Topics in Interpolation and Smoothing of Spatial Data

by

Lindsay McNeill

Thesis Presented for the Degree of
Doctor of Philosophy
in the Department of Statistical Sciences.
UNIVERSITY OF CAPE TOWN
June 1994

Abstract

Topics in Interpolation and Smoothing of Spatial Data, by Lindsay McNeill.
Thesis presented for the degree of Doctor of Philosophy in Statistical Sciences,
University of Cape Town, June 1994.

This thesis addresses a number of special topics in spatial interpolation and smoothing. The motivation for the thesis comes from two projects, one being to extend the availability of a daily rainfall model for southern Africa to sites at which little or no rainfall data is available, using data from nearby sites, and the other arising from a need to improve the species abundance estimates used to produce maps for the Southern African Bird Atlas Project in areas where the original presence/absence data is sparse.

Although problems of spatial interpolation and smoothing have been the subject of much research in recent years, leading to the development of the specialised discipline of geostatistics, these two problems have features which render the available methodology inappropriate in certain respects.

The semi-variogram plays a central role in geostