations, including:

- In general neural networks are comparable to other techniques, but because they can partition the sample space and build different functions in different portions of that space, they have advantages for modelling series with discontinuities.
- They are better than traditional methods for long-term forecast horizons, several time points ahead.
- The minimum number of training patterns required may be quite large, especially to avoid over-training.
- Data should be cleaned, scaled and deseasonalised before training.
- There is no advantage to using more than one hidden layer.
- Sigmoid activation functions (such as the logistic function) give good results.
- Specialised methods, such as a momentum term, should be used to avoid local minima.
- Holdout samples must be used in testing the model.

In general artificial neural network models tend to be at least as accurate as 'traditional' linear techniques for modelling and forecasting time series, but they are not always superior.

Medical applications

Although practical neural network applications have yet to come into widespread clinical use, a small number of enthusiastic clinicians and epidemiologists began to explore their uses as soon as computing resources allowed. This section outlines the progress with neural networks in the medical fields most relevant to this thesis. Although the most common network architecture continues to be the multilayer perceptron, once again the plethora of network techniques, fields of application and

testing methods precludes much more than a chronological listing in general categories.

Diagnostic aids

By the early 1990s neural networks had been proposed for the diagnosis of epilepsy, low back disorders, early diagnosis of heart attack, and the diagnosis of breast cancer (Lim et al., 1997).

Floyd and others (Floyd et al., 1994) trained a neural network to predict which women with suspicious mammograms would turn out to have breast cancer on biopsy. They found the network model more accurate than the radiologists who extracted the original information from the mammograms.

When Baxt (Baxt, 1995) reviewed the field in 1995, he fc 1 947 citations reporting clinical applications of neural networks in the biomedical literature. These ranged from assistance with diagnosis (from appendicitis to temporal arteritis), through imaging (radiographs to cardiac perfusion scans), to analysis of electroencephalograph and electrocardiograph wave forms, as well as outcome prediction (recovery from surgery to cancer survival). He commented on the wide variation in sophistication in the ways network accuracy was validated. Baxt's own neural networks (Baxt, 1991) trained to diagnose myocardial infarction on the basis of data available on admission, eventually reached an accuracy of 96% (compared with 80.5% for physicians given the same data) (Baxt and Skora, 1996).

Tourassi (Tourassi et al., 1996) compared neural networks with a linear classifier (using inputs identified by a linear discriminant analysis) and found similar accuracy for non-invasive diagnosis of pulmonary embolism.

Brickley and others (Brickley et al., 1996) felt they were close to completely automating the assessment of oral smears with neural networks whose training data included mean nuclear and mean cytoplasmic areas measured by image analysis. Their networks differentiated between normal mucosa and dysplastic or malignant mucosa with specificity of 0.82 and sensitivity of 0.76.

Naguib and others (Naguib et al., 1998) used data from image cytometry applied to breast cancer aspirates to train neural networks to predict lymph node metastasis. The network predictions were superior to the current standard (histological assessment).

Rietveld and others (Rietveld et al., 1999) found self-classifying neural networks could correctly classify 95% of training cases (but only 45% of test cases) when trained on tracheal breath sounds of patients with asthma and controls. They compared these with human observers' ability to classify the same sound spectra (though not with their ability to recognise the sounds themselves).

Wallace and others (Wallace et al., 2000) used neural networks trained on the reflectance spectra from pigmented skin lesions to distinguish malignant melanoma from benign naevi, and found them a substantial improvement on multivariate discriminant analysis.

Troni and others (Troni et al., 2000) assessed a commercially available neural network system for computer-assisted reading of Papanicolaou smears from the cervix. They found the neural network system slightly more accurate and slightly faster than un-assisted reading at diagnosis of *carcinoma in situ*, but concluded that it increased costs significantly (with a marginal cost for each new lesion detected in excess of US$25,000).

Finne and others (Finne et al., 2000) used neural networks trained on clinical and laboratory data to predict malignancy in prostate biopsy. The network models, tested by 'leave out one cross-validation' were superior to both the existing clinical test and to logistic regression.

Sinha and others (Sinha et al., 2001) found neural networks were better able than physicians to predict CT scan abnormalities in children with closed head injuries.

DeGroff and others used recordings from an electronic stethoscope to train a neural network to distinguish heart murmurs from normal sounds in children (DeGroff et al., 2001). Having a small training set, they used a jack-knifing technique for validation. For the optimal settings they achieved 100% sensitivities and specificities.

Hirsch and others (Hirsch et al., 2001) created network models based on patients' responses to a standardised asthma questionnaire. The questionnaire was sent by mail (with 6825 responders) and 186 were also examined by a clinician. By choosing appropriate thresholds for review, they predicted the neural network models could give above 74% true positive rates for detecting new cases of asthma.

Disease prognosis

Ebell used neural networks to predict failure to survive following in-hospital cardiopulmonary resuscitation (Ebell, 1993). The areas under receiver operating characteristic (ROC) curves were 0.765 for the network, compared to 0.717 for Ebell's own 'Prognosis After Resuscitation' score (although the standard errors overlapped). The network predictions gave 97% predictive value positive, and an Odds Ratio of 11.3 (95% CI 3.3 to 38.2) of death for patients with ANN output less than 0.001.

Doig and others (Doig et al., 1993) compared neural network models with logistic regression, predicting mortality in the Intensive Care Unit from physiological variables collected on the third day of the patient's stay in 422 cases. They found the two methods' performance identical (with the area under a receiver operating characteristic curve of 0.82) on a validation set.

Ortiz and others (Ortiz et al., 1995), though using quite a small data set (95 cases), found their neural network models gave 90% accuracy in predicting mortality in patients with cardiac failure.

Flanagan and others (Flanagan et al., 1996) compared neural networks to logistic regression in predicting survival of patients with sepsis. They found for their small sample of 173 patients that the two approaches gave similar results (around 80% accuracy), but the neural network models were less sensitive to the decision threshold.

Dybowski and others used neural networks (using a genetic algorithm) to predict outcomes in critically ill patients. They argued that neural networks (like other regression techniques) "are validated not by whether they provide a comprehensible chain of inference but by their performance. (...) a neural net (which can be regarded as a complex regression model) can be accepted in medicine with or without a detailed understanding of how it works – provided its predictive capability has been rigorously evaluated." (Dybowski et al., 1996, p1150)

Dorsey and others (Dorsey et al., 1997) compared logistic regression with neural networks for the prediction of pancreas graft survival, and found the neural networks more accurate. However, they did not test the validity of their models with a reserved test set.

Lapuerta and others (Lapuerta et al., 1998) trained neural networks to assess perioperative cardiac risk in vascular surgery patients. The neural networks gave similar accuracy but better calibration than simple logistic regression.

Zernikow (Zernikow et al., 1998) used neural networks to predict individual neonatal mortality risk (and found them better than logistic regression). They used a feed-forward fully connected three layer perceptron, with the first half of 890 preterm infants used for training and the second half used for testing. Despite the area under the ROC curve being 0.95, they concluded it was unsuitable for individual patient decisions. (Presumably they meant decisions to withdraw or withhold treatment, where they felt even greater accuracy was required.)

Edwards and others (Edwards et al., 1999) claimed that their neural network models of survival after intracerebral haemorrhage (with a training set of only 81 patients) were superior to logistic regression, but they presented only training set data, and no unseen test set.

Clermont and others (Clermont et al., 2001) compared neural networks with logistic regression for the prediction of mortality in the intensive care unit, and they too found their performance similar. Both models degraded where training data sets with fewer than 800 patients were used.

Cacciafesta and others used neural networks to predict 2-year survival in people over 81 years of age with 85% accuracy on a small (20-person) test set (Cacciafesta et al., 2001).

Buzatu and others (Buzatu et al., 2001) found neural network models based on pre-existing risk factors to have similar accuracy to the existing ('Denver') logistic regression model in predicting survival after cardiac surgery.

Trauma survival

There are a number of different scoring systems for trauma severity in widespread clinical use (reviewed by Van Camp (Van Camp and Delooz, 1998) and Fani-Saleck (Fani-Salek et al., 1999)). These are either based on the degree of physiological dysfunction (and thus used for patient triage and assessment of treatment effectiveness) or the anatomical location and severity of the injury. The two types are commonly combined to create composite indices that predict patient mortality. These are mostly used in the assessment of quality of care, allowing health care facilities to monitor their performance (for example by examining cases where survival was predicted but death occurred).

Artificial neural networks have been used in attempts to produce more accurate prognostic scores. McGonigal and others (McGonigal et al., 1993) compared a neural network model with two commonly used scores (TRISS and ASCOT) and found the neural network would have reduced the number of deaths requiring review by 40%. Although they did not compare with other scoring systems, Izenberg and others' neural networks had an overall accuracy of 91% in predicting survival, with inputs based on information available in the emergency department (Izenberg et al., 1997). DiRusso and others (DiRusso et al., 1999a, DiRusso et al., 1999b, DiRusso et al., 2000) created neural network models that were more accurate predictors of death than TRISS, then showed they were accurate predictors of survival for data sets other than the training data. They were unusual among neural network modellers, in concerning themselves with the calibration of their models' predictions – the neural network calibration was better than TRISS. Hunter and others (Hunter et al., 2000) achieved slightly better areas under receiver operating characteristic (ROC) curves than the

standard TRISS (which is based on logistic regression) even when using the same input variables. Becalick and Coats (Becalick and Coats, 2001) used a different type of neural network (a Kohonen Self-Organising Map) and found the latest United Kingdom version of the TRISS system gave superior areas under the ROC curve, but less well calibrated predictions.

Dombi and others (Dombi et al., 1995) trained neural networks on a 580-patient database to predict outcome in rib fracture, although they made no comparisons with other techniques.

Cancer prognosis

One area of medicine where neural networks have been enthusiastically applied is in cancer prognosis, whether expressed as disease progression, response to treatment, or patient survival. Accurate estimates of prognosis would be useful for individual patients for deciding between treatment options, or simply giving an answer to the inevitable question "How long have I got?" They are also useful at the institutional level, allowing a more meaningful comparison between facilities with different caseloads, or within the same facility over time.

Current prognostic models

Current cancer staging systems are based either on 'bin-model' approaches (e.g. the TNM, or 'tumour, lymph node, metastasis' classification) or statistical inference, such as logistic regression (sometimes with derived nomograms, such as the Partin nomograms used in prostate carcinoma) (Montie and Wei, 2000). They are useful for classification, teaching and (to a lesser extent) prognosis, but they are simply not

accurate enough to allow the kind of decision support that patients and clinicians need. Montie and Wei also pointed out an important practical consideration: to gain wide medical acceptance any new staging or prognostic system must not only be practical, but also supported by both the International Union Against Cancer (UICC) and the American Joint Committee on Cancer (AJCC).

In the 'bin-model' approach, all the variables are given discrete values, and the different combinations of possible values define all the states ('bins') that a patient could possibly be in. Every patient must fit into one, and only one, bin. Bostwick and Burke (Bostwick and Burke, 2001) argued that in most cases the bins are 'lumpy', with few patients in many of the bins (and thus no useful predication for those groups). The demand for discrete variables (forcing continuous variables like age to be cut into a few pieces each) decrease the accuracy these methods can attain. They concluded that inferential models (among which are numbered both neural networks and regression models) are more accurate and thus preferable.

Neural network applications in cancer prognosis have most often used the standard multilayer feedforward architecture, trained by standard backpropagation or a variant of it. However, they have varied greatly in the type of cancer, the source of data, the type of inputs, and the way the usefulness of the final models is assessed. This makes formal comparison of the results difficult.

Breast cancer

In breast cancer neural networks have mostly been used to model mortality. A number of clinical (e.g. race, and the presence of ipselateral recurrence from a previous

tumour) and pathological (e.g. histologic tumour type, nodal status, nuclear grade, and blood vessel invasion) factors have been shown to independently predict mortality in women with this cancer (Fisher et al., 2001).

Ravdin and Clark (Ravdin and Clark, 1992) used a feedforward network including information about the period of observation, and compared it with a Cox proportional hazards regression model. They felt that the superior performance of the neural network in breast cancer prognosis indicated it could detect subsets of patients where the proportionality assumption of Cox regression was violated.

Burke (Burke, 1994) used data on breast cancer from the Surveillance, Epidemiology and End-Results (SEER) registry. Starting with 44,135 cases of first-time breast cancer diagnosed in white females between 1997 and 1982, he omitted all censored cases and those with missing data, and was left with 3 subsets of about 3,500 each. He found a neural network model superior to the TNM staging system for predicting 10 year death due to the cancer. The c-index (an estimate of area under ROC curves) was 0.73 for the neural network and 0.69 for the TNM model.

Lisboa and others (Lisboa et al., 1998) used data from 1,616 patients at two large UK hospitals undergoing surgery for breast cancer. They compared a standard multilayer perceptron neural network (trained in three different ways) with Cox proportional hazards regression, and found the two techniques gave comparable results.

Lundin and others (Lundin et al., 1999) followed up nearly all the women diagnosed with breast cancer in the city Turku, Finland, between 1945 and 1984, who underwent

radical excision and axillary dissection. (They excluded patients who had adjuvant or palliative therapy, bilateral or advanced cancer at the time of diagnosis, and those who died of other causes.) They compared neural network models with logistic regression in attempts to predict survival to 5, 10 and 15 years, by areas under ROC curves. The neural networks gave larger areas (0.909 vs. 0.897 for 5 years, 0.886 vs. 0.817 for 10 years and 0.883 vs. 0.799 for 15 years for the best neural network and logistic models respectively). None of the differences, however, was statistically significant.

De Laurentiis and others (De Laurentiis et al., 1999) trained a neural network to predict relapse in breast cancer, using clinical and pathological markers from an American cancer database. They tested the network model on a series of 310 patients from Italy, and found the neural network model outperformed the TNM staging system. It was able to identify subgroups of patients with very different risks of relapse within each TNM class.

Lisboa and Wong (Lisboa and Wong,, 2001) returned to breast cancer mortality prediction, but this time arguing that the greatest use for neural networks is in fact to generate hypotheses about non-linear interactions between prognostic factors. These could then be used to improve essentially linear techniques such as logistic and Cox proportional hazards regression. They note that "medical statisticians know how to test (a) hypothesis using such models and theirs are the only results which clinicians really trust" (p 2472).

Lundin and others (Lundin et al., 2001) signalled their agreement with this argument, using neural networks as an adjuvant to Cox regression in assessing the prognostic influence of routinely performed histological grading of breast cancers in Finland.

Prostate cancer

In prostate cancer Ragde and others (Ragde et al., 1998) compared neural networks with linear regression for predicting 10 year survival (and found the networks improved accuracy by 10% compared to regression). Other applications in prostate cancer, however, have concentrated on predicting pathological stage, cancer progression, or response to treatment rather than survival. Naguib and others (Naguib et al., 1998) compared neural networks with multiple discriminant analysis for the prediction of response to treatment. They found the networks more accurate (80% vs 75%), although their sample size was very small (21 training and 20 test cases).

In another small study (20 patients with progression and 20 without), Mattfeldt and others (Mattfeldt et al., 1999) compared multilayer feedforward networks with two variants of learning vector quantisation and a linear discriminant analysis. They used histopathological and clinical data as inputs, and found the learning vector quantisation networks superior, but still only achieved about 93% accuracy.

Potter and others (Potter et al., 1999) tried an unusual neural network architecture, the 'genetically engineered neural network', with clinical and pathological inputs, to predict biochemical progression of prostate cancer. They found the networks more accurate than either standard logistic regression and Cox regression.

Han and others (Han et al., 2000) predicted biochemical recurrence (at 3 and 5 years) of clinically localised prostate cancer in men who underwent potentially curative surgery. They trained multilayer perceptron networks with pathological and clinical inputs, and found them superior to both Cox and logistic regression. The same group compared neural networks with the regression-based Partin nomograms for predicting pathologic stage from information available at the time of prostate biopsy (Han et al., 2001). In this large study (with 5744 patients) the neural network models were more accurate than the nomograms.

Predicting both pathologic stage and biochemical progression of prostate cancer, and with a 309-patient data set, Ziada and others (Ziada et al., 2001) found a four layer perceptron network gave 80% accuracy, compared to a multivariate regression analysis (which had an overall accuracy of 67%).

With a training set of 3474 Austrian men undergoing prostate biopsy, Horninger and others (Horninger et al., 2001) used multilayer perceptron networks trained on clinical and biochemical markers (including several based on Prostate-Specific Antigen (PSA) levels). They compared the risk profiles thus developed with the use of standard PSA cutoff levels, and found the neural network models 1.5 to 2 times as specific as the standard method.

Batuello and others (Batuello et al., 2001) trained a standard multilayer perceptron network to predict lymph node spread in men with clinically localised prostate cancer. Clinical, biochemical and histopathological markers were the inputs. Although they

did not compare with any other statistical technique, they did use test sets from two different institutions and found areas under ROC curves of 0.81 and 0.77 for the two.

Other cancers

Neural networks have been used in a number of other cancers. Kappen and Neijt (Kappen and Neijt, 1993) studied ovarian cancer with standard multilayer perceptron networks, and found they predicted survival better than multivariate Cox regression. Kehoe and others (Kehoe et al., 2000) also used multilayer perceptrons to predict survival in ovarian cancer, and achieved areas under ROC curves of 0.84. They did not compare their models with any other technique.

Marsh and others (Marsh et al., 1997) also did not compare their multilayer perceptron models with other techniques. They used clinical and pathological data available at the time of transplant to predict recurrence after liver transplantation in the presence of hepatocellular carcinoma. They were able to find upper and lower cut-off points (decision thresholds) for their models that gave accurate predictions, either for disease-free survival or recurrence at 1, 2 or 3 years, for 60% or more of their patients. The same group (Marsh et al., 1998) went on to show that multilayer perceptron networks were better than the TNM staging system at predicting a good result from liver transplantation for hepatocellular carcinoma.

In a small study of survival in non-small cell carcinoma of the lung (67 patients with Stage I or Stage II disease), Bellotti and others (Bellotti et al., 1997) created multilayer perceptron networks to predict recurrence and/or death from the disease.

Their models correctly classified all the cases, but they do not seem to have used any validation or test set.

Bryce and others (Bryce et al., 1998) trained multilayer perceptron networks to predict 2-year survival in patients with advanced squamous cell carcinoma of the head and neck. They used 'leave-one-out' validation rather than a separate test set, but they did compare with logistic regression. They found the neural network models more accurate than logistic regression, and also that the network models were able to use predictive variables that could not be used by logistic regression.

Kinsey and others (Kinsey et al., 1999) studied the prediction of mortality in children with acute lymphoblastic anaemia, using age, sex and white cell counts as inputs. They found the predictive accuracy, measured by areas under ROC curves, of the networks to be similar to that of a Cox proportional hazards regression model using the same inputs. They used 1271 cases for training and 300 reserved cases for testing, repeating their tests with 10 different random splits of the data.

Qureshi and others (Qureshi et al., 2000) used data from 201 patients to predict recurrence within 6 month and stage progression in Ta/T1 bladder cancer and 12-month survival in T2/T4 bladder cancer. They made self-organising map networks, and compared their accuracy with that of experienced clinicians. The accuracy of the neural network models was slightly higher than the clinicians', but the difference was only statistically significant for predicting stage progression for T1G3 tumours.

Colon cancer

In colon cancer Burke and others (Burke et al., 1997) used the American College of Surgeons' Patient Care Evaluation data set. They showed that multilayer perceptron networks could substantially out-perform the TNM staging system for 5-year survival prediction, even when given only the TNM variables as their inputs. When they included other demographic and anatomic variables, the neural network performance increased even more, with areas under ROC curves of 0.869 for the neural networks, compared with 0.737 for the TNM system. In the same paper they showed similar results for breast cancer, using the SEER data for that disease. However, in neither case did they compare with a conventional statistical technique.

Bottaci and others (Bottaci et al., 1997) also studied colon cancer. They used highly individual clinicopathological factors, including among others the rank of the surgeon, peritoneal involvement by the tumour and whether the resection was judged curative, and trained multilayer perceptron networks with data from 334 patients. They made six network models, each predicting death within a different time interval from 9 to 24 months, and found them superior to two consultant colorectal surgeons. They then applied the network trained to predict 2-year survival to unseen data from a second institution, and found the network gave 90% overall accuracy, compared with 75% for the surgeons. They did not compare their network models with conventional statistical techniques.

Finally, Snow and others (Snow et al., 2001) trained multilayer perceptron networks with two hidden layers to predict 5-year survival in colon cancer. They used the Commission on Cancer data from the (United States) National Cancer Data Base

(NCDB), although for unstated reasons they chose to work with only a random 10% sample of the 375,000 colon cancer cases, and they did not include any censored cases. Comparing the neural network predictions with a standard parametric logistic regression, and testing each on a reserved 25% test set, they found the neural network gave a ROC curve area of 0.876, and the regression 0.82. At a sensitivity to mortality of 95%, the neural network specificity was 41%, compared to 27% for the regression model.

A role for neural networks

In mathematical parlance, the cancer staging systems in widespread use today are classical linear rule-based expert systems (Niederberger, 1995). This is especially true of 'bin' models like the TNM system. These systems will always have a place in cancer care, but they do not give sufficiently accurate estimates of prognosis for individual patients.

There has been an understandable tendency, too, to aim for universal models, applicable in every setting. Perhaps the way forward will be to use newer inferential techniques to create database-specific tools, which aim not for universal applicability but for maximum local accuracy. Artificial neural networks have not proved themselves clearly and universally superior to 'traditional' statistical techniques, but they can at least be used to enhance their application. Levine (Levine, 2001) points out too that there are many more potential predictors, biochemical, histological and immunological, that are not currently recorded in cancer registry databases. Neural network models may come into their own in extracting the complex relationships that are likely to lie in these enhanced databases.

Occupational health

Artificial neural networks have as yet been little used in occupational medicine, though not for lack of potential applications.

Burge and others (Burge et al., 1999) used neural networks to distinguish occupational from non-occupational asthma, using the mean peak expiratory flow rates during days off and days at work as inputs. They found the neural network models more accurate (81% vs. 64% sensitivity, 100% vs. 98% specificity and 93% vs. 86% overall accuracy) than their own previous system based on a linear discriminant function.

Reporting earlier unpublished work Claycamp and others (Claycamp et al., 1998) concluded that neural networks are useful as complementary statistical tools to use in quantitative risk assessment. They later published a study (Claycamp et al., 2001) of the use of neural networks in assessing chronic radiation sickness amongst the workers of the Mayak Production Association in Russia. They used unsupervised neural networks to classify workers as ill or not, based on haematological indices such as leucocyte or platelet counts. They combined the neural network assessments with binary recursive decision tree methods to avoid the circular arguments otherwise involved where the radiation dose is included in the analysis.

A promising area for the application of neural networks in occupational medicine is the assessment of individual exposure to chemicals. In the past serendipitous observations of clusters of cancers in workers in the same industries have revealed strong associations between occupational exposure to chemicals and various cancers.

The earliest studies involved large groups of workers exposed for long periods of time to high levels of a small number of toxic substances, with strong correlations between occupation and exposure (Goldberg and Hemon, 1993). Nowadays, with thousands more chemicals in common use, the identification of occupational (or environmental) exposures that carry a long-term risk of disease has been described as one of the main public health problems of our era (Siemiatycki et al., 1981). However, for chemicals with lower exposures and considerable variation in exposure within the one occupation, the early techniques of assigning exposure on the basis of a simple job title or even by geographical area (e.g. ship-building towns) do not give the statistical power needed to reveal associations. Case-control methods are increasingly used, and one of the most important aspects of these has been the need for accurate assessment of exposure to individual chemicals by individual workers, sometimes many years before their recruitment to the study (Goldberg and Hemon, 1993).

In recent years assessment of exposure has made use of computerised job exposure matrices (JEMs), which are automated sets of indicators that show which exposures occur in which occupations (Siemiatycki, 1996). One such example is FINJEM (Kauppinen et al., 1998). FINJEM includes three dimensions: occupations, chemical agents and time, and produces estimates of both the probability and level of exposure for a large range of occupations and agents. The JEM approach makes it possible to partly automate exposure assessments, and also make them less subjective. More recently the JEM approach has been supplemented with job specific questionnaire modules, or JSMs, (Stewart et al., 1998) which enhance the worker's recall of exposures and provide more detailed and accurate information. As well as the general

circumstances of the job, these include questions about the exact tasks performed and their timing.

However, the actual assessment of exposure to an individual chemical agent for an individual worker still requires considerable experience and expertise on the part of the assessor (commonly an industrial hygienist, or even a panel of several people with differing expertise). It is still very time-consuming, and there is still considerable scope for inconsistency on the part of each individual assessor, and agreement between assessors is often poor (Benke, 2000). One way of improving both the speed and consistency of assessments using JEMs and JSMs might be to supplement, or even replace, the expert assessor by training an artificial neural network on previous expert assessments. This is the approach investigated in this study (see Chapter Eight).

Although not working in occupational health, Boulle and others (Boulle et al., 2001) used neural networks for a similar task to exposure assessment (i.e. replacing or supplementing expert assessors by networks trained on their own assessments). They were concerned with assigning cause of death from a standardised verbal autopsy questionnaire. They found neural networks to be more accurate than either physician review or logistic regression.

Infectious disease epidemiology and surveillance

Although there has been some interest in the non-linear and possibly even chaotic behaviour of time series of disease incidence (Kanjilal and Bhattacharya, 1999, Grenfell et al., 1995), the fields of infectious disease epidemiology and surveillance have received little attention from neural network practitioners. Hammad and others (Hammad et al., 1996) trained artificial neural networks to predict infection rates with *Schistosoma mansoni* in Egyptian school children, using demographic, clinical, laboratory and behavioural data as inputs. Training on the first year of data and predicting the second and third years' infection rates, they found the networks' sensitivity (83% vs. 66%) and positive predictive values (63% vs. 59%) superior to logistic regression (the latter value in each parenthesis).

No published applications of neural networks to the forecasting of disease rates in measles or gastroenteritis have been found, nor of their use in machine-assisted detection of outbreaks in disease surveillance data sets. The next chapter reviews the current approaches to these problems, and illustrates the ways that neural networks might offer theoretical advantages.

Criticisms and deficiencies of neural networks

Interpretation of connection weights

The commonest criticism of neural networks is that they are 'opaque', or 'black box' models. This refers to the fact that for any but the simplest of networks it is not possible to ascribe 'meaning' to the internal parameters (the connection weights) in

the way that can be done with linear and other parametric techniques. A large weight connecting a given input to the hidden layer may be an indication that the network gives great importance to that factor, but it is also quite possible that the large input weight is offset by small weights at the hidden layer. It is safest not to try to 'read' the weights at all. This is certainly a limitation, and it is the main reason neural network workers mostly concern themselves with problems where the prediction accuracy is more important than ascribing importance to individual inputs.

Conventional statistical methods are not entirely immune to this accusation, however. Bostwick and Burke (Bostwick and Burke, 2001) argue that if the phenomenon being modelled is complex, then the model must be complex, and increases in complexity reduce the transparency of both traditional statistical models and artificial neural network statistical models.

Legal concerns and clinicians' confidence

Potential legal difficulties with the use of computerised decision aids (including, but by no means limited to neural networks) have been apparent since their earliest medical applications. In 1989 *The Lancet*'s then resident legal commentator, Diana Brahams, and co-author Jeremy Wyatt (Brahams and Wyatt, 1989) raised some important points about the legal status of computerised decision support tools. It has still to be clarified in the courts whether such a tool would be considered a 'service' or a 'product' in legal terms – and thus whether the creators of the software would be considered culpable if harm came to a patient. Although there is no specific precedent, Brahams cautions doctors against reliance on such aids, especially where the reasoning process used by the machine is not clear and open to scrutiny and

challenge by a knowledgeable doctor. The doctor would likely still be regarded by the courts as a 'learned intermediary', responsible for filtering and assessing the machine's advice before applying it to the patient. Neural network models are particularly open to criticism in this regard, as it is not usually possible to tease out the actual rules by which the trained network functions.

The following year Wyatt, together with Anna Hart (Hart and Wyatt, 1990), turned more specifically to artificial neural networks. Many of their criticisms, such as the need for accurate data collection, inputs that would be easily accessible to typical users of the system, a very good 'gold standard' for training outputs, and rigorous evaluation of trained networks with data from a second source, are important reminders to researchers in this field. In addition, many of their concerns apply equally well to other decision aids, not just neural networks. But their final conclusion is valid – neural networks will find a niche where an explanation of their decision making process is not required and where sufficiently large quantities of high-quality data are available for training and evaluation.

Perhaps influenced by such concerns, creators of neural network models have commonly been circumspect in the uses they advocate. Despite high accuracy in predicting neonatal mortality, Zernikow concluded the network was unsuitable for individual patient decisions, presumably meaning decisions to withdraw or withhold treatment. Baxt (Baxt and Skora, 1996) is similarly restrained. Although his network had higher accuracy at diagnosing myocardial infarction than experienced physicians, he notes that the networks tended not to use inputs that had previously been shown to have high predictive power. Further, the network seemed to place high predictive

importance on physical findings, such as jugular venous distension and râles, which have not been shown to have high predictive power for myocardial infarction.

On the other hand, Guerriere and Detsky (Guerriere and Detsky, 1991) dispute these concerns. They point out the number of clinical advances that have been adopted after empirical success, without anyone knowing how they worked. They argue that the whole clinical training process is an apprenticeship anyway, and is strongly analogous to the way neural networks learn.

Are neural networks really superior to other techniques?

Any new technology with exciting potential will naturally attract many researchers. Then the well-known tendency of academic journals to preferentially publish papers with positive results can easily lead to an initial phase of over-optimism. Although they have had both level-headed proponents and critics, artificial neural network applications have sometimes been presented with undue optimism. In their review of business applications Wong and others (Wong et al., 1997) found that 38.5% of the papers reviewed compared neural networks with other techniques. 47 published accounts found neural networks superior to other techniques, 19 accounts found them comparable, and only 3 found their performance inferior. However Wyatt's commentary on medical applications (Wyatt, 1995) quotes a review of 3 clinical and 19 non-clinical data sets, in which neural networks clearly outperformed other statistical or decision-tree methods only once. Among the medical applications cited in this review, neural networks were commonly found superior to physicians' decisions, but a majority of comparisons with statistical techniques showed similar or

only slightly superior accuracy. Neural network advocates tend to emphasise features like improved calibration, while detractors criticise the 'black box' aspects.

Recently Sargent (Sargent, 2001) produced an important review of publications comparing neural networks with standard statistical methods such as Cox or logistic regression. He found 28 studies with sample sizes greater than 200 patients. Among these, 10 papers (36%) found neural networks superior to statistical techniques, 4 (14%) found statistical methods to be better, and the two methods gave similar results in the remaining 14 (50%). Most importantly, of the 8 studies with sample size above 5000, standard regression and neural networks gave equivalent results in all but one (and in the remaining case regression was superior). Further, he noted evidence of publication bias in favour of papers where neural networks won the competition.

In some settings there is no existing proven method, and then neural networks have a legitimate claim to establish the standard against which other techniques should be compared. Occasionally there are obvious reasons why the data invalidate the assumptions behind parametric techniques, and neural networks, with no such assumptions, deserve to be tried against the problem. Artificial neural networks have achieved great success in business applications as tools for 'data mining', finding unexpected and often non-linear relationships between inputs and outputs. The data sets to be 'mined' have generally been created for other purposes, and statisticians are rightly wary of using parametric techniques on 'fishing expeditions' into their depths. In medical applications we are on the parametric techniques' home ground, with prospectively designed data sets and problems of the type against which modern parametric techniques cut their teeth. It is not surprising that techniques such as

logistic or proportional hazards regression are often perfectly capable of extracting all the information such data sets contain.

In these cases it is appropriate, as several authors (Hart and Wyatt, 1990, Lisboa and Wong, 2001) have suggested, that neural networks be used mainly to look for non-linear relationships. Used as indicators of the maximum attainable classification accuracy, they can be used to test whether other more transparent techniques are extracting all the available information from the data set.

Chapter Three: Machine-assisted detection of outbreaks in disease surveillance data sets

Large outbreaks of water-related disease are not instantaneous disasters like earthquakes or aeroplane crashes. Contamination often continues for days or weeks, incubation periods vary, and there may be secondary transmission of the disease organism from primary cases to others by other means than water. Water related outbreaks generally evolve slowly enough that early detection of the increasing number of cases would allow useful preventive and curative measures to be applied. Prompt investigation and correction of treatment plant problems, the early introduction of a general 'boil water' alert to consumers, or alerting physicians to the presence of an outbreak could all help reduce case numbers by hundreds or even thousands, with corresponding reductions in mortality. The earlier the outbreak is detected, the smaller the eventual total case numbers, and the smaller the number of deaths. To this end, government health departments and public health laboratories around the world undertake surveillance of key organisms with outbreak potential. Astute epidemiologists and laboratory technicians perusing tabulations of recent laboratory isolates have often been the first to spot evolving outbreaks as clusters of isolates in time and/or space.

Despite these successes, smaller outbreaks probably go undetected quite often. And even at their best, laboratory-based surveillance systems may only detect increasing cases weeks, rather than days, after the initial contamination event.

There are several problems that prevent the early detection of outbreaks such as the large number of organisms with outbreak potential and the large numbers of geographic areas that need to be monitored. There is also a need for robust statistical techniques that will allow reliable detection of an outbreak as close to its onset as possible.

A typical laboratory-based system might recognise several thousand different organisms (counting all the various serotypes, etc), with a considerable range of frequencies and public health importance. The large numbers of data points and the need for more sophisticated statistical tools make it increasingly difficult for a human observer to scrutinise and analyse them all. If we are to improve the quality of water-related disease surveillance it is increasingly obvious that computers and sophisticated analytical algorithms will be needed to flag unusual events and supplement human judgement. It may also be useful to reconsider the choice of indicators to be regularly collected and analysed, going beyond the traditional dependence on laboratory isolates.

Choice of indicators

Key considerations in the choice of surveillance indicators are their sensitivity and specificity, ease of collection, and (most importantly in relation to water-borne disease) timeliness in relation to contamination events. The ideal indicator should be highly sensitive, so that it rarely misses an outbreak, however small or localised. It should also be highly specific for water-related outbreaks, and thus rarely lead to a false alarm (which can be extremely costly and damaging to the water companies and health departments, as well as disturbing to the water consumers themselves). Ideally

the information should be very easy to collect, or already collected for other reasons, so that the surveillance system adds little extra cost to the community. To allow the maximum time for intervention, the indicator should rise as quickly as possible after a contamination event. Even better, the ideal indicator would be capable of *forecasting*, so that outbreaks might be foreseen days before they happen when case numbers are still small.

Unfortunately, individual indicators tend to be at opposite ends of the spectrum for these different criteria. Laboratory isolation of organisms from disease sufferers, which is the indicator most commonly used by health departments, is highly specific, but weeks can pass between a contamination event and the reporting of increasing isolates. The contaminated water must first be distributed to the consumers (often several days for the most peripheral areas of a large city), then there is a time lag between ingesting the water and the appearance of symptoms (i.e. the incubation period for the given disease, measured in hours to days). The patients must then recognise the need for medical assistance and seek care – often another day or more. Then their medical attendants must decide that testing is necessary, and order the test. The testing procedure itself commonly takes one or two days, sometimes longer. Finally the results must be communicated to the health department and analysed. Epidemiologists have at times been embarrassed to find media reports of disease clusters occurring days ahead of their detection by routine surveillance systems.

Several candidate indicators have been suggested that are further 'upstream' in this process, and that have been shown to rise early in past outbreaks. These include absenteeism from work or schools, pharmacy sales of anti-diarrhoeal medications,

attendance at hospital casualty departments and receipt of requests for faecal analysis (rather than positive results). None of these candidate indicators has so far proved ideal – they are difficult to measure, or are not specific to diarrhoeal disease outbreaks, or at least not specific to water-related disease. Nevertheless, as each day's delay in the early stages of a city-wide outbreak may mean hundreds or even thousands more cases, the search continues for more useful indicators of disease incidence. It is also likely that some of the candidate indicators that are currently difficult to obtain will become easier in the future. For example, it is possible that even small retail pharmacies will change to a computerised stock-control and inventory system in the next few years. The fact that useful data for outbreak detection (such as the daily sales of oral rehydration salts or antidiarrhoeal medicines) would then be routinely collected for other reasons would make acquisition of the same data by surveillance systems much simpler.

Statistical Techniques for Outbreak Detection

Any statistical approach to outbreak detection (and by extension any computer-based algorithm) faces several distinct difficulties. One is the day-to-day random variation inherent in any surveillance indicator. A second problem is the tendency of many diseases, and thus also their proxy measures, to seasonal variation. Some indicators also exhibit a secular trend – a longer-term tendency to rise or fall over several years. An even more difficult problem, where the surveillance database includes previous outbreaks, is creating an algorithm that can take account of those earlier outbreaks, and not tend to include the inflated numbers from outbreak years in its assessment of the 'normal' disease rate. Although a human observer can often spot all these different types of variation, and 'see through them' to the underlying pattern, they all pose

difficulties, and sometimes insurmountable limitations, to statistical algorithms. The underlying processes we attempt to model are themselves highly complex, and models that can accurately reconstruct past data sequences are commonly very poor predictors of the future of the same series.

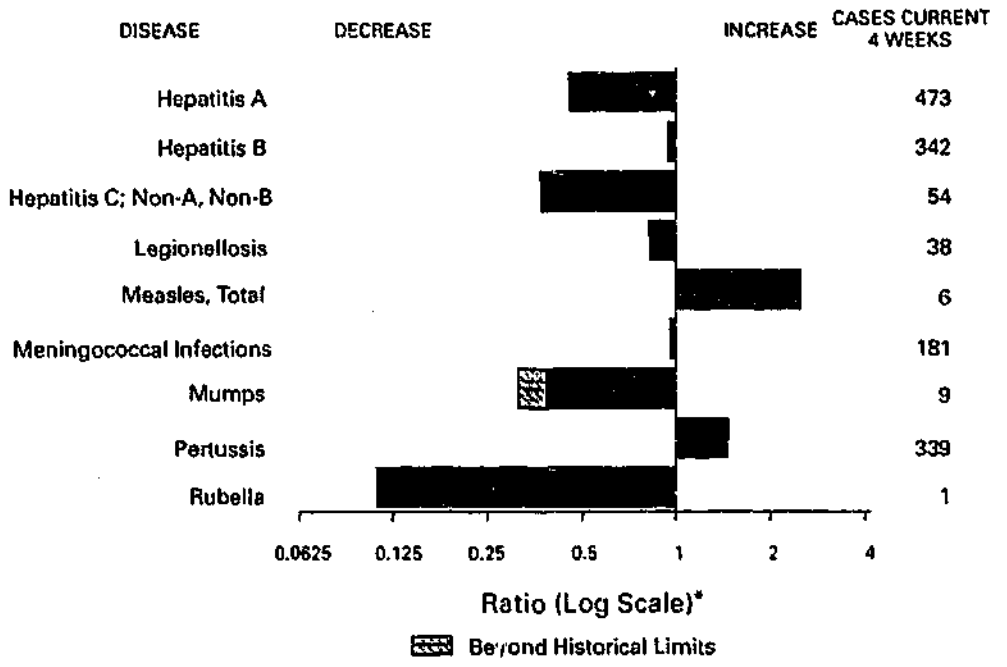
Although there is a certain amount of overlap, there are currently two main approaches to the question of machine-assisted detection of outbreaks in time series data such as disease surveillance databases. The first approach, more commonly used in current practice, seeks to answer the question "Are there more cases today than we would expect given our previous experience of the disease in question?" (Or "Has an outbreak begun?") The other approach asks the different, though related, question "How many cases are coming in the near future?" In other words "Is there an outbreak coming, and how big will it be?"

There are a number of techniques and algorithms that take the first approach – of asking whether there are there more cases being reported than would be expected, given previous experience.

Action thresholds

Simple approaches are intrinsically appealing. The simplest approach is to define a certain number of cases as an action threshold, above which further investigation or direct action will be taken (Stern and Lightfoot, 1999). This is easy where the disease is rare, and the appearance of one or a small number of cases warrants immediate action (such as poliomyelitis). But it is difficult where the disease is relatively rare, but some sporadic cases can occur without signifying an outbreak. It is also very difficult to define an action threshold where case numbers are higher, but day-to-day variation is large; a low action threshold will mean many false alarms, but a high action threshold will lead to outbreaks being detected too late for useful action. However, algorithms based on this basic idea are currently among the best in day-to-day use.

FIGURE I. Selected notifiable disease reports, United States, comparison of provisional 4-week totals ending March 10, 2001, with historical data



* Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable, and subsequent 4-week periods for the past 5 years). The point where the hatched area begins is based on the mean and two standard deviations of these 4-week totals.

Figure 3-1: A typical MMWR graph. (Source: Morbidity and Mortality Weekly Report, March 16, 2001. Vol. 50, No 10.)

The Morbidity and Mortality Weekly Review, or MMWR (Centers for Disease Control and Prevention, 2001) is a publication of the Centers for Disease Control and Prevention (CDC) of the United States of America (Figure 3-1). Each week the MMWR includes a graphical presentation of the reports of a short list of notifiable disease for the entire United States. The graph compares the reports for the four-week period just ended with a baseline number calculated from previous years. The baseline number for each disease is the mean of fifteen earlier four-week periods: the comparable four-week period in each of the preceding five years, as well as the four-week periods immediately before and after (Centers for Disease Control and Prevention, 1991). The calculation is made for a four-week period so that much of the random week-to-week variation should be smoothed out. The MMWR baseline figure

should take account of seasonal variation (by comparing with the same season in previous years), and lessen the effect of previous outbreaks (by including three months data from each of five previous years). The results are presented as ratios, with a ratio of 1.0 indicating no difference between the period just ended and the baseline. Ratios greater than 1.0 indicate an increase against the baseline (and ratios beyond historical limits are also flagged).

The MMWR graph is a useful summary, but because it relies on four-week periods to make it more robust, it is not timely. Although the effect of previous outbreaks is diluted in the baseline, it is still present and still has some influence. Because the summary applies to national data, localised outbreaks such as those associated with drinking water may be missed.

In England and Wales, The Communicable Disease Surveillance Centre (CDSC) of the Public Health Laboratory Service (PHLS) uses a similar approach, calculating a baseline for comparison with a given week, based on the six nearest weeks in the five preceding years. But the PHLS/CDSC goes further, including using the computer algorithm to scan large numbers of reports of individual organisms isolated in the various public health laboratories. The PHLS algorithm (Farrington and Beale, 1993, Farrington et al., 1996) also adjusts the baseline to allow for long-term trends by fitting a log-linear trend line to the data set. The algorithm reduces the effect of previous outbreaks in the baseline, by weighting down the influence of values from previous years that are more than one standard deviation above the general mean. Although the same algorithm is applicable to many different organisms with different weekly frequencies, the action thresholds and confidence intervals are adjusted so that

the specificity is similar over a broad range of organisms. Finally, a high threshold was chosen (a 99% confidence coefficient) in order to minimise the number of events flagged each week (i.e. only the events most likely to be real outbreaks are flagged, to make most efficient use of the human resources involved in further investigation).

The PHLS algorithm is undoubtedly useful, but it is still difficult to apply where the baseline mean varies from year to year, and it still uses the final step in the chain – the isolation of specific organisms in clinical specimens, so it is still not as timely as we might wish for the detection of water-related outbreaks. Similarly the reliance on laboratory reports of specific pathogens will not allow the detection of outbreaks where no pathogen has been detected.

The Scan statistic and related approaches

A conceptually related but statistically different approach is to ask, "Are there more disease events in the last time interval than one would expect due to chance?" (Or, "Is there unusual clustering of disease in time or space in the latest time period?") There are a number of conventional (i.e. parametric) statistical techniques that address this question. One of the best known is the Scan statistic (although it takes various forms e.g. Naus, 1965, Wallenstein, 1980, Wallenstein et al., 1989, Wallenstein et al., 1993). Calculation of the scan statistic begins with the null hypothesis that the events will be distributed in each time interval according to a defined probability distribution. In the simplest case the assumption is that the events will be evenly distributed over time. If we define a time interval 'window' of a fixed length shorter than the total time series, then the number of events in each of the possible 'windows' should be approximately equal. The number of events in a given window will vary according to the probability

distribution – the scan statistic is a measure of the probability that a given number of events seen in a given window could have occurred by chance. If a large number of events occur in the latest available time window, and it is unlikely that number could have occurred by chance, then an outbreak is flagged.

There are variants of the scan statistic, or its relatives such as Tango's index (Tango, 1984, Rayens and Kryscio, 1993), that can account for different probability distributions for the disease measure over time, and also variants that can take into account the distribution of cases in space as well as in time. They are all susceptible to the influence of previous outbreaks, and the operator adjustments needed to allow for that makes automation difficult.

The Cumulative Sum (CUSUM) chart

Proponents of the Cumulative Sum Chart (or CUSUM) suggest that rather than setting a threshold for investigation or action and waiting until it is crossed, it should be possible to detect deviations from the background disease incidence that show the indicator is approaching the threshold quite some time before the alert threshold itself is actually reached.

The CUSUM plots the accumulated diversions of a process away from a preset baseline over time (see Figure 3-2 for an example using data from Chapters Four and Five). It has its origins in the quality control of industrial manufacturing processes (Page, 1954, Page, 1961). A production line may be manufacturing a certain component with clearly defined upper and lower size limits. While the line is functioning normally, the measured size of its outputs should vary around the mean

with a predictable distribution. But as the machinery begins to wear, for example, the size of the component might begin to slowly drift away from the mean, until eventually the production line begins to produce unusable objects. The CUSUM chart is based on periodic sampling of the line's output, and plots the accumulated diversions from the mean over time. As long as the sizes vary in the usual way, the CUSUM value hovers around zero. But as soon as the process begins to deviate, the CUSUM (representing the accumulation of small errors) itself deviates rapidly. The slope of the CUSUM chart can thus indicate a small drift, and allow corrective action, long before the line's production becomes unusable.

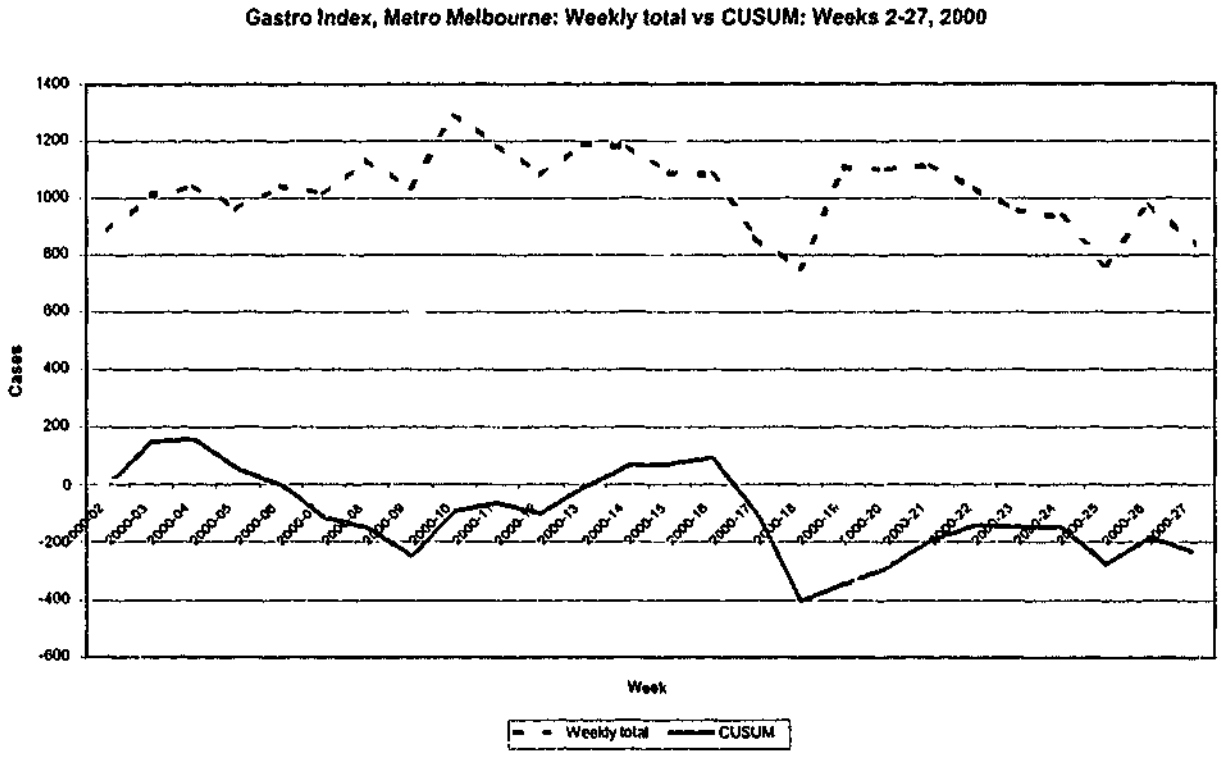


Figure 3-2: A CUSUM chart (solid line) for the total number of faecal analysis requests outside public hospitals in Melbourne, Australia, from weeks 2 to 27, 2000. The total numbers are plotted (dashed line) for comparison.

In disease surveillance (O'Brien and Christie, 1997, Hutwagner et al., 1997, Tillett and Spencer, 1982) a suitable CUSUM measurement might allow the detection of an outbreak long before the defined epidemic threshold is reached, and even before the upward trend can be discerned by eye among the random day-to-day variation. Once an outbreak is established, the same CUSUM chart might indicate more rapidly the high point and the beginning of the downward phase.

But unlike a production line, CUSUMs for disease surveillance do not have a simple baseline from which to start. The same problems apply as to other algorithms – day-to-day variation, seasonality and secular trends, and the effect of previous outbreaks. The CUSUM chart will only be as useful as the baseline calculation. Similar approaches have been suggested for this calculation, such as using a baseline calculated from the same time in previous years. The CUSUM then sums the difference between each day's incidence and the baseline for that day calculated from the same and surrounding days in preceding years.

Disease incidence is a continuous process, so it is also necessary to re-set the CUSUM chart to zero from time to time. But the optimum interval for this is not necessarily clear, and it would be difficult to automate, especially where large numbers of organisms (and thus large numbers of charts) are involved.

Figure 3-2 illustrates a CUSUM chart for the total number of faecal analysis requests outside public hospitals in Melbourne, Australia, from weeks 2 to 27, 2000. The total numbers are overplotted for comparison. Note that the rise in numbers in the first quarter is shown by the CUSUM plot to actually be less than expected when compared with the three preceding years. This graph illustrates one difficulty with CUSUM charts: the drops in week 17 and 18 are artefacts due to the coincidence of Easter and a local (ANZAC Day) public holiday, and the CUSUM chart should possibly be reset to zero from week 19. But what if an outbreak process had begun in week 16 or 17? Even without that possibility, it is clear that the re-setting of the baseline would need human intervention and judgement, adversely affecting the CUSUM's suitability as a machine-assisted algorithm in day-to-day use.

Forecasting

The second, or forecasting, approach is less well developed for disease surveillance, but is worthy of further attention. This approach is based on the perception that the development of disease outbreaks, and the pattern of endemic disease in a given population, is essentially deterministic – that there is some underlying mathematical process that can be used to describe it. Once the appropriate mathematical formula linking current disease rates to past ones has been elucidated, it should be possible to predict future disease incidence, at least into the near future. There are two ways that this might be useful. If the beginning of outbreaks themselves prove to be predictable, then the model could give prior warning of the onset of a new outbreak in time to avoid it altogether. If an outbreak turns out to be beyond the predictive power of such a model, then the model should at least give the best possible estimate of the

background rate expected, and allow rapid identification of the onset of the unforeseen epidemic. Potentially at least, these techniques might turn the bugbear of baseline calculation – the presence of past outbreaks in the time series – into an asset.

These techniques all make one basic assumption – that the factors determining current disease distribution will continue to be important into the future. Of course this is the basic assumption of all forms of prediction, but it is worth remembering that it does not always hold true.

Experience with forecasting techniques in routine disease surveillance data sets is so far limited, and results have been mixed, but the potential benefits make them worthy of further study.

Auto-Regressive, Moving Average, and Auto-regressive Integrated Moving Average (ARIMA) modelling

It might seem tempting to use conventional linear regression techniques to map the association between previous and future disease numbers. But conventional linear regression modelling techniques assume that each of the 'independent' input variables is just that – independent from all the others. Linear regression also makes assumptions about the frequency distribution of each input, generally assuming that they are normally distributed. The first assumption is almost never applicable to time series data related to disease incidence. Today's case numbers are generally closely related to yesterday's, and yesterday's to the day before, and so on.

There is however a class of statistical model that not only takes account of the correlation between adjacent elements of a disease time series, but actually uses it to make its predictions more accurate. Using two basic processes – Auto-Regressive (AR) and Moving Average (MA) filters – a time series can be generated that uses the preceding events in the series to determine the future values (Chatfield, 1996). The exact shape of the series depends on the values of a number of internal coefficients, and the technique aims to find the values for these coefficients that best replicate the observed series. The two processes are often used together, to create so-called ARMA models.

The process of estimating the coefficients, however, depends on certain assumptions about the original times series. The most important is that the series itself must be stationary (in the statistical sense, meaning that both the mean and the variance of the time series values should not change over time). This condition is commonly not satisfied, but it is often possible to create a new series which is stationary, by calculating the difference between the values at successive time points. It is sometimes even necessary to calculate second-order differences – the difference between the differences. ARMA techniques can then be used to model these 'differenced' series, and then the final results added together ('integrated') to produce a powerful technique called Auto-Regressive Integrated Moving Average (ARIMA) modelling.

Although these are potentially useful tools in disease surveillance (Choi and Thacker, 1981a, Choi and Thacker, 1981b, Stroup et al., 1988, Fernandéz-Pérez et al., 1998, Watier and Richardson, 1991), they are very much dependent on being able to achieve

stationarity, either in the original data set or in the differenced series derived from it. In real life, too, the background environment in which disease events happen may change too quickly for *any* model to capture it – a model that describes the past two years very accurately may suddenly fail tomorrow. (Doubters should ask any investors in the NASDAQ 'new-economy' share market, who saw their whole paradigm change in early 2000.)

The case for neural networks in disease forecasting

Although there are some useful tools available to the disease surveillance practitioner, none yet combines the ideal attributes of being highly accurate, timely, not needing frequent 'tweaking' by the operator, and (most importantly) able to make use of past outbreaks in the data set. Artificial neural networks have shown themselves useful in some other time series forecasting problems, and can theoretically overcome many of the problems outlined above, but until now they have not been tried in disease surveillance.

Several essential questions need to be answered. Firstly, can a disease surveillance data set be suitably prepared to emphasise its deterministic components, yet still maintaining a geographical or temporal scale that is useful to a practical surveillance and outbreak control programme? What other determinants of disease incidence and distribution can be measured and included in such a model? Can a neural network design be found that can accurately model the recent past and predict far enough ahead to enable useful disease control interventions? Finally, can such a network generalise the relationships in the data set sufficiently well to continue to make

accurate predictions for a time window beyond the end of its training data? The next three chapters of this thesis tackle these questions.

Chapter Four: Recent weather as a predictor of gastroenteritis in Melbourne, Australia

The underlying objective of the work detailed in this and the following chapter was the creation of a machine algorithm to forecast daily gastroenteritis numbers, to act as part of an early-warning system for outbreaks of water-borne disease. If a neural network could accurately forecast disease numbers, then it would give valuable time for public health authorities to take action to avoid or mitigate the outbreak.

A number of aspects of the problem need to be considered. What proxy for gastroenteritis disease can be reliably and quickly acquired every day of the year without excessive cost or new infrastructure? What set of inputs most accurately and reliably predicts future disease? On what geographical and temporal scale can a neural network distinguish the underlying deterministic element of the time series from the inevitable random element? Can predictions be made for single postcode? A Local Government Area? Only for the whole city? Can daily counts be used or only weekly ones? What other readily available variables might improve predictive accuracy?

Proxy measures of gastroenteritis

Attempts were made to acquire data relating directly to disease episodes (such as the Victorian Emergency Minimum Dataset, held by the Victorian State Government's Department of Human Services), but these data sets are still relatively new, and the number of days of reliable data available proved to be inadequate. Few Melbourne pharmacies have a computerised stock or sales system, so that pharmacy sales of anti-diarrhoeal agents, a promising potential indicator, was unavailable. Even the data set that was acquired (requests for faecal analysis according to the appropriate Medicare

item numbers) from the national Health Insurance Commission included only four years of data, which is far less than initially expected. The data actually used probably represent the minimum from which any kind of useful analysis could be done.

Figure 4-1 illustrates the metropolitan Melbourne Local Government Areas (LGA) included in this study. The figure is based on a MapMovie screenshot (see Appendix One) for the only known large outbreak in the dataset, with a list of LGA names added.

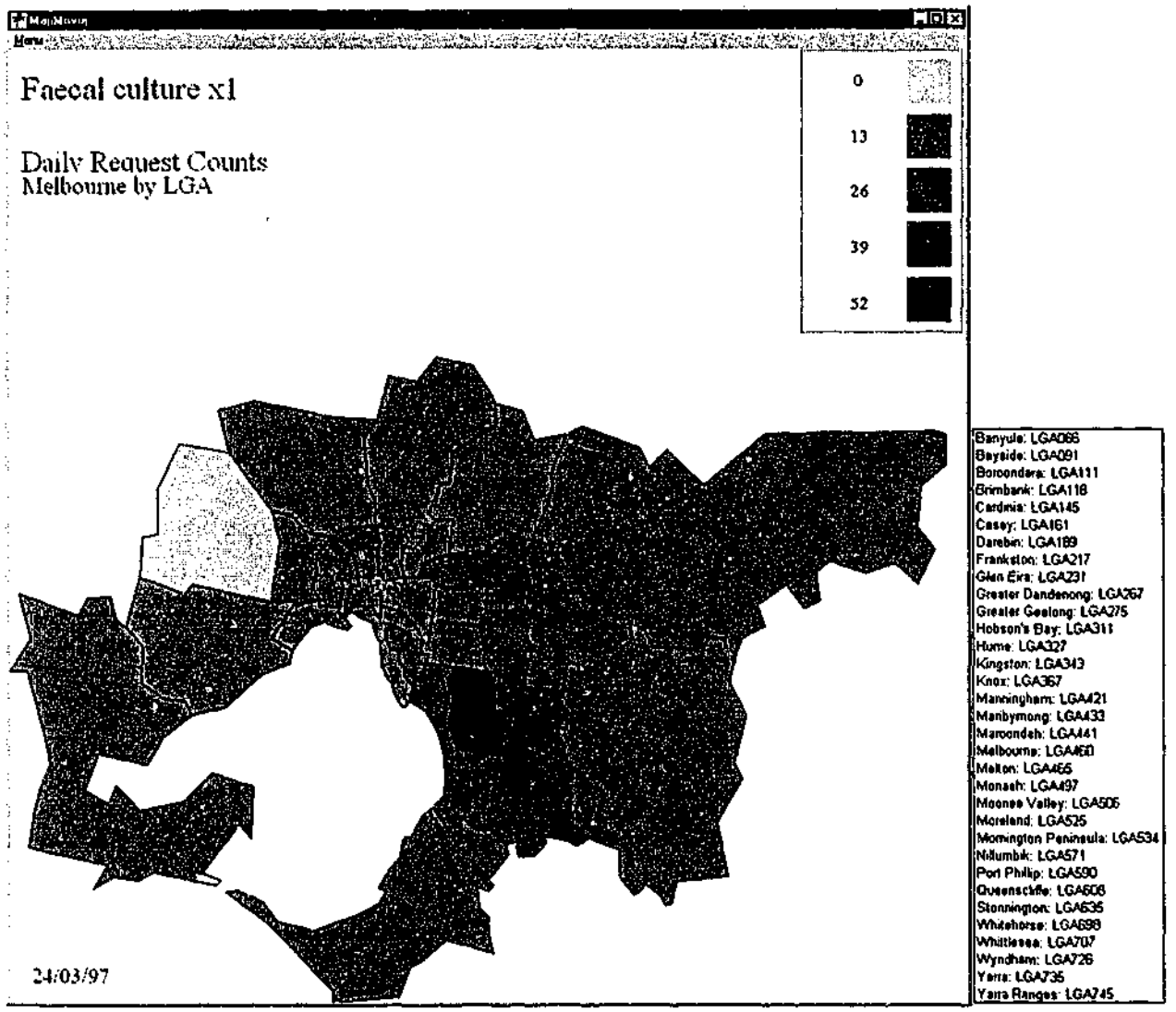


Figure 4-1: Local Government Areas (LGA) included in the study.

Other inputs to improve predictive accuracy

Having decided (mostly for the practical reasons outlined above) to use requests for faecal analysis as the proxy for gastroenteritis, the next question was what other readily-available inputs might improve the accuracy of the models. Initial analysis of the data sets suggested that the most influential predictors apart from previous requests numbers were social factors (day of the week effects and public and school holidays), which affect care-seeking behaviour and laboratory availability more than they affect disease incidence.

We hypothesised that the number of requests for faecal analysis is influenced by the true incidence of gastroenteritis, as well as availability of medical and laboratory services, and patients' health care seeking behaviour. The incidence of gastroenteritis is in turn affected by short-term weather and possibly by longer term seasonal influences, as is health-care seeking behaviour. This study aimed to explore the contribution of each of these to the ability of an artificial neural network model to forecast numbers of requests up to one week into the future.

The relationships and interactions between the factors are complex. Artificial neural network models were chosen because they make no assumptions about the independence or statistical distribution of the input variables or the linearity of the relationships between independent and dependent variables (Hinton, 1992).

Preliminary experiments

A large number of preliminary experiments were done, using different aggregations of the data (Local Government Area vs whole city, different degrees of smoothing) and

network architectures (numbers of hidden layer neurons and training parameters). The data presented here are the most illustrative.

Are the seasons more than just the weather?

A related question is the extent to which the seasonal variation evident in this and many other disease incidence data sets is simply due to short-term variation in the weather, via direct effects such as the greater growth of pathogenic organisms in food at higher ambient temperatures. Or are there truly seasonal factors independent of the weather *per se*?

Seasonal variation is a strong feature of the epidemiology of many infectious diseases. The precise patterns vary with the disease, the latitude and the hemisphere, but the consistent feature is a predictable pattern of peaks and troughs in incidence related to the time of year (Dowell, 2001, Cook et al., 1990). The presence of this seasonal variation influences research into the epidemiology of infectious diseases in several ways. Some researchers seek to eliminate the broad effects of the seasons in order to study the temporal variations caused by other influences (Pascual et al., 2000, Schwartz et al., 1997, Schwartz et al., 2000). Others are interested in the seasonal effects themselves, and seek a precise way of modelling them (Fine and Clarkson, 1982, Dowell, 2001).

Methods

The contribution of the various inputs to neural network models was assessed both separately and in combination. First a set of models was made using the faecal analysis request data with only one other input, comparing their forecasting accuracy within the same time window. In a second series of experiments a set of models was

made in which each successive model contained all the previous inputs plus one more weather or holiday input.

Data acquisition and pre-processing

Counts of requests for faecal analysis

Retrospective daily records were obtained from the Australian Health Insurance Commission (which administers Medicare, the national health insurance scheme), of the number of bills received from pathology laboratories for a defined list of Medicare item numbers. These were Item number 69336: Faecal microscopy done once during an episode of illness ('Microscopy x1' or 'M1'), Item number 69339: Faecal microscopy done twice during an episode of illness ('Microscopy x2' or 'M2'), Item number 69342: Faecal microscopy done three times during an episode of illness ('Microscopy x3' or 'M3'), Item number 69345: Faecal culture done once during an episode of illness ('Culture x1' or 'C1'), Item number 69348: Faecal culture done twice during an episode of illness ('Culture x2' or 'C2') and Item number 69351: Faecal culture done three times during an episode of illness ('Culture x3' or 'C3'). In addition two composite indicators were created, by summing the total number of cultures done per day ('Culture total') and the total number of microscopies done per day ('Microscopy total').

All pathology laboratories not associated with public hospitals were included. (Public hospital laboratories do not send accounts via the HIC). Daily counts were available for the period from July 1, 1996 to June 30, 2000 for each of these six time series.

The resulting raw time series all showed considerable day-to-day random variation, and also a strong day-of-week effect (in particular there were fewer requests on Saturdays and Sundays). Initial experiments with the raw data suggested there was too much random variation for a neural network to make useful forecasts, so the networks were trained and tested using a processed version of the original time series. The processed series were made up of the mean number of requests in the seven-day period ending on the given day. This removed the day-of-week effect, as every average included one Saturday and one Sunday. It also substantially reduced the day-to-day variation. An asymmetrical moving average was used to avoid the end-effects that would have limited forecasts of a symmetrical average. A total of 1425 days of observations were available for model creation and testing in the processed time series.

School and public holidays

Preliminary examination of these time series showed that there was a considerable decline in numbers of requests during public and school holidays, so a variable was added to represent them. For these variables a value of zero denoted no holiday, and an integer value from one to seven denoted the number of days in each seven-day moving average that were school holidays or public holidays.

Recent weather

In addition, as we hypothesised that recent weather might affect both disease incidence and care-seeking behaviour, data from the central Melbourne weather station were obtained from the Victorian Bureau of Meteorology. These comprised the daily total precipitation in millimetres, the minimum and maximum temperatures in degrees Celsius, and the wind speed (in kilometres per hour) and compass direction

at 3 p.m. The wind speed and direction were combined into a single vector, and represented as a pair of polar coordinates. Empirical tests showed that the best representation of these local weather data for network generalisation were as asymmetrical seven-day moving averages, just as for the faecal request data. A number of tests with different inputs (described in the previous chapter) confirmed the usefulness of each of the weather inputs in improving model fit.

Day of the year

To represent possible seasonal effects independent of the recent weather, the networks also received an input representing the day of the year (from 1 to 365 or 366). To avoid the artefact of an apparent major change from the end of one year to the beginning of the next, the day was converted to an angle, with a full circle being equal to the full year. Each day could then be given a unique value by calculating the sine and cosine of the day angle.

Leads and lags

Not only were all the data smoothed as asymmetric seven-day moving averages, but each type of input was presented to the networks together with some of its lags or leads. (In time series terminology the lags are the values from the same series for the preceding days, and the leads are the values for the following days. Obviously leads can only be used for training outputs, or else where the time series can be accurately predicted from other sources.)

The faecal analysis requests were thus always presented to a network as eight inputs: the day in question and seven of its lags – i.e. the values for the previous seven days.

Each of the weather inputs was always presented with fifteen of its lags, making sixteen inputs to the network. The day of the year was presented as a single pair of inputs (the sine and cosine as described above).

School and public holidays are defined years in advance, so the inputs representing these were presented with seven of their *leads*.

Henceforth in this chapter, wherever an input is referred to it should be understood to mean that day's values and the other leads or lags described above. The different inputs are commonly abbreviated using the following scheme:

Requests = Time series of requests for faecal analysis of the specified type: smoothed as an asymmetrical 7-day moving average and presented as the current day's value and seven lags.

C1, C2 or C3 = Culture x1, Culture x2 or Culture x3 during a single episode of illness.

M1, M2 or M3 = Microscopy x1, Microscopy x2 or Microscopy x3 during a single episode of illness.

Holidays (SchoolInf and PublicInf) = Two time series, based on the number of days in that day's moving average of requests which were school (SchoolInf) or public holidays (PublicInf): presented to the networks as that day's value and seven leads.

MinTemp = Minimum temperature for the 24-hour period: presented with 15 lags.

MaxTemp = Maximum temperature for the 24-hour period: presented with 15 lags.

Rainfall = Total precipitation for the 24-hour period: presented with 15 lags.

WindX = The x-axis component of polar coordinates based on a vector of wind speed and direction at 3 pm: presented with 15 lags.

WindY = The y-axis component of polar coordinates based on a vector of

		wind speed and direction at 3 pm: presented with 15 lags.
Day (SinDay and CosDay)	=	Two inputs: the Sine (SinDay) and Cosine (CosDay) of the angle representing the day of the year. These were always presented together.

Neural network architecture, training and testing

A commercial neural network package, NeuroShell2 (Ward Systems Group Inc, 2000), was used to create two separate sets of multi-layer feed-forward neural network models for each of the eight faecal analysis time series. All the networks used linear input scaling from –1 to +1, and all the hidden layer neurons used the logistic function as their activation function. All were trained using simple back-propagation of errors (Rumelhart, 1986), with learning rate and momentum terms of 0.1. Each model was trained to simultaneously predict the number of requests for the next seven leads (i.e. the number of requests the next, day, the day after, three days hence and so on up to seven days). There were thus seven output neurons in each model.

Training was done with a randomly selected 85% subset of the processed time series. The remaining patterns were used as a validation set. 'Early stopping' of training was used – for every 200 learning events (i.e. when 200 learning patterns had been presented to the network for training), the partly trained network was applied to the reserved validation set and its mean squared error (MSE) calculated. Training was stopped when 20,000 learning events had occurred without further improvement in the MSE for the validation set. To avoid over-training, the final trained network was the one that had given the lowest value for the MSE on the validation set.

The networks' ability to generalise and forecast the coming week's request numbers was assessed using the coefficient of multiple determination (R^2) for each of the outputs. There are a number of variants of R^2 – this one compares the accuracy of a single network output with a simple model in which the prediction is always the mean of all the values. It is calculated by taking the sum of the squared differences between each observed and predicted value, and dividing that by the sum of squared differences between each observed value and the mean of all observed values. The resulting ratio is subtracted from 1. Perfect forecasts would thus give an R^2 value of 1.0, while values less than 1 denote progressively poorer forecasting, and values below zero mean forecasts worse than simply predicting the mean value every time.

Models assessing single inputs

In the first series of experiments each set of inputs was added by itself to a baseline model with only the request inputs. These models thus had between eight and 24 inputs. All had 50 hidden neurons – this number was estimated by a common rule of thumb (half the sum of inputs and outputs, plus the square root of the number of training patterns (Smith, 1999)) for a typical network in this group, and then confirmed as reasonable by empirical testing.

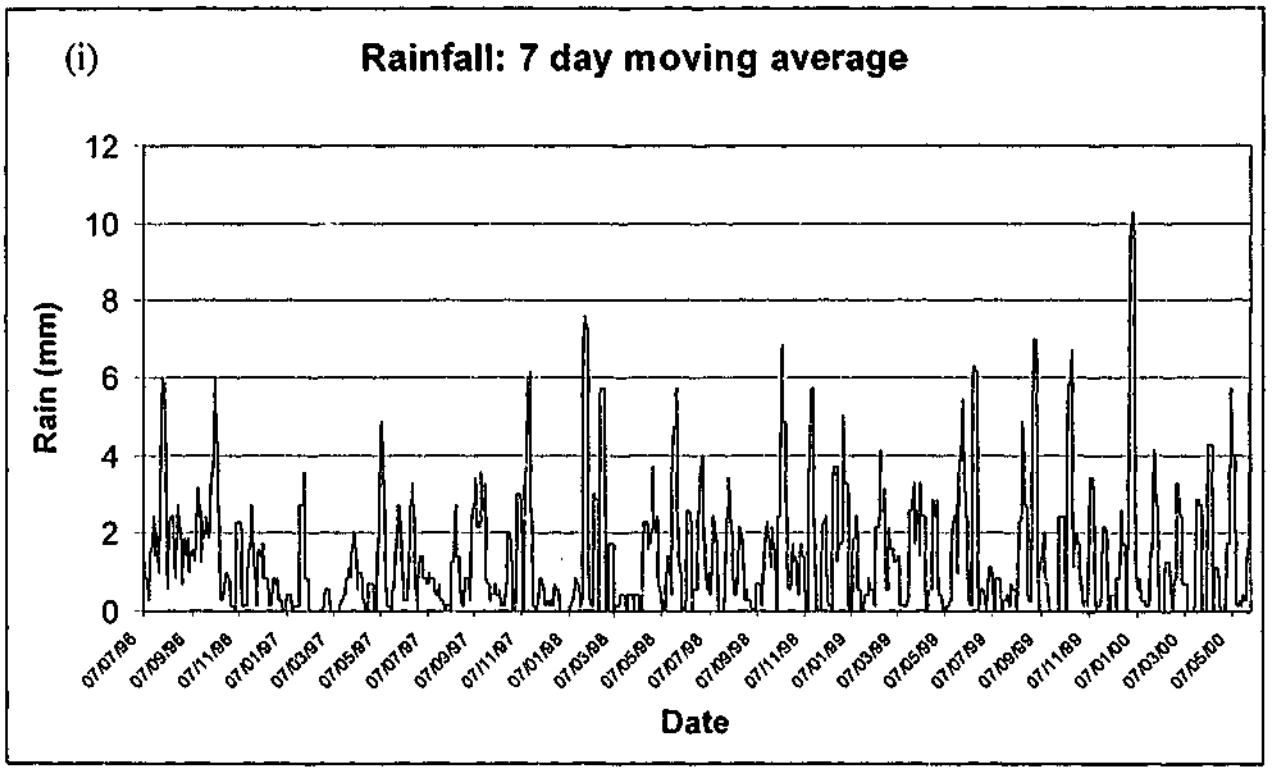
Models adding one input to the all the previous ones

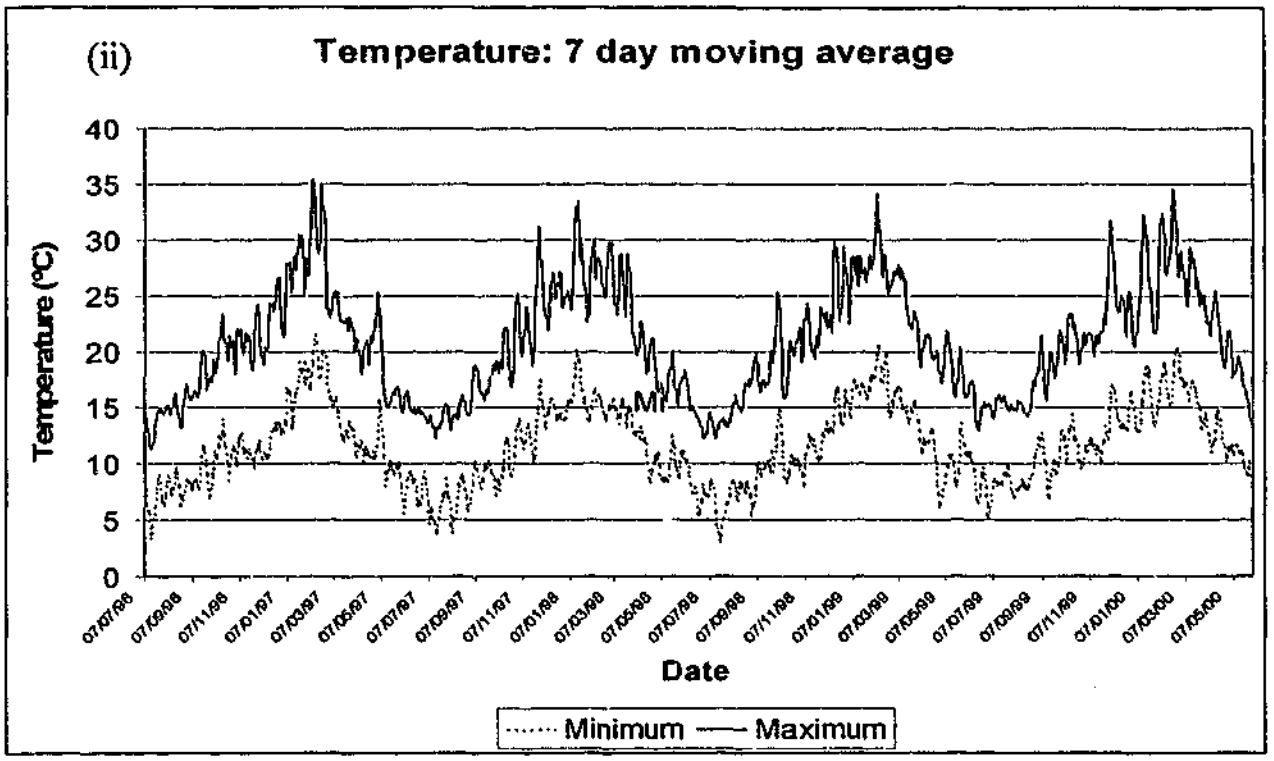
In this series of experiments sets of models were made, each with progressively more inputs than the one before. All used the faecal request numbers and holiday inputs, and the others were added in the following order: Rainfall, MinTemp, MaxTemp, WindX, WindY and then finally day of the year. The baseline model had eight inputs, and the most complex model 106. All had 70 neurons in the single hidden layer.

Results

The weather patterns for the study period were typical for the Melbourne area, with seasonal increases in maximum and minimum temperature, and variable rainfall somewhat higher in the winter months (Bureau of Meteorology Australia, 2001). The processed weather series are plotted in Figure 4-2.

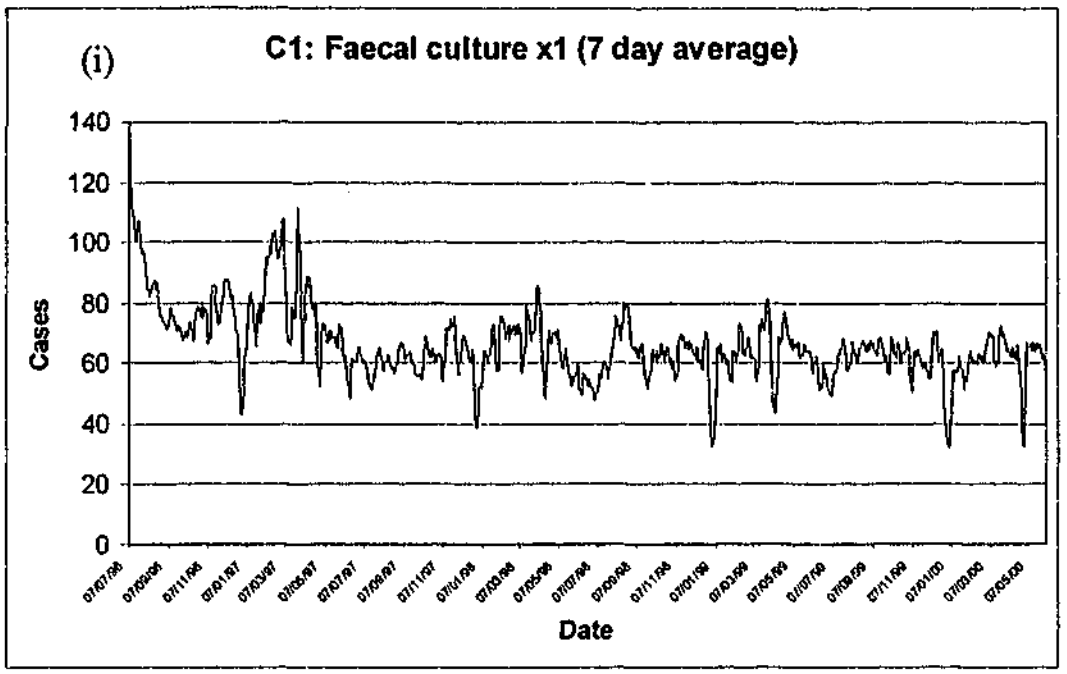
Figure 4-2 (i-ii): Smoothed time series of rainfall and temperature.

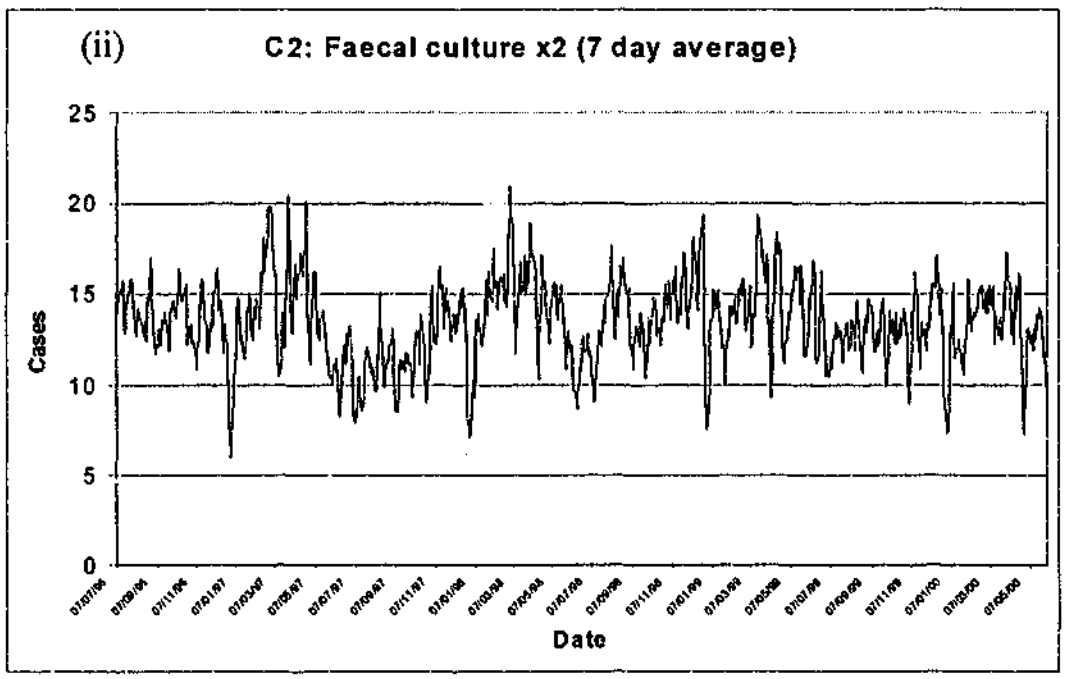


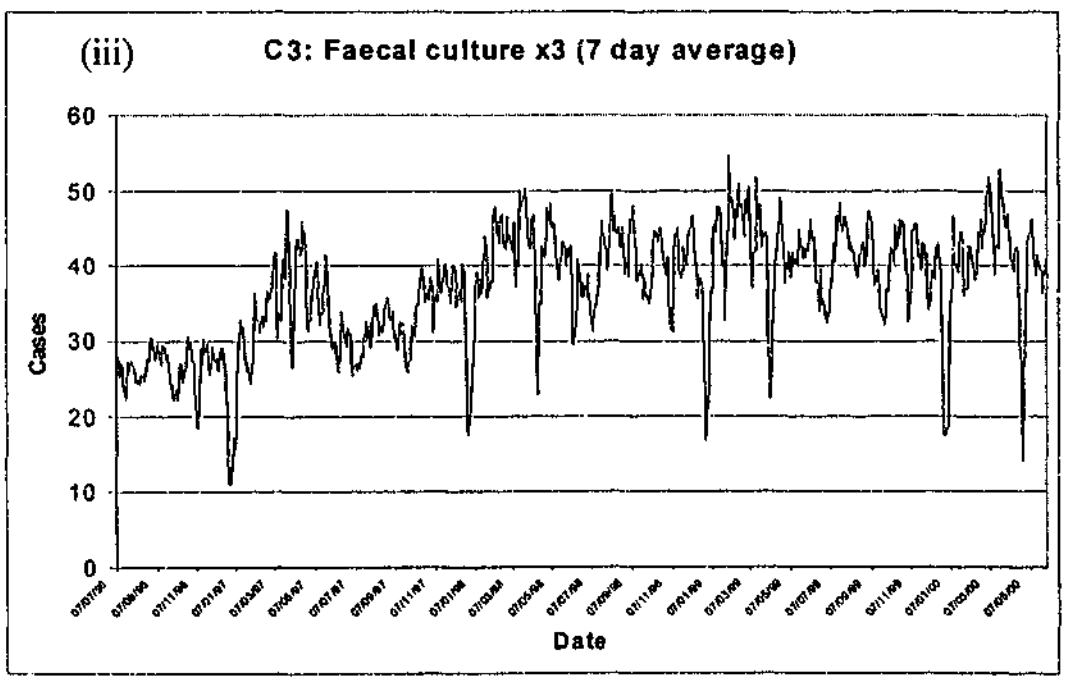


Requests for faecal analysis, plotted in Figure 4-3 as they were used in the network models (asymmetrical 7-day moving averages), show a degree of seasonal fluctuation, though partly obscured by strong effects of school and public holidays, which all cause substantial decreases in request rates. There was little evidence of rebound increases after holidays. There are no clear secular trends, although some of the series do have odd rises or falls without obvious cause.

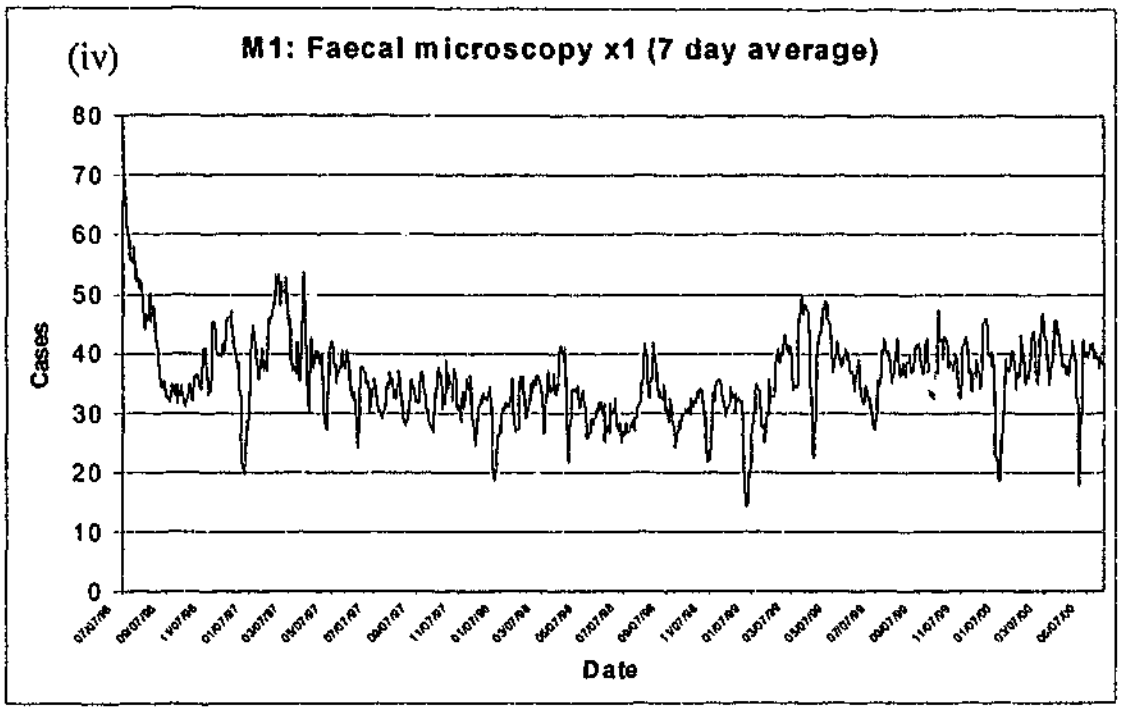
Figure 4-3(i-vi): The processed time series for the faecal analysis request indicators.



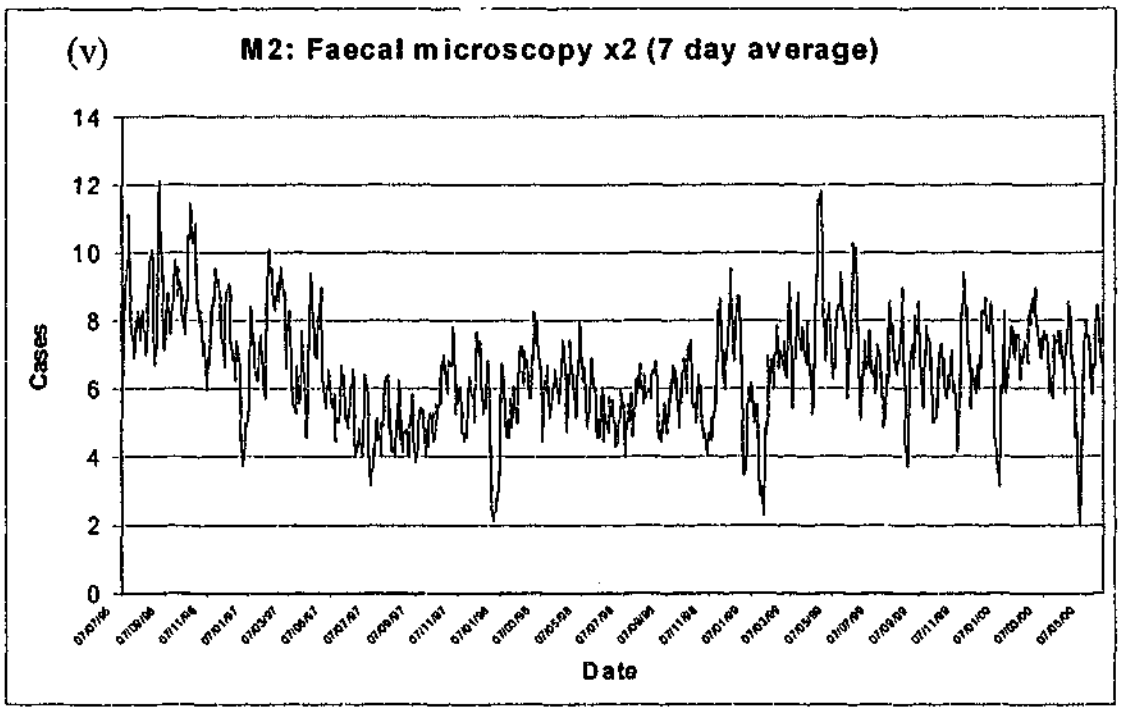




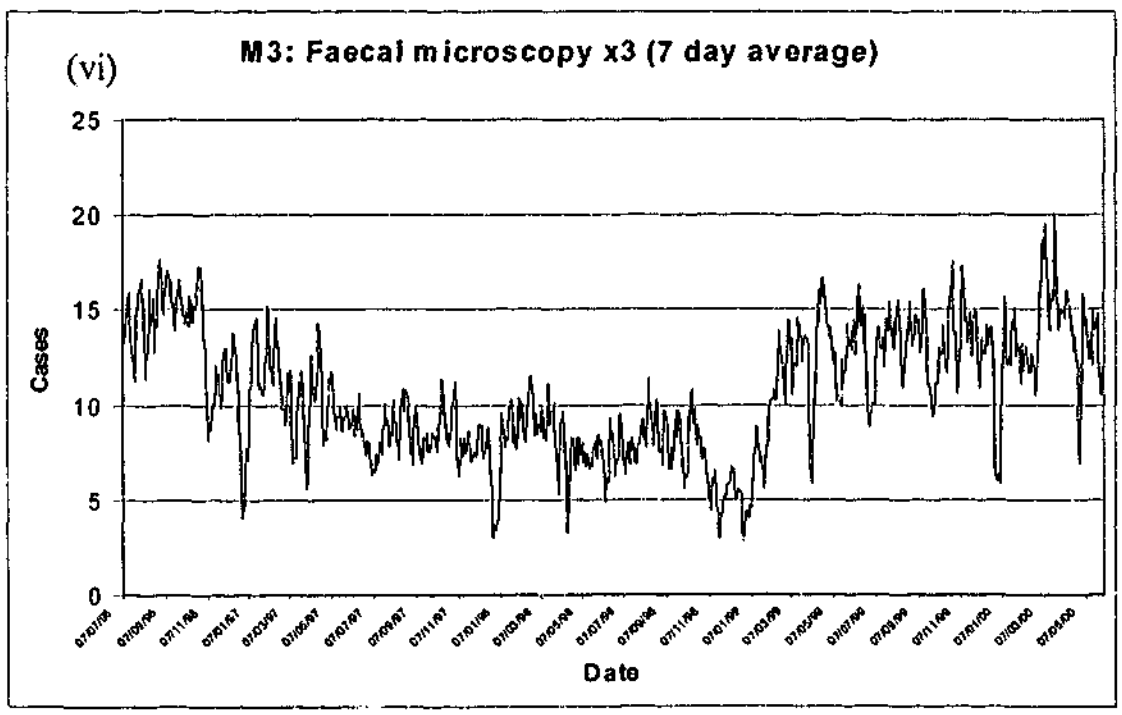
(iv)


M1: Faecal microscopy x1 (7 day average)
Cases
0
10
20
30
40
50
60
70
80
Date

(v)


M2: Faecal microscopy x2 (7 day average)
Cases
0
2
4
6
8
10
12
14
Date

(vi)


M3: Faecal microscopy x3 (7 day average)
Cases
0
5
10
15
20
25
Date

Models assessing single inputs

Table 4-1 shows the coefficient of multiple determination (R^2) calculated for the randomly selected 15% validation sets, together with the percentage difference in relation to the baseline model. Public holiday information clearly made the greatest improvement to the baseline requests model. School holiday information was the second strongest influence. Each of the weather inputs also made a moderate contribution to at least one of the models, although there were also some models in which R^2 actually decreased with the addition of MaxTemp, MinTemp, Rainfall or WindY. For the 'Total' models MaxTemp and MinTemp are the strongest contributors among the weather inputs. Except for the 'Culture total' models, Day of year rates only with the second rank of individual weather inputs.

Table 4-1: Models assessing single inputs added to the baseline model. Coefficient of multiple determination (R^2) and percentage change in relation to the baseline requests-only model, for the validation sets.

(a) Culture x1 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.93 (0.0)	0.84 (0.0)	0.77 (0.0)	0.69 (0.0)	0.64 (0.0)	0.53 (0.0)	0.43 (0.0)	0.69 (0.0)
Requests, PublicInf	0.94 (1.4)	0.90 (6.5)	0.84 (9.8)	0.81 (16.7)	0.76 (19.6)	0.67 (26.5)	0.64 (46.5)	0.79 (18.1)
Requests, SchoolInf	0.92 (-0.3)	0.85 (0.5)	0.79 (2.6)	0.73 (5.8)	0.69 (8.2)	0.59 (12.2)	0.51 (18.0)	0.73 (6.7)
Requests, Rainfall	0.92 (-0.8)	0.84 (-0.6)	0.76 (-1.2)	0.67 (-2.2)	0.61 (-4.1)	0.51 (-4.0)	0.42 (-2.5)	0.68 (-2.2)
Requests, MinTemp	0.92 (-0.8)	0.85 (0.4)	0.77 (0.7)	0.70 (1.6)	0.65 (2.5)	0.55 (3.1)	0.45 (4.5)	0.70 (1.7)
Requests, MaxTemp	0.92 (-0.5)	0.85 (1.2)	0.79 (2.3)	0.72 (4.8)	0.68 (7.5)	0.60 (12.5)	0.51 (16.8)	0.72 (6.4)
Requests, WindX	0.92 (-0.3)	0.85 (1.0)	0.78 (1.2)	0.71 (2.8)	0.66 (3.6)	0.57 (7.3)	0.48 (9.7)	0.71 (3.6)
Requests, WindY	0.92 (-0.6)	0.84 (-0.2)	0.77 (0.0)	0.69 (0.4)	0.64 (1.1)	0.53 (0.5)	0.44 (0.6)	0.69 (0.2)
Requests, Day	0.93 (0.3)	0.85 (1.0)	0.77 (0.7)	0.71 (2.9)	0.67 (6.1)	0.58 (9.5)	0.50 (15.1)	0.72 (5.1)

(b) Culture x2 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.83 (0.0)	0.63 (0.0)	0.48 (0.0)	0.37 (0.0)	0.28 (0.0)	0.12 (0.0)	0.07 (0.0)	0.40 (0.0)
Requests, PublicInf	0.86 (3.1)	0.72 (12.9)	0.60 (26.4)	0.51 (39.1)	0.42 (52.5)	0.27 (114.6)	0.25 (272.1)	0.52 (74.4)
Requests, SchoolInf	0.84 (1.6)	0.65 (1.8)	0.53 (10.3)	0.46 (25.2)	0.40 (45.7)	0.31 (150.5)	0.22 (237.9)	0.49 (67.6)
Requests, Rainfall	0.83 (-0.6)	0.63 (-1.1)	0.49 (3.1)	0.41 (11.2)	0.34 (21.9)	0.21 (66.6)	0.14 (107.0)	0.43 (29.7)
Requests, MinTemp	0.84 (0.5)	0.63 (-1.5)	0.49 (2.1)	0.38 (3.4)	0.30 (8.7)	0.16 (27.9)	0.10 (48.5)	0.41 (12.8)
Requests, MaxTemp	0.83 (-0.4)	0.63 (-1.4)	0.50 (5.4)	0.40 (9.6)	0.32 (15.2)	0.19 (53.4)	0.14 (106.1)	0.43 (26.8)
Requests, WindX	0.82 (-0.8)	0.64 (0.8)	0.48 (1.3)	0.38 (3.3)	0.30 (9.5)	0.17 (34.3)	0.11 (65.9)	0.42 (16.3)
Requests, WindY	0.84 (0.6)	0.63 (-1.1)	0.48 (0.0)	0.36 (-2.1)	0.27 (-2.8)	0.13 (6.2)	0.05 (-17.6)	0.39 (-2.4)
Requests, Day	0.84 (1.7)	0.64 (1.4)	0.51 (7.2)	0.42 (14.9)	0.33 (20.0)	0.21 (71.0)	0.13 (95.6)	0.44 (30.3)

(c) Culture x3 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.93 (0.0)	0.86 (0.0)	0.80 (0.0)	0.73 (0.0)	0.64 (0.0)	0.58 (0.0)	0.50 (0.0)	0.72 (0.0)
Requests, PublicInf	0.95 (1.5)	0.92 (6.1)	0.87 (9.4)	0.85 (16.1)	0.77 (20.7)	0.74 (27.8)	0.72 (45.7)	0.83 (18.2)
Requests, SchoolInf	0.93 (0.1)	0.88 (2.0)	0.83 (3.9)	0.77 (5.9)	0.68 (6.2)	0.63 (9.0)	0.58 (15.8)	0.76 (6.1)
Requests, Rainfall	0.92 (-1.3)	0.86 (0.0)	0.81 (1.5)	0.74 (1.3)	0.66 (4.1)	0.61 (4.7)	0.55 (10.2)	0.74 (2.9)
Requests, MinTemp	0.93 (-0.5)	0.87 (0.1)	0.79 (-1.1)	0.73 (0.5)	0.64 (-0.2)	0.58 (-0.4)	0.50 (0.4)	0.72 (-0.2)
Requests, MaxTemp	0.93 (-0.5)	0.86 (-0.2)	0.79 (-0.7)	0.74 (1.8)	0.65 (2.5)	0.60 (2.7)	0.54 (8.1)	0.73 (2.0)
Requests, WindX	0.93 (-0.2)	0.86 (-0.3)	0.80 (-0.1)	0.73 (0.0)	0.64 (0.5)	0.58 (0.3)	0.50 (1.5)	0.72 (0.3)
Requests, WindY	0.93 (-0.2)	0.87 (0.3)	0.80 (0.7)	0.74 (1.3)	0.66 (3.9)	0.61 (4.9)	0.53 (7.4)	0.74 (2.6)
Requests, Day	0.93 (0.2)	0.87 (0.6)	0.81 (1.7)	0.74 (1.2)	0.65 (2.5)	0.61 (4.2)	0.53 (7.1)	0.74 (2.5)

(d) Culture total R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.89 (0.0)	0.8 (0.0)	0.67 (0.0)	0.5 (0.0)	0.45 (0.0)	0.3 (0.0)	0.19 (0.0)	0.55 (0.0)
Requests, PublicInf	0.93 (4.0)	0.9 (13.3)	0.82 (22.6)	0.8 (41.4)	0.71 (58.2)	0.6 (95.3)	0.58 (207.9)	0.76 (63.2)
Requests, SchoolInf	0.88 (-1.2)	0.8 (3.3)	0.73 (8.4)	0.6 (17.4)	0.57 (27.0)	0.5 (53.1)	0.40 (115.2)	0.64 (31.9)
Requests, Rainfall	0.87 (-2.1)	0.8 (-0.4)	0.67 (-0.4)	0.6 (1.3)	0.46 (3.1)	0.3 (8.3)	0.22 (19.7)	0.56 (4.2)
Requests, MinTemp	0.89 (-0.3)	0.8 (0.5)	0.68 (0.8)	0.6 (3.9)	0.48 (7.1)	0.3 (10.8)	0.24 (30.2)	0.57 (7.6)
Requests, MaxTemp	0.88 (-1.0)	0.8 (0.1)	0.68 (1.2)	0.6 (5.6)	0.50 (11.7)	0.4 (23.5)	0.30 (61.9)	0.59 (14.7)
Requests, WindX	0.87 (-1.8)	0.8 (-0.3)	0.67 (0.2)	0.6 (1.7)	0.46 (2.8)	0.3 (7.6)	0.20 (6.8)	0.55 (2.4)
Requests, WindY	0.89 (-0.6)	0.8 (0.5)	0.68 (1.5)	0.6 (3.3)	0.48 (7.6)	0.4 (14.0)	0.23 (24.3)	0.57 (7.2)
Requests, Day	0.90 (1.4)	0.8 (4.3)	0.75 (11.5)	0.7 (23.0)	0.63 (40.4)	0.6 (84.1)	0.51 (172.8)	0.69 (48.2)

(e) Microscopy x1 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.91 (0.0)	0.84 (0.0)	0.76 (0.0)	0.67 (0.0)	0.61 (0.0)	0.51 (0.0)	0.40 (0.0)	0.67 (0.0)
Requests, PublicInf	0.92 (0.9)	0.88 (5.2)	0.83 (9.3)	0.78 (16.3)	0.72 (19.5)	0.65 (28.1)	0.60 (48.6)	0.77 (18.3)
Requests, SchoolInf	0.92 (1.1)	0.85 (0.8)	0.79 (3.6)	0.73 (8.1)	0.68 (11.6)	0.60 (18.3)	0.52 (28.4)	0.73 (10.3)
Requests, Rainfall	0.91 (-0.5)	0.83 (-0.7)	0.76 (-0.3)	0.67 (-0.9)	0.59 (-1.9)	0.51 (0.8)	0.42 (4.8)	0.67 (0.2)
Requests, MinTemp	0.92 (0.5)	0.84 (-0.1)	0.76 (-0.1)	0.68 (0.8)	0.63 (3.4)	0.54 (6.3)	0.47 (16.3)	0.69 (3.9)
Requests, MaxTemp	0.91 (0.0)	0.83 (-0.6)	0.76 (-0.7)	0.69 (2.2)	0.64 (6.4)	0.57 (11.9)	0.50 (25.4)	0.70 (6.4)
Requests, WindX	0.92 (1.0)	0.84 (-0.1)	0.76 (0.2)	0.67 (-0.6)	0.61 (0.7)	0.52 (2.5)	0.42 (3.8)	0.68 (1.1)
Requests, WindY	0.91 (0.2)	0.84 (-0.3)	0.77 (0.7)	0.69 (2.1)	0.64 (5.5)	0.55 (7.5)	0.46 (15.7)	0.69 (4.5)
Requests, Day	0.92 (1.0)	0.84 (0.3)	0.76 (0.2)	0.68 (1.2)	0.62 (2.6)	0.52 (3.3)	0.42 (5.3)	0.68 (2.0)

(f) Microscopy x2 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.88 (0.0)	0.78 (0.0)	0.68 (0.0)	0.59 (0.0)	0.49 (0.0)	0.39 (0.0)	0.33 (0.0)	0.59 (0.0)
Requests, PublicInf	0.89 (0.1)	0.81 (3.8)	0.75 (9.5)	0.68 (16.2)	0.61 (24.3)	0.53 (33.8)	0.48 (44.2)	0.68 (18.9)
Requests, SchoolInf	0.88 (-0.1)	0.78 (0.7)	0.71 (4.4)	0.63 (7.2)	0.54 (10.1)	0.44 (11.7)	0.39 (17.0)	0.62 (7.3)
Requests, Rainfall	0.88 (-0.8)	0.76 (-1.9)	0.69 (1.4)	0.61 (3.5)	0.53 (9.2)	0.44 (11.6)	0.37 (10.3)	0.61 (4.8)
Requests, MinTemp	0.89 (0.3)	0.78 (1.2)	0.70 (2.1)	0.61 (4.0)	0.51 (4.6)	0.43 (8.2)	0.38 (14.0)	0.61 (4.9)
Requests, MaxTemp	0.88 (-0.2)	0.78 (0.8)	0.69 (0.7)	0.60 (2.0)	0.50 (3.0)	0.43 (9.1)	0.38 (15.1)	0.61 (4.3)
Requests, WindX	0.87 (-1.3)	0.76 (-2.3)	0.68 (-0.3)	0.59 (0.3)	0.50 (1.6)	0.41 (3.3)	0.34 (2.4)	0.59 (0.5)
Requests, WindY	0.89 (0.2)	0.79 (2.4)	0.72 (5.3)	0.64 (8.5)	0.56 (14.9)	0.49 (24.4)	0.41 (23.0)	0.64 (11.3)
Requests, Day	0.88 (-0.3)	0.77 (-0.1)	0.69 (0.7)	0.59 (1.1)	0.50 (1.4)	0.40 (2.5)	0.34 (3.3)	0.60 (1.2)

(g) Microscopy x3 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.94 (0.0)	0.87 (0.0)	0.80 (0.0)	0.74 (0.0)	0.69 (0.0)	0.64 (0.0)	0.58 (0.0)	0.75 (0.0)
Requests, PublicInf	0.95 (0.1)	0.90 (3.2)	0.85 (6.6)	0.82 (10.5)	0.76 (11.2)	0.73 (13.6)	0.69 (18.9)	0.81 (9.2)
Requests, SchoolInf	0.94 (-0.1)	0.88 (1.4)	0.82 (2.5)	0.77 (3.6)	0.72 (4.0)	0.68 (5.3)	0.62 (6.3)	0.78 (3.3)
Requests, Rainfall	0.93 (-1.4)	0.87 (0.3)	0.81 (0.9)	0.75 (1.0)	0.69 (0.8)	0.66 (2.6)	0.60 (3.1)	0.76 (1.0)
Requests, MinTemp	0.93 (-1.4)	0.87 (-0.4)	0.80 (-0.1)	0.74 (0.3)	0.69 (-0.1)	0.65 (0.8)	0.59 (1.0)	0.75 (0.0)
Requests, MaxTemp	0.93 (-1.5)	0.87 (-0.4)	0.80 (-0.1)	0.74 (0.2)	0.69 (-0.1)	0.64 (0.4)	0.59 (0.6)	0.75 (-0.1)
Requests, WindX	0.94 (-0.7)	0.87 (0.0)	0.81 (0.8)	0.75 (1.2)	0.69 (0.8)	0.66 (2.3)	0.60 (3.7)	0.76 (1.2)
Requests, WindY	0.94 (-0.6)	0.88 (0.7)	0.81 (0.8)	0.75 (1.5)	0.70 (1.5)	0.66 (2.1)	0.60 (2.9)	0.76 (1.3)
Requests, Day	0.95 (0.1)	0.88 (0.5)	0.81 (0.6)	0.74 (0.0)	0.70 (1.3)	0.65 (1.7)	0.59 (1.0)	0.76 (0.8)

(h) Microscopy total R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests	0.94 (0.0)	0.87 (0.0)	0.80 (0.0)	0.72 (0.0)	0.66 (0.0)	0.58 (0.0)	0.48 (0.0)	0.72 (0.0)
Requests, PublicInf	0.95 (1.5)	0.91 (5.2)	0.87 (8.4)	0.84 (15.6)	0.79 (18.7)	0.72 (25.7)	0.69 (42.2)	0.82 (16.8)
Requests, SchoolInf	0.94 (0.1)	0.88 (1.4)	0.82 (3.5)	0.77 (6.3)	0.71 (7.2)	0.64 (10.9)	0.56 (15.9)	0.76 (6.5)
Requests, Rainfall	0.92 (-1.5)	0.86 (-1.1)	0.79 (-0.7)	0.72 (-0.9)	0.66 (0.0)	0.58 (1.7)	0.50 (3.0)	0.72 (0.1)
Requests, MinTemp	0.93 (-0.4)	0.87 (0.0)	0.80 (0.0)	0.73 (0.9)	0.68 (3.0)	0.62 (8.3)	0.56 (16.4)	0.74 (4.0)
Requests, MaxTemp	0.93 (-0.4)	0.87 (0.1)	0.81 (0.8)	0.74 (2.6)	0.70 (6.0)	0.64 (11.3)	0.59 (21.2)	0.75 (5.9)
Requests, WindX	0.93 (-1.3)	0.86 (-0.8)	0.80 (-0.1)	0.72 (-0.1)	0.67 (0.4)	0.58 (1.4)	0.49 (1.6)	0.72 (0.2)
Requests, WindY	0.94 (-0.1)	0.87 (0.2)	0.80 (0.0)	0.73 (1.5)	0.70 (5.0)	0.62 (8.0)	0.54 (11.5)	0.74 (3.7)
Requests, Day	0.94 (0.4)	0.87 (0.4)	0.81 (0.9)	0.74 (2.1)	0.68 (2.6)	0.60 (4.3)	0.51 (5.4)	0.74 (2.3)

Models adding one further set of inputs to all the previous ones

Table 4-2 shows the coefficient of multiple determination (R^2) calculated for the randomly selected 15% validation sets, together with the percentage difference for each new set of inputs in relation to the previous model.

The baseline models using only request numbers and holiday information were poor at forecasting future request numbers, even within the time window on which the networks were trained. With the addition of increasing numbers of weather inputs there was a consistent tendency to improve forecasting, and the relative improvement increased with increasing forecast range. The most complex models provided very good forecasts up to seven days in advance, with a majority achieving mean R^2 values of 0.94 or better. Models using all the weather inputs accou·ted for nearly all of the variation in future request numbers up to seven days into the future.

In six of the eight sets of models MinTemp provided the largest single increase in R^2. When added to models already containing holiday and weather information, Day of Year added relatively little to the goodness-of-fit. In several of the models it actually reduced the mean R^2 for the validation set.

Table 4-2: Models adding one input to the all the previous ones. Coefficient of multiple determination (R^2) and percentage change in relation to the previous model, for the validation sets.

(a) Culture x1 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.93 (0.	0.90 (0.0)	0.84 (0.0)	0.80 (0.0)	0.75 (0.0)	0.66 (0.0)	0.64 (0.0)	0.79 (0.0)
+SchoolInf	0.93 (0.1)	0.90 (0.4)	0.85 (1.6)	0.82 (2.2)	0.78 (3.2)	0.69 (4.7)	0.67 (4.6)	0.81 (2.4)
+Rainfall	0.9[illegible] (1.2)	0.90 (-0.2)	0.85 (-0.5)	0.82 (-0.1)	0.78 (0.4)	0.70 (1.7)	0.68 (2.2)	0.81 (0.7)
+MinTemp	0.95 (0.4)	0.93 (3.3)	0.90 (6.2)	0.88 (7.6)	0.87 (11.2)	0.82 (16.4)	0.80 (17.7)	0.88 (9.0)
+MaxTemp	0.95 (0.0)	0.93 (-0.4)	0.91 (1.0)	0.89 (1.3)	0.89 (2.5)	0.86 (4.6)	0.83 (3.7)	0.89 (1.8)
+WindX	0.96 (0.9)	0.94 (1.5)	0.93 (1.9)	0.92 (3.3)	0.94 (5.2)	0.92 (7.1)	0.92 (10.3)	0.93 (4.3)
+WindY	0.96 (0.1)	0.95 (1.1)	0.95 (2.3)	0.96 (3.7)	0.96 (2.7)	0.95 (3.3)	0.94 (2.6)	0.95 (2.3)
+SinDay, CosDay	0.96 (0.5)	0.95 (0.0)	0.95 (-0.2)	0.95 (-0.7)	0.97 (0.5)	0.95 (0.3)	0.94 (-0.1)	0.95 (0.0)

(b) Culture x2 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.85 (0.0)	0.71 (0.0)	0.60 (0.0)	0.51 (0.0)	0.42 (0.0)	0.27 (0.0)	0.24 (0.0)	0.52 (0.0)
+SchoolInf	0.86 (1.1)	0.73 (2.0)	0.63 (4.9)	0.55 (7.9)	0.48 (14.1)	0.36 (34.2)	0.32 (35.0)	0.56 (14.2)
+Rainfall	0.85 (-1.4)	0.74 (0.9)	0.68 (8.0)	0.61 (10.8)	0.56 (16.7)	0.44 (24.3)	0.41 (25.9)	0.61 (12.2)
+MinTemp	0.85 (0.9)	0.77 (4.1)	0.76 (10.9)	0.74 (20.3)	0.73 (29.7)	0.65 (47.2)	0.61 (50.0)	0.73 (23.3)
+MaxTemp	0.87 (1.4)	0.77 (0.7)	0.77 (2.0)	0.78 (5.4)	0.77 (5.8)	0.73 (12.1)	0.69 (12.3)	0.77 (5.7)
+WindX	0.86 (-0.8)	0.79 (2.6)	0.80 (3.2)	0.80 (3.4)	0.79 (3.1)	0.76 (3.7)	0.75 (8.7)	0.79 (3.4)
+WindY	0.88 (2.4)	0.85 (6.7)	0.85 (6.7)	0.86 (7.0)	0.87 (9.4)	0.85 (11.7)	0.83 (10.8)	0.85 (7.8)
+SinDay, CosDay	0.89 (1.1)	0.86 (1.4)	0.88 (4.0)	0.88 (2.6)	0.90 (4.3)	0.90 (6.2)	0.87 (5.6)	0.88 (3.6)

(c) Culture x3 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.95 (0.0)	0.92 (0.0)	0.87 (0.0)	0.85 (0.0)	0.78 (0.0)	0.75 (0.0)	0.73 (0.0)	0.84 (0.0)
+SchoolInf	0.95 (0.5)	0.92 (0.7)	0.90 (2.5)	0.88 (3.6)	0.83 (5.9)	0.80 (7.6)	0.79 (8.2)	0.87 (4.1)
+Rainfall	0.96 (0.4)	0.93 (0.9)	0.91 (1.4)	0.88 (-0.4)	0.84 (0.9)	0.81 (0.9)	0.80 (1.6)	0.87 (0.8)
+MinTemp	0.95 (-0.2)	0.94 (0.5)	0.93 (2.4)	0.92 (4.7)	0.89 (6.8)	0.88 (8.4)	0.87 (8.4)	0.91 (4.4)
+MaxTemp	0.96 (0.4)	0.94 (0.2)	0.94 (0.6)	0.93 (1.1)	0.92 (2.7)	0.90 (2.5)	0.88 (1.2)	0.92 (1.3)
+WindX	0.96 (-0.1)	0.95 (0.7)	0.95 (1.1)	0.95 (2.0)	0.95 (2.9)	0.93 (3.0)	0.92 (5.2)	0.94 (2.1)
+WindY	0.96 (0.3)	0.94 (-0.3)	0.96 (0.8)	0.95 (0.8)	0.95 (0.8)	0.94 (1.6)	0.93 (0.5)	0.95 (0.8)
+SinDay, CosDay	0.95 (-0.2)	0.95 (0.4)	0.95 (-0.4)	0.95 (-0.7)	0.95 (-0.6)	0.95 (0.6)	0.94 (1.4)	0.95 (0.1)

(d) Culture total R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.93 (0.0)	0.88 (0.0)	0.81 (0.0)	0.77 (0.0)	0.71 (0.0)	0.61 (0.0)	0.58 (0.0)	0.76 (0.0)
+SchoolInf	0.92 (-0.8)	0.87 (-0.6)	0.83 (2.0)	0.79 (3.2)	0.75 (5.3)	0.66 (7.4)	0.65 (11.1)	0.78 (3.9)
+Rainfall	0.93 (1.3)	0.90 (2.9)	0.85 (3.0)	0.82 (2.9)	0.78 (4.1)	0.68 (3.9)	0.68 (5.7)	0.81 (3.4)
+MinTemp	0.94 (1.1)	0.92 (2.7)	0.91 (6.5)	0.88 (7.9)	0.87 (11.3)	0.79 (15.9)	0.79 (14.8)	0.87 (8.6)
+MaxTemp	0.94 (-0.5)	0.91 (-1.3)	0.90 (-0.8)	0.88 (0.4)	0.87 (0.7)	0.83 (5.0)	0.83 (5.5)	0.88 (1.3)
+WindX	0.95 (0.6)	0.93 (2.1)	0.93 (3.3)	0.93 (4.7)	0.92 (6.0)	0.89 (7.3)	0.88 (6.6)	0.92 (4.4)
+WindY	0.95 (0.7)	0.94 (1.1)	0.94 (0.8)	0.95 (2.2)	0.95 (2.9)	0.92 (2.8)	0.91 (3.3)	0.94 (2.0)
+SinDay, CosDay	0.94 (-1.2)	0.93 (-1.0)	0.93 (-0.7)	0.94 (-0.8)	0.95 (0.0)	0.93 (1.8)	0.92 (1.2)	0.94 (-0.1)

(e) Microscopy x1 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.93 (0.0)	0.89 (0.0)	0.83 (0.0)	0.78 (0.0)	0.72 (0.0)	0.64 (0.0)	0.59 (0.0)	0.77 (0.0)
+SchoolInf	0.93 (0.4)	0.89 (0.2)	0.85 (2.4)	0.81 (4.5)	0.76 (5.4)	0.70 (8.7)	0.65 (9.8)	0.80 (4.5)
+Rainfall	0.93 (-0.7)	0.88 (-1.2)	0.85 (0.5)	0.82 (0.9)	0.79 (3.4)	0.73 (5.3)	0.70 (6.9)	0.81 (2.2)
+MinTemp	0.94 (1.8)	0.91 (3.3)	0.89 (5.0)	0.87 (6.1)	0.85 (8.2)	0.82 (11.8)	0.79 (14.1)	0.87 (7.2)
+MaxTemp	0.94 (-0.7)	0.91 (0.1)	0.90 (0.1)	0.89 (2.1)	0.89 (4.2)	0.86 (4.6)	0.84 (6.4)	0.89 (2.4)
+WindX	0.94 (0.8)	0.92 (1.2)	0.92 (2.7)	0.91 (2.1)	0.92 (3.1)	0.91 (5.9)	0.89 (5.8)	0.92 (3.1)
+WindY	0.95 (1.0)	0.95 (2.5)	0.95 (3.4)	0.95 (4.3)	0.96 (4.6)	0.94 (3.7)	0.93 (4.2)	0.95 (3.4)
+SinDay, CosDay	0.94 (-1.6)	0.93 (-2.2)	0.93 (-2.5)	0.92 (-3.1)	0.92 (-3.6)	0.91 (-3.3)	0.90 (-3.0)	0.92 (-2.8)

(f) Microscopy x2 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.89 (0.0)	0.81 (0.0)	0.75 (0.0)	0.68 (0.0)	0.61 (0.0)	0.53 (0.0)	0.48 (0.0)	0.68 (0.0)
+SchoolInf	0.88 (-1.2)	0.81 (-0.2)	0.75 (0.5)	0.69 (2.5)	0.64 (5.0)	0.56 (5.3)	0.52 (7.8)	0.69 (2.8)
+Rainfall	0.89 (1.3)	0.82 (1.5)	0.77 (2.6)	0.70 (0.7)	0.66 (2.3)	0.60 (7.4)	0.54 (5.0)	0.71 (3.0)
+MinTemp	0.89 (-0.2)	0.84 (1.6)	0.80 (3.6)	0.75 (7.1)	0.71 (7.1)	0.66 (8.5)	0.59 (9.2)	0.75 (5.3)
+MaxTemp	0.88 (-0.8)	0.84 (0.3)	0.81 (1.3)	0.78 (3.6)	0.76 (8.4)	0.73 (11.4)	0.66 (10.6)	0.78 (5.0)
+WindX	0.89 (1.4)	0.86 (1.9)	0.86 (6.2)	0.85 (9.1)	0.86 (12.9)	0.83 (14.2)	0.78 (18.5)	0.85 (9.2)
+WindY	0.90 (0.6)	0.89 (3.6)	0.87 (1.3)	0.88 (4.4)	0.89 (2.7)	0.88 (5.5)	0.86 (10.4)	0.88 (4.1)
+SinDay, CosDay	0.91 (1.0)	0.88 (-1.0)	0.89 (3.1)	0.91 (2.5)	0.92 (4.0)	0.91 (3.2)	0.86 (-0.2)	0.90 (1.8)

(g) Microscopy x3 R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.95 (0.0)	0.90 (0.0)	0.85 (0.0)	0.82 (0.0)	0.76 (0.0)	0.73 (0.0)	0.70 (0.0)	0.81 (0.0)
+SchoolInf	0.94 (-0.5)	0.90 (0.1)	0.86 (1.1)	0.82 (0.5)	0.78 (3.1)	0.74 (2.3)	0.72 (2.5)	0.82 (1.3)
+Rainfall	0.95 (0.7)	0.92 (1.8)	0.88 (3.3)	0.85 (3.7)	0.83 (5.4)	0.80 (7.4)	0.77 (7.2)	0.86 (4.2)
+MinTemp	0.94 (-0.8)	0.91 (-1.3)	0.87 (-2.1)	0.83 (-2.3)	0.80 (-3.1)	0.77 (-3.7)	0.75 (-2.7)	0.84 (-2.3)
+MaxTemp	0.95 (1.1)	0.93 (2.5)	0.91 (5.1)	0.89 (6.9)	0.87 (9.3)	0.87 (13.2)	0.83 (11.4)	0.89 (7.1)
+WindX	0.95 (-0.2)	0.93 (0.2)	0.92 (1.6)	0.92 (2.7)	0.91 (3.8)	0.90 (3.1)	0.86 (3.9)	0.91 (2.2)
+WindY	0.96 (0.8)	0.94 (1.2)	0.94 (2.2)	0.95 (3.6)	0.95 (4.8)	0.94 (4.9)	0.93 (7.1)	0.94 (3.5)
+SinDay, CosDay	0.96 (0.0)	0.95 (1.1)	0.95 (1.1)	0.96 (1.1)	0.95 (0.1)	0.94 (0.3)	0.93 (0.4)	0.95 (0.6)

(h) Microscopy total R^2 (Percentage change)

	Lead1 (%)	Lead2 (%)	Lead3 (%)	Lead4 (%)	Lead5 (%)	Lead6 (%)	Lead7 (%)	Mean (%)
Requests +PublicInf	0.95 (0.0)	0.91 (0.0)	0.86 (0.0)	0.83 (0.0)	0.79 (0.0)	0.73 (0.0)	0.69 (0.0)	0.82 (0.0)
+SchoolInf	0.95 (0.1)	0.92 (1.2)	0.88 (2.1)	0.86 (3.5)	0.83 (4.8)	0.77 (6.1)	0.73 (6.1)	0.85 (3.4)
+Rainfall	0.95 (0.2)	0.93 (0.7)	0.88 (0.1)	0.86 (-0.5)	0.83 (0.2)	0.77 (0.4)	0.75 (1.8)	0.85 (0.4)
+MinTemp	0.96 (1.0)	0.94 (1.8)	0.92 (4.5)	0.91 (6.3)	0.90 (8.9)	0.86 (11.6)	0.85 (13.6)	0.91 (6.8)
+MaxTemp	0.96 (0.0)	0.95 (0.4)	0.93 (1.1)	0.93 (1.7)	0.93 (2.9)	0.91 (6.0)	0.90 (5.7)	0.93 (2.5)
+WindX	0.96 (-0.2)	0.95 (0.2)	0.94 (0.7)	0.94 (1.0)	0.95 (1.8)	0.92 (0.4)	0.91 (1.0)	0.94 (0.7)
+WindY	0.97 (0.8)	0.95 (0.7)	0.95 (1.4)	0.95 (1.5)	0.96 (1.0)	0.94 (2.1)	0.93 (2.2)	0.95 (1.4)
+SinDay, CosDay	0.97 (0.1)	0.96 (0.4)	0.95 (0.1)	0.95 (0.1)	0.96 (0.1)	0.95 (1.2)	0.94 (0.9)	0.95 (0.4)

Discussion

For this urban Australian setting, these modelling exercises demonstrate that recent weather patterns, regardless of the time of year, are strong predictors of the daily number of requests for faecal analysis. Each of the weather inputs used – minimum and maximum temperature, rainfall, and wind speed and direction – had a separate and identifiable effect on model forecasting ability. Together with recent request numbers and holiday information, these weather markers accounted for nearly all the variation in request numbers up to seven days into the future. The evidence for these influences is strengthened by the consistent results across eight different time series forecasting different surrogate markers of gastroenteritis incidence. Day of year added very little to models already containing holiday and weather inputs. To the extent that day of year represents 'true' seasonal influences, it seems to be less important than the recent weather.

It is unusual to use artificial neural networks in search of predictive factors, rather than simply for predicting outputs. However, there is at least one precedent for the use of artificial neural networks in this way. Pascual and others (Pascual et al., 2000) used multilayer perceptron networks to model the effects of the weather pattern in the Pacific Ocean called the El Niño Southern Oscillation (ENSO) on rates of cholera in Bangladesh. As in this study, they found that their simple predictive models of cholera (based on the cholera time series) improved markedly when they introduced suitably lagged ENSO data. Interestingly, they only revealed that their model was a neural network in a footnote.

The 'early stopping' technique used in the training of these neural networks ensured that the improvements in model fit with new inputs were not simply due to over-training of larger models. A series of trials was done (data not shown here) in which the number of free parameters in each model was kept constant by substituting weather inputs for inputs with a constant value. The same pattern of increasing R^2 was seen in those tests.

However, this is essentially an ecological study, so it does not prove a causative relationship between the weather and disease. It is also not possible to determine how much of the variation in the requests for faecal analysis was due to variations in disease incidence, and how much was due to changes in health care seeking behaviour. The deep dips in request numbers during school and public holidays demonstrate the effect of the availability of health care and/or laboratories on requesting. The absence of rebound increases in testing after holidays suggests that patients are either treated empirically by their doctors during those periods, or treat themselves and recover at home.

The study was based on a retrospective analysis of routinely collected data, which precluded the possibility of data validation. However, the fact that these data were primarily collected for billing purposes means that both the pathology labs and the HIC would have been concerned with accuracy and timeliness.

There are a number of plausible mechanisms by which the weather might affect numbers of gastroenteritis cases. True seasonal mechanisms might include such things as the availability of seasonal fruits and vegetables, or even physiological responses to

differing day length (Dowell, 2001). Mechanisms more dependent on the weather than the season might include closer contact between people indoors during inclement weather, greater consumption of uncooked foods on hot days, or the growth of pathogenic organisms in cooked food left at room temperature on a warm day. In Melbourne there may even be reasons why the X and Y components of wind speed and direction have different effects. Warmer northerly and westerly winds reach Melbourne across the hotter inland of the Australian continent, while cooler southerly and easterly winds come more directly from the sea.

It is difficult to separate the effects of true seasonal factors from direct effects of recent weather. However, in general the most accurate models in this study contained all the weather inputs but not the day of year variable. The day of year contributed little or nothing to models already containing all the weather inputs. For gastroenteritis in Melbourne (and to the neural network modeller), the seasons do not seem to be much more than just the recent weather.

Chapter Five: Artificial neural networks and gastroenteritis surveillance

Introduction

This work follows directly from the previous study. Having established an appropriate geographical and temporal scale, and potentially useful network inputs, the next step was to create and test forecasting models. Once again 'early stopping' with a validation set was used to avoid over-training. However, this time a third portion of the data was reserved as a test set for the finished models. The test set of 180 days of patterns was reserved from the very end of the available data set, providing the nearest possible equivalent of a prospective field test for each model.

Why forecast gastroenteritis incidence?

The possibility of massive outbreaks of water-borne gastroenteritis still hangs over the major cities of the world, even where they have state-of-the-art water treatment facilities. There are organisms which are resistant to chlorination, and/or small enough to pass through filters (Hayes et al., 1989, Goldstein et al., 1996, Driedger et al., 2001), and there may be long delays between the contamination event and detection of the outbreak. An outbreak of cryptosporidiosis in Milwaukee in the USA in 1993 probably affected over 400,000 people over a period of about six weeks (MacKenzie et al., 1994).

This study investigated the possibility of a system for the earlier detection of outbreaks of water-related gastroenteritis in metropolitan Melbourne, in southeastern

Australia. The indicators of gastroenteritis were the numbers of requests for faecal analysis received by pathology laboratories. Requests for analysis are less specific indicators than confirmed isolates, but they begin to change days earlier, and it is feasible for the local health department to acquire the information very rapidly.

The statistical technique investigated was the artificial neural network. These offer some theoretical advantages for foreca ting this type of time series (Cheng and Titterington, 1994). The inner workings of a neural network, although complex, make no assumptions about the distribution or independence of the input data. In the case of time series, there is no particular requirement for either stationarity or a high degree of autocorrelation (Hinton, 1992). (Although it is generally true that a neural network will generalise better on a stationary or highly autocorrelated time series (Masters, 1995).)

Methods

Data acquisition and pre-processing

The acquisition and pre-processing of the data were described in the Methods section of Chapter Four. In summary, the different inputs are commonly abbreviated in this thesis using the following scheme:

Requests = Time series of requests for faecal analysis of the specified type: smoothed as an asymmetrical 7 day moving average and presented as the current day's value and seven lags.

C1, C2 or C3 = Culture x1, Culture x2 or Culture x3 during a single episode of

		illness.
M1, M2 or M3	=	Microscopy x1, Microscopy x2 or Microscopy x3 during a single episode of illness.
Holidays (SchoolInf and PublicInf)	=	Two time series, based on the number of days in that day's moving average of requests which were school (SchoolInf) or public holidays (PublicInf): presented to the networks as that day's value and seven leads.
MinTemp	=	Minimum temperature for the 24-hour period: presented with 15 lags.
MaxTemp	=	Maximum temperature for the 24-hour period: presented with 15 lags.
Rainfall	=	Total precipitation for the 24-hour period: presented with 15 lags.
WindX	=	The x-axis component of polar coordinates based on a vector of wind speed and direction at 3 pm: presented with 15 lags.
WindY	=	The y-axis component of polar coordinates based on a vector of wind speed and direction at 3 pm: presented with 15 lags.
Day (SinDay and CosDay)	=	Two inputs: the Sine (SinDay) and Cosine (CosDay) of the angle representing the day of the year. These were always presented together.

Neural network architecture, training and testing

A separate set of neural network models was created for each of the smoothed requests time series (four series for faecal culture and four for faecal microscopy). In each case the commercial neural network package NeuroShell2 (Ward Systems Group Inc, 2000) was used to create multi-layer feed-forward networks with one hidden layer and the logistic function as the activation function for all the hidden and output layer neurons. Each model had a different number of inputs, between 8 and 106. The model inputs were different combinations of the inputs previously shown to individually improve the fit of a simple model (see the previous chapter), ranging

from only the faecal analysis requests through to models with all the previously described inputs at the same time.

Each model had 60 neurons in the hidden layer. This number was estimated by a common rule of thumb (half the sum of inputs and outputs, plus the square root of the number of training patterns (Smith, 1999)) for a typical (medium-sized) network in this study, and then confirmed as reasonable by empirical testing. In every case the outputs of the network were the seven leads of the smoothed time series (i.e. the predicted number of faecal analysis requests for each of the next seven days).

The test set was chosen to simulate a prospective test of the trained network. The last 180 days of the processed times series (approximately 15%) were reserved before training began, and were never presented to the network during training or validation. The remaining 85% of the time series (1239 patterns) was divided randomly into a 70% training set (1053 patterns) and a 15% validation set (186 patterns). The validation sets were used in an 'early-stopping' procedure to guard against over-training (Rumelhart et al., 1994). To do this, every time 200 learning events had occurred (i.e. 200 training patterns had been presented to the network), training was stopped and the mean squared error for the validation set calculated. Training was done by standard back-propagation of errors, with learning rate and momentum terms both equal to 0.1. The final trained network was the one that gave the lowest mean squared error for the validation set.

Different aspects of the trained networks' performance were assessed in different ways. The mean squared error was calculated for the whole training set and validation

set, including all the outputs, and was thus an overall measure of the ability of the network to draw generalised patterns from the data set. The coefficient of multiple determination (R^2), on the other hand, compares the accuracy of the predictions for a single output with a trivial model that simply predicts the mean value in every case. There are several variants of R^2: here it is calculated by taking the sum of the squared differences between each observed and predicted value, and dividing that by the sum of squared differences between each observed value and the mean of all observed values. The resulting ratio is subtracted from 1, so that an R^2 value of 1.0 represents a perfect prediction, and values below zero show the model gives very poor predictions (worse than if it had simply predicted the mean number of cases each time). The values of the mean squared error and R^2 for the validation sets assess the model's ability to generalise the relationships between the inputs and the outputs within the same time window. The values for the test sets assess the ability of the trained networks to make forecasts outside the training time window. Finally, the most important assessment of the usefulness of the network model was to plot the test set forecasts against the observed values to see whether useful decisions could have been made either to begin control measures or further investigate rising levels. A graphical tool was developed to simulate the use of each network in a field situation. (StepGraph – see Appendix One and the CD inside the back cover)

Results

There was considerable day-to-day variation in request numbers, even in the processed time series. In general there were around 60 to 70 single faecal cultures per day, 10 to 15 'Cultures x2', and around 40 'Cultures x3'. Faecal microscopy showed a similar pattern, with 30 to 40 single microscopy requests, around 7 requests for 'Microscopy x2', and 10 to 15 requests for 'Microscopy x3'. Weather indicators showed typical seasonal patterns for Melbourne (Bureau of Meteorology Australia, 2001). Plots of each of the smoothed time series that provided inputs are shown in the previous chapter.

Table 5-1 lists the coefficient of multiple determination (R^2) for each of the predicted outputs, for the training, validation and test sets for each model. In each sub-table the models are listed in order of increasing mean R^2 value for the seven predictions. For each of the eight indicators (Culture x1, x2, x3 and Total, and Microscopy x1, x2, x3 and Total) the same general pattern was observed. Including information about school and public holidays improved model fit to the training, validation and test sets. The more extra weather information provided, the better the models fit both the training and validation sets (indicating that the networks generalised the relationships for the training time window). However, the more weather information included, the less well the network models tended to predict the *prospective* test data. In every case the models giving the best predictions had either requests and holiday information as their only inputs, or else incorporated only one or two of the weather inputs.

Table 5-1. Coefficient of multiple determination (R^2) for models with varying inputs. Within each category the lines are arranged in ascending order of the mean of the R^2 for all seven leads.
a) Faecal culture x1

C1 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.94	0.88	0.79	0.70	0.62	0.52	0.44	0.70
Requests, Holidays	0.96	0.92	0.88	0.83	0.78	0.74	0.68	0.83
Requests, Holidays, MaxTemp	0.95	0.93	0.89	0.85	0.80	0.75	0.72	0.84
Requests, Holidays, MaxTemp, WindX	0.95	0.93	0.90	0.87	0.84	0.80	0.77	0.87
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.97	0.97	0.96	0.96	0.97	0.96	0.96
C1 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.89	0.82	0.74	0.60	0.52	0.45	0.71
Requests, Holidays	0.96	0.92	0.89	0.83	0.77	0.72	0.66	0.82
Requests, Holidays, MaxTemp	0.95	0.92	0.90	0.87	0.80	0.77	0.74	0.85
Requests, Holidays, MaxTemp, WindX	0.94	0.92	0.90	0.87	0.82	0.80	0.77	0.86
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.96	0.93	0.94	0.94	0.94	0.94	0.92	0.94
C1 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.88	0.66	0.35	-0.06	-0.27	-0.68	-1.21	-0.05
Requests	0.86	0.78	0.64	0.51	0.33	0.17	-0.02	0.47
Requests, Holidays, MaxTemp, WindX	0.89	0.84	0.80	0.79	0.74	0.69	0.59	0.76
Requests, Holidays, MaxTemp	0.92	0.84	0.80	0.77	0.75	0.68	0.62	0.77
Requests, Holidays	0.92	0.86	0.83	0.78	0.75	0.72	0.64	0.79

b) Faecal culture x2

C2 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.86	0.74	0.59	0.46	0.35	0.23	0.14	0.48
Requests, Holidays	0.89	0.78	0.69	0.60	0.52	0.44	0.36	0.61
Requests, Holidays, Rainfall	0.88	0.77	0.70	0.62	0.53	0.45	0.34	0.61
Requests, Holidays, Rainfall, MaxTemp	0.89	0.83	0.77	0.72	0.69	0.64	0.60	0.73
Requests, Holidays, Rainfall, MaxTemp, WindX	0.90	0.84	0.79	0.74	0.72	0.67	0.63	0.75
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.94	0.93	0.94	0.95	0.95	0.95	0.94	0.94
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.94
C2 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.86	0.74	0.62	0.53	0.39	0.29	0.21	0.52
Requests, Holidays, Rainfall	0.87	0.77	0.70	0.62	0.52	0.46	0.36	0.61
Requests, Holidays	0.88	0.77	0.69	0.61	0.52	0.47	0.37	0.62
Requests, Holidays, Rainfall, MaxTemp	0.86	0.80	0.76	0.71	0.66	0.62	0.55	0.71
Requests, Holidays, Rainfall, MaxTemp, WindX	0.86	0.81	0.79	0.75	0.74	0.67	0.62	0.75
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.88	0.88	0.89	0.91	0.90	0.88	0.84	0.88
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.89	0.87	0.88	0.90	0.90	0.91	0.89	0.89

C2 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.64	0.25	-0.13	-0.36	-0.45	-0.38	-0.31	-0.10
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.66	0.37	0.13	-0.08	-0.21	-0.18	-0.15	0.08
Requests, Holidays, Rainfall, MaxTemp	0.76	0.59	0.39	0.17	0.18	0.04	0.17	0.33
Requests	0.80	0.63	0.50	0.38	0.22	0.03	-0.08	0.36
Requests, Holidays, Rainfall, MaxTemp, WindX	0.82	0.69	0.51	0.36	0.20	0.12	0.14	0.41
Requests, Holidays, Rainfall	0.81	0.61	0.57	0.48	0.38	0.25	0.17	0.47
Requests, Holidays	0.83	0.66	0.59	0.52	0.44	0.32	0.30	0.52

c) Faecal culture x3

C3 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.94	0.89	0.82	0.74	0.66	0.57	0.48	0.73
Requests, Holidays, Rainfall	0.96	0.93	0.91	0.88	0.87	0.83	0.81	0.88
Requests, Holidays	0.96	0.93	0.91	0.89	0.86	0.84	0.80	0.89
Requests, Holidays, Rainfall, MaxTemp, WindY	0.95	0.93	0.92	0.91	0.90	0.88	0.85	0.91
Requests, Holidays, Rainfall, MaxTemp	0.96	0.94	0.93	0.91	0.90	0.88	0.85	0.91
Requests, Holidays, Rainfall, MaxTemp, WindY, Day	0.97	0.96	0.95	0.96	0.96	0.96	0.94	0.96
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97

C3 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.89	0.84	0.78	0.71	0.66	0.54	0.77
Requests, Holidays, Rainfall	0.95	0.93	0.91	0.89	0.88	0.85	0.81	0.89
Requests, Holidays	0.96	0.93	0.91	0.90	0.88	0.86	0.81	0.89
Requests, Holidays, Rainfall, MaxTemp, WindY	0.94	0.92	0.91	0.90	0.90	0.88	0.82	0.90
Requests, Holidays, Rainfall, MaxTemp	0.95	0.93	0.92	0.91	0.90	0.88	0.84	0.90
Requests, Holidays, Rainfall, MaxTemp, WindY, Day	0.96	0.95	0.94	0.93	0.93	0.93	0.92	0.94
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.96	0.95	0.95	0.95	0.96	0.95	0.93	0.95
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96
C3 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.92	0.82	0.71	0.57	0.38	0.17	-0.07	0.50
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.89	0.75	0.73	0.69	0.62	0.57	0.55	0.69
Requests, Holidays, Rainfall, MaxTemp, WindY, Day	0.92	0.86	0.81	0.72	0.65	0.59	0.55	0.73
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY	0.92	0.87	0.84	0.79	0.76	0.70	0.61	0.79
Requests, Holidays	0.93	0.88	0.84	0.81	0.76	0.69	0.65	0.79
Requests, Holidays, Rainfall	0.93	0.87	0.84	0.83	0.75	0.69	0.65	0.79
Requests, Holidays, Rainfall, MaxTemp, WindY	0.92	0.86	0.83	0.82	0.76	0.76	0.69	0.81
Requests, Holidays, Rainfall, MaxTemp	0.92	0.86	0.82	0.80	0.79	0.75	0.70	0.81

d) Faecal microscopy x1

M1 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.94	0.86	0.78	0.68	0.59	0.49	0.41	0.68
Requests, Holidays	0.95	0.90	0.85	0.80	0.74	0.69	0.64	0.80
Requests, Holidays, MaxTemp	0.94	0.90	0.86	0.81	0.76	0.69	0.67	0.80
Requests, Holidays, MaxTemp, WindY	0.95	0.91	0.88	0.84	0.80	0.76	0.73	0.84
Requests, Holidays, MaxTemp, WindY, MinTemp	0.96	0.94	0.93	0.92	0.89	0.88	0.87	0.91
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.96	0.97	0.97	0.97	0.98	0.97	0.97
M1 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.93	0.85	0.74	0.64	0.50	0.41	0.35	0.63
Requests, Holidays	0.94	0.89	0.81	0.75	0.67	0.60	0.58	0.75
Requests, Holidays, MaxTemp	0.93	0.89	0.83	0.79	0.72	0.65	0.65	0.78
Requests, Holidays, MaxTemp, WindY	0.93	0.89	0.84	0.82	0.76	0.71	0.69	0.80
Requests, Holidays, MaxTemp, WindY, MinTemp	0.93	0.90	0.88	0.88	0.85	0.82	0.82	0.87
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.95	0.93	0.93	0.93	0.94	0.92	0.90	0.93

M1 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.89	0.77	0.64	0.44	0.24	0.03	-0.15	0.41
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.82	0.71	0.56	0.37	0.25	0.25	0.28	0.46
Requests, Holidays, MaxTemp, WindY, MinTemp	0.91	0.85	0.79	0.72	0.60	0.51	0.45	0.69
Requests, Holidays	0.93	0.85	0.79	0.70	0.62	0.59	0.49	0.71
Requests, Holidays, MaxTemp	0.91	0.85	0.81	0.74	0.69	0.59	0.54	0.73
Requests, Holidays, MaxTemp, WindY	0.89	0.84	0.81	0.77	0.71	0.65	0.56	0.75

e) Faecal microscopy x2

M2 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.89	0.80	0.68	0.56	0.46	0.36	0.29	0.58
Requests, Holidays	0.90	0.83	0.74	0.65	0.57	0.52	0.45	0.67
Requests, Holidays, MaxTemp	0.90	0.83	0.75	0.67	0.59	0.50	0.46	0.67
Requests, Holidays, MinTemp, Rainfall	0.90	0.83	0.76	0.70	0.61	0.55	0.51	0.69
Requests, Holidays, WindY	0.91	0.84	0.78	0.71	0.66	0.57	0.53	0.72
Requests, Holidays, WindY, MinTemp	0.93	0.89	0.85	0.86	0.83	0.80	0.77	0.85
Requests, Holidays, WindY, MinTemp, Rainfall	0.93	0.89	0.87	0.85	0.84	0.82	0.80	0.86
Requests, Holidays, WindY, MinTemp, Rainfall, MaxTemp	0.93	0.93	0.92	0.93	0.93	0.92	0.90	0.92
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.94

M2 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.88	0.80	0.69	0.60	0.51	0.43	0.38	0.61
Requests, Holidays	0.89	0.81	0.71	0.63	0.55	0.51	0.47	0.65
Requests, Holidays, MaxTemp	0.88	0.81	0.72	0.64	0.59	0.52	0.48	0.66
Requests, Holidays, WindY	0.88	0.81	0.71	0.63	0.60	0.55	0.54	0.67
Requests, Holidays, MinTemp, Rainfall	0.88	0.80	0.73	0.64	0.60	0.55	0.51	0.67
Requests, Holidays, WindY, MinTemp	0.90	0.84	0.77	0.74	0.74	0.73	0.71	0.78
Requests, Holidays, WindY, MinTemp, Rainfall, MaxTemp	0.89	0.85	0.83	0.82	0.84	0.84	0.79	0.83
Requests, Holidays, WindY, MinTemp, Rainfall	0.93	0.89	0.87	0.85	0.84	0.82	0.80	0.86
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.89	0.87	0.89	0.88	0.90	0.90	0.87	0.89
M2 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.60	0.15	-0.38	-0.68	-0.89	-1.15	-1.20	-0.50
Requests, Holidays, WindY, MinTemp, Rainfall, MaxTemp	0.66	0.17	-0.19	-0.56	-0.94	-1.06	-0.93	-0.41
Requests, Holidays, WindY, MinTemp	0.81	0.60	0.27	0.03	-0.32	-0.71	-0.92	-0.04
Requests, Holidays, WindY, MinTemp, Rainfall	0.80	0.54	0.27	0.03	-0.14	-0.42	-0.34	0.10
Requests	0.84	0.68	0.52	0.33	0.11	-0.13	-0.30	0.29
Requests, Holidays, WindY	0.83	0.74	0.59	0.36	0.27	0.15	0.06	0.43
Requests, Holidays, MinTemp, Rainfall	0.83	0.66	0.59	0.48	0.40	0.30	0.25	0.50
Requests, Holidays, MaxTemp	0.84	0.72	0.64	0.56	0.43	0.32	0.27	0.54
Requests, Holidays	0.85	0.72	0.64	0.57	0.47	0.34	0.33	0.56

f) Faecal microscopy x3

M3 Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.90	0.85	0.79	0.72	0.65	0.59	0.78
Requests, Holidays	0.96	0.92	0.89	0.85	0.81	0.77	0.73	0.85
Requests, Holidays, WindY	0.95	0.92	0.89	0.86	0.82	0.80	0.76	0.86
Requests, Holidays, MaxTemp	0.96	0.93	0.90	0.88	0.85	0.82	0.78	0.87
Requests, Holidays, MaxTemp, WindX	0.96	0.94	0.92	0.91	0.89	0.87	0.85	0.91
Requests, Holidays, WindY, WindX	0.96	0.94	0.94	0.93	0.92	0.91	0.89	0.93
Requests, Holidays, WindY, WindX, Rainfall	0.97	0.95	0.95	0.95	0.94	0.94	0.92	0.94
Requests, Holidays, WindY, WindX, Rainfall, Day	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97

M3 Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.89	0.82	0.75	0.69	0.64	0.58	0.76
Requests, Holidays	0.95	0.91	0.85	0.80	0.78	0.72	0.69	0.81
Requests, Holidays, WindY	0.95	0.91	0.86	0.82	0.78	0.73	0.69	0.82
Requests, Holidays, MaxTemp	0.95	0.91	0.86	0.83	0.80	0.76	0.72	0.83
Requests, Holidays, MaxTemp, WindX	0.95	0.92	0.88	0.85	0.84	0.81	0.78	0.86
Requests, Holidays, WindY, WindX	0.95	0.94	0.91	0.89	0.88	0.85	0.82	0.89
Requests, Holidays, WindY, WindX, Rainfall	0.95	0.94	0.92	0.90	0.90	0.88	0.86	0.91
Requests, Holidays, WindY, WindX, Rainfall, Day	0.95	0.95	0.94	0.94	0.94	0.93	0.91	0.94
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.96	0.95	0.94	0.95	0.96	0.95	0.92	0.95

M3 Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.63	0.41	0.14	-0.36	-0.68	-0.93	-0.83	-0.23
Requests, Holidays, WindY, WindX, Rainfall	0.77	0.55	0.26	-0.07	-0.40	-0.57	-0.62	-0.01
Requests, Holidays, WindY, WindX	0.80	0.60	0.35	0.16	-0.13	-0.44	-0.65	0.10
Requests, Holidays, WindY, WindX, Rainfall, Day	0.76	0.57	0.33	0.12	-0.07	-0.22	-0.22	0.18
Requests	0.87	0.74	0.58	0.43	0.23	0.07	-0.14	0.40
Requests, Holidays, MaxTemp	0.83	0.72	0.56	0.38	0.19	0.11	0.02	0.40
Requests, Holidays, WindY	0.80	0.64	0.55	0.39	0.19	0.16	0.16	0.41
Requests, Holidays, MaxTemp, WindX	0.85	0.74	0.61	0.44	0.27	0.16	0.01	0.44
Requests, Holidays	0.88	0.77	0.64	0.56	0.39	0.32	0.21	0.54

g) Culture total

Culture total Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.92	0.83	0.71	0.58	0.47	0.35	0.24	0.59
Requests, Holidays	0.95	0.91	0.86	0.81	0.77	0.72	0.67	0.81
Requests, Holidays, MaxTemp	0.94	0.91	0.88	0.84	0.80	0.77	0.73	0.84
Requests, Holidays, MaxTemp, WindY	0.95	0.92	0.89	0.85	0.82	0.79	0.74	0.85
Requests, Holidays, Day, MaxTemp	0.95	0.92	0.89	0.86	0.85	0.82	0.79	0.87
Requests, Holidays, Day, MaxTemp	0.95	0.92	0.89	0.86	0.85	0.82	0.79	0.87
Requests, Holidays, Day	0.96	0.93	0.90	0.87	0.85	0.82	0.79	0.87
Requests, Holidays, Day, WindY	0.96	0.93	0.90	0.89	0.87	0.86	0.83	0.89
Requests, Holidays, Day, Rainfall	0.96	0.93	0.91	0.89	0.87	0.85	0.83	0.89
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96

Culture total Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.93	0.84	0.75	0.64	0.48	0.40	0.29	0.62
Requests, Holidays	0.95	0.91	0.87	0.83	0.80	0.75	0.69	0.83
Requests, Holidays, MaxTemp	0.93	0.91	0.89	0.86	0.83	0.80	0.76	0.85
Requests, Holidays, MaxTemp, WindY	0.94	0.91	0.90	0.87	0.83	0.80	0.75	0.86
Requests, Holidays, Day	0.95	0.91	0.90	0.88	0.84	0.84	0.81	0.88
Requests, Holidays, Day, MaxTemp	0.95	0.91	0.90	0.88	0.85	0.83	0.81	0.88
Requests, Holidays, Day, MaxTemp	0.95	0.91	0.90	0.88	0.85	0.83	0.81	0.88
Requests, Holidays, Day, WindY	0.95	0.91	0.90	0.89	0.87	0.84	0.82	0.88
Requests, Holidays, Day, Rainfall	0.95	0.92	0.90	0.87	0.87	0.84	0.82	0.88
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.96	0.93	0.93	0.94	0.94	0.93	0.91	0.93

Culture total Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.91	0.82	0.71	0.59	0.44	0.28	0.10	0.55
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.91	0.82	0.75	0.69	0.61	0.50	0.48	0.68
Requests, Holidays, Day, MaxTemp	0.91	0.83	0.79	0.75	0.68	0.58	0.51	0.72
Requests, Holidays, Day, MaxTemp	0.91	0.83	0.79	0.75	0.68	0.58	0.51	0.72
Requests, Holidays, Day, Rainfall	0.92	0.86	0.78	0.75	0.71	0.66	0.60	0.76
Requests, Holidays, Day, WindY	0.91	0.85	0.82	0.76	0.72	0.65	0.62	0.76
Requests, Holidays, Day	0.94	0.88	0.84	0.79	0.74	0.74	0.73	0.81
Requests, Holidays, MaxTemp, WindY	0.93	0.88	0.85	0.83	0.82	0.75	0.73	0.83
Requests, Holidays	0.95	0.90	0.87	0.84	0.82	0.79	0.74	0.84
Requests, Holidays, MaxTemp	0.93	0.89	0.87	0.84	0.85	0.78	0.78	0.85

h) Microscopy total

Microscopy total Training sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.89	0.82	0.74	0.66	0.57	0.49	0.73
Requests, Holidays	0.96	0.93	0.90	0.86	0.82	0.79	0.75	0.86
Requests, Holidays, WindY	0.96	0.94	0.90	0.87	0.84	0.80	0.77	0.87
Requests, Holidays, MaxTemp	0.95	0.93	0.91	0.88	0.85	0.81	0.78	0.87
Requests, Holidays, WindX	0.96	0.94	0.91	0.88	0.85	0.82	0.78	0.88
Requests, Holidays, MinTemp	0.96	0.94	0.91	0.89	0.86	0.84	0.81	0.89
Requests, Holidays, MinTemp, WindY	0.97	0.95	0.94	0.91	0.92	0.91	0.88	0.92
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.97	0.96	0.96	0.96	0.96	0.96	0.95	0.96

Microscopy total Validation sets:	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.95	0.89	0.83	0.75	0.65	0.58	0.52	0.74
Requests, Holidays	0.96	0.92	0.89	0.85	0.81	0.76	0.74	0.85
Requests, Holidays, WindY	0.95	0.93	0.90	0.86	0.81	0.76	0.75	0.85
Requests, Holidays, WindX	0.95	0.92	0.89	0.86	0.82	0.77	0.75	0.85
Requests, Holidays, MaxTemp	0.94	0.93	0.90	0.87	0.84	0.79	0.78	0.86
Requests, Holidays, MinTemp	0.95	0.92	0.91	0.88	0.85	0.81	0.79	0.87
Requests, Holidays, MinTemp, WindY	0.95	0.93	0.92	0.89	0.88	0.84	0.83	0.89
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.95	0.95	0.94	0.93	0.94	0.92	0.91	0.93

Microscopy total Test sets: (Last-180 days)	Lead1	Lead2	Lead3	Lead4	Lead5	Lead6	Lead7	Mean
Requests	0.91	0.80	0.66	0.49	0.29	0.04	-0.16	0.43
Requests, Holidays, Rainfall, MinTemp, MaxTemp, WindX, WindY, Day	0.89	0.80	0.73	0.66	0.56	0.49	0.50	0.66
Requests, Holidays, MinTemp, WindY	0.91	0.85	0.81	0.70	0.64	0.54	0.47	0.70
Requests, Holidays, WindY	0.92	0.84	0.77	0.71	0.67	0.54	0.49	0.71
Requests, Holidays, WindX	0.92	0.87	0.81	0.76	0.67	0.59	0.49	0.73
Requests, Holidays	0.93	0.87	0.80	0.77	0.67	0.62	0.59	0.75
Requests, Holidays, MinTemp	0.93	0.88	0.83	0.77	0.71	0.62	0.56	0.76
Requests, Holidays, MaxTemp	0.93	0.89	0.85	0.82	0.78	0.69	0.64	0.80

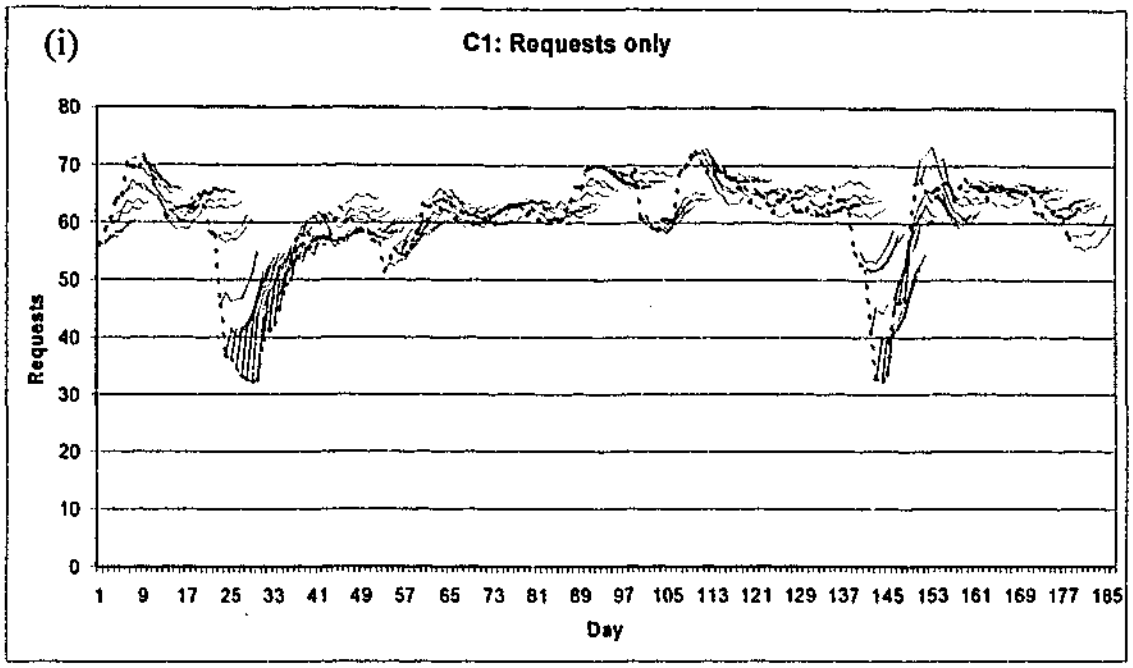
The series with the highest the daily counts generally gave the best results, both for large and small models. The very best model was for 'Culture total', using requests, holidays and maximum temperature, which had a mean R^2 for all seven outputs of 0.85 when applied to the prospective test set (see Table 5-1(g), page 143). The model for 'Culture x3' using requests, holidays, rainfall and maximum temperature had a mean R^2 of 0.81 for the test set (see Table 5-1(c), page 137). Third best was the model for 'Microscopy total' using requests, holidays and maximum temperature, with a mean R^2 of 0.80 for the test set (Table 5-1(h), page 146).

Figure 5-1 plots model predictions against the smoothed actual data for the test sets. In each graph the x-axis is time in days, beginning on December 4, 1999. The y-axes are the smoothed numbers of requests. The heavier dotted line is the observed requests time series, and the thinner solid lines join the seven predictions for each day. The two deepest dips in the actual request numbers correspond to the Christmas/New Year holiday for 1999/2000, and Easter 2000, which coincided that year with a local public holiday (ANZAC Day). The smaller dip starting around Day 97 corresponds to the first school term holidays. (This is best seen in the models of Culture Total and Microscopy Total.)

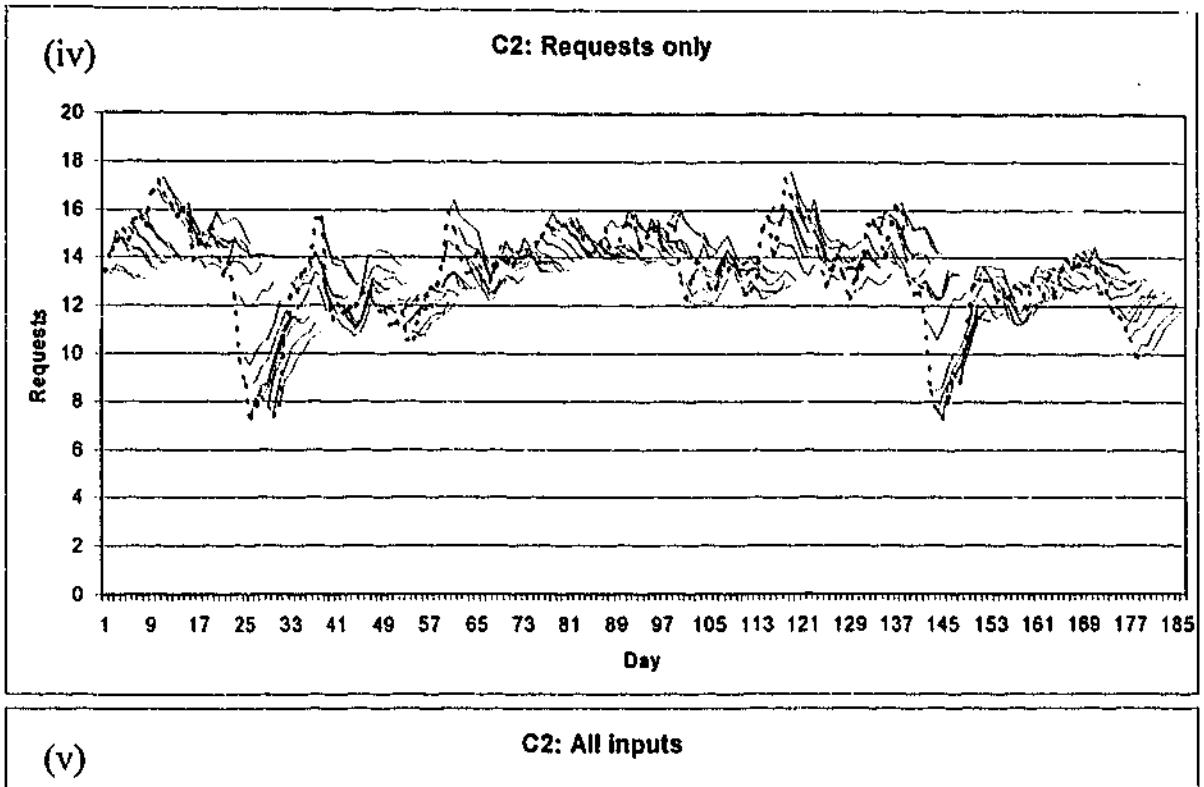
Figure 5-1 includes three examples for each set of experiments: the simplest model using only requests, the most complex model using all the available inputs, and the model with the best mean R^2. The other models are all available on the CD inside the back cover using the StepGraph program. The StepGraph program also gives a better indication of how the models would appear to a surveillance manager in a field situation – see Appendix One for details.

In general the simplest models were least impressive, tending to predict that the current trend at each point would continue, and thus missing virtually all the turning points. On the other hand the most complex models 'invented' many turning points and outbreaks that never occurred. The best models (and especially those for the total numbers of cultures and microscopies) made no severe false turns. (See Figure 5-1(xxi), page 156, and (xxiv), page 157.) A surveillance manager armed with both the 'Culture total' and 'Microscopy total' models would have found them useful tools during the six month test period.

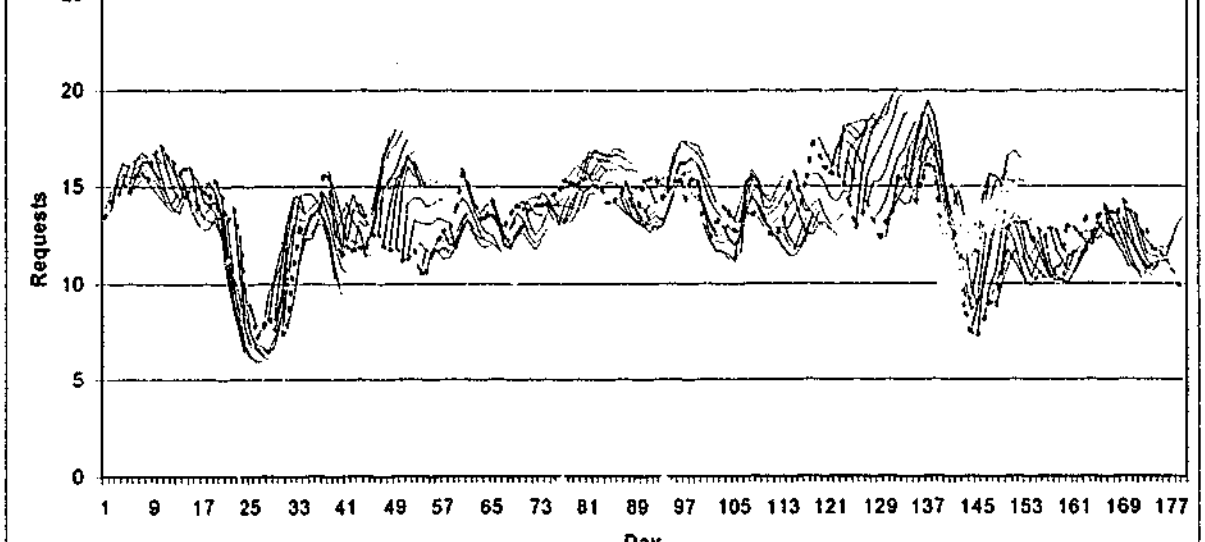
Figure 5-1 (i-xxiv): Network forecasts for the reserved test set based on the last 180 days of the time series. The heavier dotted lines represent the actual values of these smoothed series, and the fine solid lines join the seven predictions of the network at each point.



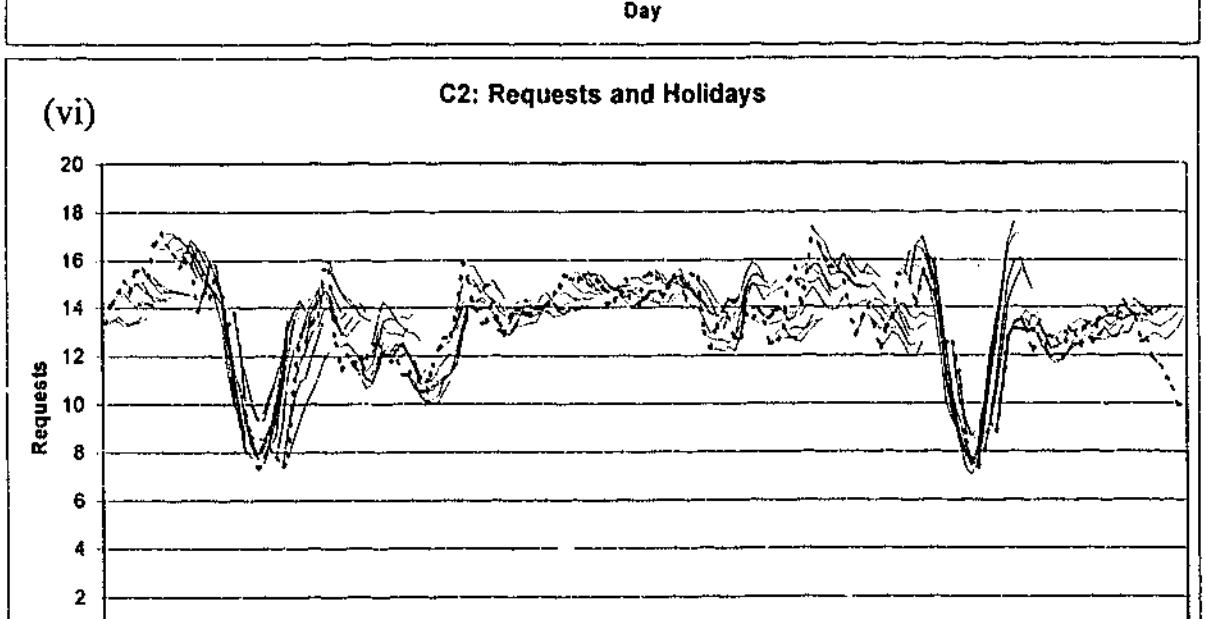
(iv)


C2: Requests only
Requests
0
2
4
6
8
10
12
14
16
18
20
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177 185
Day

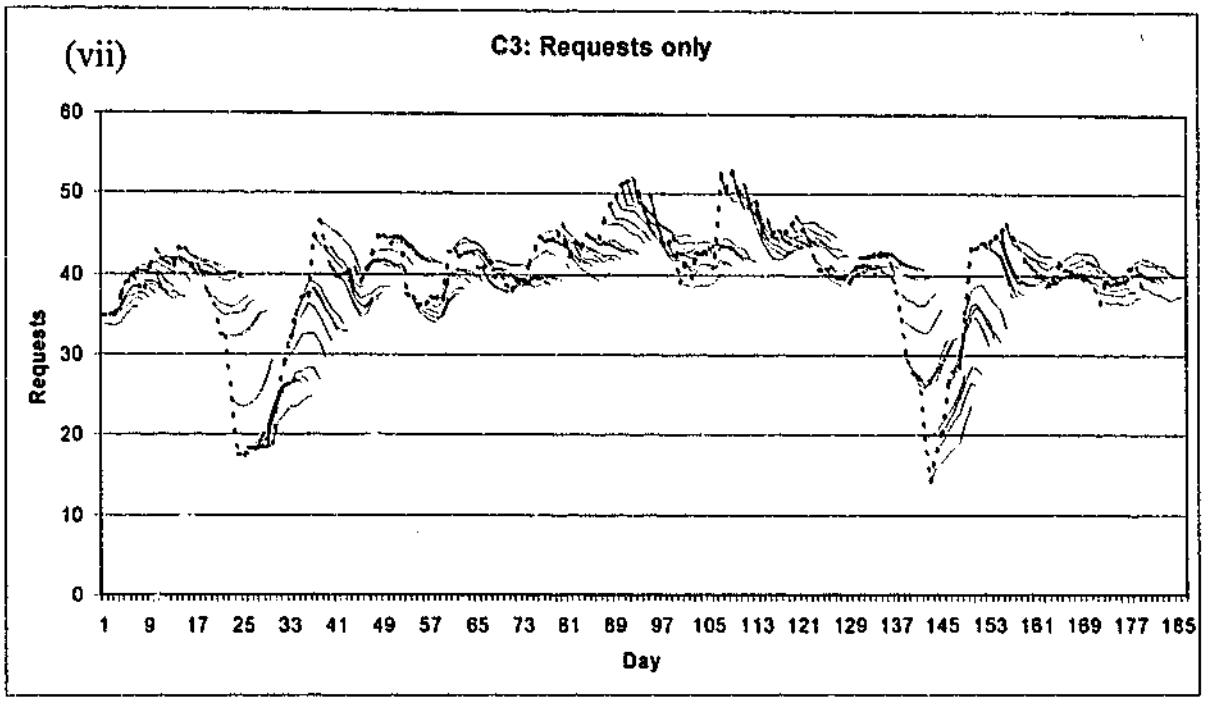
(v)


C2: All inputs
Requests
0
5
10
15
20
25
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

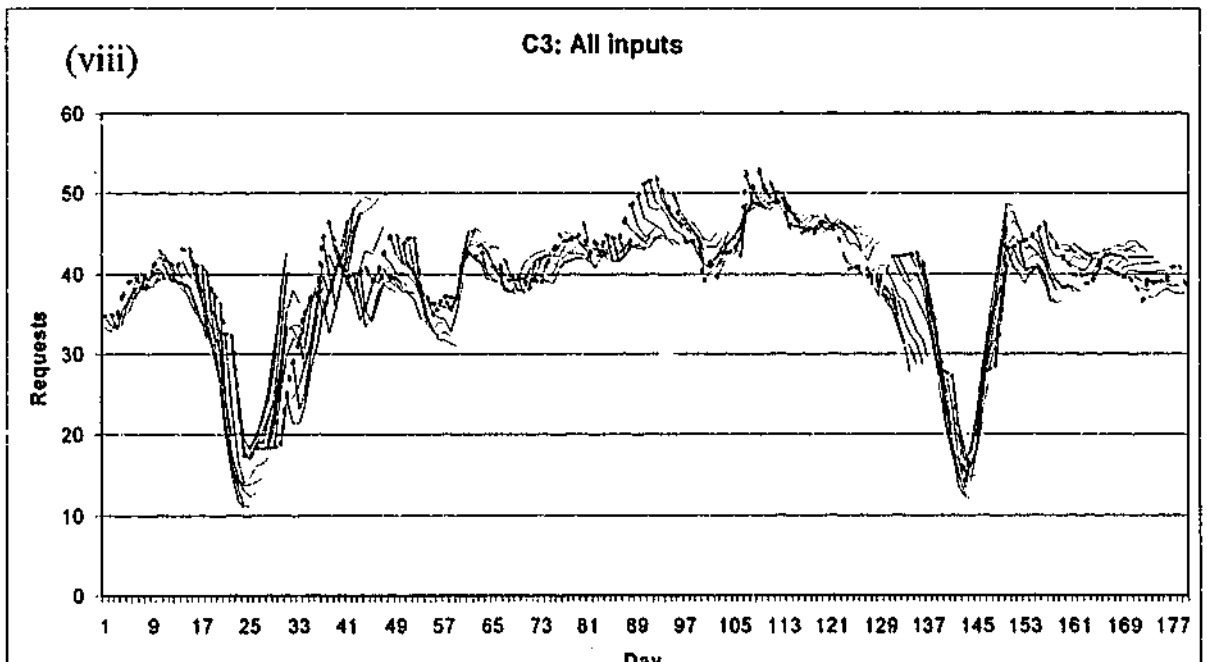
(vi)


C2: Requests and Holidays
Requests
0
2
4
6
8
10
12
14
16
18
20
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

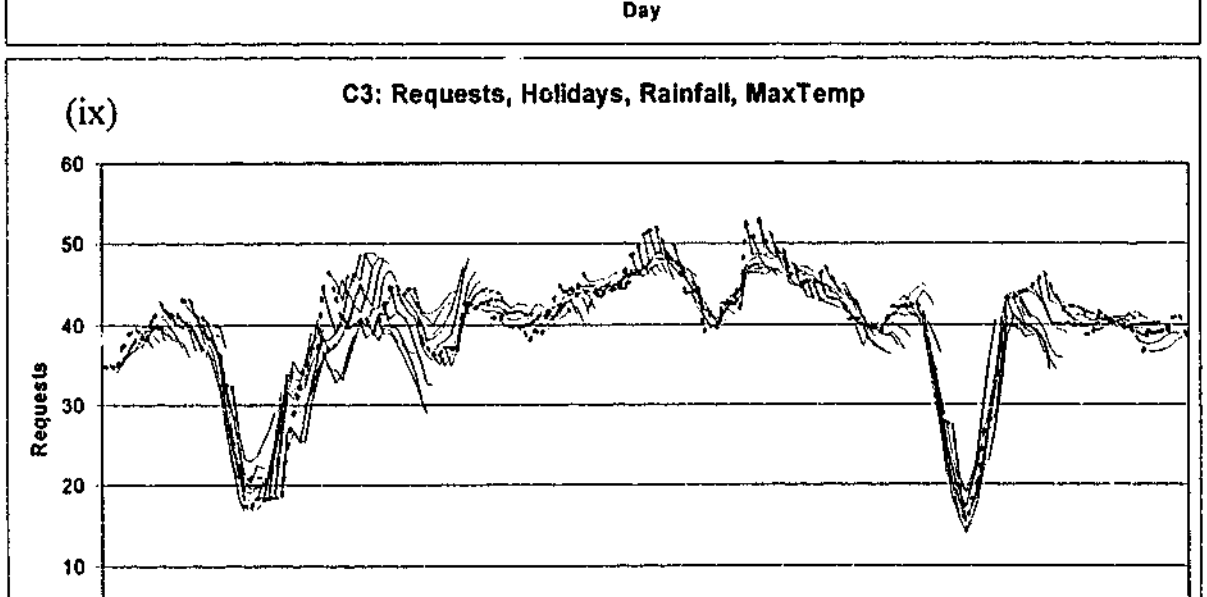
(vii)


C3: Requests only
Requests
0
10
20
30
40
50
60
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177 185
Day

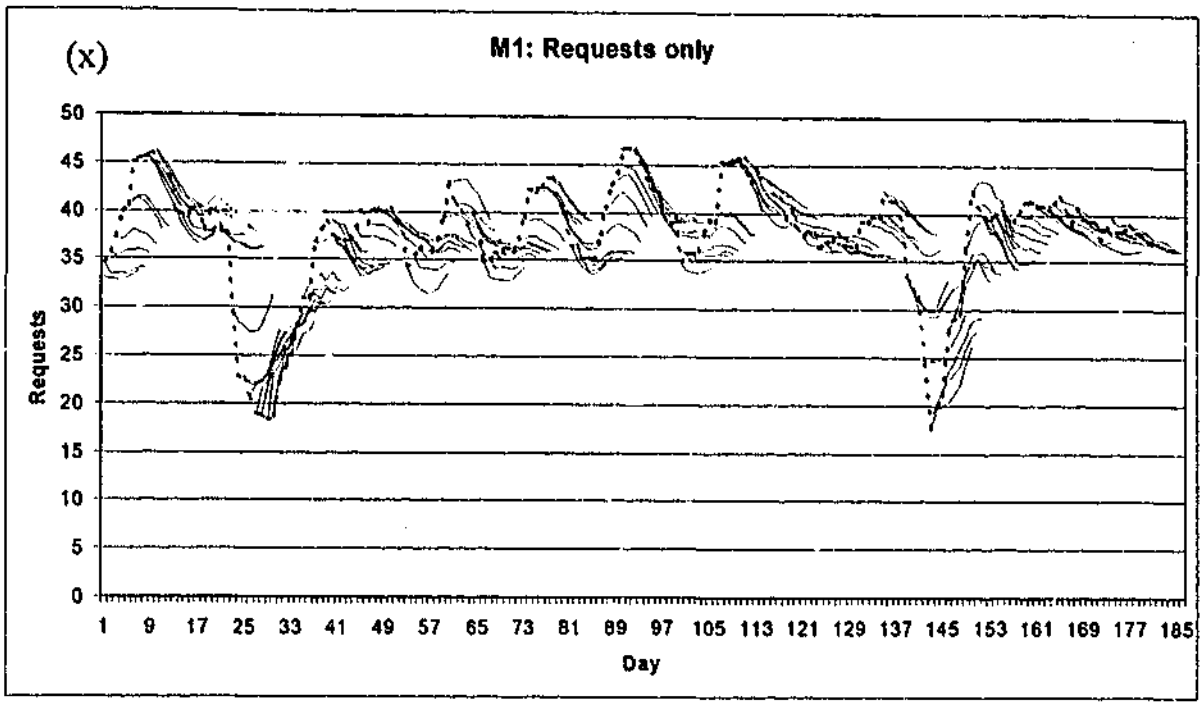
(viii)


C3: All inputs
Requests
0
10
20
30
40
50
60
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

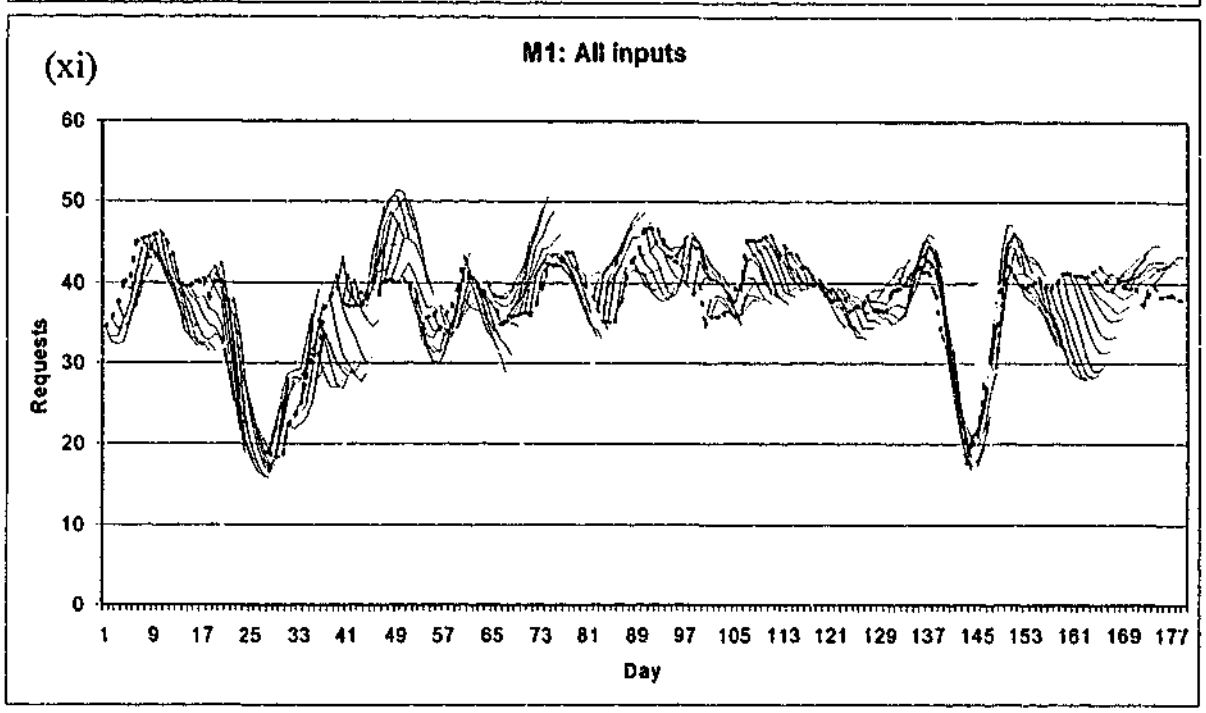
(ix)


C3: Requests, Holidays, Rainfall, MaxTemp
Requests
0
10
20
30
40
50
60
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

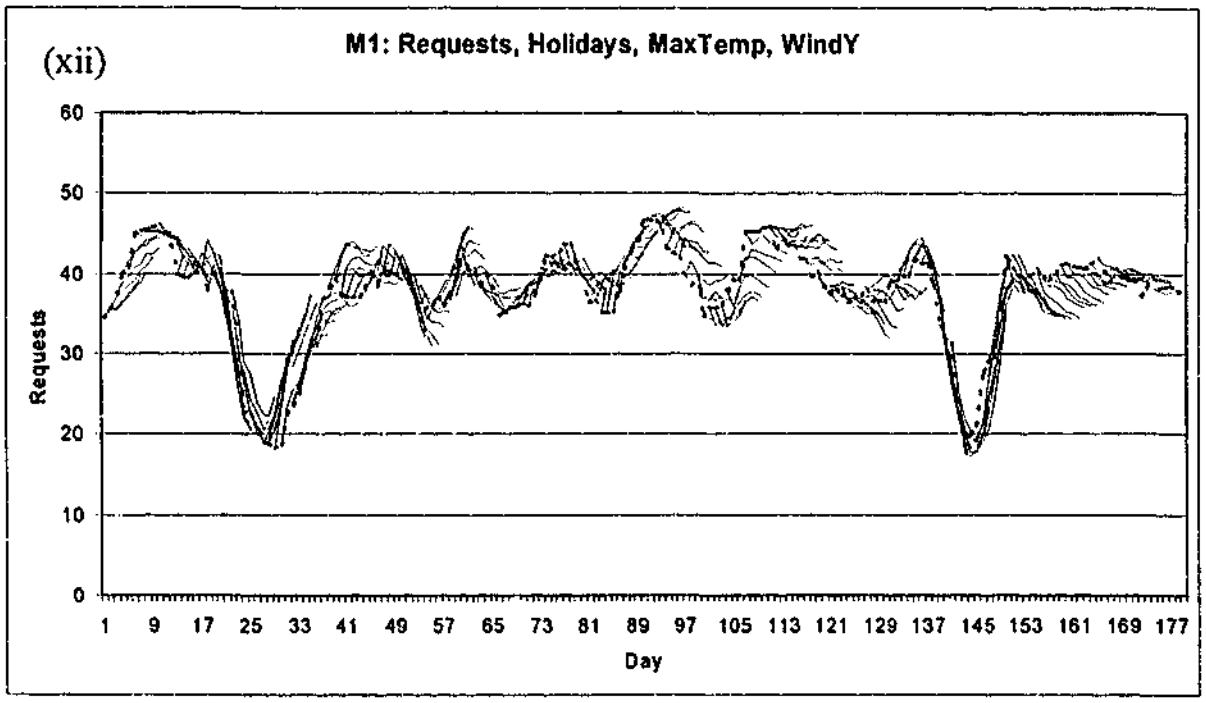
(x)


M1: Requests only
Requests
0
5
10
15
20
25
30
35
40
45
50
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177 185
Day

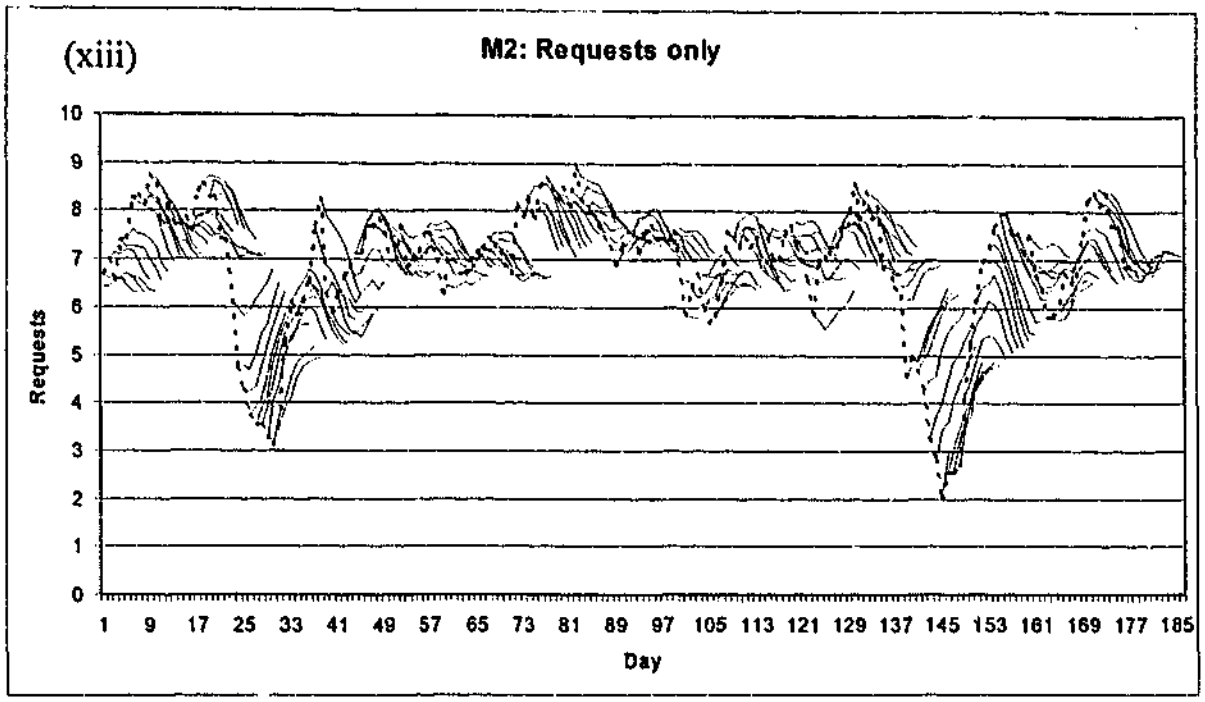
(xi)


M1: All inputs
Requests
0
10
20
30
40
50
60
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

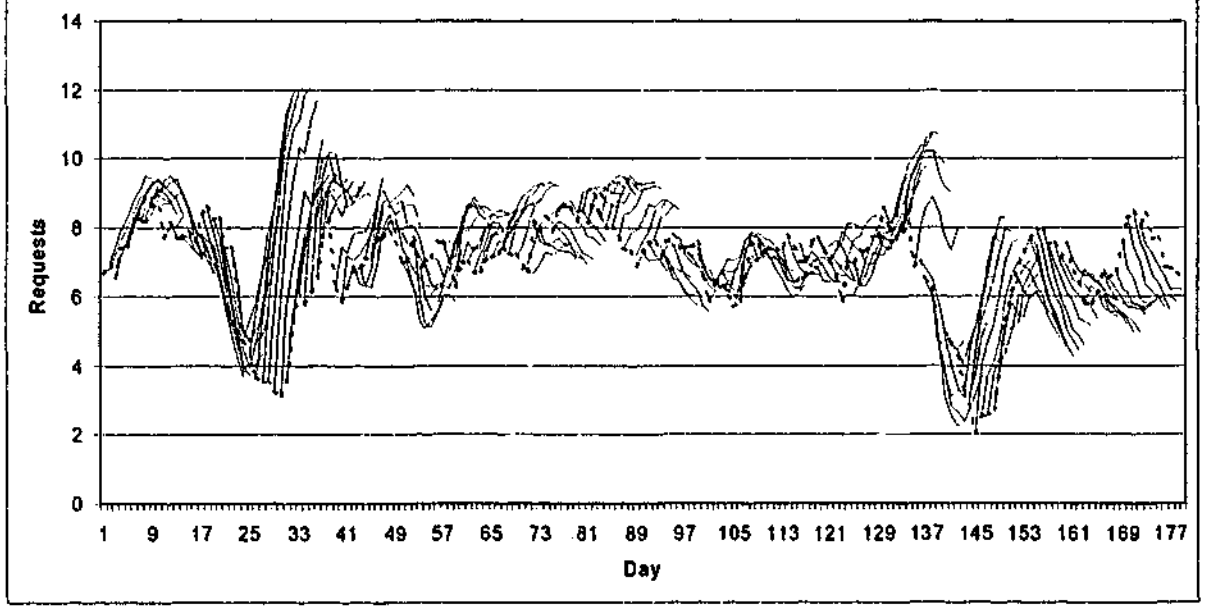
(xii)


M1: Requests, Holidays, MaxTemp, WindY
Requests
0
10
20
30
40
50
60
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

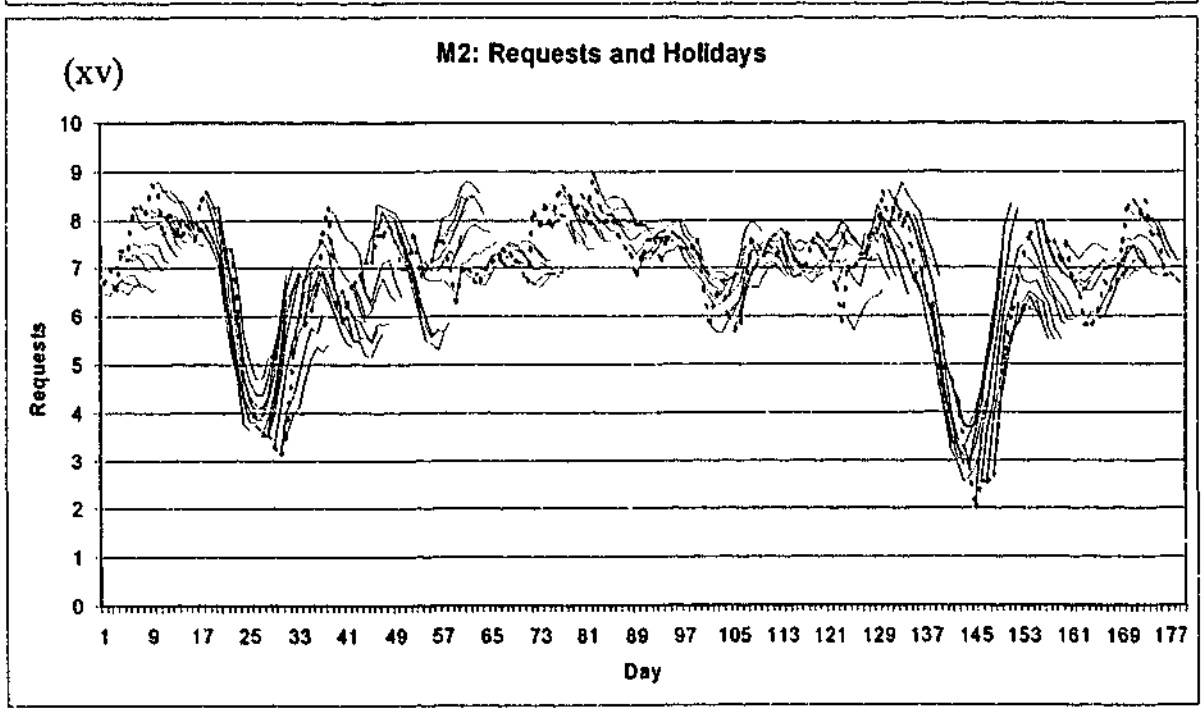
(xiii)


M2: Requests only
Requests
0
1
2
3
4
5
6
7
8
9
10
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177 185
Day

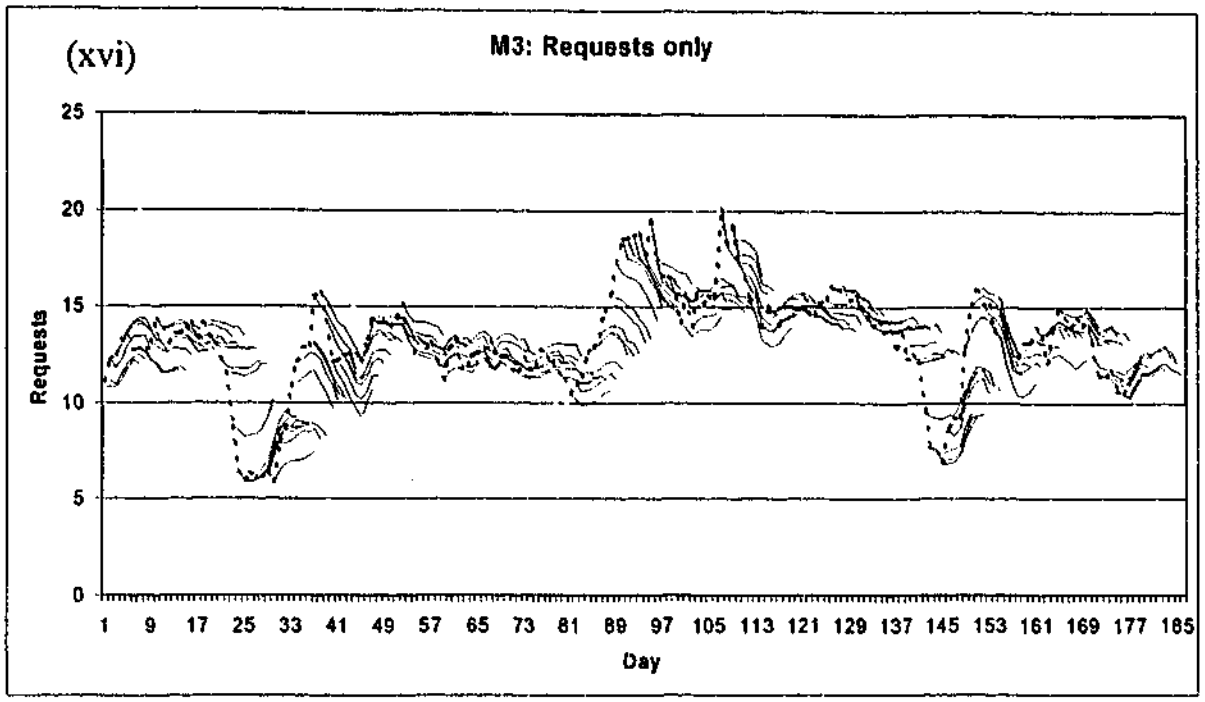
(xiv)


M2: All inputs
Requests
0
2
4
6
8
10
12
14
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

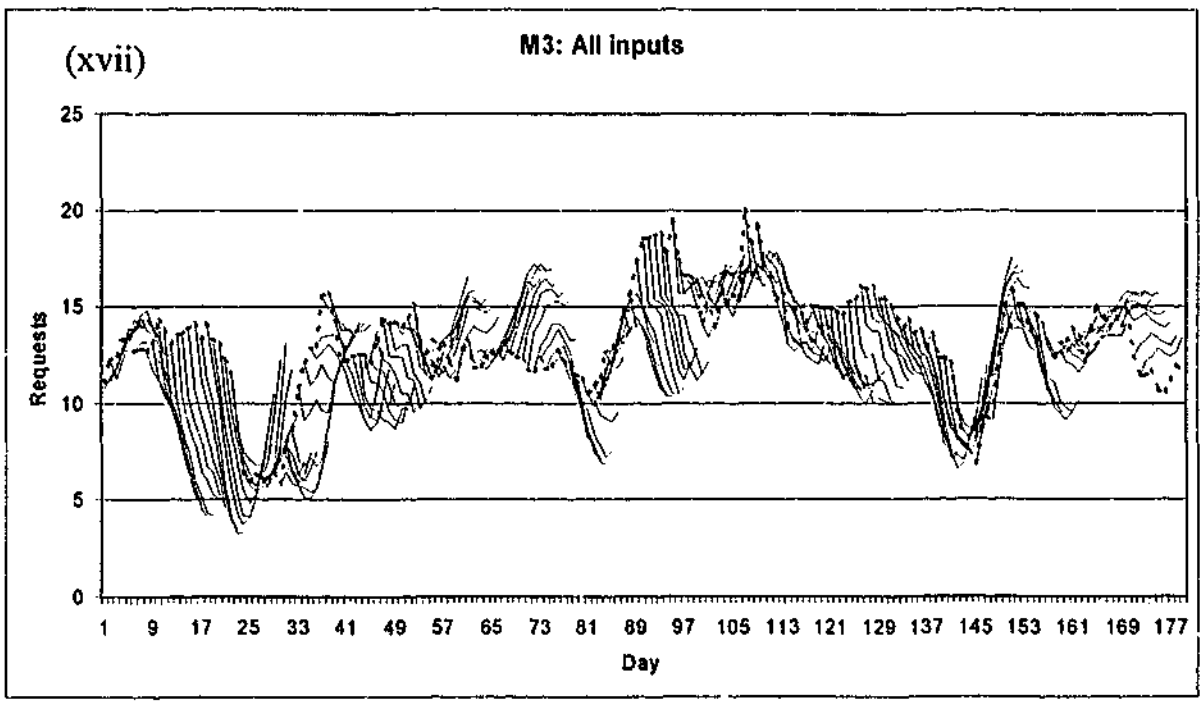
(xv)


M2: Requests and Holidays
Requests
0
1
2
3
4
5
6
7
8
9
10
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177
Day

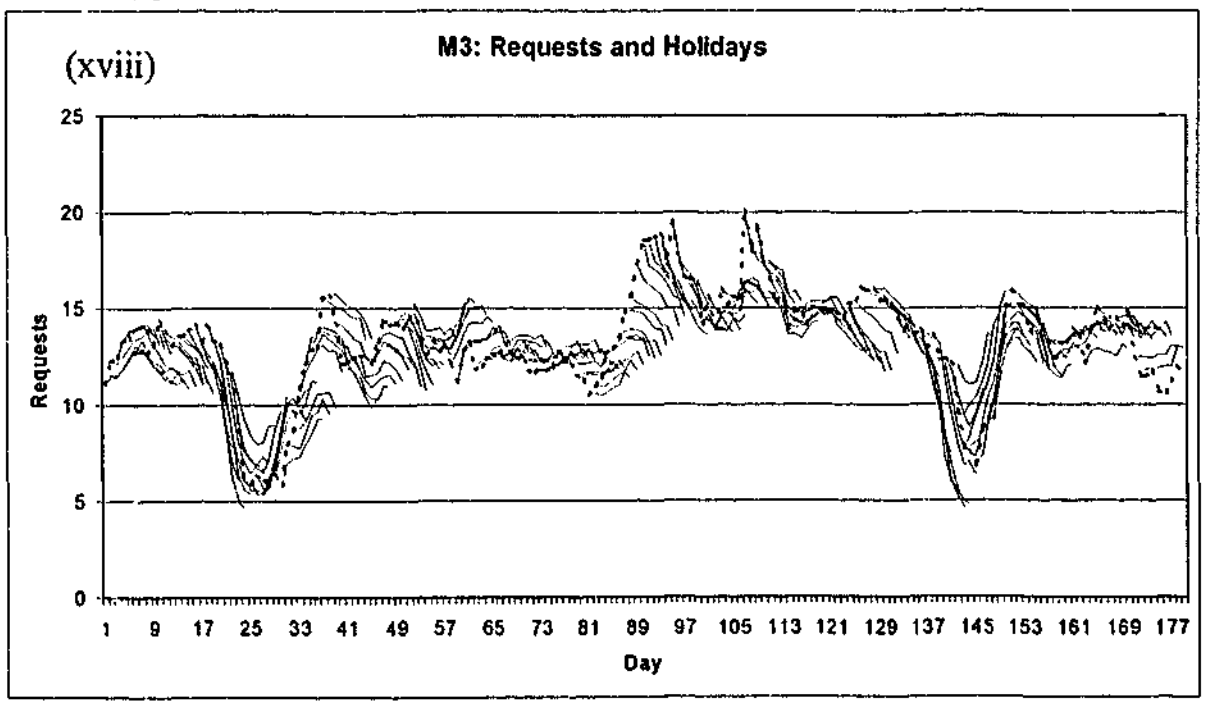
(xvi)



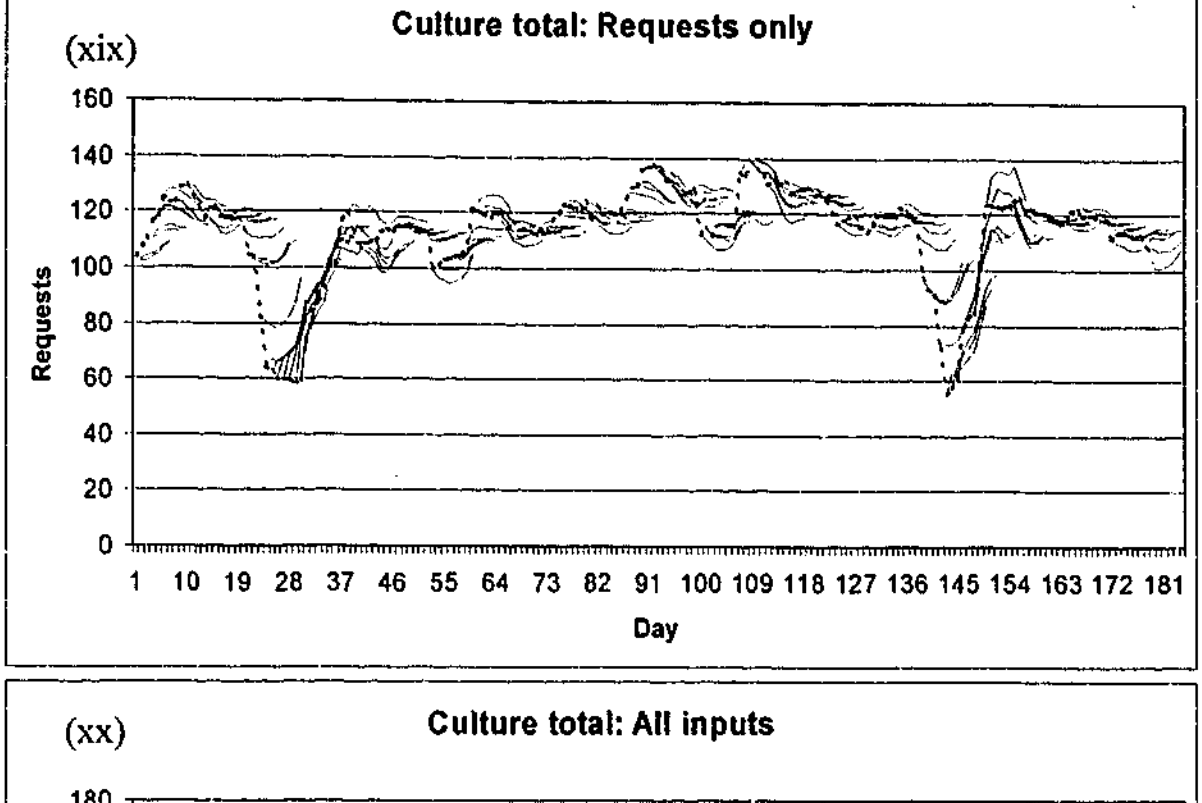
(xvii)



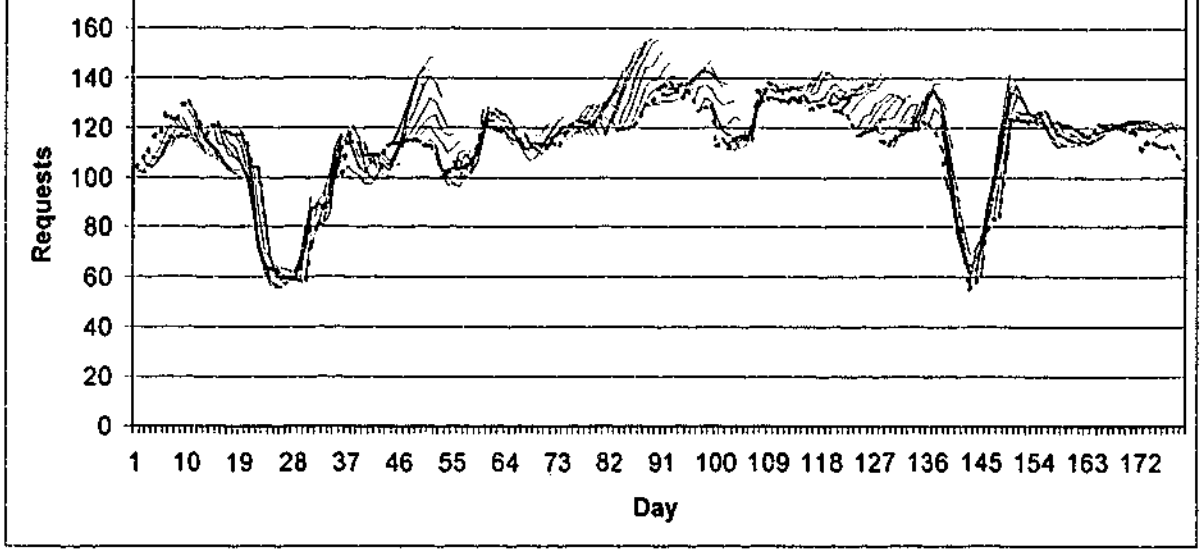
(xviii)



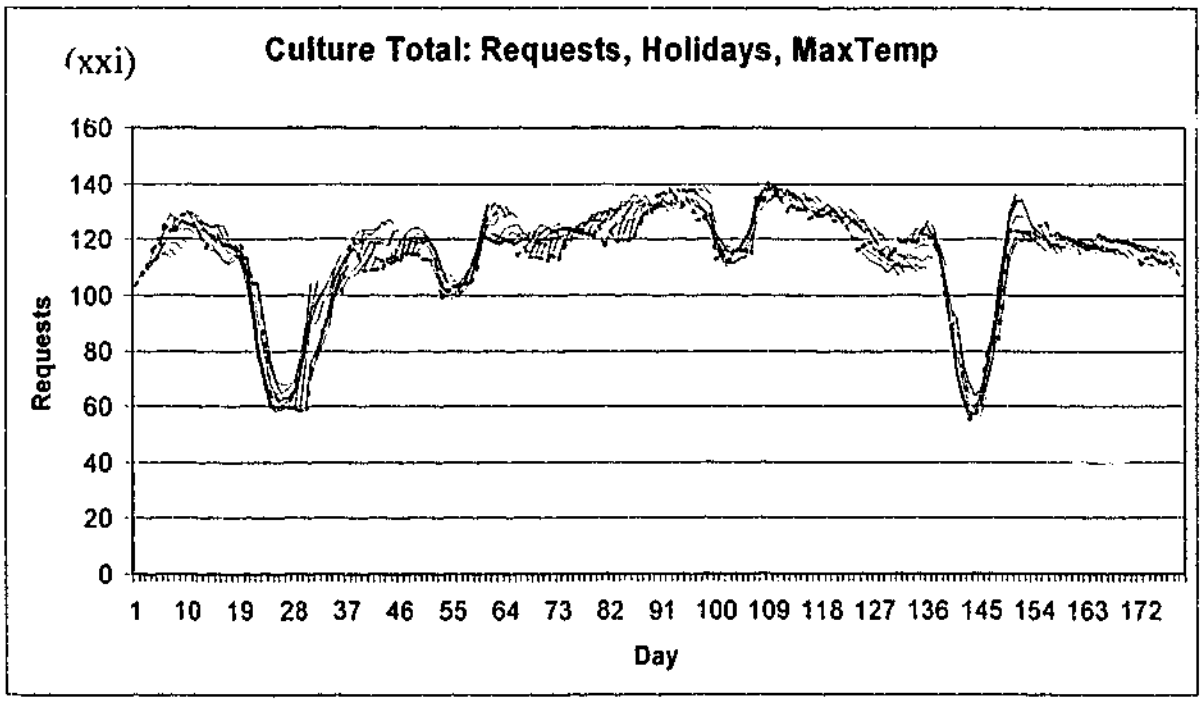
(xix)


Culture total: Requests only
Requests
0
20
40
60
80
100
120
140
160
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181
Day

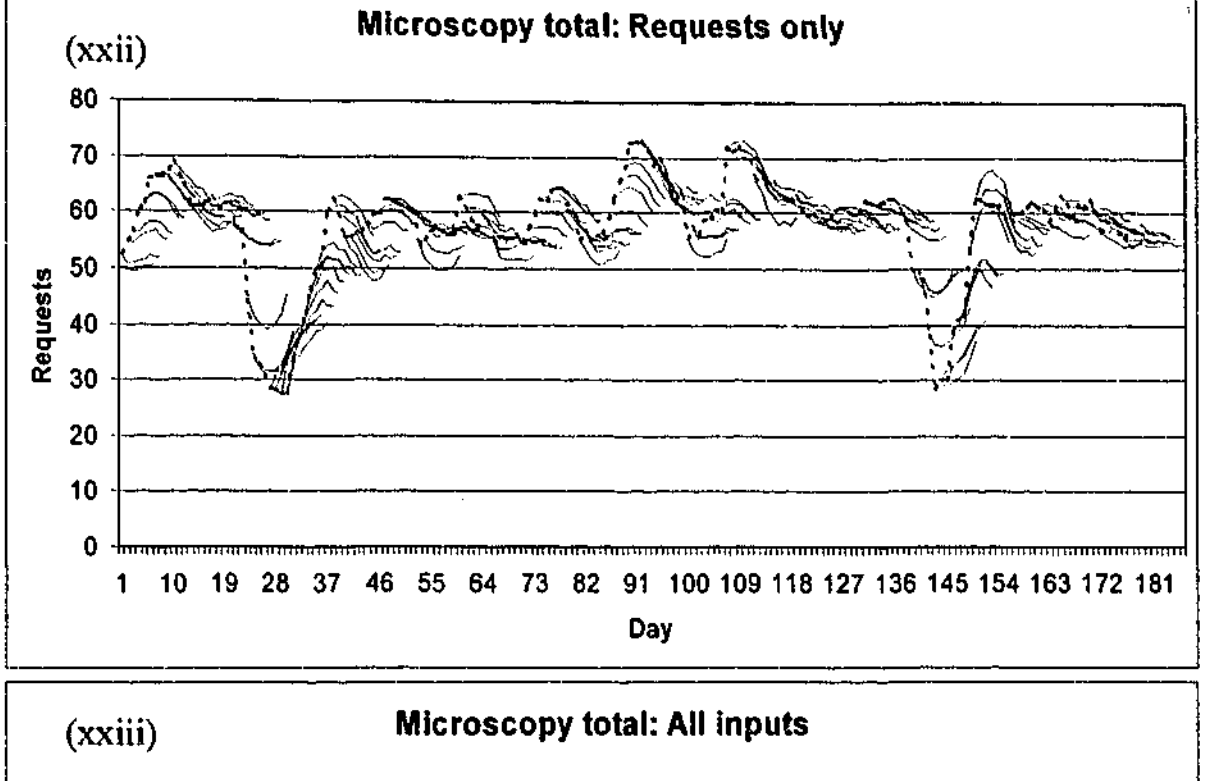
(xx)


Culture total: All inputs
Requests
0
20
40
60
80
100
120
140
160
180
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172
Day

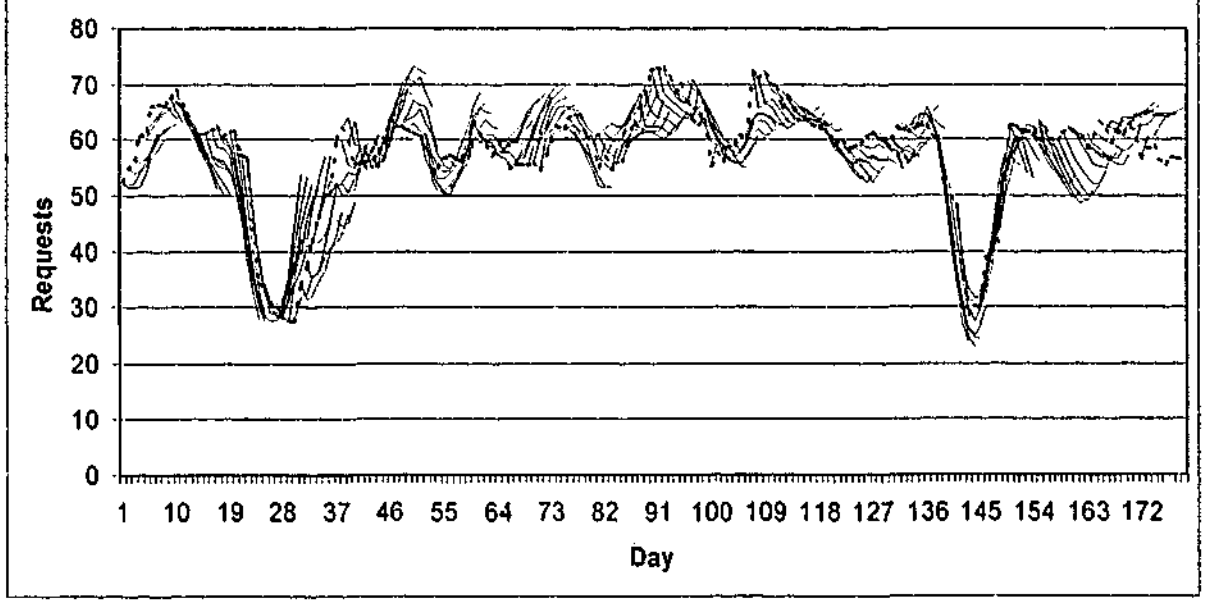
(xxi)


Culture Total: Requests, Holidays, MaxTemp
Requests
0
20
40
60
80
100
120
140
160
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172
Day

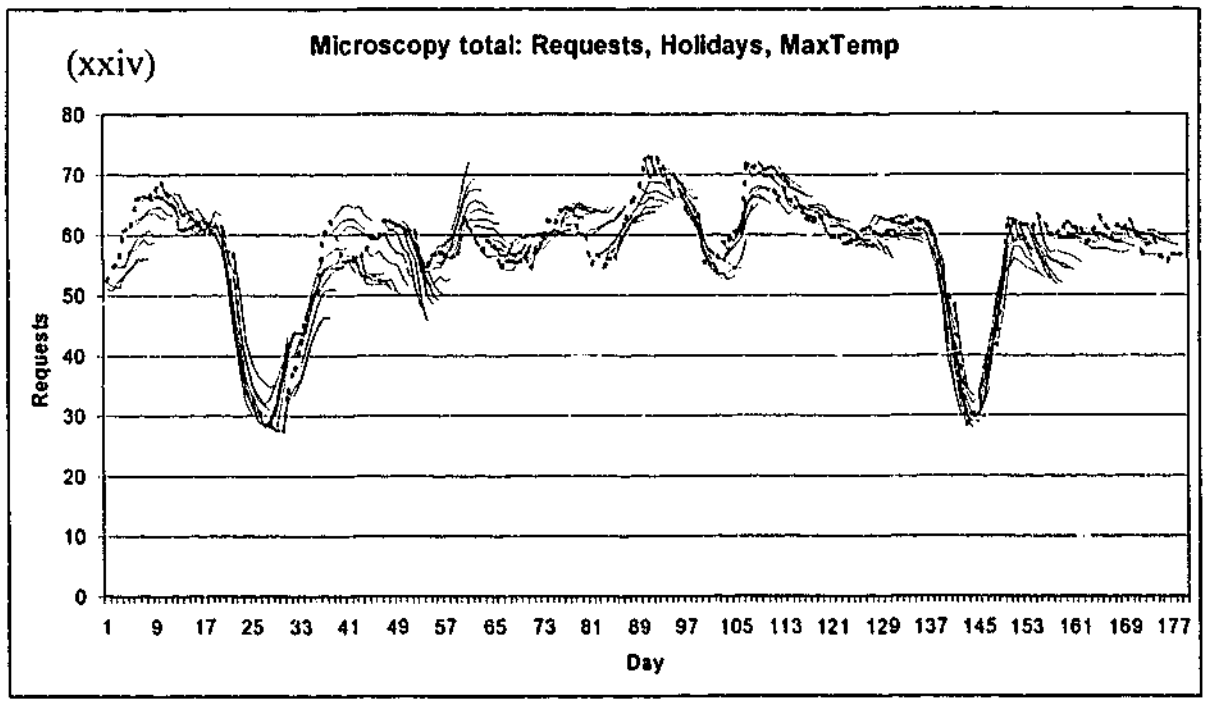
(xxii)



(xxiii)



(xxiv)



Discussion

This study demonstrates that, at a potentially useful scale in time and space (i.e. daily reports for a whole city), artificial neural network models can accurately generalise the relationships between disease, weather and social events in the recent past, and future requests for faecal analysis. For some of the models the same relationships hold well enough into the future to allow useful forecasts beyond the end of the training time period.

This is the first report of its type. Surveillance systems currently in use in industrialised countries either do not attempt computer-assisted detection of outbreaks, or else concentrate on essentially retrospective methods such as the 'Farrington' algorithm (Farrington et al., 1996), MMWR technique (Centers for Disease Control and Prevention, 1991), CUSUMs (O'Brien and Christie, 1997) or scan statistics (Wallenstein et al., 1989). These assess the likelihood that the latest set of observations deviates from the expected more than could be explained by the play of chance. They do not attempt to forecast even into the near future, and all tend to be hampered rather than enhanced by the presence of past outbreaks in the data set. The neural network modelling exercises described here attempt to make use of the past variation in disease incidence, as no weighting of previous outbreak years should be required for a neural network.

The ability of a relatively non-specific 'upstream' measure such as requests for faecal analysis to mirror true changes in gastroenteritis incidence, or to detect outbreaks caused by microbial contamination of drinking water, has not been formally tested here. However the suggestion that unexplained increases in total request numbers

should lead to further investigation is uncontroversial. In practice more than one network might be used (e.g. one each for total culture and total microscopies). Rising request numbers not predicted by the networks would lead first to a consideration of the geographical distribution of the increase, and then perhaps to field investigations. For example an increase spread across the distribution zone of a single water catchment would lead to investigation of the source water. A program such as MapMovie would facilitate such an analysis. (See Appendix One and the enclosed CD for details of MapMovie, together with MapMovies of the data from this chapter.)

Only one relatively large outbreak is known to have occurred during the study period, and that was a very brief foodborne outbreak in late March 1997 in which several hundred cases were related to a single food outlet and a single day of contamination (Andrews et al., 1997). As this outbreak occurred early in the data set, it was always included in the training sets, and never in the prospective test sets taken from the end of the series. The potential ability of a neural network model to incorporate aspects of past *water-borne* outbreaks into forecasts could thus not be formally tested in this study, although a number of smaller food-borne outbreaks are known to have occurred during the study period. (The localised nature of the 1997 outbreak is clearly visible in the MapMovie sequences covering that period. The full data set can be viewed from the CD in the back cover using the copy of MapMovie supplied.)

'Early stopping' – stopping the training of each network at the point of minimum error on the validation set – avoids the acknowledged potential problem of network over-training (Rumelhart et al., 1994). The high R^2 values for every indicator and every one of the seven outputs in the validation sets strongly suggests that these

networks are truly extracting generalised relationships between recent request counts, weather and public holiday events, and the request numbers in the coming week. However, only in the relatively simple models do the same relationships apply outside the time window in which they were derived. The more complex models gave misleading forecasts when applied to a true prospective test set, but some of the simplest were very accurate.

This is possibly because of the relatively small training sets available. Although the networks were trained on all the data the Health Insurance Commission was able to supply, these did not by any means include all the possible combinations of weather, holiday and disease patterns that can occur in Melbourne. With larger sets of training data it is possible that more weather inputs could be added to the models with improved prospective accuracy. Larger training sets would also allow for larger test sets – at least a year would be required to properly assess the full seasonal range of the models.

Nevertheless, this study produced forecasting models reliable enough to warrant formal field-testing. It established that neural network models incorporating recent weather events, public holidays and recent request numbers can derive generalised relationships over appropriate temporal (one week ahead) and geographic (city-wide) scales. It also emphasises the importance of suitable pre-processing of the time series to smooth the random components of the series without losing the underlying patterns. With larger data sets and further research this approach may yield a valuable new addition to the surveillance epidemiologist's toolkit.

Chapter Six: Modelling measles in Mozambique with artificial neural networks

Introduction

This chapter reports experiments aiming to forecast numbers of measles cases by province in the southern African country of Mozambique.

Why choose this disease and this data set?

There are no published accounts of the application of artificial neural networks to the forecasting of measles. As discussed in the literature review above, there is currently no reliable or accurate method for forecasting outbreaks.

Predicting measles cases would assist control efforts

Measles is a highly contagious viral disease affecting only human beings (Cutts et al., 1999, Chin, 2000). This means that the only way a susceptible person can acquire the disease is by contact with another person who either has the disease or is late in the incubation period. Although it may never be possible to predict exactly which child will have contact with the measles virus in a given week, in a highly affected country it is likely that the evolution of a measles outbreak is never a completely random process (Anderson and May, 1991, Grenfell et al., 1995). It has long been assumed that (at the right scale in space and time) a time series of measles reports is composed of a strongly deterministic component together with an overlying random component (Hamer, 1906, Grenfell et al., 1995). It may thus be possible to model the relationship between past and future measles case numbers and make predictions into the near future.

From the perspective of an infectious disease epidemiologist working on the surveillance and control of measles this is more than an academic question. If it were possible to predict which province was next in line for an outbreak, and how big that outbreak was likely to be, both preventive and curative actions could be directed more efficiently. With six or eight weeks advance warning of an impending outbreak, vaccine stocks could be reviewed and replenished, and a catch up vaccination campaign launched to pre-empt the outbreak. Even if the warning period were too short to allow prevention of the outbreak, it would still be possible to mobilise resources to tackle the caseload. Perennially scarce drug supplies (such as vitamin A, oral rehydration salts, and antibiotics) and/or medical teams could be redirected to the soon-to-be-affected areas. Clinical staff could be warned to raise their alertness, and perhaps even given refresher training if necessary.

Measles in Mozambique is typical of Africa, but this data set is unusually complete

Mozambique's experience of measles is typical for sub-Saharan Africa, where measles is essentially a disease of young and school-age children (Cutts et al., 1994, World Health Organization, 1998). At the national level a measles outbreak occurs almost annually, but any of the provinces may pass several years with few or no cases.

The fact that this data set exists, with ten years of weekly data for every district, is fairly unusual for an African country. Since 1985 Mozambique has had a national infectious disease surveillance system (based on the *Boletim Epidemiológico*

Semanal, or *BES*) which includes weekly totals of measles cases seen by each district. The data from 1989 to the present are computerised and stored in dBase IV format.

I was closely involved with the collection and use of the data set during his years as a clinician in several rural and urban settings in Mozambique, and then as a provincial epidemiologist in the west-central province of Manica. I wrote the computer software used to record and analyse the disease surveillance data throughout the country since 1994, and converted the earlier data from an obsolete database format.

Measles may be predictable

It has been recognised for at least a century that there are theoretical reasons why the spread of measles might be predictable (Hamer, 1906, Fine and Clarkson, 1982, Anderson and May, 1991, Grenfell et al., 1995). Measles has no non-human reservoir, and no other means of transmission than personal contact. Despite global efforts at control (Cutts et al., 1994, Cutts et al., 1999, World Health Organization, 1998) it remains a very common disease in Africa (Dgedge et al., 2001), which also means it is almost entirely a disease of children there. Population movement in Mozambique is relatively slow and movement between provinces is limited to a few main transport routes. Apart from the provinces bordering other countries, virtually every child with measles recorded in this data set contracted measles by contact with another infected child in Mozambique. It seems likely that the measles dataset is composed of a strongly deterministic component related to the previous case numbers in the same and neighbouring provinces, overlaid with a completely random component. At different temporal and geographical scales one or other of the components tends to predominate. (For example, aggregating the data by year or for the whole country provides obvious patterns, while weekly case numbers for a single district often seem

essentially random). The objective of this work was to find a way to emphasise and model the deterministic component at a scale that would still allow useful control measures to be taken.

Neural networks have theoretical advantages

Among the available statistical tools, artificial neural networks offer a good theoretical prospect of modelling the complex patterns of spread of measles in an endemic area (Cheng and Titterington, 1994). The internal workings of a neural network make no assumptions about the stationarity of the time series (Hinton, 1992, Masters, 1995). (This is not to say that a neural network will not be adversely affected by non-stationarity, but it has a better chance of surmounting that obstacle than a technique that explicitly demands stationarity.) Further, the essential non-linearity of a neural network means that it can potentially detect more complex relationships between the variables in the model than a conventional parametric algorithm (Cheng and Titterington, 1994).

From the beginning, however, the measles time series for Mozambique give some clear warnings to the analyst. Up to 1993 the country was riven by war, and both the number and location of reporting health facilities varied. After the peace accord and elections (in 1993/94) the number of health facilities increased, as did the proportion sited in remote rural areas.

Methods

Preliminary experiments with the data

Many initial attempts have been made in this research to forecast disease numbers for single districts. A large number of experiments were performed with the raw data, using different numbers of district inputs and different lags of the time series. A number of different smoothing schemes were attempted, and also aggregation of the data at different geographical and temporal scales (e.g. by province, groups of provinces, national level, by week, month or quarter). In all around 50 separate series of experiments were conducted. The results presented here are the most illustrative.

Data used for this study

After pre-processing, 542 weeks of data, from week 16 in 1989 to week 34 in 1999, were available for use in model creation and testing. Figure 6-1 shows the distribution of cases by province. All the data were obtained from the Epidemiology department of the Mozambican Ministry of Health. The raw time series came directly from the national database of Weekly Epidemiological Bulletins (*Boletim Epidemiológico Semanal*) generated by the routine disease surveillance system since the 1980s. The ministry maintains a separate database for each province, in which a single record represents the numbers of cases and deaths of a number of important infectious diseases for a single district for a single epidemiological reporting week. The reporting weeks are numbered from the beginning of each year, and each runs from Sunday to Saturday.

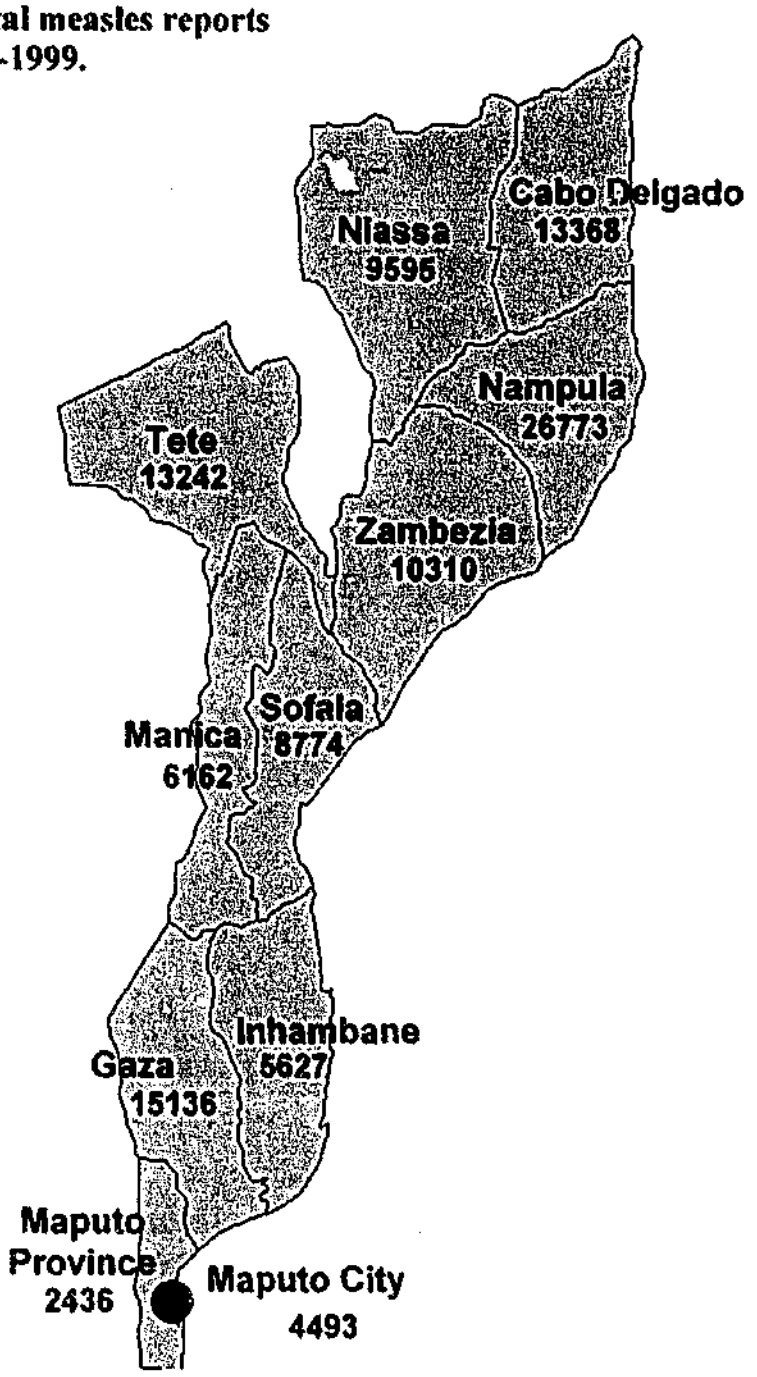


Figure 6-1: Total measles reports in Mozambique, 1989 to 1999, by province.

Pre-processing of the data

The raw time series data sets for individual districts have a strong random component, and it is my experience that there have been many occasions where data recorded for a district for a single week actually represent the accumulation of several weeks' observations. Conversely, there are gaps in the series which do not truly represent zero cases. Together with the fact that the main epidemiological surveillance and

outbreak response decisions are generally made at the level of the provincial health directorate, it was most appropriate to attempt to model measles reports for whole provinces rather than single districts. (Maputo City, the national capital, has the status of a province within the disease surveillance system.)

Measles outbreaks usually evolve over several months, and week-to-week variation even at provincial level is relatively large. It was decided to smooth the provincial time series with a type of moving average filter. To avoid the end effects that are inherent in symmetrical moving average filters and limit their usefulness for near-future forecasting, the filter used was simply the mean number of cases in the six-week period ending that week.

Network architecture, inputs and training

Two distinct approaches were taken to the creation and testing of the neural network models. In **Approach 1** about fifteen percent (70 weeks) of the original data set was randomly selected and reserved before training began. In **Approach 2** the last 70 weeks were reserved as a single contiguous block. Approach 1, using a randomly selected test set, provides a good test of the network's ability to 'learn' the features of the data set as a whole. But Approach 2, using a test set based on the last 70 weeks, is a more realistic test of how the network would perform in the true field situation. In both approaches the remaining data after the test set was reserved was further divided into a training set and a randomly selected validation set of 72 patterns. The validation set was used during training to determine the optimum generalisation, but not to adapt the network connection weights. (See below, and also the section of Chapter Two titled "Use a validation set for 'early stopping' of training" for details of this technique.)

All the data used for training and prediction were the new time series resulting from the pre-processing. A separate multi-layered feed-forward neural network was created for each province, using the NeuroShell2 package (Ward Systems Group Inc, 2000). Each network had 96 inputs, being the weekly count and its first six lags for each province and the national total, plus the rate of change (velocity) over the past four weeks. Each network had one hidden layer of 250 nodes, and as outputs the next eight leads for the smoothed weekly measles count for that province. (The first lead is the smoothed number of cases one week in the future, the second lead the case numbers two weeks ahead, and so on.) The optimum number of nodes in the hidden layer was arrived at by estimating with a common rule-of-thumb (half the sum of inputs and outputs, plus the square root of the number of patterns in the training file (Smith, 1999)), and then empirically tested by the creation of a number of models with greater and lesser numbers of hidden layer nodes.

The data were further pre-processed in the software package by linear scaling of all inputs to a range from –1 to +1. Each network was initialised with random weights for all connections between –0.3 and +0.3.

Model training followed a typical methodology (Rumelhart, 1986, Masters, 1993). Each network was trained by back-propagation of errors, using learning rate and momentum terms each of 0.1. The mean squared error (of the model output compared with the observed case numbers) for the whole training set was used as the training criterion. In each case 72 patterns (about 15% of the initial data set) were randomly selected and removed from the training set before training began. This subset of the data was used for validation after every 200 training patterns had been presented to

the network (each one presentation being a single 'learning event'). Training was stopped when 300,000 learning events had occurred without further improvement in the error on the validation set. In an attempt to maximise the ability of the trained network to generalise, the final network configuration used was the one corresponding to the lowest error ever recorded for the validation set.

Assessing the models

There are several ways in which the usefulness of a neural network trained to forecast disease incidence might be assessed, and each approach assesses a different aspect of the network's performance. We can assess the overall ability of the network to model the relationship between the inputs and outputs by calculating the mean squared error (MSE) for the training, validation or test data sets, taking all the eight outputs as a whole (Masters, 1993). With large neural networks there is always a danger that the network will simply 'memorise' the training data, and make inaccurate predictions for any data set it has not previously encountered. The mean squared error for the validation set is a good measure of the extent to which the trained network has generalised, and the network is considered optimally trained when the error for the validation set is at its lowest. We can assess the goodness of fit of the network model by calculating a version of R^2, the coefficient of multiple determination, for each of the separate outputs (Hosmer and Lemeshow, 2000). This R^2 is a comparison between the network model and a simple benchmark in which every prediction is the mean of all the observed values. It is calculated by taking the sum of the squares of the differences between the actual value and the predicted one, dividing by the sum of the squares of the differences between the actual values and the mean, and subtracting the result from 1. The closer the R^2 is to 1, the better the model's predictions (Altman,

1991). R^2 values below zero indicate that the model predictions were worse than could be attained by simply predicting the mean number of cases every week.

Perhaps more importantly (to assess the usefulness of the model in the field), we can plot the predictions it makes at each point along the time series against the actual values of the time series for that time point. This will show whether the model tends to accurately predict important turning points such as the onset or peak of an outbreak, which may be almost as useful to the surveillance epidemiologist as perfectly correct predictions of case numbers.

Results

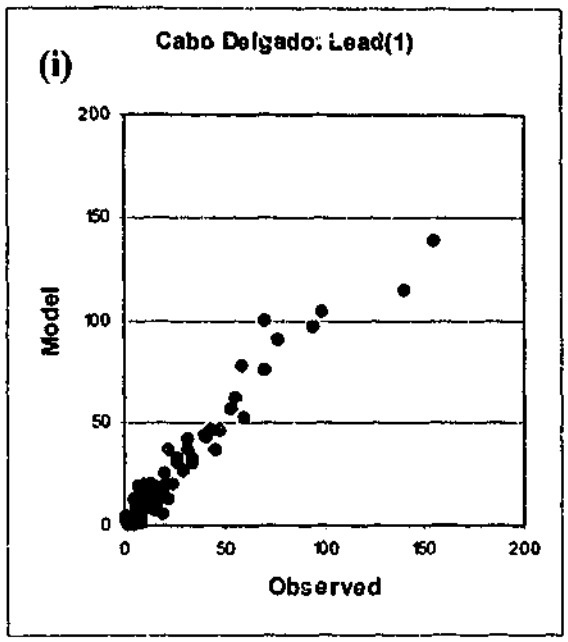
Table 6-1 lists the mean squared error for each trained network. As expected, the error for the validation set levelled off or began to rise some time before the error on the training data. The errors were generally lower for the models with randomly selected test sets than for those tested on the last 70 weeks. The mean squared errors for the validation data sets were good for both approaches, showing that the neural networks have truly learnt a generalised relationship between past and future measles reports rather than simply 'memorising' the training data.

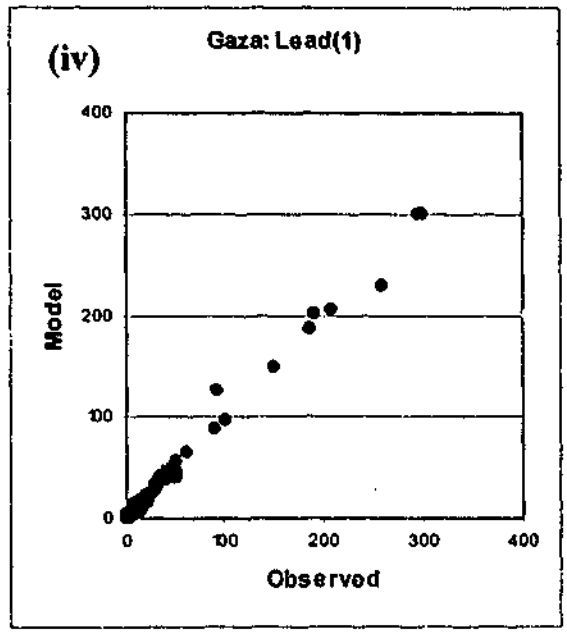
Table 6-1: Mean squared error for each trained network.

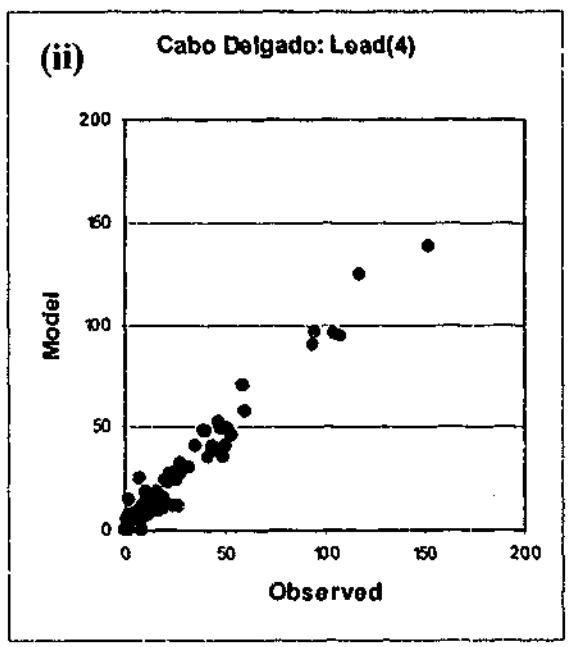
	Approach 1: Randomly-selected test set:		**Approach 2: Last-70 week test set:**	
	Training	Validation	Training	Validation
Cabo Delgado	0.0066	0.0122	0.0078	0.0106
Gaza	0.0004	0.0029	0.0001	0.0015
Inhambane	0.0006	0.0009	0.0001	0.0010
Manica	0.0007	0.0049	0.0015	0.0048
Maputo City	0.0025	0.0071	0.0014	0.0048
Maputo Prov.	0.0008	0.0067	0.0007	0.0034
Nampula	0.0012	0.0056	0.0012	0.0100
Niassa	0.0084	0.0205	0.0028	0.0214
Sofala	0.0026	0.0052	0.0013	0.0039
Tete	0.0007	0.0040	0.0004	0.0040
Zambezia	0.0004	0.0074	<0.0001	0.0088

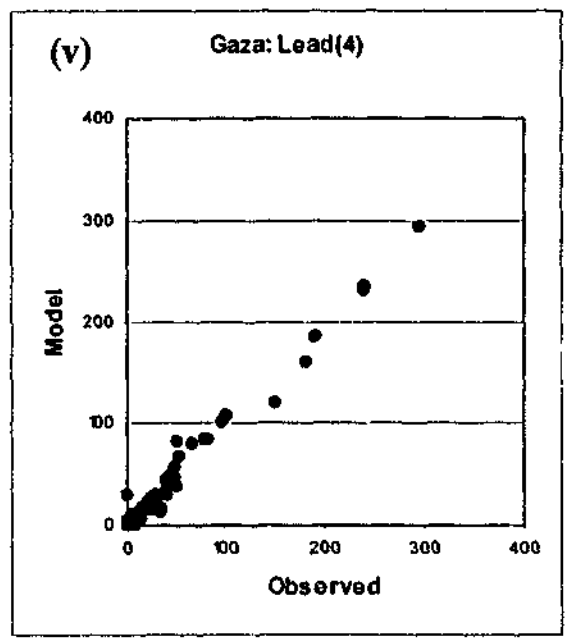
Figure 6-2 demonstrates the close fit between the trained model and the training data (although this does not tell about generalisation).

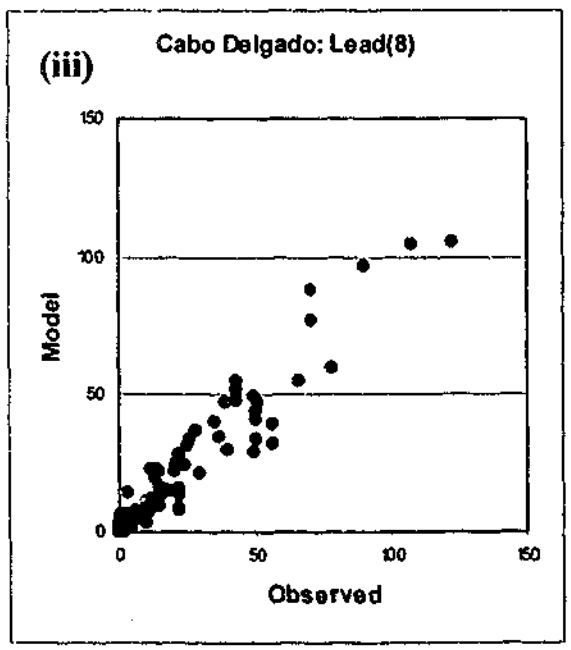
Figure 6-2 (i-xxxiii): Scatter plots (observed vs. model prediction for 1, 4 and 8 weeks ahead) for the test set in Approach 1 (where the test set was randomly selected from the same time window as the training set).

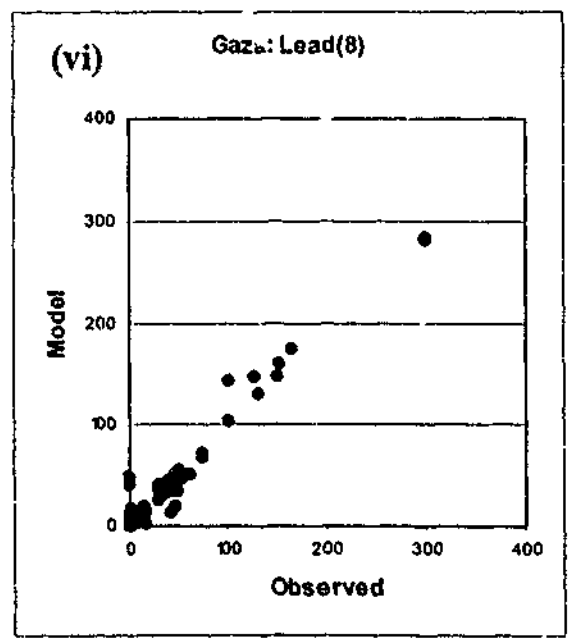


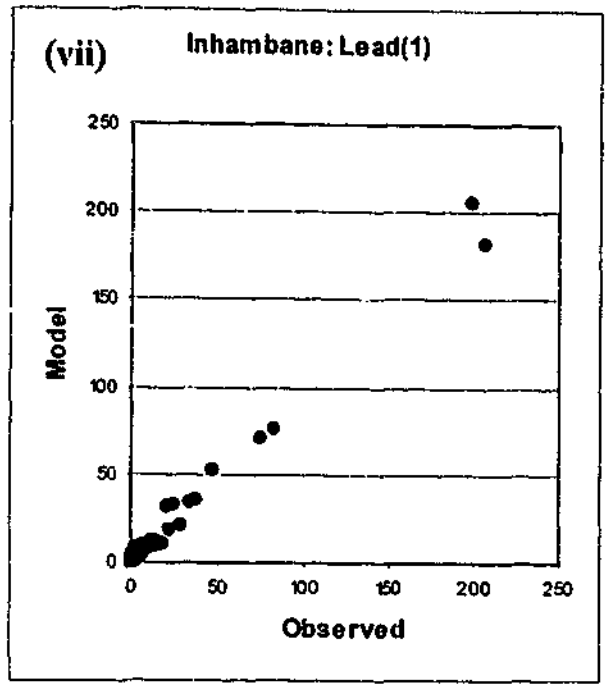


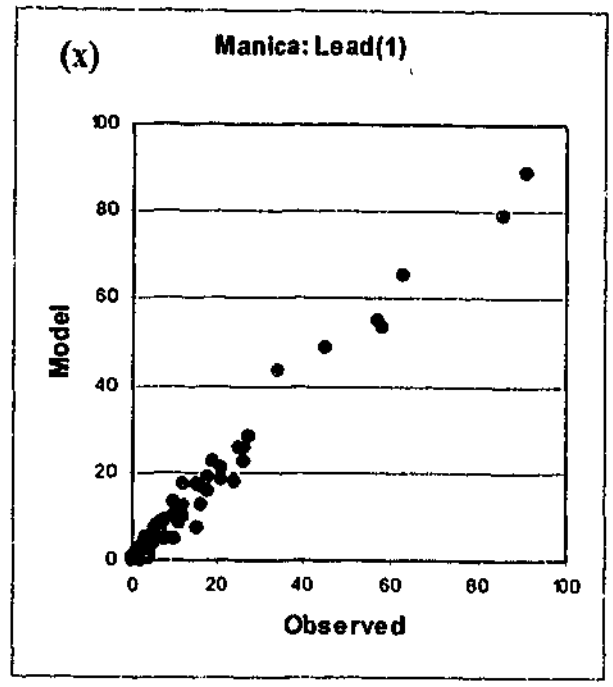


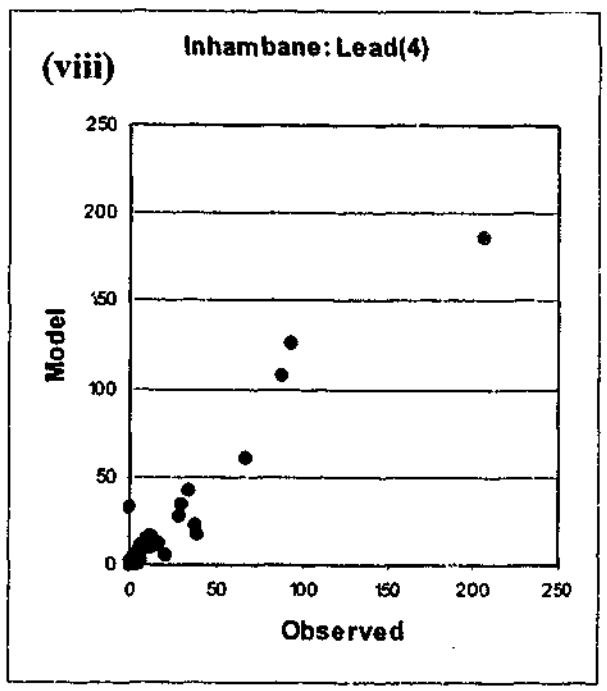


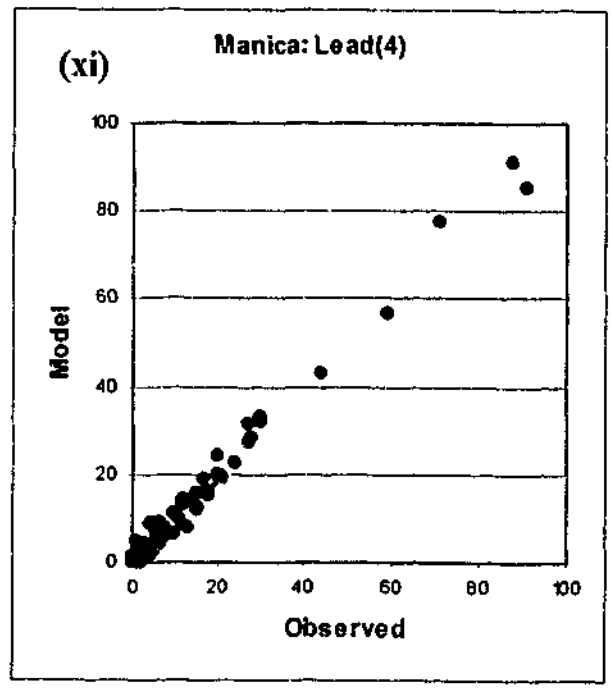


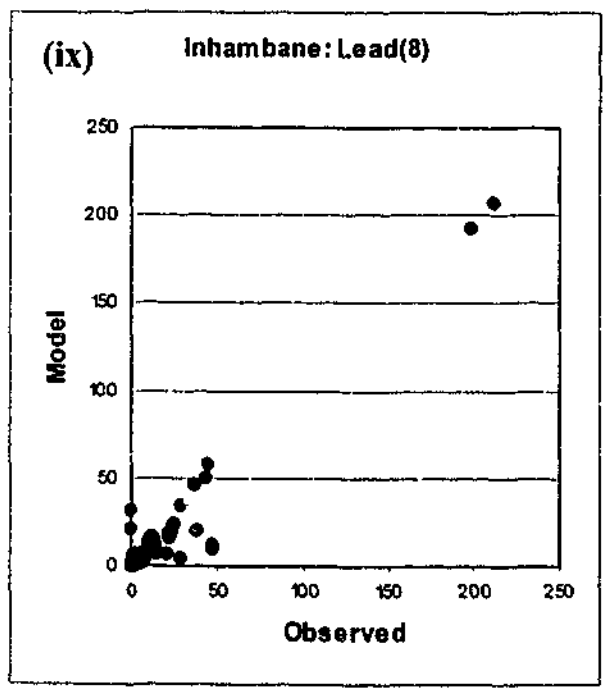


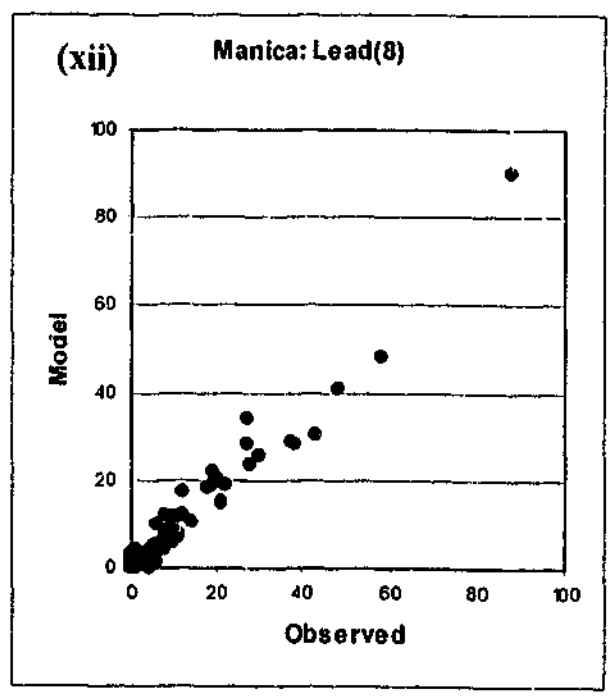

(vii)
Inhambane: Lead(1)
Model
Observed
0
50
100
150
200
250


(x)
Manica: Lead(1)
Model
Observed
0
20
40
60
80
100

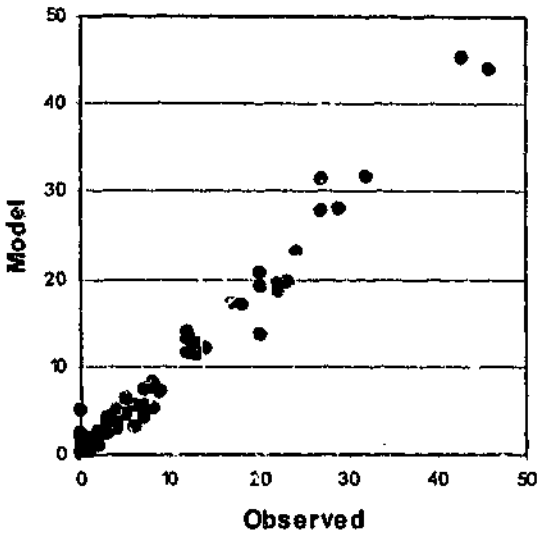

(viii)
Inhambane: Lead(4)
Model
Observed
0
50
100
150
200
250


(xi)
Manica: Lead(4)
Model
Observed
0
20
40
60
80
100

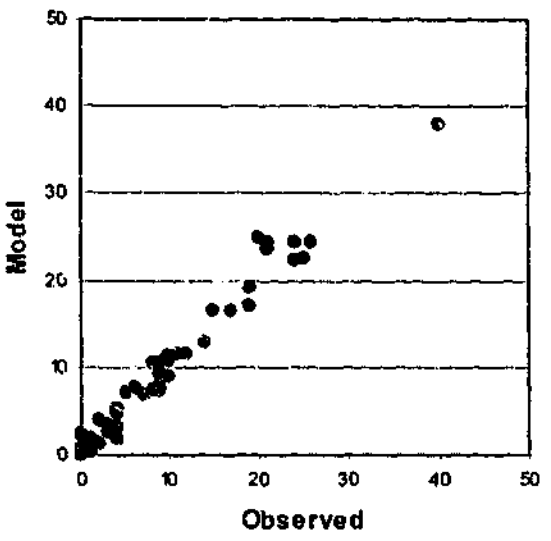

(ix)
Inhambane: Lead(8)
Model
Observed
0
50
100
150
200
250


(xii)
Manica: Lead(8)
Model
Observed
0
20
40
60
80
100

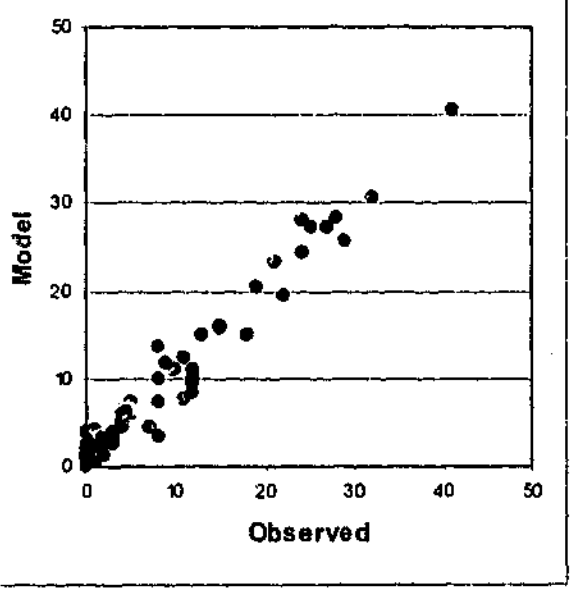
(xiii) Maputo City: Lead(1)



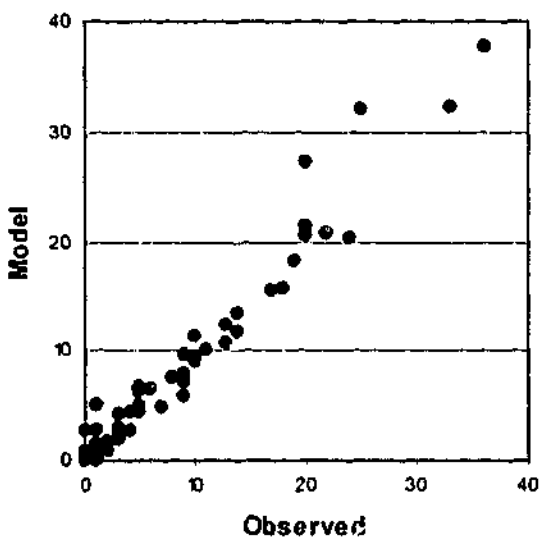
(xvi) Maputo Province: Lead(1)



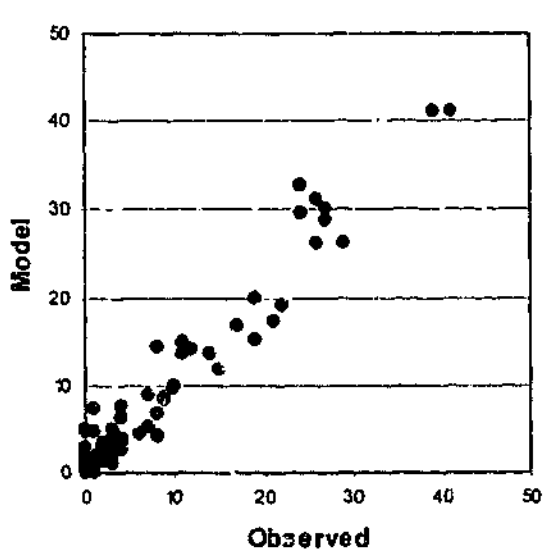
(xiv) Maputo City: Lead(4)



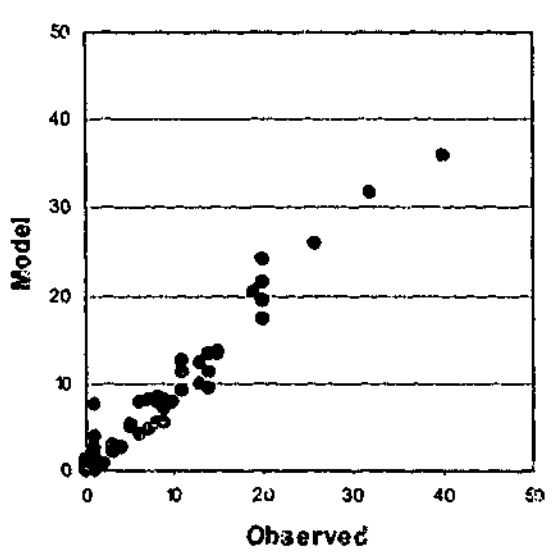
(xvii) Maputo Province: Lead(4)

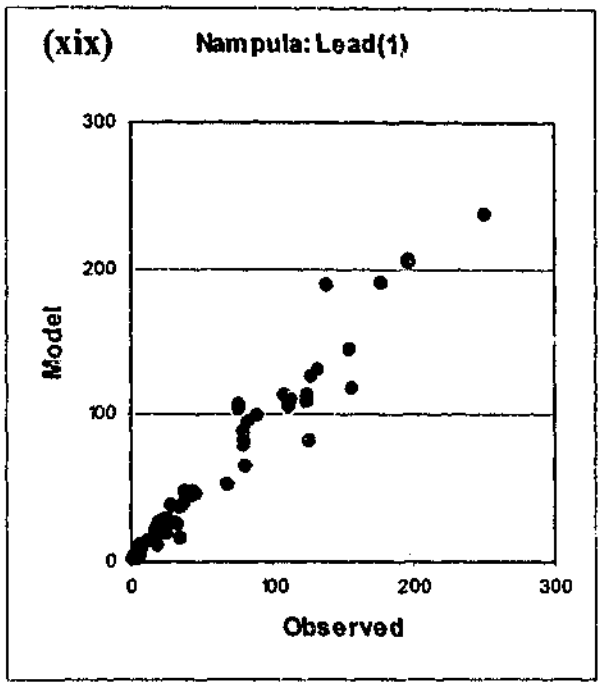


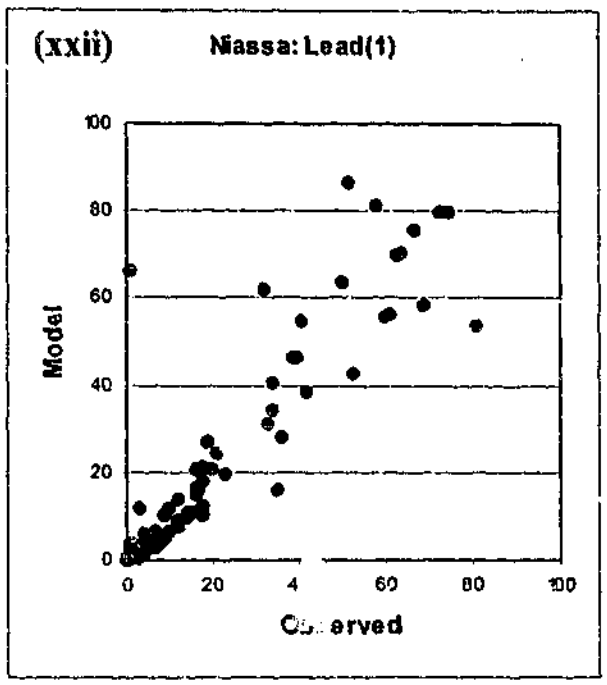
(xv) Maputo City: Lead(8)

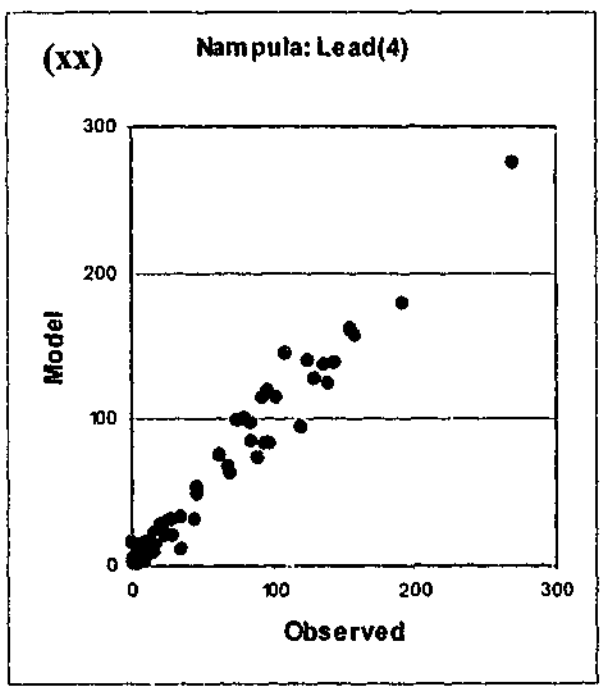


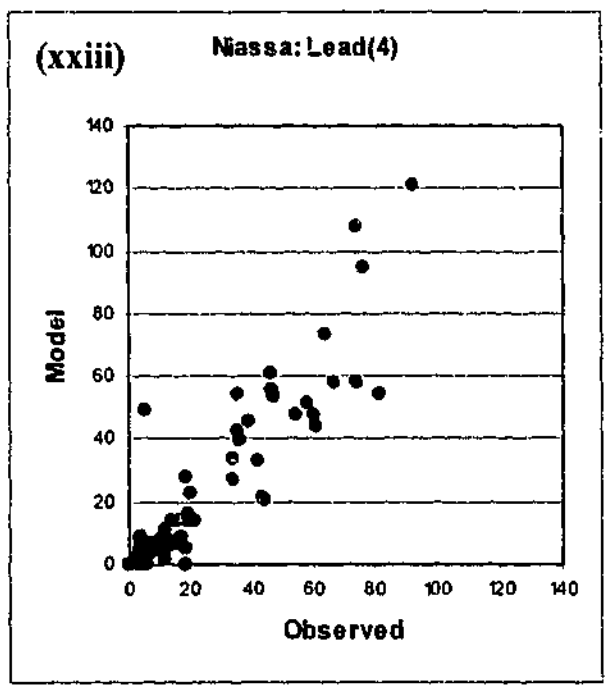
(xviii) Maputo Province: Lead(8)

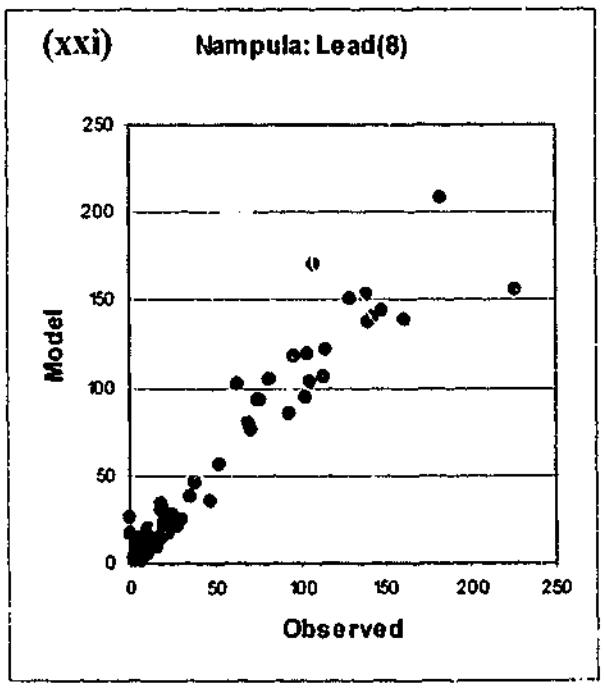


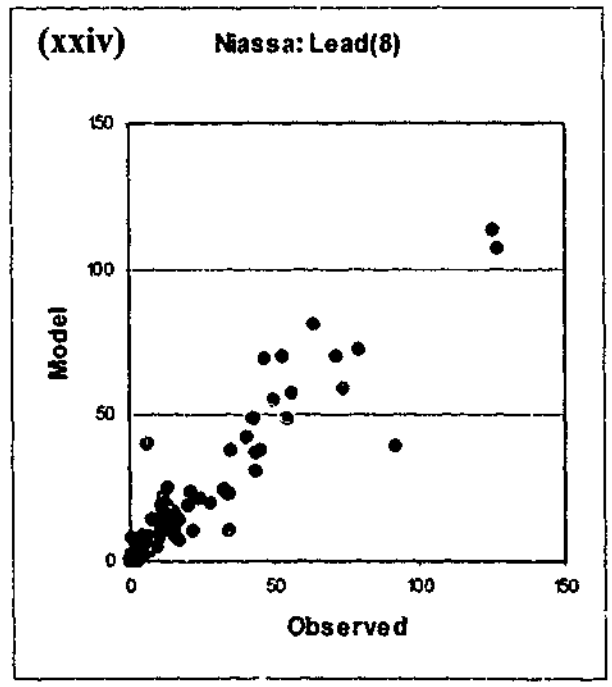

(xix)
Nampula: Lead(1)
Model
Observed
0
100
200
300


(xxii)
Niassa: Lead(1)
Model
0
20
40
60
80
100

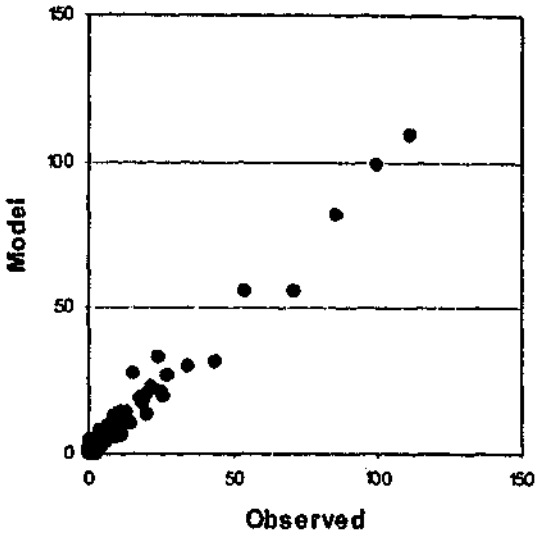

(xx)
Nampula: Lead(4)
Model
Observed
0
100
200
300


(xxiii)
Niassa: Lead(4)
Model
Observed
0
20
40
60
80
100
120
140

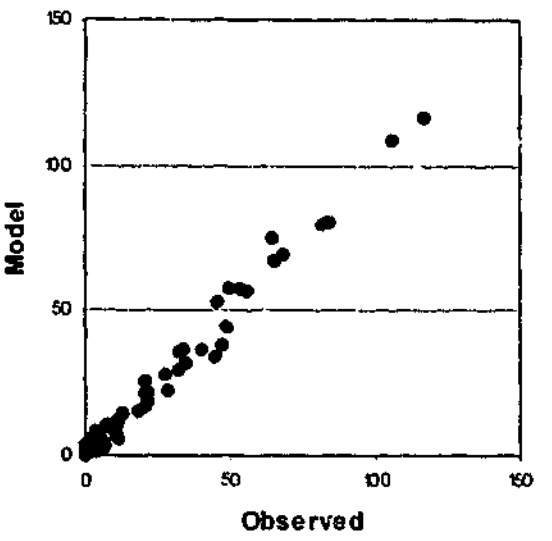

(xxi)
Nampula: Lead(8)
Model
Observed
0
50
100
150
200
250


(xxiv)
Niassa: Lead(8)
Model
Observed
0
50
100
150

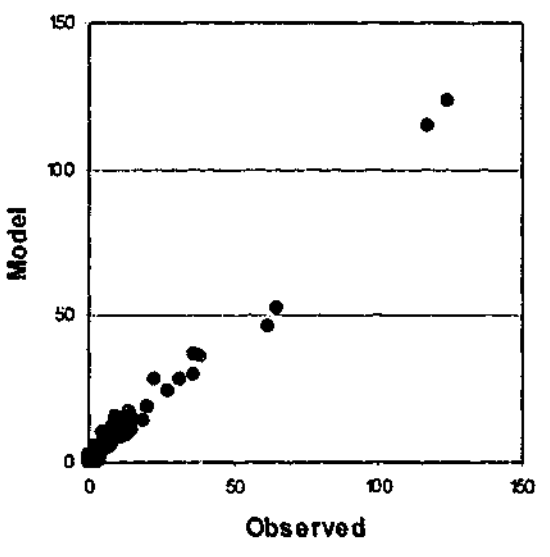
(xxv) Sofala: Lead(1)



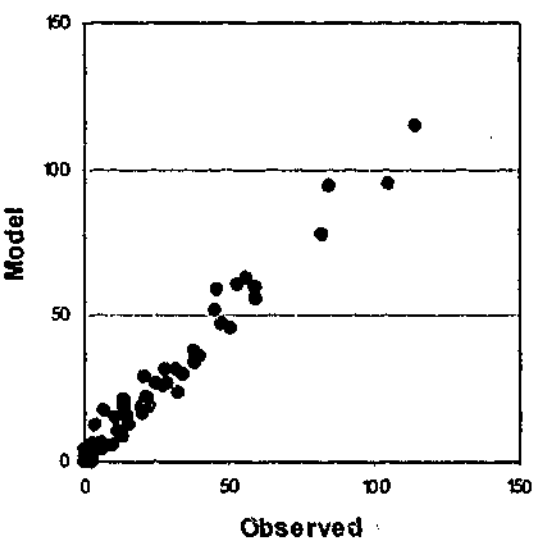
(xxviii) Tete: Lead(1)



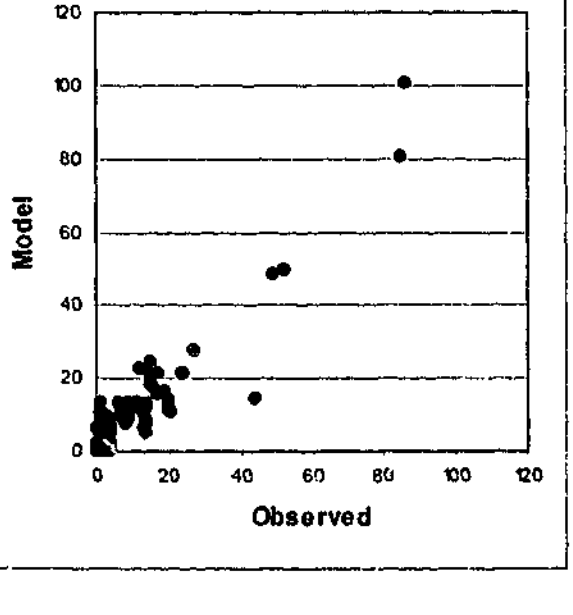
(xxvi) Sofala: Lead(4)



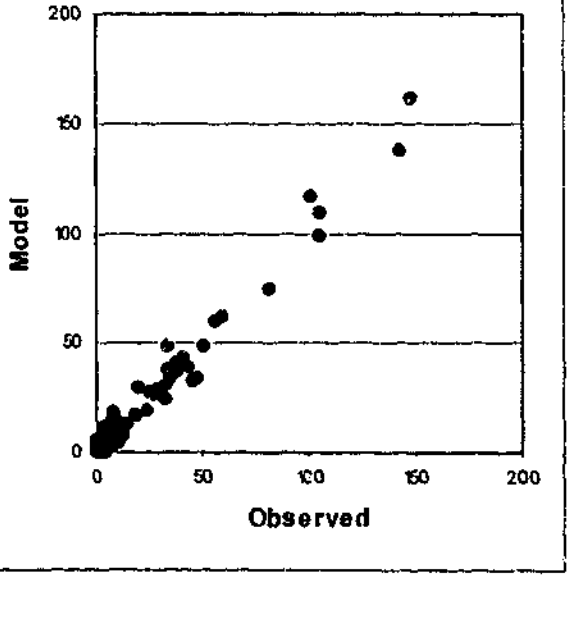
(xxix) Tete: Lead(4)



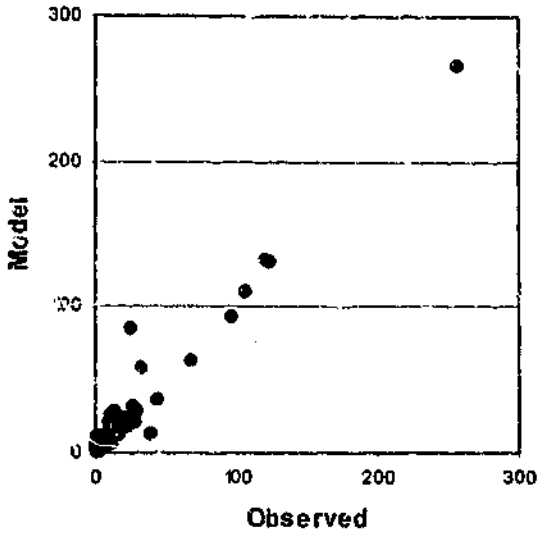
(xxvii) Sofala: Lead(8)



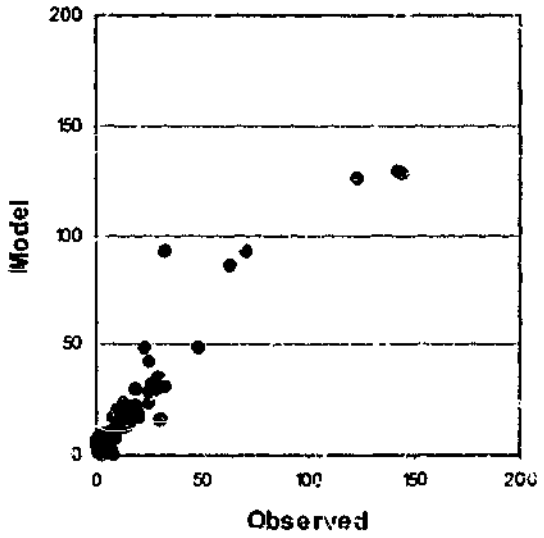
(xxx) Tete: Lead(8)



(xxxi) Zambezia: Lead(1)



(xxxii) Zambezia: Lead(4)



(xxxiii) Zambezia: Lead(8)

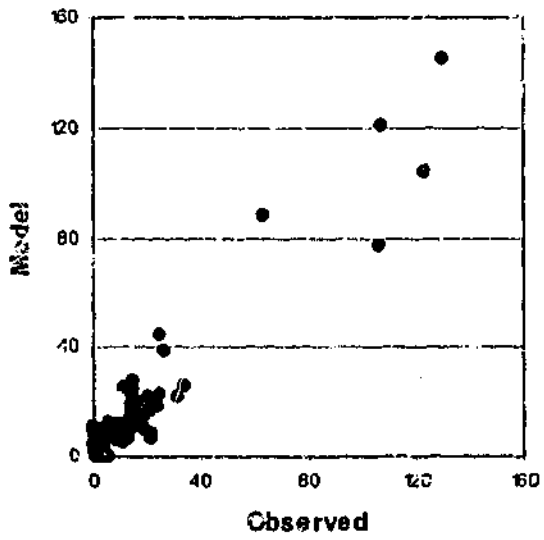


Table 6-2 shows the R^2 for each of the different predicted leads, by province, for Approach 1 (the randomly selected test sets), and Figure 6-3 shows scatter plots of model predictions versus observed case numbers for the same models. There is very good correlation between the actual and predicted values across all the provinces and all the prediction times up to eight weeks.

Table 6-2: R^2 (coefficient of multiple determination) for each of the different predicted leads, by province, for randomly selected test sets (Approach 1).

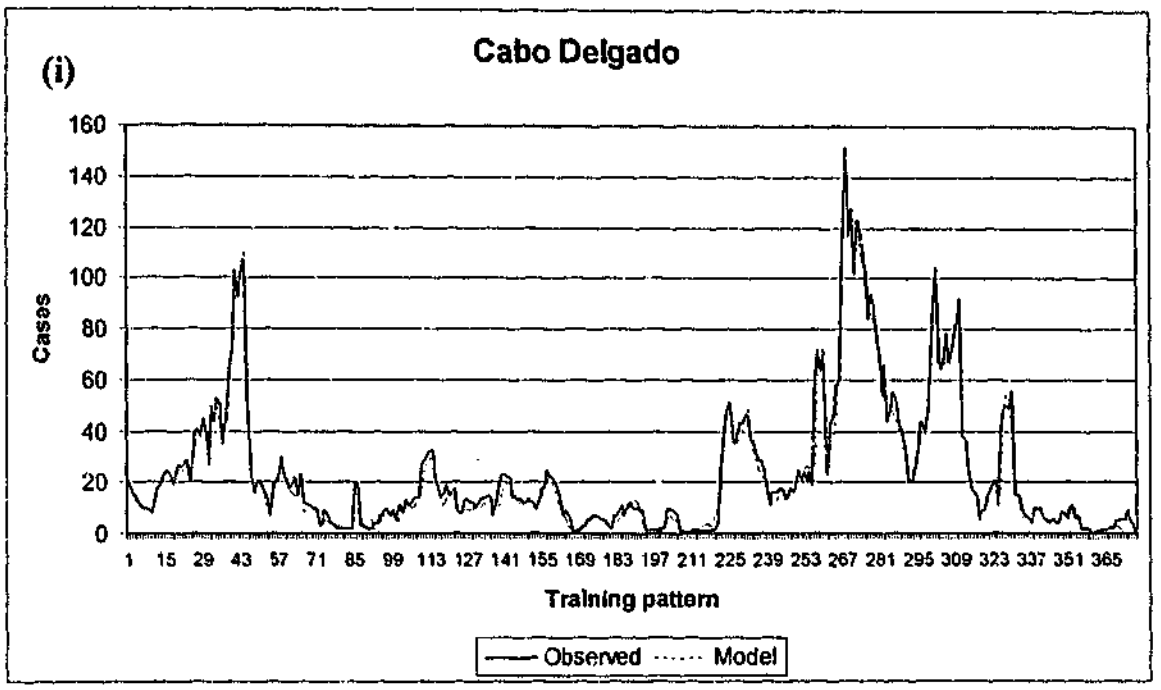
(a) R^2 for training set (Approach 1, where the test set was randomly selected):

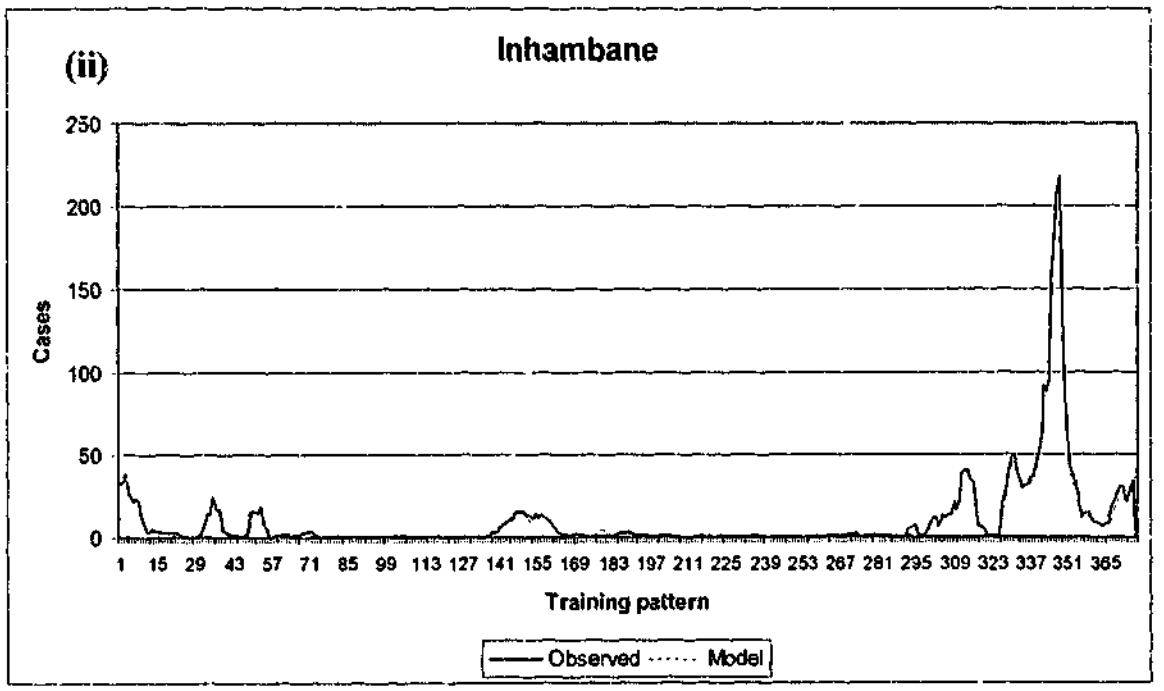
	Lead(1)	Lead(2)	Lead(3)	Lead(4)	Lead(5)	Lead(6)	Lead(7)	Lead(8)
Cabo Delgado	0.972	0.967	0.960	0.963	0.967	0.954	0.964	0.954
Gaza	0.998	0.998	0.998	0.998	0.997	0.996	0.996	0.995
Inhambane	0.993	0.991	0.990	0.985	0.981	0.978	0.977	0.974
Manica	0.996	0.995	0.991	0.996	0.996	0.993	0.995	0.990
Maputo City	0.994	0.993	0.992	0.990	0.989	0.989	0.989	0.987
Maputo Prov.	0.996	0.992	0.995	0.992	0.993	0.994	0.993	0.992
Nampula	0.997	0.996	0.995	0.996	0.994	0.995	0.995	0.995
Niassa	0.942	0.902	0.914	0.878	0.904	0.876	0.912	0.901
Sofala	0.992	0.990	0.988	0.979	0.963	0.950	0.935	0.911
Tete	0.997	0.996	0.996	0.995	0.996	0.995	0.996	0.995
Zambezia	0.996	0.985	0.996	0.991	0.992	0.984	0.984	0.987

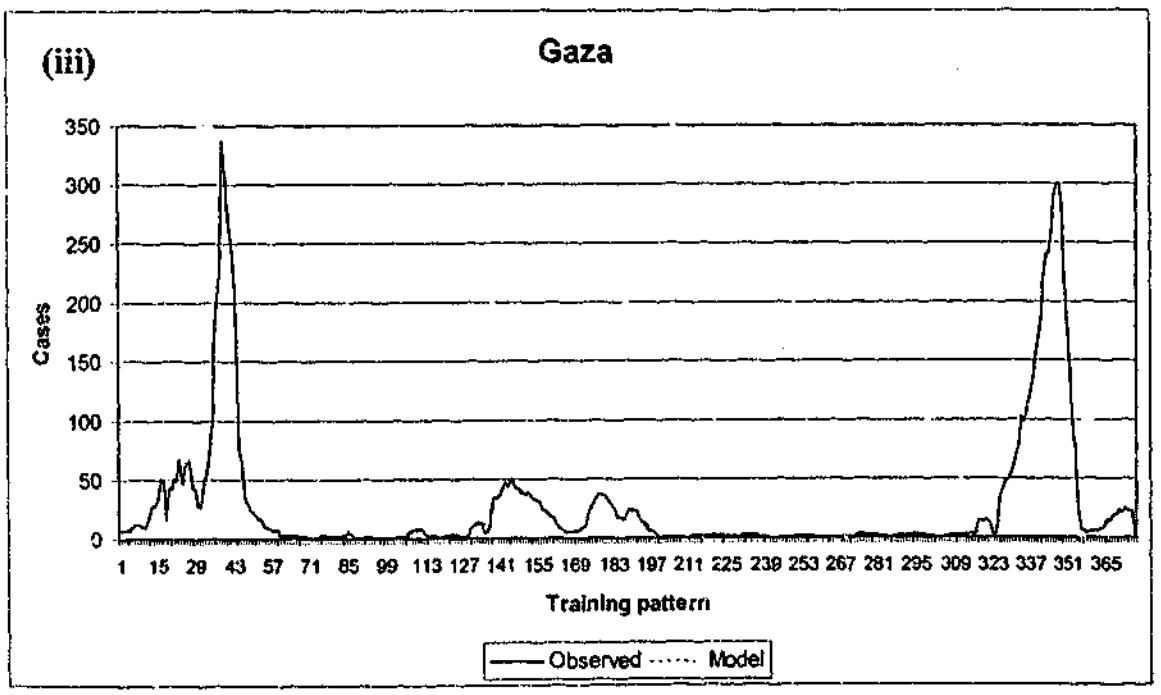
(b) R^2 for the test set (Approach 1, where the test set was randomly selected):

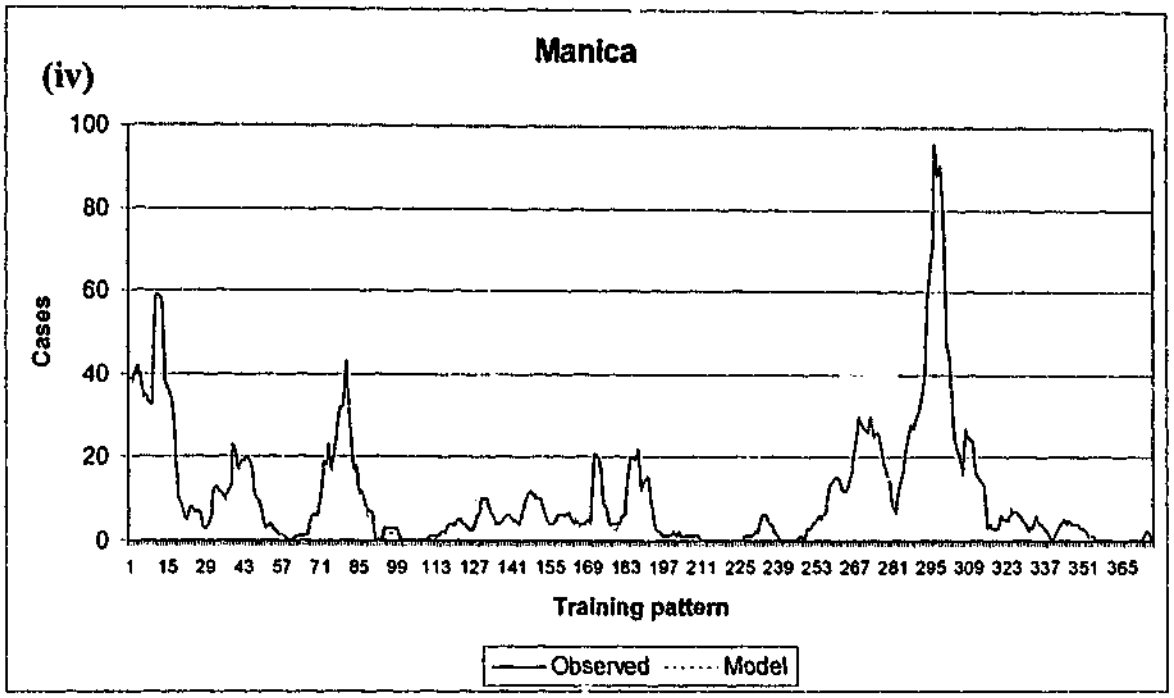
	Lead(1)	Lead(2)	Lead(3)	Lead(4)	Lead(5)	Lead(6)	Lead(7)	Lead(8)
Cabo Delgado	0.945	0.938	0.968	0.958	0.905	0.875	0.876	0.904
Gaza	0.992	0.986	0.981	0.982	0.986	0.977	0.977	0.977
Inhambane	0.986	0.980	0.988	0.952	0.927	0.922	0.935	0.944
Manica	0.982	0.975	0.980	0.985	0.976	0.956	0.964	0.944
Maputo City	0.977	0.969	0.968	0.967	0.964	0.964	0.971	0.943
Maputo Prov.	0.979	0.969	0.977	0.962	0.967	0.977	0.955	0.960
Nampula	0.961	0.961	0.975	0.972	0.963	0.949	0.956	0.922
Niassa	0.749	0.849	0.750	0.769	0.644	0.883	0.780	0.361
Sofala	0.972	0.971	0.963	0.978	0.980	0.947	0.950	0.874
Tete	0.983	0.970	0.965	0.974	0.961	0.966	0.956	0.975
Zambezia	0.939	0.938	0.940	0.880	0.972	0.936	0.924	0.905

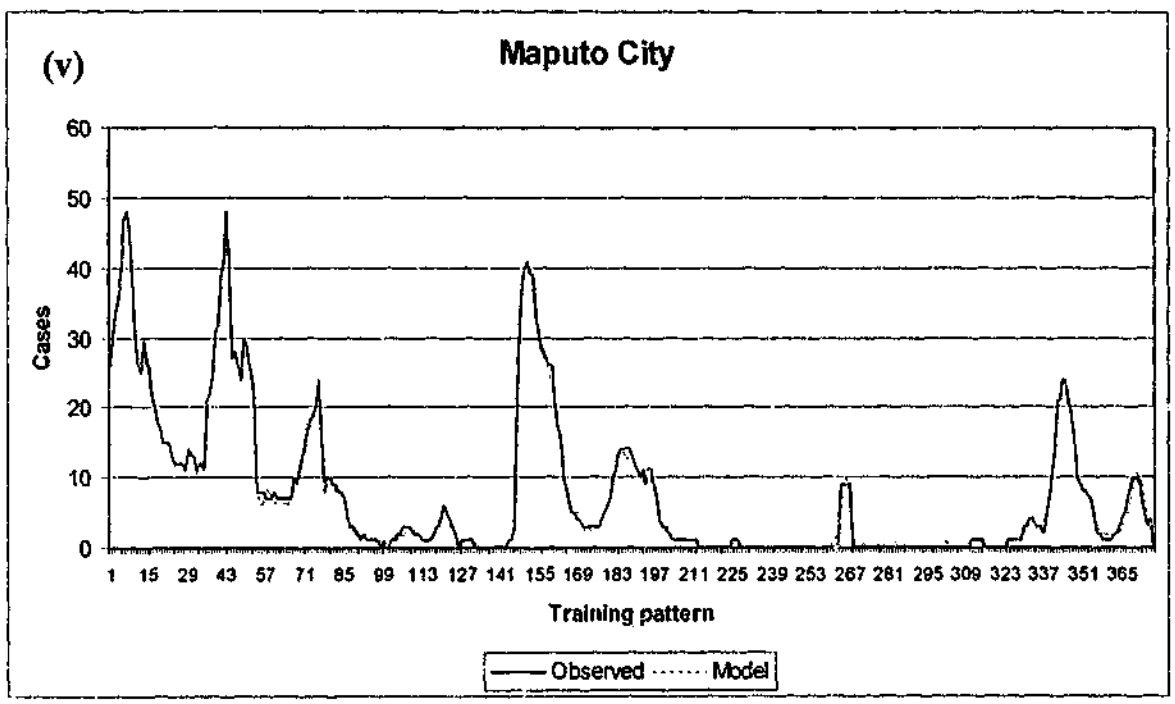
Figure 6-3 (i-xi): Plots of the training time series by province, with predictions for the next point in the time series overplotted.

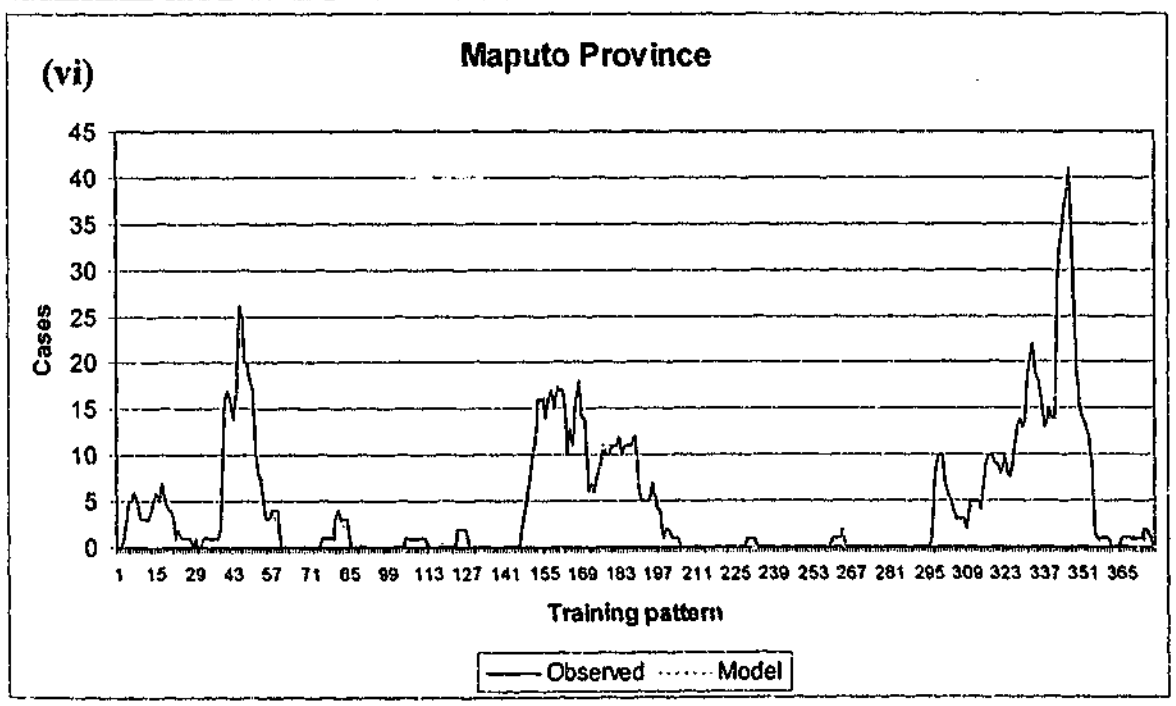




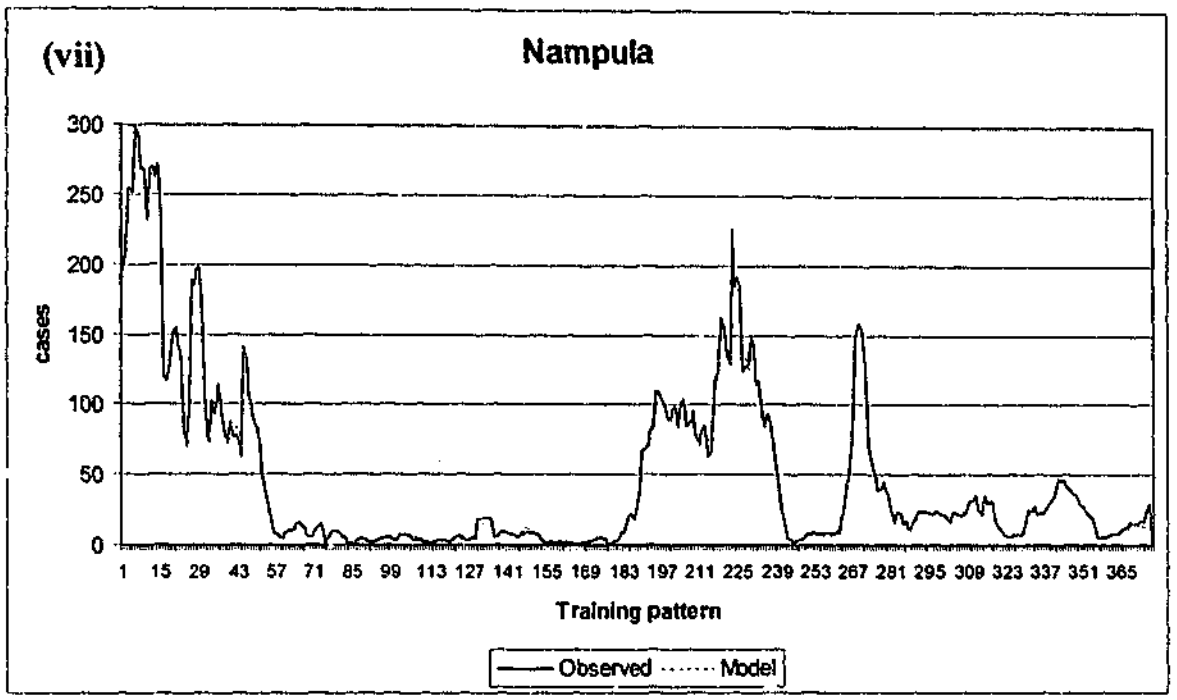



(iv)
Manica
Cases
100
80
60
40
20
0
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model

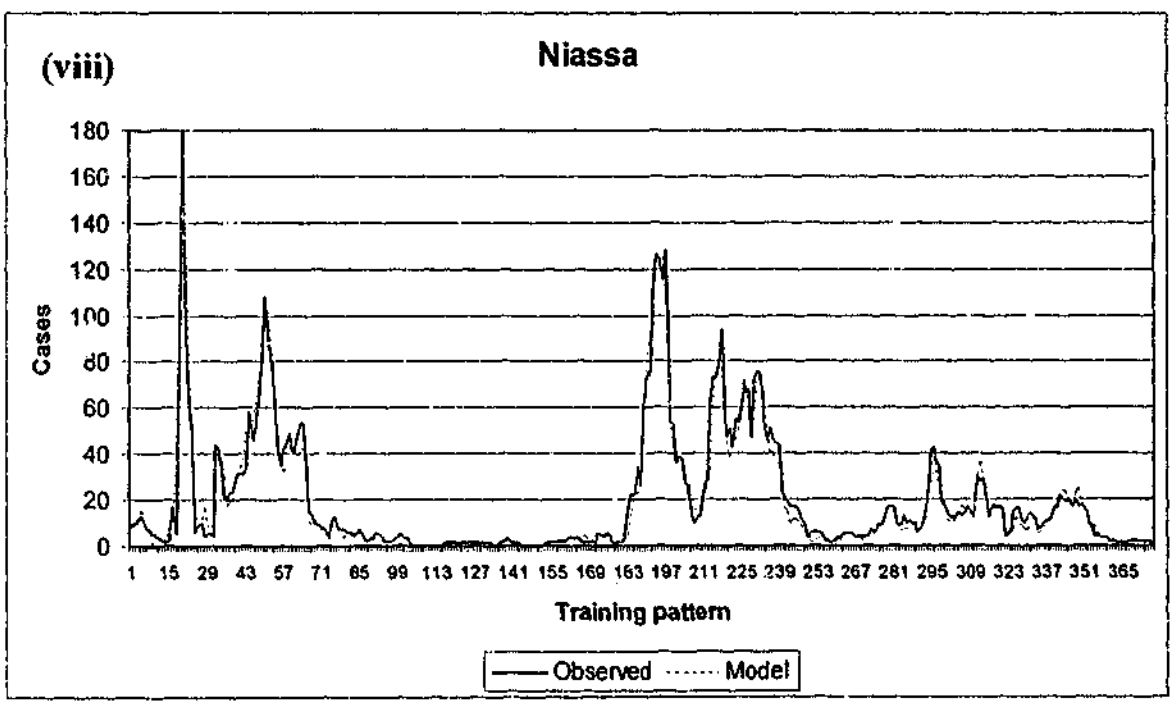

(v)
Maputo City
Cases
60
50
40
30
20
10
0
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model


(vi)
Maputo Province
Cases
45
40
35
30
25
20
15
10
5
0
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model

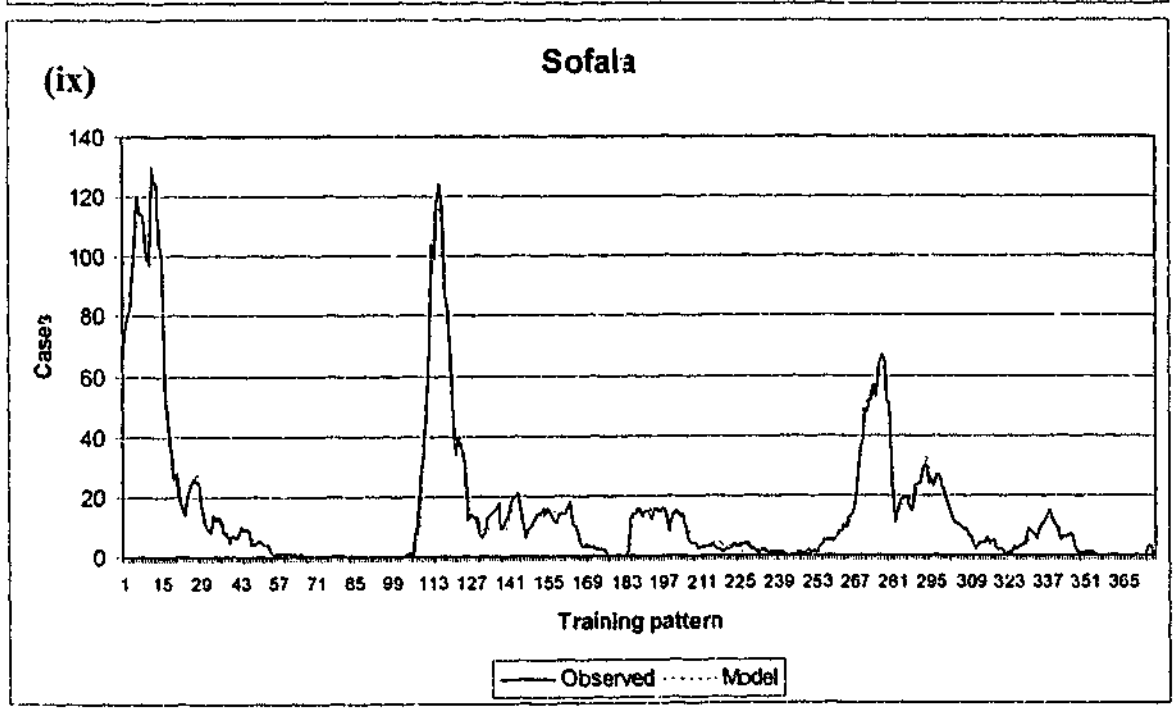
(vii)

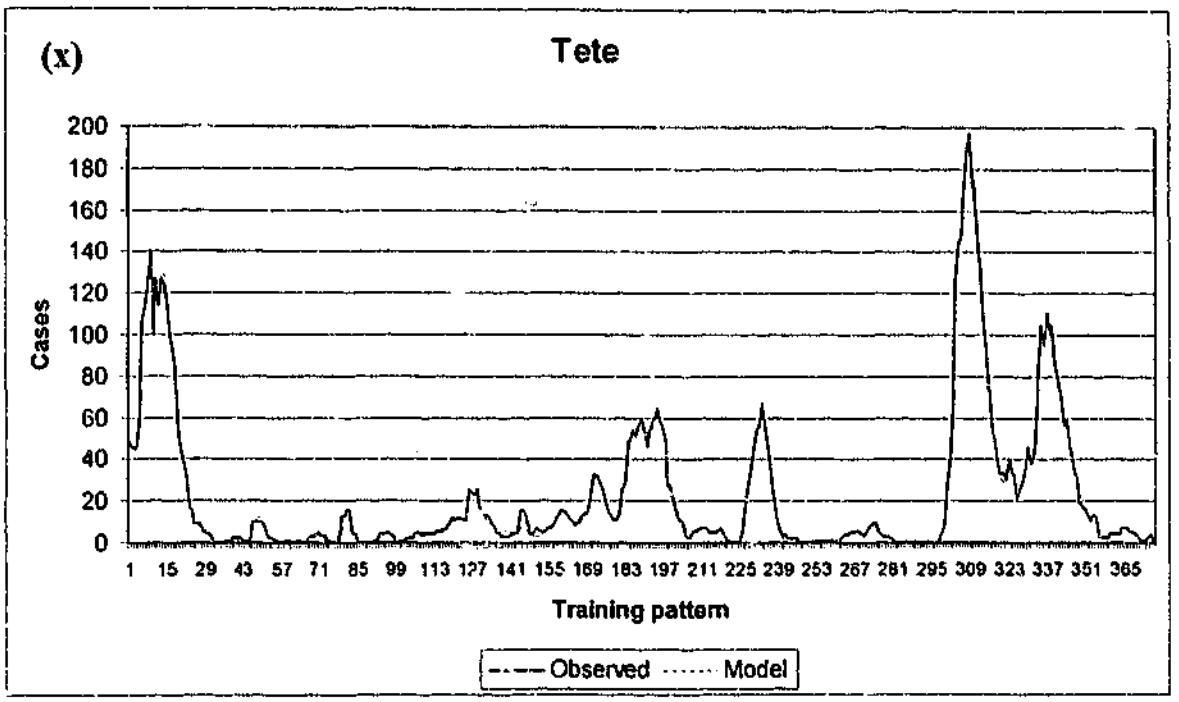

Nampula
300
250
200
150
100
50
0
cases
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model

(viii)


Niassa
180
160
140
120
100
80
60
40
20
0
Cases
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model

(ix)


Sofala
140
120
100
80
60
40
20
0
Cases
1 15 29 43 57 71 85 99 113 127 141 155 169 183 197 211 225 239 253 267 281 295 309 323 337 351 365
Training pattern
Observed
Model



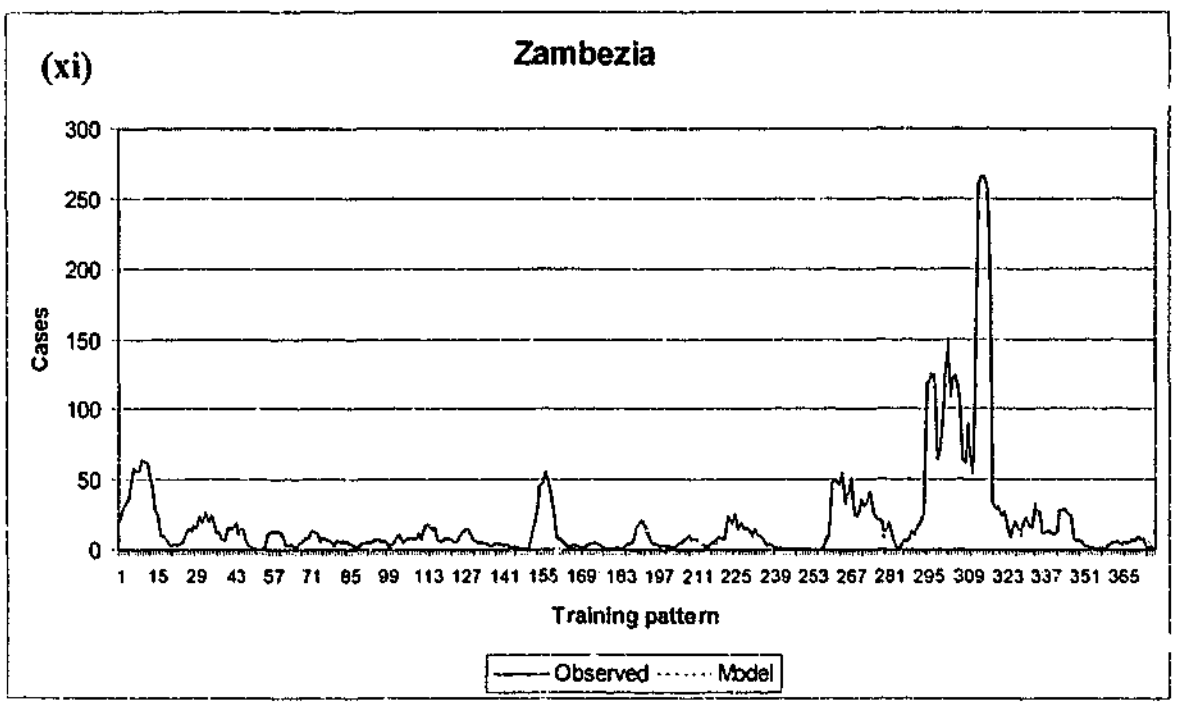


Table 6-3 shows, however, that the predictive accuracy of these networks deteriorated markedly when the test set was the last 70 weeks (Approach 2). In many cases the values of the R^2 fall below zero, indicating that a trivial model that simply predicted the mean number for every case would have done better than the neural network. Each graph in Figure 6-4 plots the actual values for the time series (the heavy line) together with the model's predictions at each point. The lighter lines join the model predictions

for the eight weeks following each time point. They represent the predictions that would have been available to a surveillance epidemiologist at each point.

Table 6-3: R^2 (coefficient of multiple determination) for each of the different predicted leads, by province, for the last-70-weeks test sets (Approach 2).

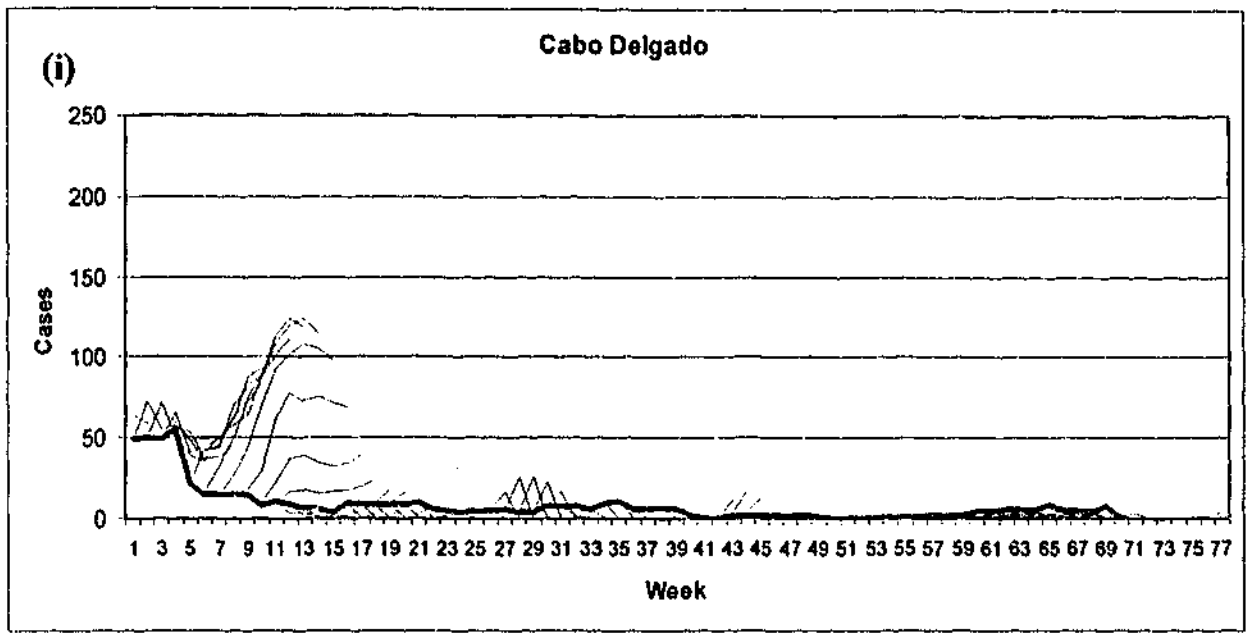
(a) R^2 for the training set (Approach 2, where the test set was the last 70 weeks of the time series):

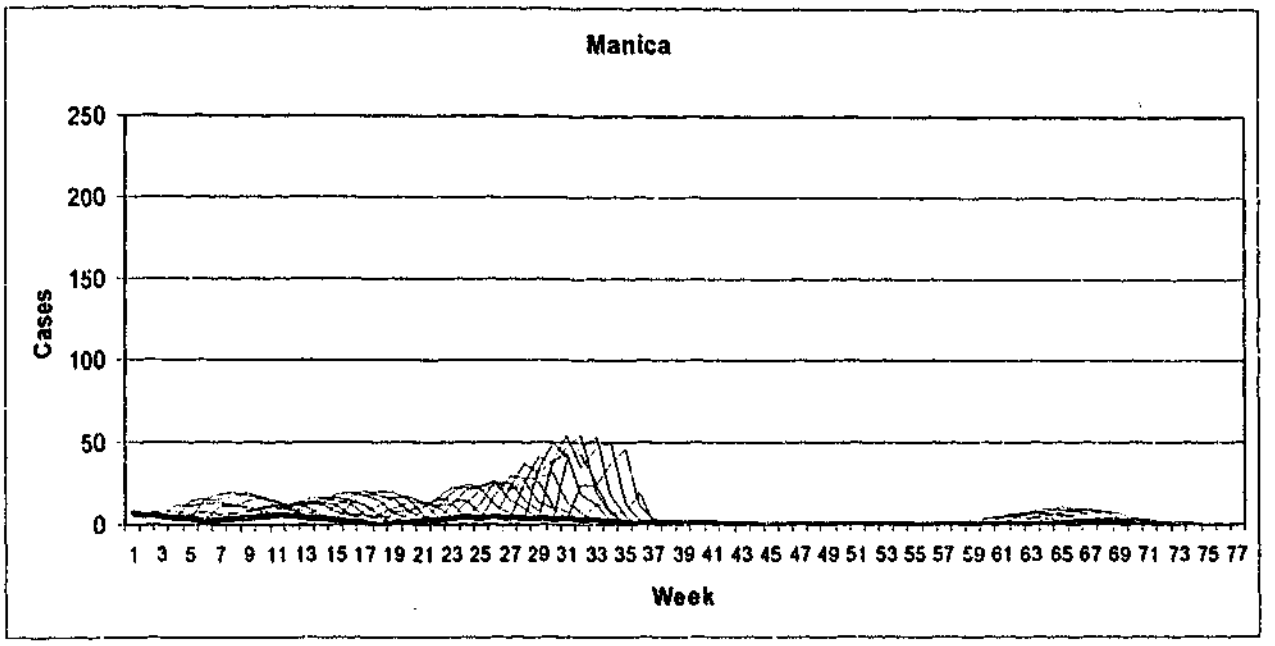
	Lead(1)	Lead(2)	Lead(3)	Lead(4)	Lead(5)	Lead(6)	Lead(7)	Lead(8)
Cabo Delgado	0.954	0.953	0.919	0.890	0.853	0.817	0.797	0.788
Gaza	0.978	0.961	0.959	0.885	0.886	0.720	0.633	0.616
Inhambane	0.900	0.865	0.797	0.698	0.509	0.488	0.207	0.327
Manica	0.937	0.880	0.900	0.874	0.847	0.967	0.980	0.982
Maputo City	0.940	0.643	0.512	0.318	0.138	0.002	0.083	-0.132
Maputo Prov.	0.891	0.873	0.817	0.760	0.426	0.192	0.003	-0.115
Nampula	0.980	0.981	0.978	0.981	0.980	0.980	0.977	0.930
Niassa	0.920	0.819	0.792	0.503	0.391	0.306	0.269	-0.041
Sofala	0.990	0.992	0.989	0.971	0.953	0.934	0.900	0.757
Tete	0.978	0.975	0.965	0.910	0.873	0.859	0.847	0.728
Zambezia	0.833	0.707	0.818	0.878	0.878	0.889	0.939	0.873

(b) R^2 for the test set (Approach 2, where the test set was the last 70 weeks of the time series):

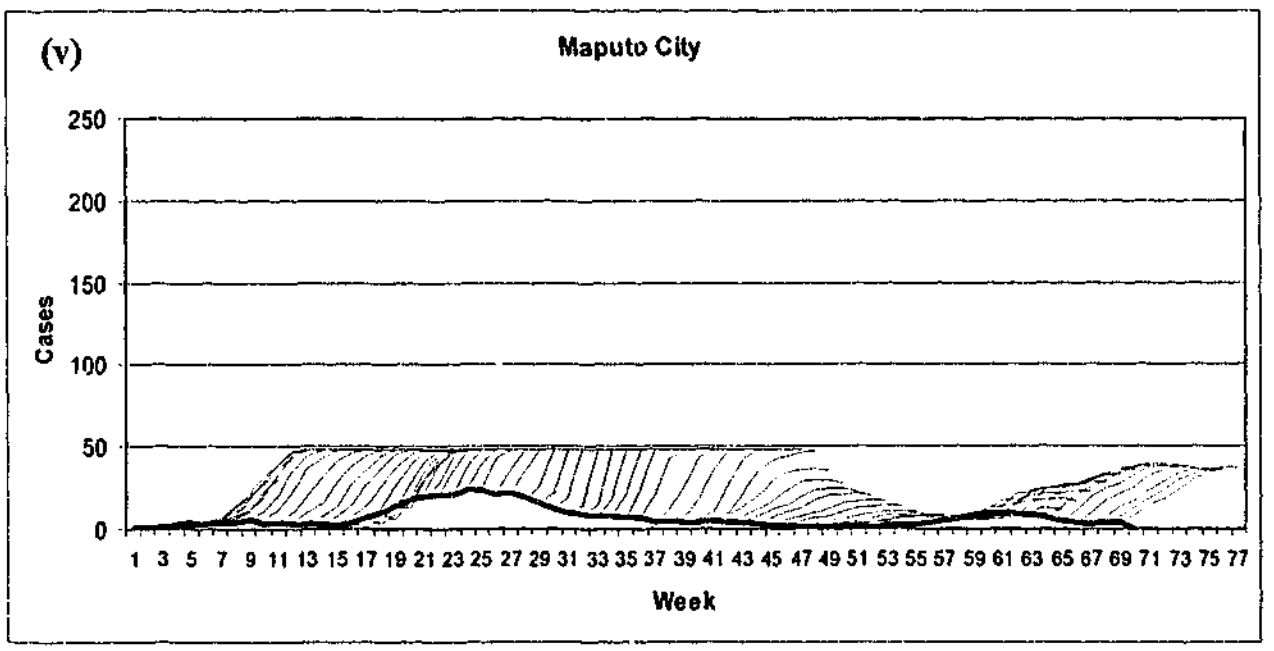
	Lead(1)	Lead(2)	Lead(3)	Lead(4)	Lead(5)	Lead(6)	Lead(7)	Lead(8)
Cabo Delgado	0.288	0.167	-2.313	-6.861	-30.45?	-56.209	-71.273	-82.148
Gaza	0.935	0.883	0.878	0.647	0.656	0.149	-0.110	-0.156
Inhambane	0.825	0.765	0.645	0.474	0.145	0.107	-0.386	-0.177
Manica	-23.781	-55.237	-44.846	-61.881	-76.861	-9.372	-1.281	0.211
Maputo City	-0.248	-7.083	-10.062	-14.659	-18.698	-21.355	-22.863	-23.246
Maputo Prov.	0.683	0.624	0.463	0.294	-0.695	-1.360	-1.897	-2.217
Nampula	-1.357	-1.533	-2.245	-1.193	-0.912	-0.731	-0.686	-9.402
Niassa	-0.894	-12.578	-14.659	-52.792	-65.117	-74.554	-80.638	-115.616
Sofala	-0.559	-0.409	-0.384	-6.797	-12.230	-17.292	-28.740	-73.824
Tete	0.793	0.768	0.684	0.134	-0.269	-0.414	-0.511	-1.696
Zambezia	-16.481	-29.775	-17.304	-10.685	-10.629	-8.962	-4.299	-11.443

Figure 6-4 (i-xi): Measles time series for Approach 2 (using the last 70 weeks of the series as the withheld test set), with model predictions at each point. The heavy lines are the observed reports, and the lighter lines join the eight predictions for each time point.

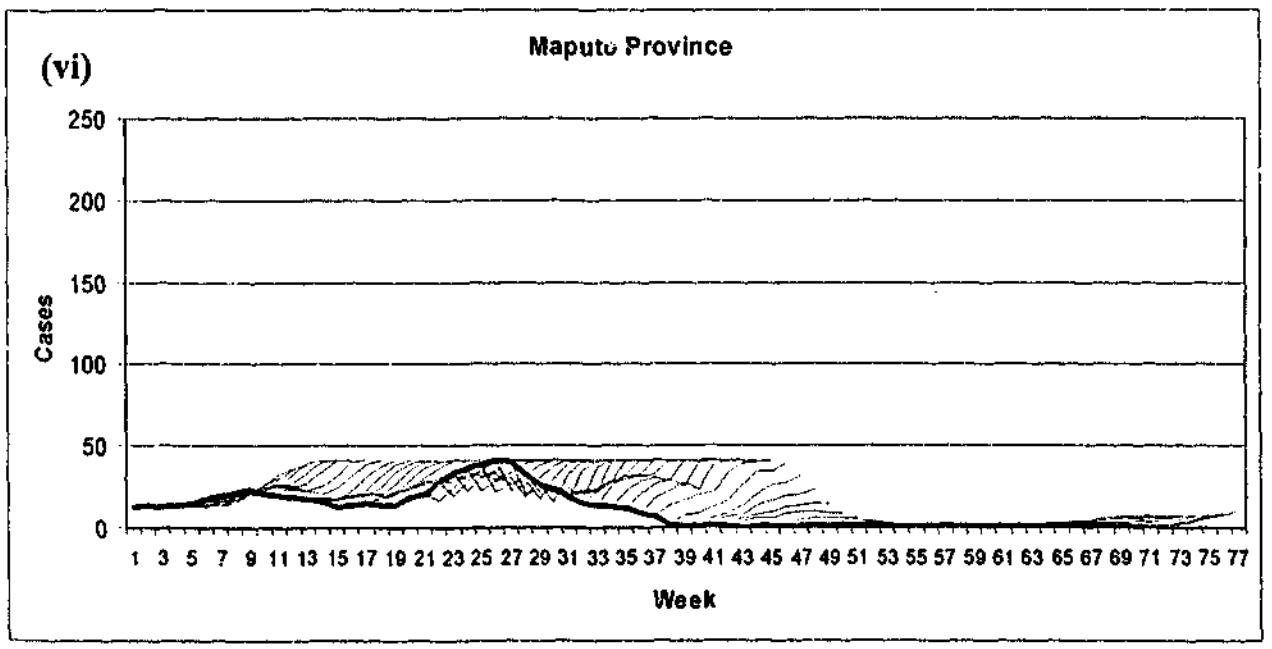



Manica
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

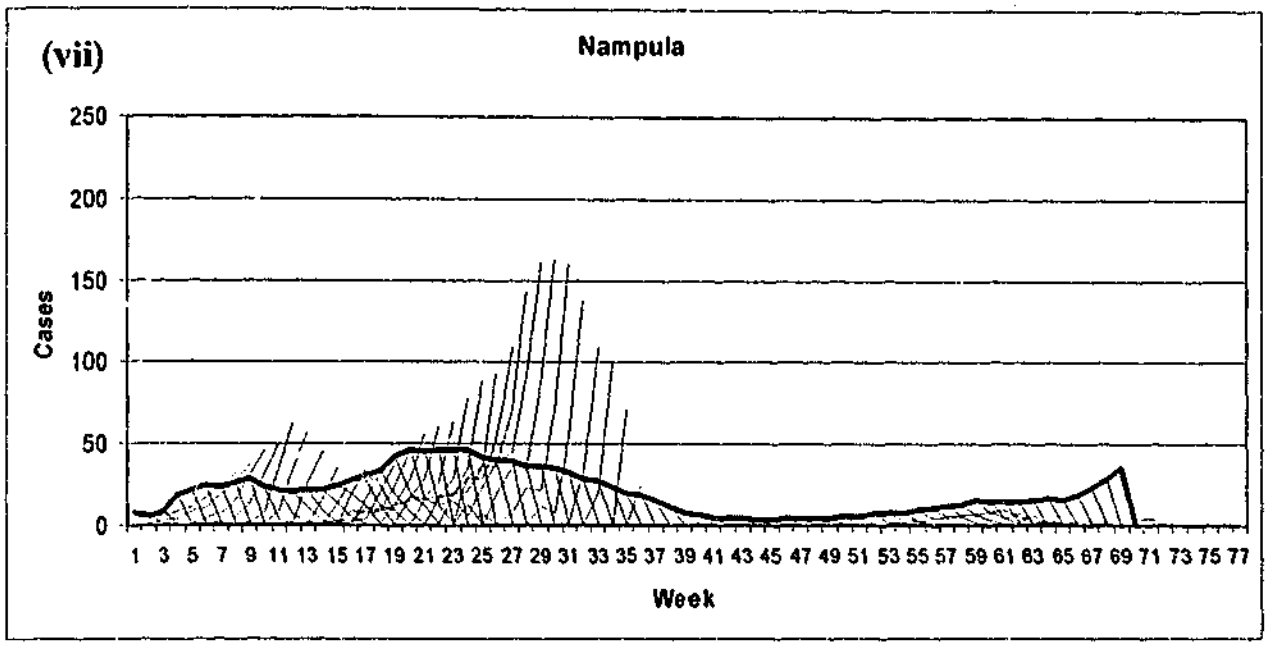
(v)


Maputo City
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

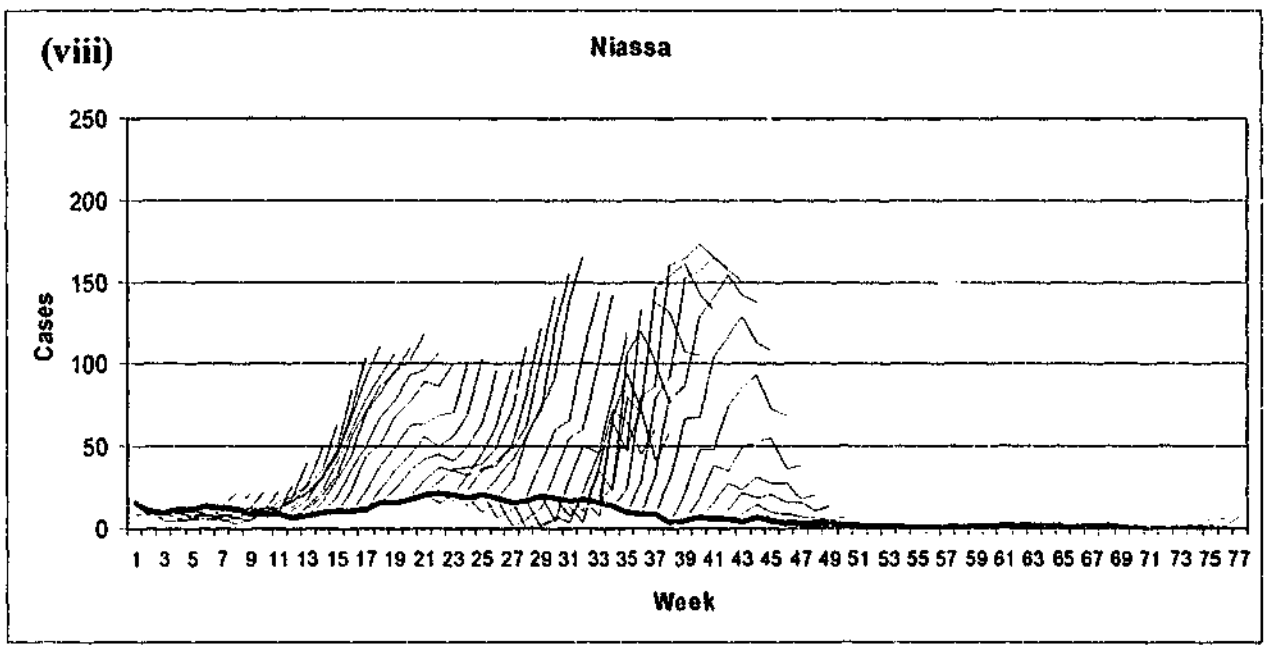
(vi)


Maputo Province
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

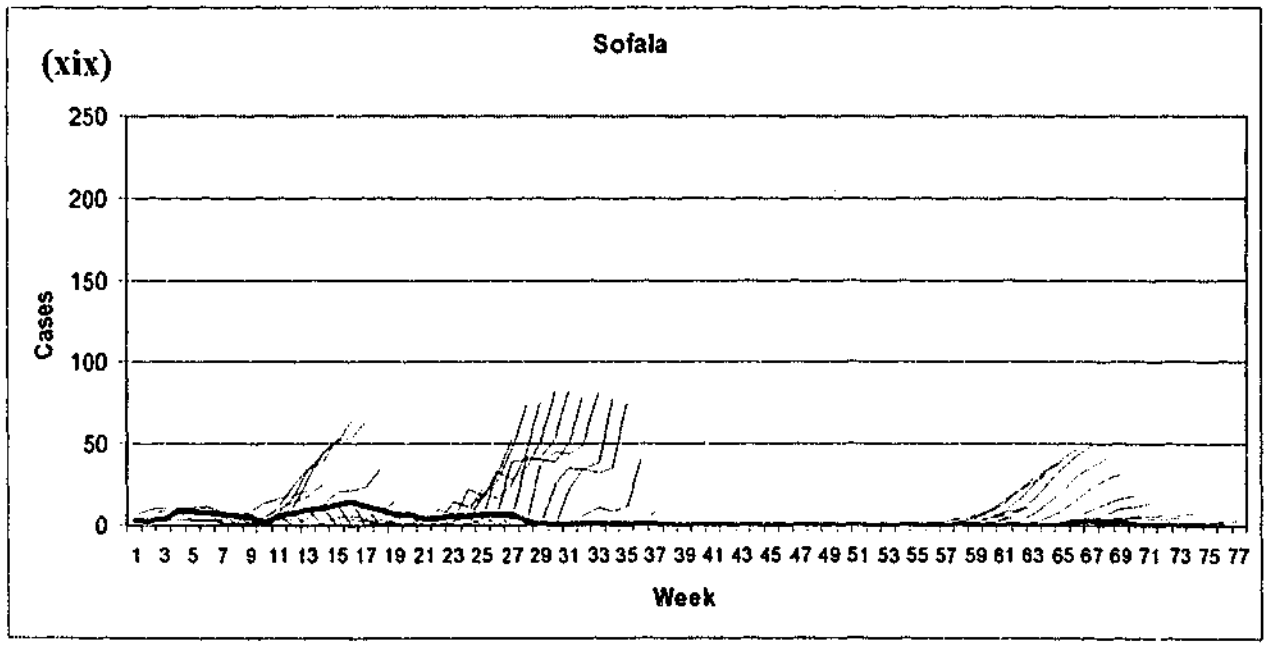
(vii)

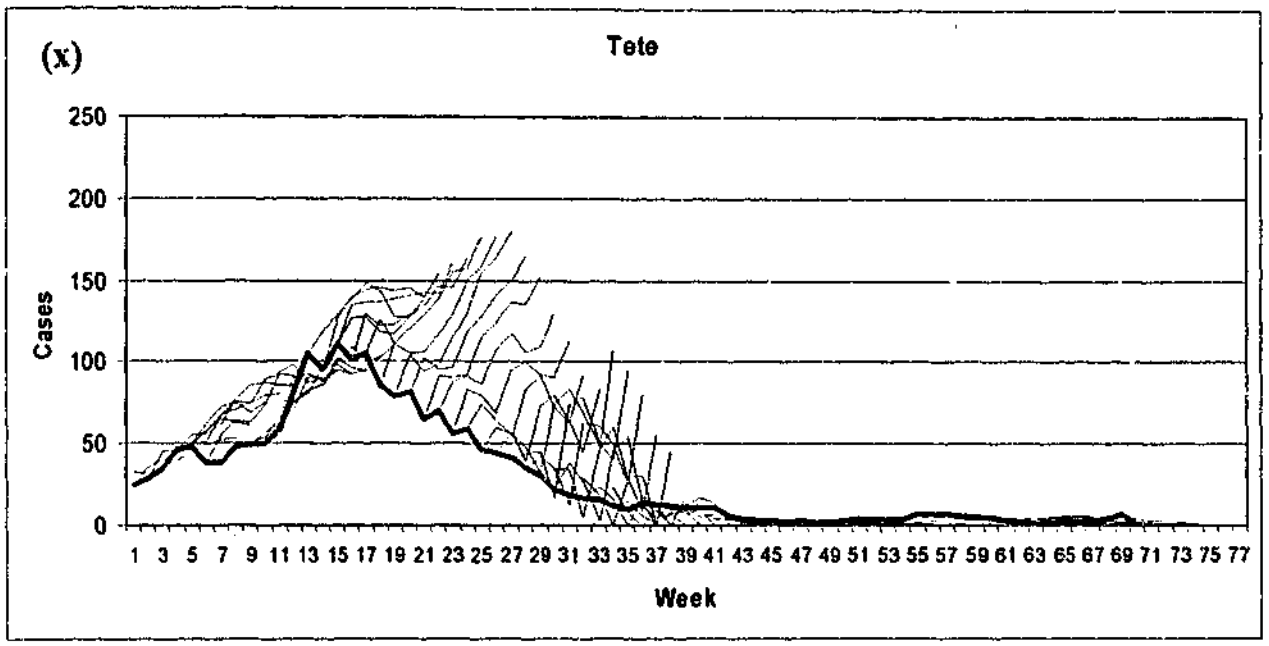

Nampula
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

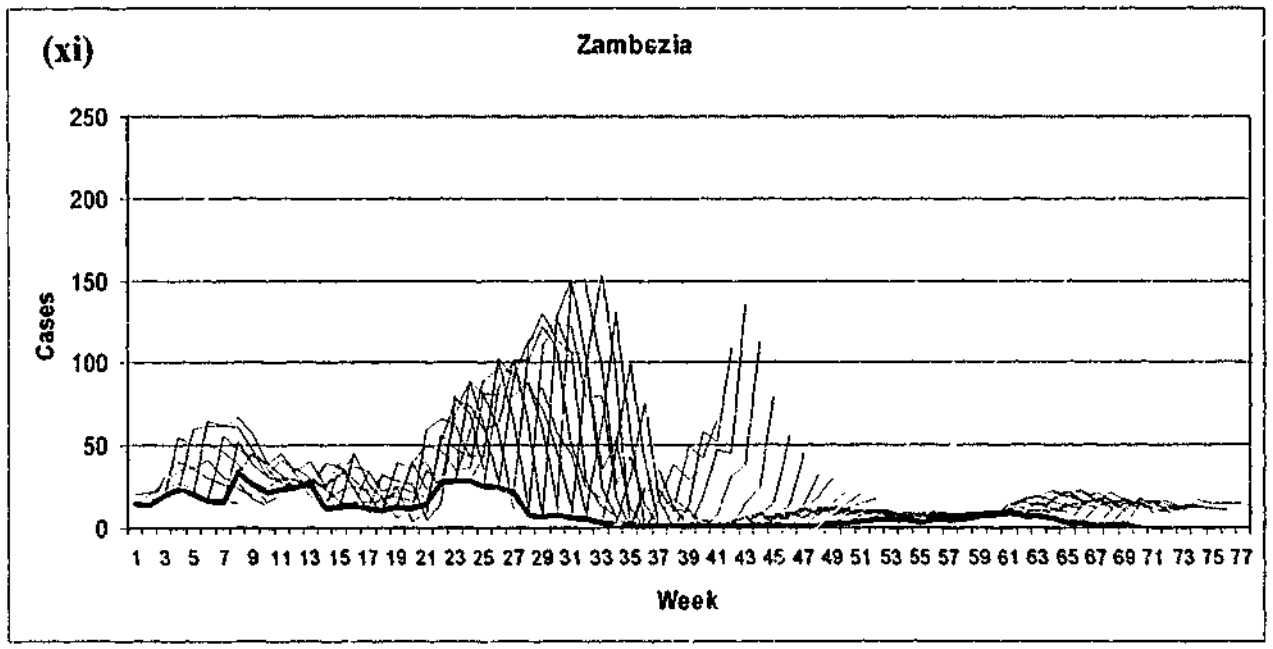
(viii)


Niassa
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

(xix)


Sofala
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week


(x)
Tete
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week


(xi)
Zambezia
250
200
150
100
50
0
Cases
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77
Week

Discussion

Three essential elements are required for successful forecasting of disease incidence. First there must truly be a 'signal' among the noise of the time series data - there must be a consistent relationship, however complex, between earlier events and future ones. Prior knowledge of measles transmission suggests that this should be the case. Secondly, the statistical modelling technique used must be capable of accurately modelling that relationship. Thirdly, and most importantly for the surveillance practitioner in the field, the same relationship between the past and future behaviour of the disease must continue to hold into at least the near future. This work with measles time series from Mozambique demonstrates that the first two requirements can be satisfied. There is a strong deterministic component to the weekly measles reporting rate at provincial level. With suitable pre-processing to accentuate the signal at the appropriate scale, and *within a given time window*, a simple artificial neural network can accurately model the distribution of measles cases up to at least eight weeks in advance.

These models do not, however, provide accurate forecasts beyond the end of the time window from which the training set was chosen, where they would be of practical use to surveillance practitioners. In the tests of their true forecasting ability the networks tended to 'overpredict', either overestimating the size of outbreaks that did occur, or predicting outbreaks that never happened.

There are a number of possible reasons for this. It is possible that there is insufficient information available to the network. Mozambique is a long, narrow country with several neighbours and a good deal of population movement across its borders. No

information about the measles events in these neighbouring countries was included in these models, but it is likely that many of the outbreaks recorded were begun when measles virus was inadvertently imported from a neighbouring country. Similarly, vaccination rates are known to affect measles outbreak frequency and size, but vaccination coverage rates for Mozambique are available only as yearly aggregates for these provinces, and could not be incorporated in these models. Even local weather patterns, which can be extreme in Mozambique, may influence measles distribution by changing transport and other means of people-mixing, and these were not available to these models.

Paradoxically, another potential problem with these networks is the amount of information they *do* include, and the complexity of the models they create. Each of these neural networks has 96 x 250 + 250 x 8 (i.e. 26,000) connections, each with a weight that may be altered during training. It is this degree of complexity that allows the network to model each outbreak, (and empirical tests of smaller networks did not produce similar training ability). At the same time, it is unlikely that the trained network will see again an exactly similar pattern of events, and this may be the limiting factor in the use of neural networks in the surveillance of outbreak-prone diseases. The challenge for neural networks in measles is finding a temporal and/or geographic scale at which there is some likelihood that patterns will repeat themselves in the future, but which still allows useful and timely preventive measures to be taken before the forecast outbreaks occur.

Chapter Seven: Extracting prognostic information from cancer registry and health care utilisation data: Artificial neural networks and Cox proportional hazards models

Introduction

'Bin model' cancer staging systems such as the Tumour-Lymph Node-Metastasis (TNM) system offer universal applicability, and are useful in many ways. Their usefulness as prognostic tools for individual patients, however, leaves something to be desired. Several studies have shown that alternative approaches based on inferential statistical techniques (such as the Cox proportional hazards regression or artificial neural networks (Burke et al., 1997)) can provide more accurate individual estimates of the probability of survival for newly-diagnosed cancer patients.

Cancer registries gather population-based data on patients within defined geographical areas. A newly diagnosed cancer patient in the registry's catchment area might logically expect the local cancer registry to be the most appropriate source of prognostic information for survival and other outcomes for patients in a similar setting. Health workers caring for cancer patients could also look to the local cancer registry to compare outcomes in their own practices with the regional norms.

But how much prognostic information does a particular cancer registry contain, and what is the best way to extract that prognostic information? This study examines the National Cancer Institute's Surveillance Epidemiology and End Results-Medicare

(SEER-Medicare) data set (Potosky et al., 1993) on carcinoma of the colon. In an endeavour to extract the maximum available prognostic information it compares prognostic models created by artificial neural networks with models made using the Cox proportional hazards approach.

Components of predictive accuracy

Burke and others (Burke et al., 1997) have suggested three components to predictive accuracy in cancer outcomes: the amount and quality of the original data, the predictive power of the prognostic factors within the data set, and the modelling technique's ability to capture the power of those prognostic factors. The SEER-Medicare dataset contains as good quality data as can be expected outside an individual study. This study thus considers both the predictive power of the SEER-Medicare variables and the relative abilities of artificial neural network and Cox proportional hazards models to extract the prognostic relationships.

Universal versus local applicability

Existing prognostic approaches have striven for universal applicability. Perhaps the availability of 'local' data sets that still cover many thousands of patients means that locally created (or at least locally tuned) models will offer more relevant and more accurate information to people inside their catchment. The catchment of the SEER-Medicare database covers tens of millions of people.

This study thus also considers Burke's third point, asking whether the local factors in the SEER-Medicare database of operable colon cancer in over-64s add sufficient accuracy to make a useable new model for individual prognosis. There is an existing and reliable statistical technique (the Cox proportional hazards regression technique),

so we felt it important to compare the neural network models with that. To our knowledge this is the first time such a study has been done in relation to colon cancer in the SEER-Medicare database.

Curiously, given Wyatt and Altman's (Wyatt and Altman, 1995) logical arguments about prognostic modelling in medicine, authors in this field have given little attention to the 'calibration' of their models' predictions. (In simple terms, do 50% of patients given a 50% chance of survival actually survive, and so on?) We felt it important to include an assessment of model calibration in this study.

Artificial neural networks versus Cox regression

Although their name reflects their origins in neuropsychology, in this kind of application artificial neural networks are best regarded as a form of non-linear regression algorithm. Unlike linear, parametric regression techniques, they make no assumptions about the distribution of individual input variables, and interactions between variables do not invalidate their outputs (Rumelhart et al., 1994, Hinton, 1992, Cross et al., 1995).

Both neural networks and Cox regression have been proposed for the creation of prognostic models. Neural networks have been compared with TNM staging (Burke et al., 1997) and with logistic regression (Duh et al., 1998). Neural networks predicting death from colorectal carcinoma on data from one UK institution achieved 90% overall accuracy when applied to data from a different institution (Bottaci et al., 1997).

The Cox regression technique can make use of 'censored' data for patients lost to follow-up. It also has the advantage of providing directly interpretable estimates of the strength of relationship between each prognostic variable and survival outcome. The difficulty reading the relative contribution of each input means that neural networks are commonly regarded as 'black boxes'. This potentially affects their usefulness to clinicians, not least for medico-legal reasons (Brahams and Wyatt, 1989). However, neural networks are able to model complex and non-linear relationships. It has been suggested that the major role of neural networks in prognostic modelling is to look for non-linear relationships as indicators of the maximum attainable classification accuracy, to test whether other more transparent techniques are extracting all the available information from the data set. (Hart and Wyatt, 1990)

Adjuvant chemotherapy as a predictor

Since 1990, adjuvant chemotherapy with 5-fluorouracil (5FU)-based regimens after surgical resection of the tumour has been the standard of care for patients with node-positive colon cancer (Moertel et al., 1990, Moertel et al., 1995, National Institutes of Health, 1990). Our analysis focuses on individuals with node-positive colon cancer diagnosed between 1992 and 1996, all of whom have undergone potentially curative surgical resection. Treatment with adjuvant chemotherapy is one of the prognostic variables included in the models.

Objective of this study

This study thus aimed to discover whether accurate 'locally applicable' models (although potentially applicable to a catchment population of some millions) can be made using information available to colon cancer patients at or shortly after diagnosis.

Neural network models are compared with Cox proportional hazards regression in order to determine whether the latter approach captures all the available prognostic information.

Methods

Data

We used a database developed by Potosky and colleagues in 1993 (Potosky et al., 1993) in which the files of patients with cancer from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registry were linked with their Medicare claims files. The SEER database provides information on tumour histology, location, stage of disease, and individual demographics, as well as primary surgical and radiation treatment and survival, for cancers occurring in geographically diverse regions covering approximately 14% of the US population. The population of the SEER counties is reasonably representative of the US population, with differences in affluence, education and urban/rural status rarely greater than 5% (Nattinger et al., 1997). The Medicare claims files contain extensive diagnostic, treatment, and cost data for patients covered by Medicare, including information on chemotherapy.

We identified all individuals in the linked SEER-Medicare database with a first diagnosis of primary node-positive colon cancer in the period from 1992 to 1996 and aged 65 years or greater. None of the patients in our sample was diagnosed at autopsy or by death certificate. All subjects had undergone surgical resection of the tumour. The sample selection process has been described in detail elsewhere (Sundararajan et al., 2001, Sundararajan et al., (in press)).

From the Medicare claims data we identified patients who had received 5-FU either in their physician's office or at an outpatient department within a hospital. Those who began 5FU within 120 days of their cancer diagnosis were classified as receiving adjuvant 5FU treatment.

We used the Deyo (Deyo et al., 1992) adaptation of the Charlson (Charlson et al., 1987) comorbidity index to assess the prevalence of comorbid disease in our cohort. This is based on history of myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disease, peptic ulcer disease, liver disease, diabetes, hemiplegia, renal disease, or AIDS.

For those who died by the end of follow-up on April 15, 1999, we determined survival time as the interval from the cancer diagnosis date to the Medicare date of death. Those surviving past April 15, 1999, were classified as censored (alive at the end of follow-up) and contributed the time interval from diagnosis date to the end of follow-up to the survival analysis.

The model inputs were based on potentially prognostic parameters available at or shortly after diagnosis. Separate models were made to predict survival to 2 years and 5 years. All 4463 cases were followed either to death or 2 years; of these, 2615 cases had been followed for 5 years.

Before each modelling exercise the data were split randomly into two subsets containing 85% and 15% of the individual records. The larger subset (the 'training set') was used for fitting the Cox model and training the neural network models. The smaller subset (the 'test set') was then used to evaluate each completed model. This process was repeated five times for each time interval, using different random divisions to create training and test sets.

Cox models

The Cox proportional hazards models were estimated using maximum likelihood as implemented in the PROC PHREG procedure in SAS 8.2 (SAS, 1999). Covariates were chosen based on their clinical availability and potential prognostic value: age, gender, race, tumour grade, number of lymph nodes involved with tumour, extent of invasion (classified according to the TNM staging system), comorbid disease status, treatment with 5FU, site of registry, and residence in an urban area. Both the 2-year and 5-year models used the same covariates.

For each of the five random splits of the data, the model parameters estimated with the training set (85%) were applied to the covariates of each individual in the test set to provide a survival curve for each individual. The predicted probabilities of surviving to 2 years and to 5 years were then extracted from the survival curves derived from the baseline statement.

Neural network models

Standard multi-layer feed-forward neural networks were created using a commercial program, NeuroShell2 (Ward Systems Group Inc, 2000). All had one hidden layer and a single output (the network's estimate of survival probability for each patient). All

were trained by back-propagation of errors, using learning rate and momentum terms of 0.1 each. To avoid over-training, an 'early stopping' technique was used. To do this, validation sets were randomly drawn from the training set, equal in size to the test sets. Training was stopped when 30,000 training patterns had been presented without improvement in the mean squared error for the validation set. The 'trained' network was taken to be that which had given the lowest mean squared error for the cross-validation set.

In search of non-linear relationships between the variables, a larger number of inputs was used in the neural network models than the Cox models. As well as all the Cox model inputs, network inputs included histological grade, type of surgery, and type of radiotherapy. The age at diagnosis was standardised by subtracting the mean for the data set from each value and then dividing by the standard deviation. Categorical inputs were represented by a dummy variable for each category, bringing the total to 51 inputs. Unlike the Cox models, the number of days of follow-up was not used at any stage in model building or testing. The optimum number of hidden layer neurons was estimated using a standard rule-of-thumb (half the sum of inputs and outputs, plus the square root of the number of training patterns (Smith, 1999)), so the two-year models had 82 hidden neurons and the five year models had 69.

Assessing the models

All the models were assessed and compared in three ways. In the first, the model's estimates of the probability of survival were converted to 0 or 1 using differing cut-off points between 0 and 1 in increments of 0.1. (These are better known in the neural network community as 'decision thresholds'.) For each of these increments we

calculated the percentage of patients in the test set whose survival or death would have been correctly predicted by the model. Models were compared using the percentage correct at the optimum cut-off point.

Next the overall accuracy of each model was assessed by plotting its receiver operating characteristic (ROC) curve (DeLeo and Rosenfeld, 2001) using a commercial statistics package (SPSS) (SPSS Inc, 1999) and comparing the areas under the curves.

Finally, the 'calibration' of each model was assessed by dividing the patients into strata based on their predicted probabilities of survival (0-0.19, 0.2 – 0.39, and so on to 0.8-1.0), and calculating the percentage of patients in each stratum who actually survived (Wyatt and Spiegelhalter, 1990, Wyatt and Altman, 1995). The basic question is, "Did 20 percent of those given a 20 percent probability of surviving by the model actually survive, did 50 percent of those given a 50 percent probability survive, and so on?" The calibration assessed each model as it was applied to its own reserved test set.

Results

Table 7-1 shows the distribution of patient characteristics for the complete 2 year and 5 year data sets. Table 7-2 lists the numbers of deaths in each randomly-assigned test set for the 2 year and 5 year modelling exercises.

The two modelling techniques gave very similar results for each of the assessment methods. For 2 year survival, the overall mean percentage correct was 76.2% for the neural networks (range 74.3 to 79.0) compared to 76.8% for the Cox models (range 74.3 to 79.4). For 5 year survival, the overall mean percentage correct was slightly lower: 69.3% for the neural networks (range 65.8 to 70.9) and 70.1% for the Cox models (range 67.3 to 71.7). The optimum cut-off point (threshold) for each model was either 0.5 or 0.6 in every case.

Table 7-3 compares the areas under receiver operating characteristic (ROC) curves for the two modelling techniques. Once again, they gave quite similar results, with areas around 73% in each case. The Cox models had slightly better areas (1.2 and 1.3 percentage points respectively for the means of 2 year and 5 year predictions), but the asymptotic 95% confidence intervals overlapped widely in every case.

Table 7-1: Demographic and clinical characteristics of the Two year and Five year data sets.

Characteristics	Two year	Five year
Total	4463	2615
Survived, N(%)	3290 (74)	995 (38)
Age, mean (sd)	77 (7)	77 (7)
Race, N(%):		
Non-Hispanic White	3733 (84)	2209 (84)
Non-Hispanic Black	348 (8)	197 (8)
Hispanic	158 (4)	85 (3)
Other	224 (5)	124 (5)
Gender, male, N (%)	1922 (43)	1109 (42)
Urban residence (pop. ≥250,000) N (%)	3693 (83)	2188 (84)
Nodes involved with tumour, N(%):		
1	1599 (36)	859 (33)
2	901 (20)	495 (19)
3+	1818 (41)	1153 (44)
Unknown	143 (3)	108 (4)
Extent of disease, N(%):		
T1	78 (2)	38 (1)
T2	338 (8)	172 (7)
T3	3698 (83)	2145 (82)
T4	349 (8)	260 (10)
Charlson comorbidity score, mean (sd)	0.5 (0.9)	0.6 (0.9)
Adjuvant chemotherapy with 5FU, N (%)	2367 (53)	1276 (49)

Table 7-2: Mortality by test set: Number who died in each test set for Two and Five year models.

	Number each test set	in Deaths In Test1	Deaths In Test2	Deaths in Test3	Deaths in Test4	Deaths in Test5
Two year	670	489	499	512	498	494
Five year	392	159	139	137	157	147

Table 7-3: Areas under ROC curves for predicting survival.

	Two-year survival				Five-year survival			
	Neural Networks		Cox Regression		Neural Networks		Cox Regression	
	Area	(95% CI)	Area	(95% CI)	Area	(95% CI)	Area	(95% CI)
Test1	**0.743**	(0.701 - 0.785)	**0.742**	(0.701 - 0.784)	**0.679**	(0.625 - 0.734)	**0.702**	(0.650 - 0.754)
Test2	**0.699**	(0.652 - 0.745)	**0.719**	(0.674 - 0.763)	**0.752**	(0.703 - 0.800)	**0.760**	(0.712 - 0.809)
Test3	**0.749**	(0.705 - 0.793)	**0.764**	(0.722 - 0.807)	**0.745**	(0.695 - 0.795)	**0.745**	(0.694 - 0.795)
Test4	**0.749**	(0.708 - 0.791)	**0.758**	(0.718 - 0.799)	**0.755**	(0.707 - 0.804)	**0.773**	(0.726 - 0.821)
Test5	**0.721**	(0.678 - 0.764)	**0.735**	(0.691 - 0.778)	**0.700**	(0.646 - 0.754)	**0.715**	(0.662 - 0.767)

Model calibration was also quite similar for the two modelling techniques. Mean calibration data for all the models are presented in Table 7-4. For a well calibrated model, the percentage who actually survived in each stratum should be close to the mid-point of the stratum. Both the neural network and Cox models were quite well calibrated, especially for survival predictions above 0.19.

Table 7-4: Model calibration for two and five-year survival predictions: mean results of the five tests.

	Two-year survival				Five-year survival			
	Neural networks		Cox Regression		Neural networks		Cox Regression	
Prediction Range	Mean number (%) Surviving		Mean number (%) Surviving		Mean number (%) Surviving		Mean number (%) Surviving	
0.0-0.19	1.6	(20.0)	0.6	(12.0)	14.4	(16.5)	14.8	(16.3)
0.2-0.39	12	(38.2)	7.8	(39.8)	39.4	(28.8)	29.2	(26.4)
0.4-0.59	47.2	(54.4)	40.0	(48.9)	52.6	(50.4)	52.8	(46.6)
0.6-0.79	164.2	(71.0)	187.8	(70.9)	32.4	(61.1)	40.8	(71.6)
0.8-1.0	273.4	(87.5)	239.4	(89.6)	9.0	(88.2)	3.6	(81.8)

The similarity in overall accuracy was reflected by general agreement for individual patients. For the optimum cut-off points (decision thresholds), there were only a small number of patients whose survival was correctly predicted by one type of model but

not the other. Table 7-5 presents the mean number and percentage of concordant and non-concordant predictions for a cut-off (threshold) of 0.5.

Table 7-5: Concordance between the Neural network and Cox model predictions: Means of all five tests.

	Two-year		Five-year	
	Number	(%)	Number	(%)
Both correct	495.4	(73.9)	251.0	(64.0)
Neither correct	140.6	(21.0)	98.6	(25.2)
Cox only	18.4	(2.7)	20.8	(5.3)
Network only	15.6	(2.3)	21.6	(5.5)
Total	670	(100.0)	392	(100.0)

Discussion

This study shows that the SEER-Medicare database does contain prognostic information, allowing the creation of models that give moderately accurate predictions of survival to 2 years and 5 years for individual patients. Mathematical models made by both the Cox proportional hazards and artificial neural network approaches were able to correctly predict the survival of about 75% of the patients in randomly selected test sets, and the predictions were all well calibrated. The models used information that is routinely and reliably available at or soon after the time of diagnosis. None of the models gave 100 percent accuracy, so there could be other contributory factors, apart from chance alone, that are not captured by this data set.

Neural networks have been found superior to the TNM staging system when applied to colorectal carcinoma (the American College of Surgeons Patient Care Evaluation

data set) and to breast carcinoma (the SEER breast carcinoma data set) (Burke et al., 1997). They have been compared with logistic regression by area under ROC curves for hepatic disorders, using data from the Fallon Community Health Plan (Duh et al., 1998). Neural networks trained to predict death from colorectal carcinoma under 12 months on data from one UK institution achieved 90% overall accuracy when applied to data from a different institution (Bottaci et al., 1997). These neural networks were not restricted to routinely collected data from a cancer registry, the forecasting period was short (12 months) and the test sets were relatively small, so the results are not directly comparable with this study.

Although they were given more inputs than the Cox models, the neural networks did not, on average, perform better on any of the assessments. The minor differences are in favour of the Cox models, but none approach statistical significance or represent a clinically useful improvement in prediction accuracy. The use of a validation set to avoid network over-training means that the neural networks effectively had access to less training data, which may explain the Cox models' slight edge in the test sets. However, the marked similarity of performance suggests that there are no important non-linearities in these data, and that the Cox models are probably capturing as much prognostic information as exists. Levine (Levine, 2001) has suggested that cancer registries might usefully incorporate more detailed data on their patients, including biochemical and immunological parameters, and that neural networks may be valuable for teasing out the complex relationships between such variables and survival.

The models were tested against randomly selected subsets of the original data, but have not been tested prospectively with new data. Neither have they been tested against data from other cancer registries, but that was never the intention. It is quite possible that there are important differences between geographic regions, not to mention countries, which limit the creation of universally applicable models. Nevertheless, this dataset does cover a large and well-defined area, and both types of models provide quite well calibrated predictions, so they might be of use to many treatment facilities.

Wyatt and Altman (Wyatt and Altman, 1995) have summarised the challenge facing the developers of prognostic models. We believe that at least the Cox models in this study satisfy most of their criteria. For clinical credibility they include all clinically relevant data, and the data are obtainable with high reliability. There are no arbitrary thresholds for continuous variables, the model's structure is clear and does not transgress any of the method's assumptions, and it is relatively simple to calculate the model's prediction for a given patient. The probabilities the models generate are well calibrated, improving their applicability in, for example, weighting the caseload of a treatment facility to allow more realistic comparison of treatment outcomes.

Others have compared neural networks and Cox models for a number of prognostic problems (Kates et al., 1999, Ohno-Machado, 1997), including survival after diagnosis of other tumour types from the SEER data set (Burke et al., 1997), but none have assessed survival from colon cancer in this way. Previous studies have tended to emphasise the competition between the two modelling approaches, while saying little about the nature of the data set itself.

This study shows that the SEER-Medicare data set does contain a substantial amount of prognostic information, able to accurately predict the individual prognosis of 75% of the cases in a test set. This percentage may not be high enough for individual use, but the very well-calibrated models made from this data set do accurately predict the prognosis of groups of cases. They would be of potential use to health facilities wanting to compare their results over time or with others. The similar performance of the two types of model shows that there are no complex or non-linear relationships between the available data that would make a neural network approach more accurate than the more transparent Cox regression.

Chapter Eight: Using Artificial Neural Networks and Job Specific Modules to assess Occupational Exposure

Introduction

In the past the discovery of associations between industrial exposure to chemicals and human cancer was limited to situations where large numbers of workers were exposed to high levels of highly carcinogenic agents. Nowadays there are many more chemicals in the workplace and the general environment but, paradoxically, lower exposures and improved working environments make it very difficult to elucidate the causal connections.

In occupational epidemiology, rare diseases such as cancer can only be studied in large numbers in case-control studies. The challenge in such studies is to assess chemical exposure for a wide range of jobs for which no measured values are available. It is extremely important to accurately determine lifetime exposure to the putative carcinogen. Incorrect assignment of the degree of exposure introduces a bias that may either mask a true association or suggest one where there is actually none.

Expert assessment methods using Job Exposure Matrices and Job-Specific Modules

There have been some very useful developments over the past decade. The current best practice for exposure assessment is the use of an expert assessment method (Siemiatycki et al., 1981), with job specific questionnaires or modules (JSMs) (Stewart et al., 1998). These modules contain a series of questions about the

frequency and intensity of specific tasks. Each subject in the study provides a brief lifetime occupational history. For those jobs with potential to the occupational exposures of interest, further questions are asked using the JSMs. Expert hygienists then review the answers and estimate the probability of exposure to the chemicals of interest and the frequency, level and duration of such exposure (Siemiatycki, 1996, Siemiatycki et al., 1981). For example, a subject might have been a truck driver all his life with one company. He would be allocated the driver JSM which asks detailed questions about his driving patterns, refuelling behaviours and so on. The expert reviews these answers to allocate whether the subject is exposed to benzene, and the level, frequency, and duration of such exposure.

The Job Exposure Matrix (JEM) approach helps both automate and standardise the assignment of exposures to occupational categories. The newer Job-Specific Module (JSM) technique allows the expert assessor to differentiate more accurately between exposures experienced by different workers within the same job category.

Speed, consistency and expertise remain difficult areas

However the step from questionnaire to assessment is time consuming and requires considerable expertise on the part of the expert. In addition there are still subjective components to the assessment that can lead to substantial misclassification, so it is difficult to guarantee that every subject in the study has been assessed in a consistent way, and comparisons between different assessors (and thus between studies) are fraught with difficulty.

The New South Wales (NSW) Non-Hodgkin's Lymphoma (NHL) Study, whose data were used in this thesis, was a case-control study that recruited more than 650 cases and 650 controls to investigate multiple potential risk factors for NHL. A large number of different JSMs were used to collect data on occupational exposure. A single occupational hygienist read each completed questionnaire and assigned exposure probabilities and levels using a JEM. Even just the driver JSM used in this thesis took the hygienist at least ten minutes for each subject, amounting to many hours of work for all the relevant subjects.

Objective of this study

This study explored the possibility of using artificial neural networks to assess exposure to benzene in a JSM for drivers which was used in a community-based study. An expert hygienist assessed benzene exposure using the answers to the JSM and his ratings were taken as the correct exposure assessments. Artificial neural networks were then used to model the patterns of answers which resulted in the given benzene exposure assessments.

Methods

The New South Wales (NSW) Non-Hodgkin's Lymphoma (NHL) study was a collaborative community based case-control study that recruited more than 650 cases and 650 controls matched by age and sex, to investigate immunological, infectious, occupational (e.g. benzene, solvents, metals, pesticides and organic dusts) and environmental risk factors to NHL. Following completion of a self-administered questionnaire and job calendar, reported job histories were reviewed by the study

hygienist. The hygienist then allocated relevant JSMs to subjects. Subjects were interviewed using a computer assisted telephone interview, which involved the use of up to 5 JSMs per subject.

A total of 44 JSMs were used in the NHL study, which were based on approximately 80 modules originally developed for the National Cancer Institute, Bethesda, Maryland, USA (Stewart et al., 1998). The JSMs used in the NSW NHL study were modified for Australian industry and occupation conditions. Figure 8-1 depicts the Driver module.

The expert assessor assessed exposure based upon the answers to specific questions in the modules, community based JEMs (FINJEM) (Kauppinen et al., 1998), specially developed exposure matrices, published and unpublished literature, occupational/hygiene texts and advice from a network of expert government, corporate and consultant hygienists. Exposure to benzene was then assessed as 'no exposure', or 'probable exposure'. A secondary assessment was made for exposed cases, to decide if the exposure was intermittent.

Training data

A total of 189 driver JSMs were assessed by an experienced industrial hygienist, and the probability of occupational exposure to benzene determined. The questionnaire data from the driver JSM were processed before presentation to the neural network. Some of the free text questions were omitted. Most of the multiple choice and Yes/No questions were converted to a series of binary inputs. Some, which had few differing responses, were converted to binary inputs. Numeric answers were standardised by

subtracting the mean and dividing by the standard deviation. The final data set had 37 inputs.

Overall benzene exposure, expressed by the expert as 'none' or 'probable', and intermittent benzene exposure, expressed as true or false, were each represented by single variables.

Figure 8-1: The Driver Job-Specific Module (JSM) used to assess exposure.

```
DRIVER (DR)

DR1. What kind of vehicle did you usually drive?  Was it . . . (SELECT ONE)
a. A semi-trailer?
b. A tray truck?
c. A bus?
d. A taxi or limosine?
e. Some other type of vehicle  (DESCRIBE)

DR2. What kind of fuel did the vehicle use? (SELECT ONE)
PETROL
DIESEL
LPG
SOMETHING ELSE (SPECIFY)
DK

DR3. On how many weeks of the year did you fill the fuel tank yourself?

   And, on average, how many hours on one of those weeks?

DR4 INSTRUCTIONS: IF DR1 = d, END MODULE.  IF DR1 = c, GO TO DR4.
OTHERWISE, GO TO DR5
DR4. What kind of bus did you usually drive? (SELECT ONE)
School
Municipal, city, or local
Long-distance
Some other type? (SPECIFY)

DR5 INSTRUCTIONS: IF DR1 = c OR e, END MODULE. OTHERWISE, GO TO DR5
DR5. Did you haul one or more products . . . (SELECT ONE)
For a single company, or
For several different companies, or
For a local, state, or federal government?
OTHER (DESCRIBE)

DR7.  What percentage of your total time did you work on trucks carrying the
following cargo?  (Record percentage for each of following)
a. Chemicals
b. Grains/fruits/vegetables
c. Live animals
d. Petroleum products
e. Wood products/timber

IF DR7a + b + c + d + e < 10%, GO TO DR19

DR8 INSTRUCTIONS: IF DR7a (CHEMICALS) > 80%, ASK DR8.  OTHERWISE, GO TO
DR10.
```

DR8. What were the specific chemicals you hauled?
RECORD VERBATIM

DR9. Were these chemicals mainly . . (SELECT ONE)
1. Liquids?
2. Gases?
3. Powders or dry products?

DR10. How was the cargo you carried in your truck or trailer packaged?
(SELECT ALL THAT APPLY)
IN PAPER OR BURLAP BAGS
IN OTHER CLOSED CONTAINERS SUCH AS BOXES, DRUMS, OR PLASTIC BAGS OR WRAP
IN LOOSE LIQUID FORM
GASES
IN SOME OTHER FORM OR CONTAINER (SPECIFY)
IT WASN'T PACKAGED (TIMBER, SAND, GRAIN)
DK HOW PACKAGED

DR11. Were you involved in loading your truck?
YES
NO (GO TO DR15)
DK (GO TO DR15)

DR12. How was the loading usually done? By ... (SELECT ONE)
1. Hand, handcart or hand truck, or other mechanical device
2. Forklift truck
3. Conveyor belt
4. Hose
5. Front end loader
6. Some other method (SPECIFY)
8. DK

DR13 INSTRUCTIONS: IF DR12 = 4, ASK DR13. OTHERWISE, GO TO DR14
DR13. When you loaded the truck using a hose, was the loading usually done
... (SELECT ONE)
At the top of the tanker with an open dome
At the top of the tanker with a closed dome
Underneath the tanker
Some other way (DESCRIBE)
DK

DR14. On how many weeks of the year did you load the truck this way?

And, on average, how many hours on one of those weeks?

DR15. Were you involved in unloading your truck?
YES
NO (GO TO DR19)
DK (GO TO DR19)

DR16. How was the unloading usually done? By . . . (SELECT ONE)
1. Hand, handcart or hand truck, or other mechanical device
2. Forklift truck
3. Conveyor belt
4. Hose
5. Front end loader
6. Some other method (SPECIFY)
8. DK

DR17 INSTRUCTIONS: IF DR16 = 4, ASK DR17. OTHERWISE, GO TO DR18
DR17. When you unloaded the truck using a hose, was the unloading usually
done . . . (SELECT ONE)

At the top of the tanker with an open dome
At the top of the tanker with a closed dome
Underneath the tanker
Some other way (DESCRIBE)
DK

```
DR18. On how many weeks of the year did you unload the truck this way?

   And, on average, how many hours on one of those weeks?

DR19. Did you perform any maintenance work on your truck?
YES, ALL
YES, SOMETIMES
NO
DK
END MODULE
```

Creating and training the neural networks

A commercial neural network program, NeuroShell2 (Ward Systems Group Inc, 2000), was used to train Feed-forward networks with one hidden layer by the back-propagation of errors. In each case starting weights were randomly assigned between −0.3 and +0.3, the activation functions for the hidden and output layers were all logistic, and momentum and learning rate terms of 0.1 each were used.

A separate set of networks was created for each of the different outputs: overall benzene exposure, and intermittent benzene exposure. In neural network parlance the set of answers to the JSM of one subject about one job, combined with the expert's assessment of benzene exposure, is referred to as a 'pattern'. Because the number of patterns was small the allocation of unusual cases to the validation or test sets might have a large effect on the test result, so five different networks were created for each output. Each used different random divisions of the data into 133 patterns used for training, 28 patterns used for validation, and 28 patterns used for final testing. The number of hidden neurons (30) was calculated using a common rule of thumb (half the sum of inputs and outputs, plus the square root of the number of training patterns (Smith, 1999)).

The training criterion was the mean squared error for the entire training set. As is common with the back-propagation training algorithm, training was stopped when 40,000 patterns had been presented to the network (i.e. the whole training set had been presented more than 200 times) without further improvement in the mean squared error for the validation set. To avoid over-training, the final 'trained' network was the version that had given the lowest mean squared error for the validation set.

Interpreting and testing the neural networks

Each network was tested by applying it to the reserved test set of 28 patterns. The output of this type of neural network is a continuous variable with a value between 0 and 1, and can be interpreted as the network's assessment of the probability that a given pattern should be classified as positive (i.e. should have a true value of 1). To compare with the all-or-nothing assignments given by the expert assessor, one approach is for the modeller to choose a cutoff point below which the network output will be interpreted as zero. (These values are more commonly called decision thresholds among the neural network community.) Values equal to or greater than the cutoff are considered to be 1. The percentage of correct assignments on the test set can then be used as one assessment of network accuracy. Where the prevalence of the positive state is low, it is valid to ask whether the neural network is better than a trivial model that simply predicts a zero for every case. (This 'default accuracy' equals one minus the prevalence of positives.) These were also assessed.

Rather than using an arbitrary cutoff (decision threshold) a more accurate assessment of a diagnostic test such as this can be made by plotting a receiver-operating characteristic (ROC) curve (DeLeo, 1993). By plotting false negatives against false

positives this effectively considers a range of possible cutoff points, and the greater the area under the ROC curve the more accurate the test. ROC curves were plotted for each of the networks applied to the reserved test sets. ROC curves were plotted using SPSS (SPSS Inc, 1999).

A third way of assessing the network accuracy is to experiment with cutoff points (decision thresholds) in search of levels that give very good positive or negative predictive values. If the network is always correct when it assesses an exposure as negative, for example, then the expert assessor need only look at those questionnaires given positive or equivocal ratings by the network. This could lead to substantial savings in time and effort.

Results

Table 8-1 shows the overall accuracy of each set of neural networks when applied to the five different test sets of 28 reserved cases. For example, when the 4th network trained for overall benzene exposure was applied to its test set, 95% of the network's assessments agreed with the expert's original assessment. For overall benzene exposure the mean number of accurate assessments was 25.2/28 or 90%. For intermittent benzene exposure the average correct assignments was 26/28 (93%). If the optimum cut-off point was used for each network instead of the arbitrary choice of 0.5, the percentage correct improved slightly to 93% for benzene exposure (cutoffs 0.3, 0.4 or 0.5), and 94% for intermittent benzene exposure (cutoff 0.9).

Table 8-1: Accuracy of the neural networks applied to the 28 reserved test patterns: Percentage correct using a cutoff (threshold) of 0.5, compared with a 'default zero' model.

Benzene exposure	Network 1	Network 2	Network 3	Network 4	Network 5	Mean
Overall	96	89	79	96	89	90.0
Intermittent	89	86	100	96	93	92.9
Overall default[a]	68	64	64	71	71	67.9
Intermittent default[a]	89	89	100	96	93	93.6

[a] Percentage correct for a simple model predicting zero for every case.

Tables 8-2 and 8-3 show the areas under ROC curves (and nonparametric 95% confidence intervals [CI]) for each of the networks applied to the reserved test sets. A perfect match would give an area of exactly one; the mean area for the five networks assigning overall benzene exposure status was 0.93, and for intermittent benzene the mean was 0.82.

All the networks assessing overall benzene exposure performed better than the trivial 'zero only' model. One of the networks assessing intermittent exposure performed less well, and the others had the same overall accuracy as the 'zero only' model.

Table 8-2: Overall benzene exposure: Areas under receiver-operating characteristic (ROC) curves and number of positives in each test set.

	Area	95% CI		Number Positive
Network 1	0.98	0.94-	1.00	9
Network 2	0.96	0.88-	1.00	10
Network 3	0.84	0.68-	1.00	10
Network 4	1.00	1.00-	1.00	8
Network 5	0.88	0.75-	1.00	8
Mean	0.93	0.85-	1.02	

Table 8-3: Intermittent benzene exposure: Areas under receiver-operating characteristic (ROC) curves and number of positives in each test set.

	Area	95% CI		Number Positive
Network 1	0.73	0.56-	0.91	3
Network 2	0.80	0.64-	0.96	3
Network 3	Cannot be calculated[b]			0
Network 4	1.00	1.00-	1.00	1
Network 5	0.73	0.56-	0.90	2
Mean	0.82	0.69-	0.94	

[b] there were no positive assignments in this test set

Tables 8-4 and 8-5 show the sensitivity, specificity, positive and negative predictive values, and the percentage of cases correctly identified as negative. For a decision threshold (cutoff) of 0.3 all but one of the networks assigning overall benzene exposure gave a negative predictive value of 100%. The only false negatives in all the tests were two petrol tanker drivers. Notably, there were only three tanker drivers in the whole data set, so the one poorly-performing network had seen only one such

example in its training set. These networks would have correctly identified between 54% and 71% of the drivers as **not** exposed to benzene. Similarly, for a threshold cutoff of 0.04 (reflecting the lower prevalence of the positive state in this data set (DeLeo, 1993)) all the five networks assessing intermittent benzene exposure gave negative predictive values of 100%. These networks would have correctly identified between 43% and 61% of the drivers as not exposed.

Table 8-4: Overall benzene exposure: Sensitivity, Specificity, Positive and Negative Predictive Values, (as percentages) and percentage of cases correctly identified as negative, with a decision threshold (cutoff) of 0.3.

	Network 1	Network 2	Network 3	Network 4	Network 5
Sensitivity	100	100	80	100	100
Specificity	95	94	83	100	80
Positive Predictive Value	90	91	73	100	67
Negative Predictive Value	100	100	88	100	100
Negatives correctly identified	64	61	54	71	57

Table 8-5: Intermittent benzene exposure: Sensitivity, Specificity, Positive and Negative Predictive Values (as percentages), and percentage of cases correctly identified as negative, with a decision threshold (cutoff) of 0.03.

	Network 1	Network 2	Network 3	Network 4	Network 5
Sensitivity	100	100	[c]	100	100
Specificity	68	56	57	56	46
Positive Predictive Value	27	21	0	8	13
Negative Predictive Value	100	100	100	100	100
Negatives correctly identified	61	50	57	54	43

[c] Division by zero – there were no positives in this test set.

Discussion

This study shows that simple multi-layer fed-forward neural networks trained by back-propagation of errors can extract useful information from a standardised driver's JSM, and simulate an expert assessment of benzene exposure with relative accuracy. Even this preliminary study produced neural networks that could accurately and reliably say who had **not** been exposed, potentially decreasing the expert assessor's workload for future studies by as much as 60% for this JSM.

Although this is in keeping with studies in other fields, we are not aware of other studies of this kind in occupational health. Claycamp and others (Claycamp et al., 2001) have reported the use of neural networks in a study amongst the Mayak Production Association workers in Russia, but this did not involve exposure assessment. Earlier unpublished work by Claycamp and others (Claycamp et al., 1998), concluded that networks are useful as a complementary statistical tool to use in quantitative risk assessment.

A limitation with this study is the very small number of training examples available. There was considerable variation in overall network accuracy between the different random allocations to training, validation and test sets. This, and the tendency of one network to incorrectly assign tanker drivers to the 'unexposed' category, highlights the fact that the network training set must include multiple examples of every possible pattern if the network is to correctly categorise similar patterns in the future.

The neural networks trained for this study are not necessarily portable to other studies – if the questions in the modules were changed it would be necessary to train new

networks. Use of the networks outside the original study is however feasible if the future study has the same exposures in the primary hypotheses.

Even this small study illustrates three potential uses of artificial neural networks together with JSMs to make the assessment of occupational chemical exposures quicker, more consistent, less subjective, and less resource intensive. Firstly, the expert may use the outputs of a trained network to check the consistency of his/her assessments. Second, by setting the decision threshold cutoff to a level that gives a negative predictive value of 100%, the network could be used a filter, eliminating more than half of new records by reliably assessing them as negative. This could mean a substantial reduction in workload for the expert assessor. Finally, with larger data sets for training, it should be possible to train networks to near-complete accuracy for both positive and negative assessments.

We are optimistic that it should be possible to supplement and even replace some of the work of expert assessors with artificial neural networks trained to simulate the experts' own decision-making processes.

Chapter Nine: Conclusions.

The main findings of this research

The experiments described in this thesis cast light into some hitherto unexplored corners of modern epidemiology. Some promising avenues have been discovered for further investigation. And like any voyage of exploration, even when a new route to success is not discovered, it can still be an achievement to simply put up a warning sign to stop others going down an enticing but ultimately disappointing *cul-de-sac*.

Gastroenteritis in urban Australia.

Main findings

This study showed that there are strong deterministic elements in the relationship between the social calendar, previous requests for faecal analysis, and recent weather, and the future numbers of requests. Like other early-warning proxies for gastroenteritis disease, the number of requests for faecal analysis sacrifices specificity in favour of rapidity. The indicator obviously mixes disease (itself caused by multiple organisms), care-seeking behaviour, doctors' level of clinical suspicion, and laboratory availability. In its favour, though, is its sensitivity plus the fact that rises in this indicator could occur a week or more sooner than rises in laboratory-confirmed cases of specific pathogens.

Although each of the weather inputs was shown to have a separate and identifiable influence on the prediction of future requests, it seems that the patterns in the data are too complex for all the possible effects to be captured in only four years worth of daily data. Nevertheless simpler models using fewer weather inputs gave accurate

predictions for a prospective test over nearly six months. As the first application of artificial neural networks to this problem, these are very encouraging results.

Future research

These positive results suggest that this may be a very useful technique, at least for the early detection of city-wide outbreaks (such as water-borne gastroenteritis, which was the original inspiration for the study). Further studies are needed to confirm the results for different time windows. Larger sets of training data may be available in future, which may allow more of the weather inputs to be included in the models with improvement in their prospective predictive accuracy. The same method should be tried for data sets from other cities, and for other diseases.

Measles in Mozambique

Main findings

This study shows that there is indeed a strong deterministic element in the relationship between recent measles case numbers in Mozambique and future cases for a useful distance into the future. Once the data have been suitably aggregated (to provincial level) and smoothed (using a six-week asymmetric moving average), a feed-forward multilayer neural network can accurately model future case numbers up to eight weeks ahead *for a given time window.*

However, the relationship is very complex, requiring a large network with many inputs. More importantly, the same relationship identified for a particular set of training data does not hold into the future. This severely limits the usefulness of this approach as a practical tool for surveillance practitioners. The data set available was

never going to be easy one to model. The years it covers include the latter part of Mozambique's civil war, the unstable period of the peace negotiations and the national elections that followed, and then a period of rapid health service expansion. But the inescapable fact for the measles modeller is that real-life data sets for highly endemic countries are unlikely ever to be better than this one. What is more, the changing conditions in Africa mean that past data sets may never be highly indicative of the future. The main conclusion must be that this initially promising approach has turned out not to be of any practical use to the measles surveillance practitioner.

Future research

The failure of this approach to create a tool of day-to-day use to measles surveillance practitioners is disappointing. (As Robert Burns wrote, *The best laid schemes o' mice an' men/ Gang aft agley,/ An' lea'e us nought but grief an' pain/ For promis'd joy.* (Burns, 1786)) This does not necessarily mean, however, that neural networks have nothing at all to offer in the forecasting of measles cases. There is still room to explore different representations of the data. Perhaps the problem could be turned into a simpler classification exercise by attempting to predict the rate of change in case numbers as general categories rather than exact numbers. The challenge will be great, as the quality of data available will always be relatively poor – but any advance that hastens the end of measles as a public health problem in Africa is worth the effort.

Cancer prognosis in the United States

Main findings

This study illustrates a novel approach to the creation of prognostic models. It deliberately makes no attempt to build a one-size-fits-all model to be universally

applicable. Rather, this study asks whether a model based on a large but (relatively) local database such as the SEER-Medicare data set might accurately predict survival for future patients with similar disease and personal characteristics.

In addition, rather than asking simply which technique – Cox regression or neural networks – gives a better result, we used the neural networks to explore the potential non-linearity and complexity of the relationships between the inputs and survival. We tested various random splits of the original data into training and test sets, to assess the effect of randomisation on the out-of-sample accuracy of the final models. We tested the 'calibration' of our models and found them also to be highly accurate in this regard.

The end result is firstly an answer to our principle question. The SEER-Medicare dataset does contain a substa ıtial amount of prognostic information. There do not seem to be substantial non-linearities or complexity in the data, so we tend to favour the Cox model because its internal workings are more 'transparent'. But by comparing techniques in this way we are more confident that the predictions of this model are as accurate as possible for this data set.

Future research

There is scope for a good deal of further research in this area. This study and others like it suggest that all new prognostic models might usefully include a neural network analysis in their construction. This is not to say that neural networks will always provide more accurate models than parametric techniques (and the greater 'transparency' of parametric techniques makes them in some ways more attractive). But at the very least neural network models can provide an indication of how much prognostic information there is in the data set under investigation, to guide the creation of the more transparent models. Further, the ability of neural networks to find subtle and complex non-linear relationships between inputs and outputs is a cause for optimism, and might allow future researchers to use inputs which are currently not included in prognostic models (or even in cancer registries) to create ever more accurate predictive models.

Further, this kind of modelling exercise is relatively straightforward. It would be possible to create similar models for other geographically-based cancer data sets. There are undoubted advantages to universal models like the TNM staging system, but they can be supplemented by locally-applicable models such as those explored in this study.

Occupational exposure to possible carcinogens

Main findings

Occupational epidemiologists eagerly seek more accurate, more consistent and less time-consuming methods for exposure assessment. Even with this small data set it was possible to model much of the reasoning behind the assessment of benzene

exposure in drivers. The model was not sufficiently accurate to replace the assessor completely. But with simple modifications, such as a pre-network test that new questionnaires are similar to the training data, it could be put into use for deciding which drivers did not have exposure (and thus which questionnaires do *not* need close assessment by the expert).

The artificial neural network approach explored in this chapter offers a way of at least reducing the workload of the assessor, and making the assessments more consistent. If the same JSM were to be used in future studies together with the model created here it would go some way to allowing more meaningful comparisons between the different studies.

Future research

Future research directions in occupational health research are clear: the encouraging results from the small number of driver questionnaires should be confirmed and expanded in three ways. First, before the trained neural networks can be applied with confidence to other studies, it will be necessary to validate their performance, particularly on data from other studies in which the assessments are made in the conventional way.

Second, the same methodology could be applied to other Job-Specific Modules in the same study (teachers, health workers, farmers and many others), to see if artificial neural network models could also provide useful information to expert assessors, and possibly make their work less onerous and more consistent.

Finally if new network models can be created from much larger data sets, perhaps incorporating data from several studies using the same JSMs, they may prove much more accurate than the current one, so that there is a possibility of eventually replacing the expert assessor or panel altogether.

Implications for epidemiologists

Artificial neural networks are not about to overtake all the existing methods for classification or forecasting currently used by epidemiologists. Indeed, they have possibly been 'oversold' in the past, leading to understandable scepticism. But the usefulness of artificial neural network modelling in so many of the growth areas of modern epidemiology, evidenced by this thesis and the many other studies in the field, suggests that neural networks should become a greater part of the research 'toolkit' for many different kinds of epidemiological study. In some cases neural networks will probably be best used as an adjunct to existing statistical methods (or perhaps in hybrid techniques). In other areas, especially in 'mining' data sets not initially designed for statistical analysis, their intrinsic non-linearity and lack of assumptions about the structure of the data may yet prove to be of real practical worth. In infectious disease surveillance the models produced in this study are good enough to consider field trials.

Summary

This thesis explores potential uses for artificial neural networks for some of the key areas of modern epidemiology.

In the process of building the gastroenteritis models the importance of recent weather in modelling rates of requests for faecal analysis was emphasised, using neural networks as a non-parametric statistical tool. The best of the resulting models indicate the considerable promise this technique holds for the surveillance of common diseases.

In forecasting measles in Mozambique the models were limited – probably more because of limitations in the data than of the method.

In cancer prognosis the role of the networks was as part of a novel paradigm of geographically localised models based on and applicable to a single (albeit large) cancer registry. The goal is a model that extracts all available prognostic information from the registry data set, providing accurate prognosis for newly-diagnosed cancer patients.

In occupational health neural networks were shown to be useful in 'learning' the rules by which an expert assessor assigned benzene exposure in drivers. This leads, at least, to a way of making exposure assessments more consistent and less time-consuming. Future models may become more accurate.

Artificial neural networks deserve a place in the statistical 'tool-kit' of modern epidemiology.

References

Altman, D. G. (1991) *Practical statistics for medical research,* Chapman & Hall, London.

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans - Dynamics and Control,* Oxford University Press, New York.

Andrews, R., Carnie, J. and Tallis, G. (1997) *Surveillance of notifiable disease in Victoria 1997,* Department of Human Services, Public Health and Development Division, Melbourne.

Astion, M., Wener, M., Thomas, R., Hunder, G. and Bloch, D. (1993) Overtraining in neural networks that interpret clinical data, *Clinical Chemistry,* **39,** 1998-2004.

Baba, N. and Kozaki, M. (1992) An intelligent forecasting system of stock price using neural networks, *IJCNN International Joint Conference on Neural Networks (Cat. No.92CH3114-6). IEEE. 1992, 371-7 vol.1. New York, New York, USA.*

Batuello, J. T., Gamito, E. J., Crawford, E. D., Han, M., Partin, A. W., McLeod, D. G. and O'Donnell, C. (2001) Artificial neural network model for the assessment of lymph node spread in patients with clinically localized prostate cancer, *Urology,* **57,** 481-5.

Baxt, W. G. (1991) Use of an artificial neural network for the diagnosis of myocardial infarction, *Annals of Internal Medicine,* **115,** 843-8.

Baxt, W. G. (1995) Application of artificial neural networks to clinical medicine, *Lancet,* **346,** 1135-38.

Baxt, W. G. and Skora, J. (1996) Prospective validation of artificial neural network trained to identify acute myocardial infection, *Lancet,* **347,** 12-15.

Becalick, D. C. and Coats, T. J. (2001) Comparison of Artificial Intelligence Techniques with UKTRISS for Estimating Probability of Survival after Trauma, *Journal of Trauma-Injury Infection & Critical Care,* **51,** 123-133.

Bellman, R. (1961) *Adaptive control processes: a guided tour,* Princeton University Press, Princeton, New Jersey.

Bellotti, M., Elsner, B., De Lima, A. P., Esteva, H. and Marchevsky, A. M. (1997) Neural networks as a prognostic tool for patients with non-small cell carcinoma of the lung, *Modern Pathology,* **10,** 1221-7.

Benke, G. P. (2000) *Retrospective assessment of occupational exposures by job exposure matrices and expert hygiene panels* (Ph.D. thesis, Department of Epidemiology and Preventive Medicine, Monash University, Melbourne.)

Berne, R. M. and Levy, M. N. (1998) *Physiology,* Mosby, St. Louis, Missouri.

Bostwick, D. G. and Burke, H. B. (2001) Prediction of individual patient outcome in cancer: comparison of artificial neural networks and Kaplan-Meier methods, *Cancer,* **91,** 1643-6.

Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W. R., Macintyre, I. M. C., Duthie, G. S. and Monson, J. R. T. (1997) Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions, *Lancet,* **350,** 469-72.

Boulle, A., Chandramohan, D. and Weller, P. (2001) A case study of using artificial neural networks for classifying cause of death from verbal autopsy, *International Journal of Epidemiology,* **30,** 515-520.

Brahams, D. and Wyatt, J. (1989) Decision aids and the law, *Lancet,* **i,** 632-34.

Breiman, L. (1994) Comment on: Neural networks: A review from a statistical perspective, *Statistical Science,* **9,** 38-42.

Brickley, M. R., Cowpe, J. G. and Shepherd, J. P. (1996) Performance of a computer simulated neural network trained to categorise normal, premalignant and malignant oral smears, *Journal of Oral Pathology and Medicine,* **25,** 424-8.

Brierley, P. D. and Batty, W. J. (1997) Electric load modelling with neural networks: an insight into the black box, *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems (ICONIP '97), Dunedin, 24-28 November 1997,* **Vol. 2,** 1326-1329.

Brierley, P. D. and Batty, W. J. (1998) Neural data mining and modelling for electric load prediction, *Engineering Applications of Neural Networks (EANN98), (Gibraltar, 10-12th June 1998).*

Bryce, T. J., Dewhirst, M. W., Floyd, C. E., Hars, V. and Brizel, D. M. (1998) Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head neck, *International Journal of Radiation Oncology, Biology, Physics,* **41,** 339-45.

Bureau of Meteorology Australia (2001) *http://www.bom.gov.au/climate/averages/tables/cw_086071.shtml* (Climate averages for Melbourne regional office. Access 27/8/2001).

Burge, P. S., Pantin, C. F. A., Newton, D. T., Gannon, P. F. G., Bright, P., Belcher, J., McCoach, J., Baldwin, D. R. and Burge, C. B. S. G. (1999) Development of an expert system for the interpretation of serial peak expiratory flow measurements in the diagnosis of occupational asthma, *Occupational and Environmental Health,* **56,** 758-764.

Burke, H. B. (1994) Artificial neural networks for cancer research: Outcome prediction, *Seminars in Surgical Oncology,* **10,** 73-9.

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., Marks, J. R., Winchester, D. P. and Bostwick, D. G. (1997) Artificial neural networks improve the accuracy of cancer survival prediction, *Cancer*, **79,** 857-62.

Burns, R. (1786) *To a mouse, on turning up her nest with the plough, November, 1785*, Representative Poetry On-line, Web Development Group, Information Technology Services, University of Toronto Library. *http://www.library.utoronto.ca/utel/rp/poems/burns6.html*. Access 15/1/2002.

Buzatu, D. A., Taylor, K. K., Peret, D. C., Darsey, J. A. and Lang, N. P. (2001) The determination of cardiac surgical risk using artificial neural networks, *Journal of Surgical Research*, **95,** 61-6.

Cacciafesta, M., Campana, F., Piccirillo, G., Cicconetti, P , Trani, I., Leonetti-Luparini, R., Marigliano, V. and Verico, P. (2001) Neural network analysis in predicting 2-year survival in elderly people: a new statistical-mathematical approach, *Archives of Gerontology and Geriatrics*, **32,** 35-44.

Caire, P., Hatabian, G. and Muller, C. (1992) Progress in forecasting by Neural Networks, *1992 IEEE Joint conference on Neural Networks*, **2,** 540-545.

Centers for Disease Control and Prevention *http://www.cdc.gov/mmwr* (The home page of the Morbidity and Mortality Weekly Report.) Access 27/8/2001.

Centers for Disease Control and Prevention (1991) Current trends update: Graphic method for presentation of notifiable disease data - United States, 1990, *Morbidity and Mortality Weekly Report*, **40,** 124-125.

Chakraborty, K., Mehrotra, K., Mohan, C. K. and Ranka, S. (1992) Forecasting the behaviour of multivariate time series using neural networks, *Neural Networks*, **5,** 961-970.

Chan, D. Y. C. and Prager, D. (1991) Analysis of time series by neural networks, *1991 IEEE Joint conference on Neural Networks,* **1,** 355-360.

Chang, G-C., Luh, J-J., Liao, G-D., Lai, J-S., Cheng, C-K., Kuo, B-L. and Kuo, T-S. (1997) A neuro-control system for the knee joint position control with quadriceps stimulation, *IEEE Transactions on Rehabilitation Engineering,* **5,** 2-11.

Charles, J. (1998) AI and law enforcement, *IEEE Intelligent Systems,* **13,** 77-80.

Charlson, M. E., Pompei, P., Ales, K. L. and MacKenzie, C. R. (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation, *Journal of Chronic Diseases,* **40,** 373-83.

Chatfield, C. (1996) *The analysis of time series: An introduction,* Chapman & Hall, London.

Cheng, B. and Titterington, D. M. (1994) Neural networks: A review from a statistical perspective, *Statistical Science,* **9,** 2-54.

Chin, J. (Ed.) (2000) *Control of communicable diseases manual,* American Public Health Association, Washington.

Choi, K. and Thacker, S. B. (1981a) An evaluation of influenza mortality surveillance, 1962-1979. I. Time series forecasts of expected pneumonia and influenza deaths, *American Journal of Epidemiology,* **113,** 215-26.

Choi, K. and Thacker, S. B. (1981b) An evaluation of influenza mortality surveillance, 1962-1979. II. percentage of pneumonia and influenza deaths as an indicator of influenza activity, *American Journal of Epidemiology,* **113,** 227-235.

Claycamp, H., Sussman, N. and Marcia, O. (1998) Artificial neural networks for quantitative risk assessment and hazard identification, *American Industrial Hygiene Conference and Exposition, (Atlanta, Georgia)* 2.

Claycamp, H. G., Sussman, N. B., Okladnikova, N. D., Azizova, T. V., Pesternikova, V. S., Sumina, M. V. and Teplyakov, II (2001) Classification of chronic radiation sickness cases using neural networks and classification trees, *Health Physics,* **81,** 522-9.

Clermont, G., Angus, D. C., DiRusso, S. M., Griffin. M. and Linde-Zwirble, W. T. (2001) Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models, *Critical Care Medicine,* **29,** 291-296.

Cook, S. M., Glass, R. I., LeBaron, C. W. and Ho, M-S. (1990) Global seasonality of rotavirus infections, *Bulletin of the World Health Organization,* **68,** 171-77.

Cottrell, M., Girard, B., Girard, Y., Mangeas, M. and Muller, C. (1995) Neural Modeling for time series: A statistical stepwise method for weight elimination, *IEEE Transactions on Neural Networks,* **6,** 1355-64.

Cross, S. S., Harrison, R. F. and Kennedy, R. L. (1995) Introduction to neural networks, *Lancet,* **346,** 1075-9.

Cutts, F. T., Henao-Restrepo, A. and Olive, J. M. (1999) Measles elimination: progress and challenges, *Vaccine,* **17,** S47-52.

Cutts, F. T., Monteiro, O., Tabard, P. and Cliff, J. (1994) Measles control in Maputo, Mozambique, using a single dose of Schwarz vaccine at age 9 months, *Bulletin of the World Health Organization,* **72,** 227-31.

Dayhoff, J. E. and DeLeo, J. M. (2001) Artificial neural networks - Opening the black box, *Cancer,* **91,** 1615-35.

De Laurentiis, M., De Placido, S., Bianco, A. R., Clark, G. M. and Ravdin, R. M. (1999) A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients, *Clinical Cancer Research,* **5,** 4133-9.

de Olivera, K. A., Vannucci, A. and da Silva, E. C. (2000) Using neural networks to forecast chaotic time series, *Physica A*, **284,** 393-404.

DeGroff, C. G., Bhatikar, S., Hertzberg, J., Shandas, R., Valdes-Cruz, L. and Mahajan, R. L. (2001) Artificial neural network-based method of screening heart murmurs in children, *Circulation*, **103,** 2711-6.

DeLeo, J. M. (1993a) The receiver operating characteristic function as a tool for uncertainty management in artificial neural network decision-making, *Proceedings Second International Symposium on Uncertainty Modeling and Analysis, April 1993*,141-4.

DeLeo, J. M. (1993b) Receiver operating characteristic laboratory (ROCLAB): Software for developing decision strategies that account for uncertainty, *Proceedings Second International Symposium on Uncertainty Modeling and Analysis*, 318-25.

DeLeo, J. M. and Rosenfeld, S. J. (2001) Essential roles for receiver operating characteristic (ROC) methodology in classifier neural network applications, *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, **4,** 2730-1.

Deyo, R. A., Cherkin, D. C. and Ciol, M. A. (1992) Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases, *Journal of Clinical Epidemiology*, **45,** 613-9.

Dgedge, M., Novoa, A., Macassa, G., Sacarlal, J., Black, J., Michaud, C. and Cliff, J. (2001) The burden of disease in Maputo City, Mozambique: registered and autopsied deaths in 1994, *Bulletin of the World Health Organization*, **79,** 546-52.

DiRusso, S., Sullivan, T., Kamath, R., Holly, C., Cuff, S., Siegel, B. and Savino, J. (1999a) An artificial neural network model for prediction of survival in trauma patients: Co-morbid conditions do not improve model performance, *Critical Care Medicine,* **27,** 179A.

DiRusso, S. M., Kealey, G. P., Sullivan, T. H., Wibbenmeyer, L., Morgan, L., Schmidt, S., Burr, M. and Savino, J. A. (1999b) An artificial neural network as a model for prediction of survival in trauma patients: External validation with a dissonant data set, *Critical Care Medicine,* **27,** A153.

DiRusso, S. M., Sullivan, T., Holly, C., Cuff, S. N. and Savino, J. (2000) An artificial neural network as a model for prediction of survival in trauma patients: Validation for a regional trauma area, *Journal of Trauma-Injury Infection & Critical Care,* **49,** 212-20; discussion 220-3.

Doig, G. S., Inman, K. J., Sibbald, W. J., Martin, C. M. and Robertson, J. M. (1993) Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression, *Proceedings of the Annual Symposium on Computer Applications in Medical Care,* 361-5.

Dombi, G. W., Nandi, P., Saxe, J. M., Ledgerwood, A. M. and Lucas, C. E. (1995) Prediction of rib fracture injury outcome by an artificial neural network, *Journal of Trauma-Injury Infection & Critical Care,* **39,** 915-921.

Dorsey, S. G., Waltz, C. F., Brosch, L., Connerney, I., Schweitzer, E. J. and Bartlett, S. T. (1997) A neural network model for predicting pancreas transplant graft outcome, *Diabetes Care,* **20,** 1128-33.

Dostál, P. (1999) Neural network or ARIMA Model?, *Nostradamus '99,*

Dowell, S. F. (2001) Seasonal variation in host susceptibility and cycles of certain infectious diseases, *Emerging Infectious Diseases,* **7,** 369-374.

Driedger, A. M., Rennecker, J. L. and Marinas, B. J. (2001) Inactivation of Cryptosporidium parvum oocysts with ozone and monochloramine at low temperature, *Water Research,* **35,** 41-8.

Duh, M-S., Walker, A. M., Pagano, M. and Kronlund, K. (1998) Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example, *American Journal of Epidemiology,* **147,** 407-13.

Dybowski, R., Weller, P., Chang, R. and Gant, V. (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, *Lancet,* **347,** 1146-50.

Ebell, M. H. (1993) Artificial neural networks for predicting failure to survive following in-hospital cardiopulmonary resuscitation, *Journal of Family Practice,* **36,** 297-303.

Edwards, D. F., Hollingsworth, H., Zazulia, A. R. and Diringer, M. N. (1999) Artificial neural networks improve the prediction of mortality in intracerebral hemorrhage, *Neurology,* **53,** 351-7.

Ellis, D. and Morgan, N. (1999) Size matters: an empirical study of neural network training for large vocabulary continuous speech recognition, *Proceedings, 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing,* **2,** 1013 -1016.

Fani-Salek, M. H., Totten, V. Y. and Terezakis, S. A. (1999) Trauma scoring systems explained, *Emergency Medicine,* **11,** 155-166.

Farrington, C. P., Andrews, N. J., Beale, A. D. and Catchpole, M. A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease, *Journal of the Royal Statistical Society,* **159,** 547-63.

Farrington, C. P. and Beale, A. D. (1993) Computer-aided detection of temporal clusters of organisms reported to the Communicable Disease Surveillance Centre, *Communicable Disease Report CDR Review,* **3 Review No 6,** R78-R82.

Fernandéz-Pérez, C., Tejada, J. and Carrasco, M. (1998) Multivariate time series analysis in nosocomial infection surveillance: a case study, *International Journal of Epidemiology,* **27,** 282-288.

Fine, P. E. M. and Clarkson, J. A. (1982) Measles in England and Wales - 1: An analysis of factors underlying seasonal patterns, *International Journal of Epidemiology,* **11,** 5-14.

Finne, P., Finne, R., Auvinen, A., Juusela, H., Aro, J., Maattanen, L., Hakama, M., Rannikko, S., Tammela, T. L. and Stenman, U. (2000) Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network, *Urology,* **56,** 418-22.

Fischer, M. M. (1996) Computational Neural Networks: A new paradigm for spatial analysis, *1st International Conference on GeoComputation.*

Fisher, E. R., Anderson, S., Tan-Chiu, E., Fisher, B., Eaton, L. and Wolmark, N. (2001) Fifteen-year prognostic discriminants for invasive breast carcinoma, *Cancer,* **91,** 167[illegible]7.

Flanagan, J. R., Pittet, D., Li, N., Thievent, B., Suter, P. M. and Wenzel, R. P. (1996) Predicting survival of patients with sepsis by use of regression and neural network models, *Clinical Performance and Quality Health Care,* **4,** 96-103.

Fletcher, I., Adgar, A., Cox, C. S. and Boehme, T. J. (2001) Neural network applications in the water industry, *DERA/IEE Workshop Intelligent Sensor Processing (Ref No 01/050) IEE 2001*, 16/1-6.

Floyd, C. E., Jr., Lo, J. Y., Yun, A. J., Sullivan, D. C. and Kornguth, P. J. (1994) Prediction of breast cancer malignancy using an artificial neural network, *Cancer*, **74**, 2944-8.

Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural networks and the bias/valance dilemma, *Neural Computation*, **4**, 1-58.

Goldberg, M. and Hemon, D. (1993) Occupational epidemiology and assessment of exposure, *International Journal of Epidemiology*, **22**, S5-9.

Goldstein, S. T., Juranek, D. D., Ravenholt, O., Hightower, A. W., Martin, D. G., Mesnik, J. L., Griffiths, S. D., Bryant, A. J., Reich, R. R. and Herwaldt, B. L. (1996) Cryptosporidiosis: An Outbreak Associated with Drinking Water Despite State-of-the-Art Water Treatment, *Annals of Internal Medicine*, **124**, 459-68.

Gonzalez, L. P. and Arnaldo, C. M. (1993) Classification of drug-induced behaviors using a multi-layer feed-forward neural network, *Computer Methods and Programs in Biomedicine*, **40**, 167-73.

Green, D. and Swets, J. (1966) *Signal detection theory and psychophysics*, John Wiley and Sons, Inc., New York.

Grenfell, B. T., Kleczkowski, A., Gilligan, C. A. and Bolker, B. M. (1995) Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases, *Statistical Methods in Medical Research*, **4**, 160-83.

Guerriere, M. R. and Detsky, A. S. (1991) Neural networks: what are they? [letter; comment], *Annals of Internal Medicine*, **115**, 906-7.

Guyton, A. C. and Hall, J. E. (1996) *Textbook of medical physiology,* W.B. Saunders Co., Philadelphia.

Hamer, W. H. (1906) The Milroy lectures on epidemic disease in England - the evidence of variability and persistency of type. Lecture III, *Lancet,* 733-39.

Hammad, T. A., Abdel-Wahab, M. F., DeClaris, N., El-Sahly, A., El-Kady, N. and Strickland, G. T. (1996) Comparative evaluation of the use of artificial neural networks for modelling the epidemiology of schistosomiasis mansoni, *Transactions of the Royal Society of Tropical Medicine and Hygiene,* **90,** 372-6.

Han, M., Snow, P. B., Brandt, J. M. and Partin, A. W. (2001) Evaluation of artificial neural networks for the prediction of pathologic stage in prostate carcinoma, *Cancer,* **91,** 1661-6.

Han, M., Snow, P. B., Epstein, J. I., Chan, T. Y., Jones, K. A., Walsh, P. C. and Partin, A. W. (2000) A neural network predicts progression for men with Gleason score 3+4 versus 4+3 tumors after radical prostatectomy, *Urology,* **56.**

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology,* **143,** 29-36.

Hart, A. and Wyatt, J. (1990) Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks, *Medical Informatics,* **15,** 229-36.

Hayes, E. B., Matte, T. D., O'Brien, T. R., McKinley, T. W., Logsdon, G. S., Rose, J. B., Ungar, B. L. P., Word, D. M., Pinsky, P. F., Cummings, M. L., Wilson, M. A., Long, E. G., Hurwitz, E. S. and Juranek, D. D. (1989) Large community outbreak of Cryptosporidiosis due to contamination of a filtered public water supply, *New England Journal of Medicine,* **320,** 1372-6.

Hebb, D. O. (1949) *The organization of behavior: A neuropsychological theory,* Wiley, New York.

Hill, T., Marquez, L., O'Connor, M. and Remus, W. (1994) Artificial neural network models for forecasting and decision making, *International Journal of Forecasting,* **10,** 5-15.

Hinton, G. E. (1992) How neural networks learn from experience, *Scientific American,* **267,** 105-9.

Hirsch, S., Shapiro, J. L., Turega, M. A., Frank, T. L., Niven, R. M. and Frank, P. I. (2001) Using artificial neural networks to screen for a population for asthma, *Annals of Epidemiology,* **11,** 369-76.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences of the United States of America,* **79,** 2554-2558.

Hopfield, J. J. (1999) Brain, neural networks, and computation, *Reviews of Modern Physics,* **71,** S431-7.

Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators, *Neural Networks,* **2,** 359-66.

Horninger, W., Bartsch, G., Snow, P. B., Brandt, J. M. and Partin, A. W. (2001) The problem of cutoff levels in a screened population: appropriateness of informing screenees about their risk of having prostate carcinoma, *Cancer,* **91,** 1667-72.

Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression,* John Wiley & Sons Inc., New York.

Hunter, A., Kennedy, L., Henry, J. and Ferguson, I. (2000) Application of neural networks and sensitivity analysis to improved prediction of trauma survival, *Computer Methods and Programs in Biomedicine,* **62,** 11-9.

Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L. and Martin, S. M. (1997) Using laboratory-based surveillance data for prevention: An algorithm for detecting *Salmonella* outbreaks, *Emerging Infectious Diseases,* **3,** 395-400.

Inprise Corporation (2000), Borland C++ Builder 5.0, Build 12.34, Update Pack 1 (Computer program), Inprise Corporation.

InstallShield Software Corporation (1999), InstallShield Express 2.12 (Computer program) InstallShield Software Corporation.

Izenberg, S. D., Williams, M. D. and Luterman, A. (1997) Prediction of trauma mortality using a neural network, *American Surgeon,* **63,** 275-81.

Joppe, A., Cardon, H. R. A. and Bioch, J. C. (1990) A neural network for solving the travelling salesman problem on the basis of city adjacency in the tour, *INNC 90 Paris. International Neural Network Conference. Kluwer. 1990, 254-7 vol.1. Dordrecht, Netherlands.*

Kanjilal, P. P. and Bhattacharya, J. (1999) Revisited measles and chickenpox dynamics through orthogonal transformation, *Journal of Theoretical Biology,* **197,** 163-74.

Kappen, H. J. and Neijt, J. P. (1993) Neural network analysis to predict treatment outcome, *Annals of Oncology,* **4,** S31-S34.

Karnin, E. D. (1990) A simple procedure for pruning back-propagation trained neural networks, *IEEE Transactions on Neural Networks,* **1,** 239-42.

Kates, R. E., Berger, U., Ulm, B. K., Harbeck, N., Graeff, H. and Schmitt, M. (1999) Performance of neural nets, CART, and Cox models for censored survival

data, *1999 Third International Conference on Knowledge-Based Intelligent Information Engineering Systems Proceedings (Cat No 99TH8410) IEEE 1999*, 309-12.

Kauppinen, T., Toikkanen, J. and Pukkala, E. (1998) From cross-tabulations to multipurpose exposure information systems: a new job-exposure matrix, *American Journal of Industrial Medicine,* **33,** 409-17.

Kehoe, S., Lowe, D., Powell, J. E. and Vincente, B. (2000) Artificial neural networks and survival prediction in ovarian carcinoma, *European Journal of Gynaecological Oncology,* **6,** 583-4.

Kinsey, S. E., Groves, D. J., Smye, S. W., Richards, S. M., Chessells, J. M., Eden, O. B. and Bailey, C. C. (1999) A comparison of Cox regression and neural networks for risk stratification in cases of acute lymphoblastic leukaemia in children, *British Journal of Haematology,* **104,** 198-200.

Korn, F., Pagel, B-U. and Faloutsos, C. (2001) On the dimensionality curse and self-similarity blessing, *IEEE Transactions on Knowledge and Data Engineering,* **13,** 96-111.

Kwok, T. and Smith, K. A. (2000) Experimental analysis of chaotic neural network models for combinatorial optimization under a unifying framework, *Neural Networks,*, 731-744.

Lapuerta, P., L'Italien, G. J., Paul, S., Hendel, R. C., Leppo, J. A., Fleisher, L. A., Cohen, M. C., Eagle, K. A. and Giugliano, R. P. (1998) Neural network assessment of perioperative cardiac risk in vascular surgery patients, *Medical Decision Making,* **18,** 70-5.

Le Cun, Y. (1985) Une procedure d'apprentissage pour reseau a seuil asymetrique. [A procedure for training a network with asymmetric threshold.], *Cognitiva,* **85,** 599-604.

Levine, R. F. (2001) Clinical problems, computational solutions, *Cancer,* **91,** 1595-602.

Lezos, G., Tull, M., Havlicek, J. and Sluss, J. (1999) Predicting the future with the appropriate embedding dimension and time lag, *IJCNN '99 International Joint Conference on Neural Networks,* **4,** 2509 -2513.

Lim, C. P., Harrison, R. F. and Kennedy, R. L. (1997) Application of autonomous neural network systems to medical pattern classification tasks, *Artificial Intelligence in Medicine,* **11,** 215-39.

Lisboa, P. J. G. and Wong, H. (2001) Are neural networks best used to help logistic regression? An example from breast cancer survival analysis, *Proceedings IJCNN '01 International Joint Conference on Neural Networks, 2001,* **4,** 2472 -2477.

Lisboa, P. J. G., Wong, H., Vellido, A., Kirby, S. P. J., Harris, P. and Swindell, R. (1998) Survival of breast cancer patients following surgery: a detailed assessment of the multi-layer perceptron and Cox's proportional hazard model, *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227). IEEE, 1998, 112-16 vol.1. New York, New York, USA.*

Loff, B. and Black, J. (1998) Human rights and epidemiology, *Lancet,* **352,** 1153.

London School of Economics & Political Science (2002) *Charles Booth Online Archive, http://booth.lse.ac.uk/* Web site of the Library of the London School of Economics & Political Science. Access 7/1/2002.

Lozowski, A., Miller, D. A. and Zurada, J. M. (1996) Dynamics of error backpropagation learning with pruning in the weight space, *Proceedings IEEE International Symposium on Circuits and Systems,* **3,** 449-452.

Lundin, J., Lundin, M., Holli, K., Kataja, V., Elomaa, L., Pylkanen, L., Turpeenniemi-Hujanen, T. and H., J. (2001) Omission of histologic grading from clinical decision making may result in overdose of adjuvant therapies in breast cancer: results from a nationwide study, *Journal of Clinical Oncology,* **19,** 28-36.

Lundin, M., Lundin, J., B Burke, H. B., Toikkanen, S., Pylkkanen, L. and Joensuu, H. (1999) Artificial neural networks applied to survival prediction in breast cancer, *Oncology,* **57,** 281-286.

MacKenzie, W. R., Hoxie, N. J., Proctor, M. E., Gradus, M. S., Blair, K. A., Peterson, D. E., Kazmierczak, J. J., Addiss, D. G., Fox, K. R., Rose, J. B. and Davis, J. P. (1994) A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply, *New England Journal of Medicine,* **331,** 161-7.

Marsh, J. W., Dvorchik, I. and Iwatsuki, S. (1998) Liver transplantation in the treatment of hepatocellular carcinoma, *Journal of Hepato-Biliary-Pancreatic Surgery,* 1998, 24-8.

Marsh, J. W., Dvorchik, I., Subotin, M., Balan, V., Rakela, V., Popechitelev, E. P., Subbotin, V., Casavilla, V., Carr, B. I., Fung, J. J. and Iwatsuki, S. (1997) The prediction of risk of recurrence and time to recurrence of hepatocellular carcinoma after orthotopic liver transplantation: A pilot study, *Hepatology,* **26,** 444-450.

Marzban, C. and Stumpf, G. J. (1998) *A neural network for tornado prediction based on Doppler Radar-derived attributes,* (Technical report, University of Oklahoma, USA).

Masters, T. (1993) *Practical neural network recipes in C++,* Academic Press, Boston.

Masters, T. (1995) *Neural, novel & hybrid algorithms for time series prediction,* John Wiley and Sons, New York.

Mattfeldt, T., Kestler, H. A., Hautmann, R. and Gottfried, H. W. (1999) Prediction of prostatic cancer progression after radical prostatectomy using artificial neural networks: a feasibility study, *BJU International,* **84,** 316-23.

McCulloch, W. S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics,* **5,** 115-33.

McGonigal, M. D., Cole, J., Schwab, C. W., Kauder, D. R., Rotondo, M. F. and Angood, P. B. (1993) A new approach to probability of survival scoring for trauma quality assurance, *Journal of Trauma-Injury Infection & Critical Care,* **34,** 863-8; discussion 868-70.

Minsky, M. L. and Papert, S. (1969) *Perceptrons,* MIT Press, Cambridge, Massachusetts, USA.

Mittal, G. S. and Zhang, J. (2001) Artificial neural network for the prediction of temperature, moisture and fat contents in meatballs during deep-fat frying, *International Journal of Food Science and Technology,* **36,** 489-97.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H. et al. (1990) Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma, *New England Journal of Medicine,* **322,** 352-8.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Tangen, C. M., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H. et al. (1995) Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report, *Annals of Internal Medicine,* **122,** 321-6.

Montie, J. E. and Wei, J. T. (2000) Artificial neural networks for prostate carcinoma risk assessment: An overview, *Cancer,* **88,** 2655-60.

Morgan, N. and Bourlard, H. (1990) Continuous speech recognition using multilayer perceptrons with hidden Markov models, *ICASSP-90, 1990 International Conference on Acoustics, Speech, and Signal Processing, 1990,* **1,** 413 -416.

Naguib, R. N. G., Robinson, M. C., Neal, D. E. and Hamdy, F. C. (1998) Neural network analysis of combined conventional and experimental prognostic markers in prostate cancer: a pilot study, *British Journal of Cancer,* **78,** 246-50.

National Institutes of Health (1990) NIH consensus conference. Adjuvant therapy for patients with colon and rectal cancer, *JAMA,* **264,** 1444-50.

Nattinger, A. B., McAuliffe, T. L. and Schapira, M. M. (1997) Generalizability of the surveillance, epidemiology, and end results registry population: factors relevant to epidemiologic and health care research, *Journal of Clinical Epidemiology,* **50,** 939-45.

Naus, J. I. (1965) The distribution of the size of the maximum cluster of points on a line, *American Statistical Association Journal,* 532-8.

Neelakantan, T. R., Brion, G. M. and Lingireddy, S. (2001) Neural network modelling of cryptosporidium and giardia concentrations in the Delaware river, USA, *Water Science and Technology,* **43.**

Niederberger, C. S. (1995) This month in Investigative Urology. Commentary on the use of neural networks in clinical urology, *Journal of Urology,* **153,** 1362.

Norman, G. and Streiner, D. (2000) *Biostatistics: the bare essentials,* B.C. Becker Inc., Hamilton.

O'Brien, S. J. and Christie, P. (1997) Do CuSums have a role in routine communicable disease surveillance?, *Public Health,* **111,** 255-258.

Ohno-Machado, L. (1997) A comparison of Cox proportional hazards and artificial neural network models for medical prognosis, *Computers in Biology and Medicine,* **27,** 55-65.

Ortiz, J., Ghefter, C. G., Silva, C. E. and Sabbatini, R. M. (1995) One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data, *Journal of the American College of Cardiology,* **26,** 1586 93.

Pacut, A. and Czajka, A. (2001) Recognition of human signatures, *Proceedings of IJCNN'01. International Joint Conference on Neural Networks. (Cat. No.01CH37222). IEEE, 2001, 1560-4 vol.2. Piscataway, New Jersey, USA.*

Page, E. S. (1954) Continuous inspection schemes, *Biometrika,* **41,** 100-115.

Page, E. S. (1961) Cumulative sum charts, *Technometrics,* **3,** 1-9.

Parker, D. B. (1985) *Learning logic: casting the cortex of the human brain in silicon. Technical Report TR-47,* Centre for Computational Research in Economics and Management, MIT, Cambridge, Massachusetts, USA.

Pascual, M., Rodó, X., Ellner, S. P., Colwell, R. and Bouma, M. J. (2000) Cholera dynamics and El Niño-Southern Oscillation, *Science,* **289,** 1766-1769.

Potosky, A. L., Riley, G. F., Lubitz, J. D., Mentnech, R. M. and Kessler, L. G. (1993) Potential for cancer related health services research using a linked Medicare-tumor registry database, *Medical Care,* **31,** 732-48.

Potter, S. R., Miller, M. C., Mangold, L. A., Jones, K. A., Epstein, J. I., Veltri, R. W. and Partin, A. W. (1999) Genetically engineered neural networks for predicting prostate cancer progression after radical prostatectomy, *Urology,* **54,** 791-5.

Qureshi, K. N., Naguib, R. N., Hamdy, F. C., Neal, D. E. and Mellon, J. K. (2000) Neural network analysis of clinicopathological and molecular markers in bladder cancer, *Journal of Urology,* **163,** 630-3.

Ragde, H., Elagamal, A-A. A., Snow, P. B., Brandt, J., Bartolucci, A. A., Nadir, B. S. and Korb, L. J. (1998) Ten-year disease free survival after transperineal sonography-guided iodine-125 brachytherapy with or without 45-gray external beam irradiation in the treatment of patients with clinically localized, low to high Gleason grade prostate carcinoma, *Cancer,* **83,** 989-1001.

Ravdin, P. M. and Clark, G. M. (1992) A practical application of neural network analysis for predicting outcome of individual breast cancer patients, *Breast Cancer Research and Treatment,* **22,** 285-93.

Rayens, M. K. and Kryscio, R. J. (1993) Properties of Tango's index for detecting clustering in time, *Statistics in Medicine,* **12,** 1813-1827.

Reed, R. (1993) Pruning algorithms - a survey, *IEEE Transactions on Neural Networks,* **4,** 740-7.

Remus, W. and O'Connor, M. (2001) Neural networks for time series forecasting, In *Principles of forecasting: A handbook for researchers and practitioners.* (Ed, Armstrong, J. S.) Klewer Academic Publishers, Norwell, Massachusetts.

Riedmiller, M. (1994a) Advanced supervised learning in multi-layer perceptrons - from backpropagation to adaptive learning algorithms, *Computer Standards and Interfaces,* **5**.

Riedmiller, M. (1994b) *Rprop - Description and implementation details: technical report, January 1994.* Institut für Logik, Komplexität und Deduktionssyteme, University of Karlsruhe, Germany.

Riedmiller, M. and Braun, H. (1993) A Direct adaptive method for faster backpropagation learning: The RPROP algorithm, *Proceedings of the IEEE International Conference on Neural Networks, 1993.*

Rietveld, S., Oud, M. and Dooijes, E. H. (1999) Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural networks, *Computers and Biomedical Research,* **32,** 440-8.

Rosenblatt, F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review,* **65,** 386-408.

Rothman, K. J., Adami, H-O. and Trichopoulos, D. (1998) Should the mission of epidemiology include the eradication of poverty?, *Lancet,* **352,** 810-13.

Rumelhart, D. E. (1986) Learning representations by back-propagating errors, *Nature,* **323,** 533-6.

Rumelhart, D. E., Widrow, B. and Lehr, M. A. (1994) The basic ideas in neural networks, *Communications of the ACM,* **37,** 87-92.

Sackett, D. L., Haynes, R. B., Guyatt, G. H. and Tugwell, P. (1991) In *Clinical Epidemiology: a basic science for clinical medicine,* Little, Brown and Company, Boston, 117-9.

Sargent, D. J. (2001) Comparison of artificial neural networks with other statistical approaches, *Cancer,* **91,** 1636-42.

Sarle, W. S. (1994a) Neural Network Implementation in SAS software, *Proceedings of the Nineteenth Annual SAS Users Group International Conference, 1994.*

Sarle, W. S. (1994b) Neural networks and statistical models, *Proceedings of the Nineteenth Annual SAS Users Group International Conference, 1994.*

SAS Institute Inc. (1999), SAS v8.2 (Computer program) SAS Institute Inc, Cary North Carolina USA.

Schwartz, J., Levin, R. and Goldstein, R. (2000) Drinking water turbidity and gastrointestinal illness in the elderly of Philadelphia, *Journal of Epidemiology and Community Health,* **54,** 45-51.

Schwartz, J., Levin, R. and Hodge, K. (1997) Drinking water turbidity and pediatric hospital use for gastrointestinal illness in Philadelphia, *Epidemiology,* **8,** 615-20.

Shamma, S. A. (1987) *Neural networks for speech processing and recognition SRC TR 87-114.* Technical report, Department of Electrical Engineering and Systems Research center, University of Maryland, USA.

Siemiatycki, J. (1996) Exposure assessment in community based studies of occupational cancer, *Occupational Hygiene,* **3,** 41-58.

Siemiatycki, J., Day, N. E., Fabry, J. and Cooper, J. A. (1981) Discovering carcinogens in the occupational environment: a novel epidemiologic approach, *Journal of the National Cancer Institute,* **66,** 217-25.

Sinha, M., Kennedy, C. S. and Ramundo, M. L. (2001) Artificial neural network predicts CT scan abnormalities in pediatric patients with closed head injury, *Journal of Trauma-Injury Infection & Critical Care,* **50,** 308-12.

Smith, K., Palaniswami, M. and Krishnamoorthy, M. (1996) A hybrid neural approach to combinatorial optimization, *Computer and Operations Research,* **23,** 597-610.

Smith, K. A. (1999a) *Introduction to neural networks and data mining for business applications,* Eruditions, Emerald.

Smith, K. A. (1999b) Neural networks for combinatorial optimization: A review of more than a decade of research, *INFORMS Journal on Computing,* **11,** 15-34.

Smith, K. A. and Gupta, J. N. D. (2000) Neural networks in business: techniques and applications for the operations researcher, *Computer and Operations Research,* **27,** 1203-1044.

Snow, P. B., Kerr, D. J., Brandt, J. M. and Rodvold, D. M. (2001) Neural network and regression predictions of 5-year survival after colon carcinoma treatment, *Cancer,* **91,** 1673-8.

SPSS Inc. (1999), SPSS for Windows v10.0.5 (Computer program) SPSS Inc., Chicago. Illinois.

Stern, L. and Lightfoot, D. (1999) Automated outbreak detection: a quantitative retrospective analysis, *Epidemiology and Infection,* **122,** 103-110.

Stewart, P. A., Stewart, W. F., Siemiatycki, J., Heineman, E. F. and Dosemeci, M. (1998) Questionnaires for collecting detailed occupational information for community-based case control studies, *American Industrial Hygiene Association Journal,* **59,** 39-44.

Stroup, D. F., Thacker, S. B. and Herndon, J. L. (1988) Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962-1983, *Statistics in Medicine,* **7,** 1045-1059.

Sundararajan, V., Grann, V. R., Jacobson, J. S., Ahsan, H. and Neugut, A. I. (2001) Variations in the use of adjuvant chemotherapy for node-positive colon cancer in the elderly: a population-based study, *Cancer,* **7,** 213-8.

Sundararajan, V., Mitra, N., Jacobson, J. S., Grann, V. R., Heitjan, D. F. and Neugut, A. I. (in press) Survival associated with 5-Fluorouracil-based adjuvant chemotherapy among elderly patients with node-positive colon cancer, *Annals of Internal Medicine.*

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems, *Science,* **240,** 1285-93.

Tango, T. (1984) The detection of disease clustering in time, *Biometrics,* **40,** 15-26.

Taylor, J. G. (1997) Neural computation: the historical background, In *Handbook of Neural Computation* Oxford University Press, Oxford, UK.

Tillett, H. E. and Spencer, I. L. (1982) Influenza surveillance in England and Wales using routine statistics. Development of 'cusum' graphs to compare 12 previous winters and to monitor the 1980/81 winter, *Journal of Hygiene,* **88,** 83-94.

Tourassi, G. S., Floyd, C. E. and Coleman, E. (1996) Improved noninvasive diagnosis of acute pulmonary embolism with optimally selected clinical and chest radiographic findings, *Academic Radiology,* **3,** 1012-18.

Troni, G. M., Cipparrone, I., Cariaggi, M. P., Ciatto, S., Miccinesi, G., Zappa, M. and Confortini, M. (2000) Detection of false-negative Pap smears using the PAPNET system, *Tumori,* **86,** 455-457.

Van Camp, L. A. and Delooz, H. H. (1998) Current trauma scoring systems and their applications, *European Journal of Emergency Medicine,* **5,** 341-53.

Venayagamoorthy, G. K., Moonasar, V. and Sandrasegaran, K. (1998) Voice recognition using neural networks, *Proceedings of the 1998 South African*

Symposium on Communications and Signal Processing-COMSIG '98 (Cat. No. 98EX214). IEEE. 1998, 29-32.

Wallace, V. P., Bamber, J. C., Crawford, D. C., Ott, R. J. and Mortimer, P. S. (2000) Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminant analysis and artificial neural networks, *Physics in Medicine & Biology,* **45,** 2859-71.

Wallenstein, S. (1980) A test for the detection of clustering over time, *American Journal of Epidemiology,* **111,** 367-372.

Wallenstein, S., Gould, M. S. and Kleinman, M. (1989) Use of the scan statistic to detect time-space clustering, *American Journal of Epidemiology,* **130,** 1057-64.

Wallenstein, S., Naus, N. and Glaz, J. (1993) Power of the scan statistic for detection of clustering, *Statistics in Medicine,* **12,** 1829-1843.

Ward Systems Group Inc (2000),Ward Systems Group Inc, Frederick Maryland.

Warner, B. and Misra, M. (1996) Understanding neural networks as statistical tools, *American Statistician,* **50,** 284-293.

Watier, L. and Richardson, S. (1991) A time series construction of an alert threshold with application to *S. Bovismorbificans* in France, *Statistics in Medicine,* **10,** 1493-1509.

Wong, B. K., Bodnovich, T. A. and Selvi, Y. (1997) Neural network applications in business: A review and analysis of the literature (1988-95), *Decision Support Systems,* **19,** 301-320.

World Health Organization (1998) Measles: Progress towards global control and regional elimination, 1990-1998, *Weekly Epidemiological Record,* **73,** 389-394.

Wu, B. (1995) Model-free forecasting for nonlinear time series (with application to exchange rates), *Computational Statistics and Data Analysis,* **19,** 433-459.

Wu, C. H. (1997) Artificial neural networks for molecular sequence analysis, *Computers and Chemistry,* **21,** 237-56.

Wu, C. H., Zhao, S. and Chen, H-L. (1996) A protein class database organized with ProSite protein groups and PIR superfamilies, *Journal of Computational Biology,* **3,** 547-61.

Wyatt, J. (1995) Nervous about artificial neural networks? [letter; comment], *Lancet,* **346,** 1175-7.

Wyatt, J. and Spiegelhalter, D. (1990) Evaluating medical expert systems: what to test and how?, *Medical Informatics,* **15,** 205-17.

Wyatt, J. C. and Altman, D. G. (1995) Commentary: Prognostic models: clinically useful or quickly forgotten?, *British Medical Journal,* **311,** 1539-1541.

Zernikow, B., Holtmannspoetter, K., Michel, E., Pielemeier, W., Hornschuh, F., Westermann, A. and Hennecke, K. H. (1998) Artificial neural network for risk assessment in preterm neonates, *Archives of Disease in Childhood Fetal and Neonatal Edition,* **79,** F129-34.

Zhang, G., Patuwo, B. E. and Hu, M. Y. (1998) Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting,* **14,** 35-62.

Zhang, X. and Thearling, K. (1994) *Non-linear time-series prediction by system by systematic data exploration on a massively parallel computer*, Santa Fe Institute Technical report 94-07-045. Santa Fe Institute, New Mexico, USA.

Ziada, A. M., Lisle, T. C., Snow, P. B., Levine, R. F., Miller, G. and Crawford, E. D. (2001) Impact of different variables on the outcome of patients with clinically confined prostate carcinoma, *Cancer,* **91,** 1653-60.

Appendix One: Computer tools for graphical presentation and analysis of time-series data.

Background

For time-series studies (such as those presented in Chapters Four, Five and Six) where there are variations both in time and by geographical location, it can be difficult for the investigator to clearly see the inherent patterns in the data sets using existing computer tools. A program is needed that displays both the geographical and temporal dimensions at the same time.

Similarly, a tool that presents the network's current predictions alongside the evolving time series would be useful both for the assessment of new models and the deployment of network models in the field.

Neither of these tools is currently available, nor are they easy to produce with current statistical or graphics software. To remedy this I created two new computer programs, MapMovie and StepGraph. Both programs are included in the CD inside the back cover of this thesis, together with relevant raw data sets and network outputs from the studies presented in Chapters Three, Four and Five. This appendix explains how they work, and how to install and use the software on the disk.

Using the CD

ThesisDemo.exe runs automatically

As soon as the CD is inserted into the appropriate drive, a small program called ThesisDemo.exe runs automatically. (It may take a minute or more to get going on an older computer. If it does not start at all it can be run directly from the CD using the Windows Explorer or the Windows Run menu command.) ThesisDemo is a shell program making it easy to install the other two programs and then use them to view the relevant data sets. To exit ThesisDemo, press the Escape key, or click on the button labelled 'Close' in the bottom right hand corner.

The data and the ThesisDemo program itself are not installed to the hard disk. To re-start the ThesisDemo program, you can remove the CD from the drive and re-insert it, or else use the Windows Start menu 'Run' option to run the program called ThesisDemo.exe from the CD's root directory. Or else find that file and run it using the Windows Explorer or My Computer programs. If you wish to run ThesisDemo often, you can make a shortcut to it on the desktop – although the CD must be inserted each time for it to run.

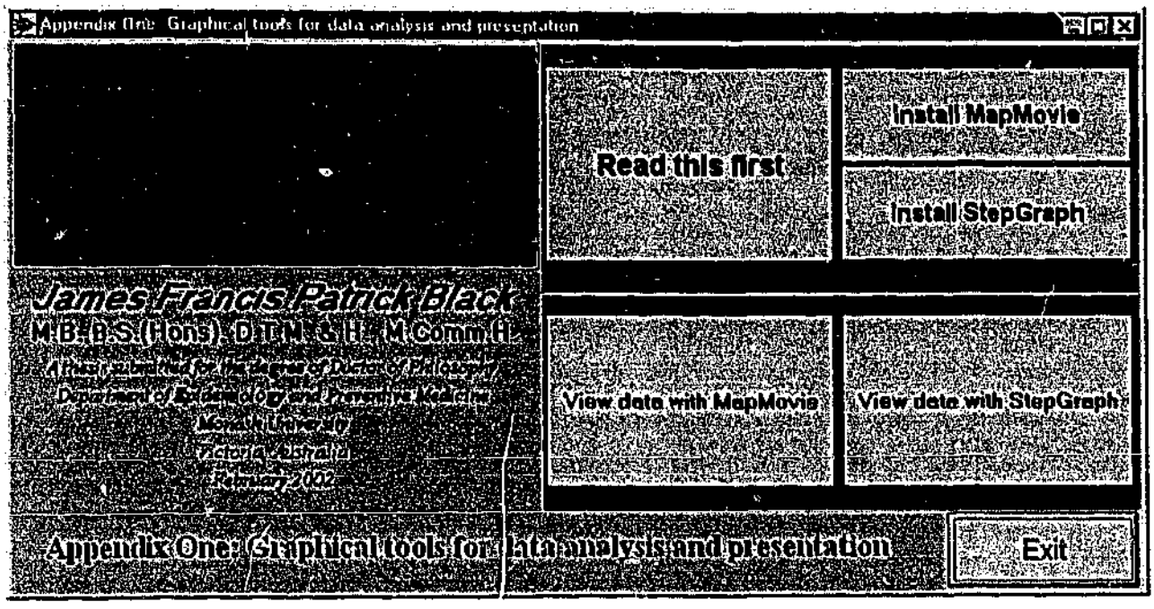


Figure A-1: The ThesisDemo main screen

The button labelled 'Read this first' opens a memo file with information about the use of the CD and the two programs. It also warns that there is no on-line help file for either program, and directs the user to these instructions.

Installing MapMovie and StepGraph

MapMovie and StepGraph each have their own separate installation programs. Install them one at a time by clicking on the relevant buttons in the upper right-hand corner of the main ThesisDemo screen. Follow the on-screen instructions to complete the installation, choosing a suitable location on the hard disk for the program files. Wait for the first installation to complete before beginning the next.

As part of the installation of each program the necessary parts of the Borland Database Engine will also be automatically installed to the hard disk. The installation programs will also make some changes to the Windows Registry.

Uninstalling MapMovie and StepGraph

Both MapMovie and StepGraph can be uninstalled at any time using the Windows Control Panel. Open the Windows *Start Menu*, then *Settings*, then *Control Panel*. Double-click the *Add/Remove Programs* icon, and find the program you wish to uninstall on the list.

Viewing the data sets

To view geographical and temporal distribution of disease cases with MapMovie, click the button labelled 'View data with MapMovie'. To view the evolution of cases and network predictions for faecal analysis data sets in Melbourne, click the button labelled 'View data with StepGraph'. A new small window will appear, with a button for each of the available data sets. Click on a button to open a new instance of MapMovie or StepGraph featuring that data set. You can open as many different data sets at the same time as your computer's memory will allow.

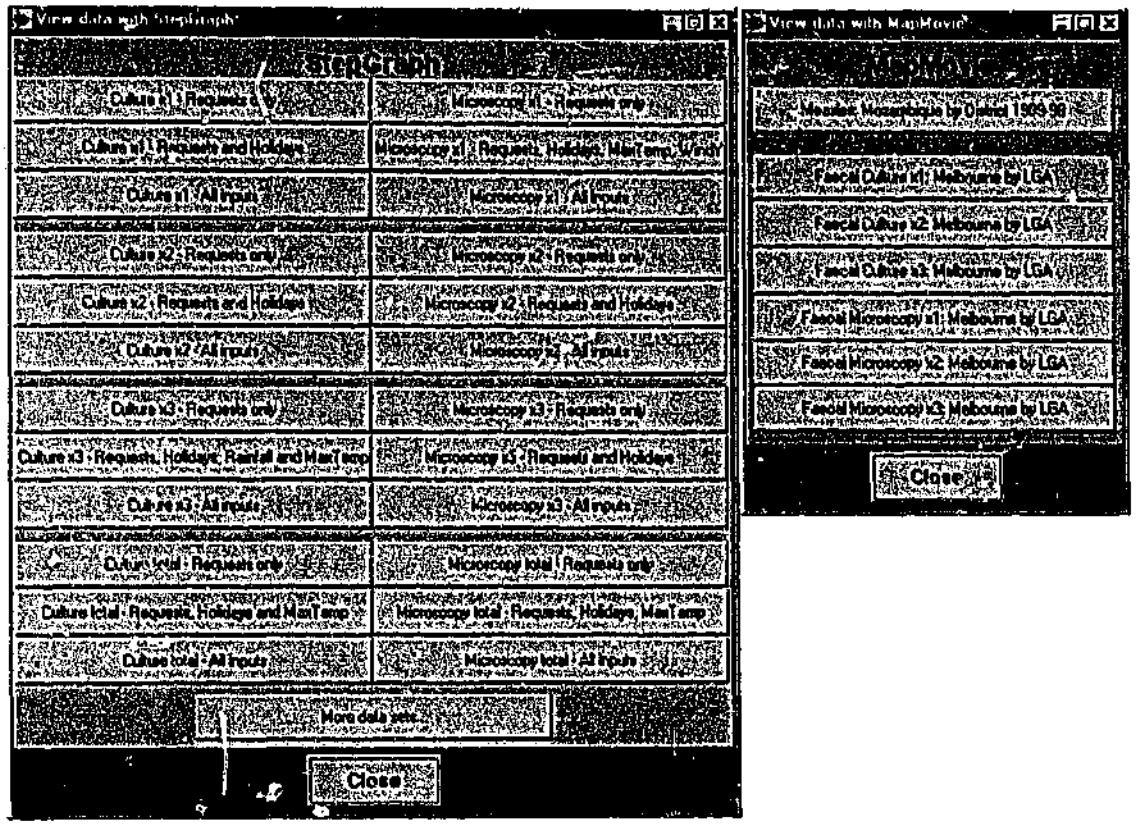


Figure A-2: The data viewing menu windows.

To view a data set not on the list, click the button marked 'More data sets' and choose the file using the dialog box. All the data files are in the CD, in folders called x:\StepGraph\Data and x:\MapMovie\Data (where x is the name of the CD drive).

MapMovie

MapMovie is a rapid scanning tool that allows the user to view the distribution of disease numbers or rates by geographical location as well as over time. The program steps through a suitably prepared time series database one time period per step, and re-draws a map of the disease distribution for each time point. The user can step forward or backward individually through the maps, search for a particular time point or disease pattern, or else run the maps in sequence automatically (creating the 'map movie' effect). The maps can be made using blue shading to indicate the number of cases in each map region, or else using a scale of deciles. (Blue shading was chosen to make it easy for colour-blind users to see the differences.) MapMovie can also create graphs of the time series for any of the individual areas.

Using MapMovie

Running MapMovie from the ThesisDemo menu opens the relevant data set directly, and MapMovie begins by creating a map for the first point in the database. A separate control panel also appears, allowing the user to step through the maps or create a moving series.

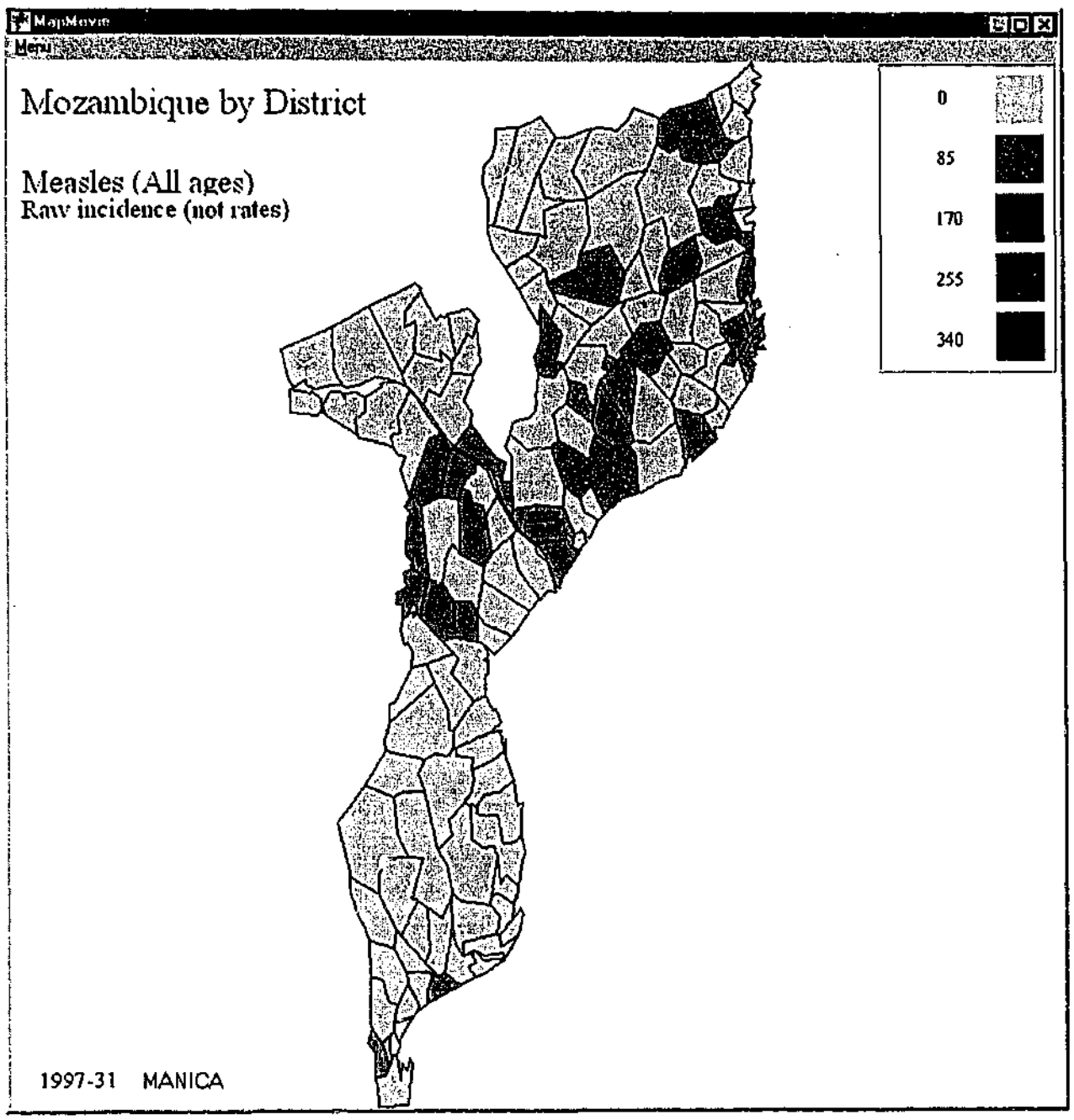


Figure A-3: The MapMovie main screen.

Running MapMovie without any command-line parameters (i.e. away from the ThesisDemo menu system) first opens a file-selection dialog box. Select a file with the extension MCF from anywhere on the computer, and MapMovie will open the appropriate data files and draw a map for the first point in the time series.

The MapMovie control panel

The 'Position in data set' slider bar at the top of the control panel indicates the current position in the time series. By dragging the slider button with the left mouse button

you can search through the database for a particular time point, watching the maps change as you do.

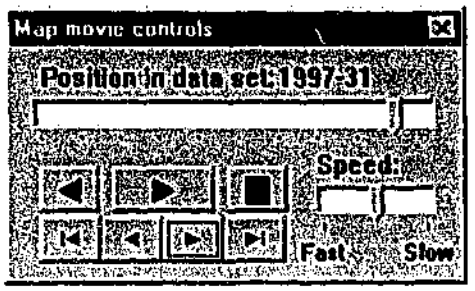


Figure A-4: The MapMovie control panel

The top layer of three buttons labelled with triangles and a square control the 'movie' mode, allowing the user to start a movie in forward or reverse order, or to stop a movie at any point. The 'Speed' slider bar on the bottom right regulates the time interval between successive images in 'movie' mode. Drag the slider button to the right to slow the rate (down to one frame per second) or to the left to increase the rate. The maximum speed depends on the memory and processor of the computer.

The bottom layer of buttons marked with triangles control the 'single step' mode, allowing the user to step forward or backward through the database by single steps, or to jump directly to the beginning or end.

The main MapMovie menu

MapMovie has a main menu that allows the user to open the MapMovie control panel, hide or show the legend and titles, print the map, make the map window smaller or larger, and create graphs for a specific map region. The menu also allows the user to map by deciles of the data rather than by blue shading.

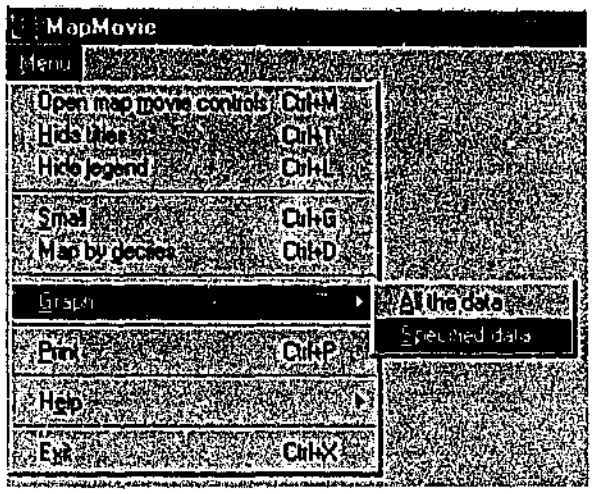


Figure A-5: The MapMovie main menu

The MapMovie context menus

Provided the monitor is set to 'True colour', whenever the cursor lies over a map region the region name will appear in the bottom left hand corner of the main MapMovie screen. A context menu will then be available if you click over the region with the right mouse button.

Figure A-6: Context menu available when the cursor is over a map region.

From this context menu you can create a graph for this region, either displaying all the data, or a selected number of time series points centred around the current point.

The MapMovie graph menu

MapMovie graphs appear in a separate window, with its own menu. With the graph menu the user can choose to display the graph in a large or small window, print it, and show or hide the current map point.

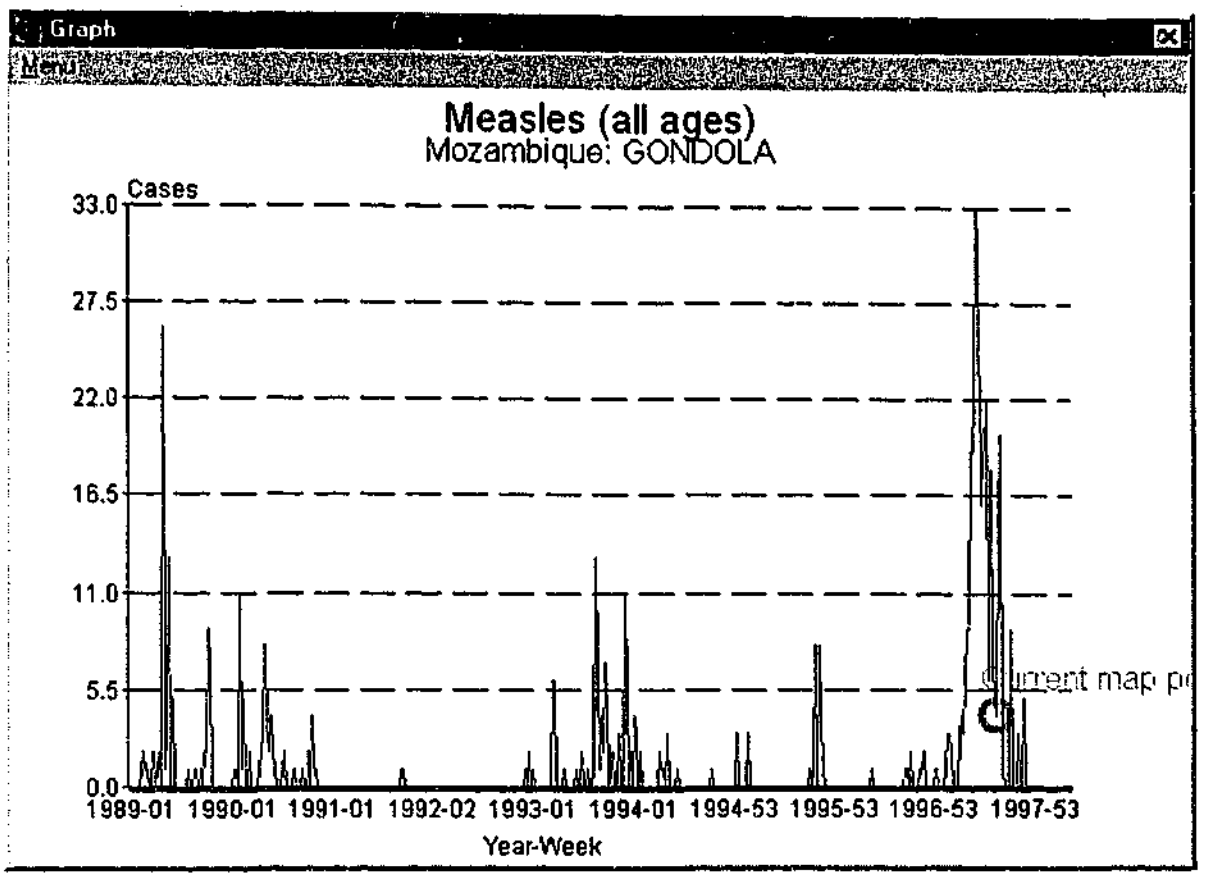


Figure A-7: The MapMovie graph screen

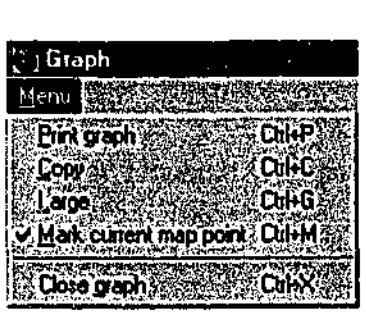


Figure A-8: The MapMovie graph menu

MapMovie is a 'beta' release

Although MapMovie has been successfully installed on a number of different computers running Windows 95, 98 and NT 4.0, it is best thought of as a beta test version. It is not a 'polished' program, and there are a number of known issues, which will be addressed in later versions:

1) The map does not automatically re-draw itself when the MapMovie window gets focus. For example, if you make and then close a graph, the portion of the map that

was covered by the graph does not reappear. Click to move forward or backwards one frame and the map will be redrawn.

2) To copy a map or graph to another document, click on the main map window, then press Alt-PrintScreen. The map (in fact the whole window) will be copied to the Windows clipboard. To use it in a document or a graphics program, just use the 'Paste' function of the other software to draw it in as a bitmap image.

3) If the monitor is not set to 'True colour' (or is not capable of showing it), then moving the mouse pointer over the map will not accurately tell the name of the indicated region. Sometimes it will not say anything, and sometimes it will give wrong names. In that case, you can still create graphs for a single region by choosing 'Graphs' on the main menu.

MapMovie file structure

The file structure of MapMovie was inspired by EpiMap, a freeware mapping program developed by the Centers for Disease Control and Prevention and the World Health Organization. Like EpiMap, MapMovie uses three separate files for each map: one with the extension MCF, one with JMP, and one in the standard dBase IV DBF format. (The example files 'Example.jmp', 'Example.dbf' and 'Example.mcf' files installed in the program files directory can be used as templates to create new maps and data files. The JMP and MCF files are plain ASCII text, so they can be modified in any text editor.)

The JMP file contains the map boundaries, and is re-used for any map for the same region. This is a format specific to MapMovie, although conversion of existing maps from EpiMap BND format is possible. The DBF file has one field for each region in

the JMP file, plus one field to hold the time variable. The MCF file contains the information needed to link the data and boundary files, plus formatting and title information.

StepGraph

StepGraph illustrates the way neural network time series predictions might be used in practice. The y-axis represents the case numbers and the x-axis the time points.

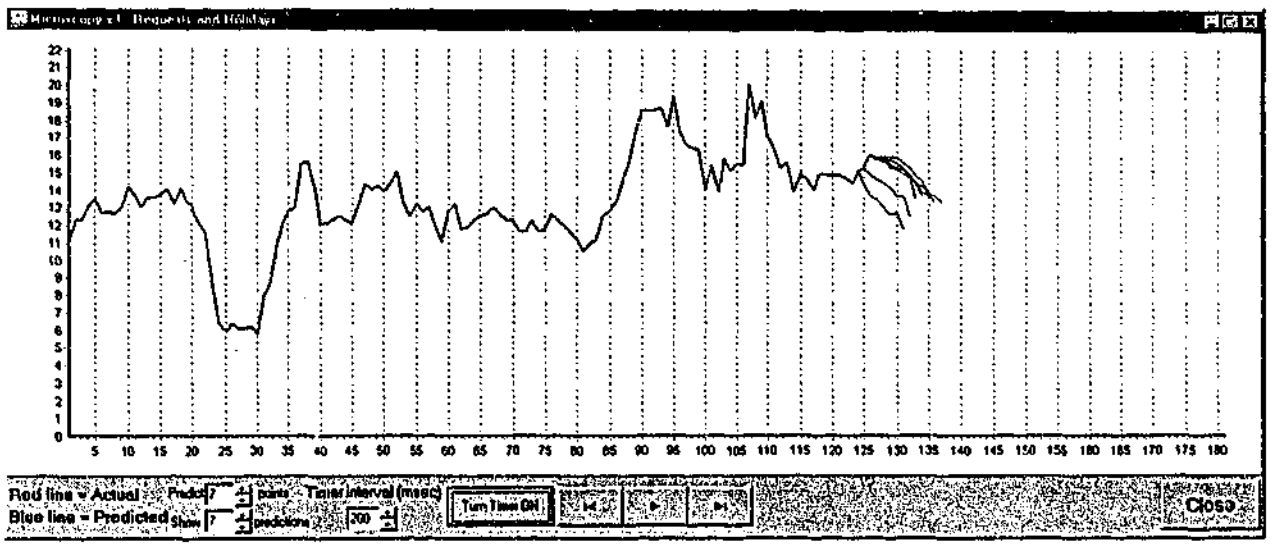


Figure A-9: The StepGraph main screen

As the program moves through the time series a red line indicates the actual number of cases, while blue lines indicate the network's predictions at each time point.

Using StepGraph

If the StepGraph program is run without any command line parameters, the user is prompted to choose a database file to open. Alternatively, the name of a suitable (dBase IV DBF format) file on the command line opens that file automatically when StepGraph opens.

In the bottom centre of the StepGraph screen are three buttons labelled with triangles. These allow the user to jump to the beginning of the data set, or the end, or step forward one time step per click.

The button labelled 'Turn timer ON' switches StepGraph into 'timer' mode, and it automatically steps through the data from that point. The button label changes to 'Turn Timer OFF', and can be used to stop the timer at any point.

The user can specify how many time point predictions are displayed (1 to 7), using the edit box and spinner labelled 'Predict ... points'. The number of sets of predictions retained on the screen can also be altered using the edit box and spinner labelled 'Show ... predictions'.

The timer interval in timer mode can be changed using the edit box and spinner labelled 'Timer interval (msec)'. Within the computer's capabilities, the time interval between steps will be the value of this edit box in milliseconds – the higher the number, the slower the steps.

Close StepGraph by pressing the Escape key, or clicking the button labelled 'Close'.

StepGraph file structure

Each StepGraph requires only one data file, in dBase IV DBF format. The first field contains the actual case numbers, and the other fields the network predictions in order. The file itself must be in the correct time series order.

StepGraph is a beta release

Like MapMovie, StepGraph should run under any Windows 32-bit operating system. It has been tested on several computers running Windows 95, 98 and NT 4.0, and there are no known 'bugs', but it is not a polished program, and is best regarded as a 'beta' version.

Copyright

The data pertaining to requests for faecal analysis i. Melbourne are based on data owned by the Australian Health Insurance Commission, and the copyright remains with that organisation.

The data pertaining to measles in Mozambique remain the property of the Mozambique Ministry of Health, who retain the copyright to these data.

Technical specifications

All the programs on the CD were created using the Borland C++ Builder Professional, versions 2 to 5. (Inprise Corporation, 2000). They were developed in the C++ programming language, using the Borland standard Visual Component Library, the Borland Database Engine and the additional TChart component, all of which are

licensed for royalty-free distribution. The programs are optimised for Pentium processors, and compiled to run under any Windows 32-bit operating system (they have been tested on computers running Windows 95, 98 and NT, but they should also run successfully under later Windows versions). Installation to the user's computer is done using InstallShield Express version 2.12 (InstallShield Software Corporation, 1999).

Appendix Two: Computing resources used

Hardware

The neural network modelling and most of the other computing for this thesis was done on a networked workstation with a single 733 MHz Intel Pentium III processor 256 MB of Random Access Memory and a 20 GB hard disk.

Software

The operating system was Windows NT workstation version 4.0 (Build 1381, Service Pack 6) (Microsoft Corporation, Redmond, Washington, USA).

All the neural network models in this thesis were constructed and trained using NeuroShell 2 Release 4.0 (Ward Systems Group, Frederick, Maryland, USA).

All areas under receiver operating characteristics curves were calculated using SPSS for Windows Release 10.0.7 (27 Nov 1999) Standard Version. (SPSS Inc, Chicago, Illinois, USA)

The Cox proportional hazards models were fitted using SAS Version 8.2 (SAS Institute Inc., Cary North Carolina, USA).

Spreadsheet and word-processing functions were performed with Microsoft Office 97 SR-2. (Microsoft Corporation, Redmond, Washington, USA).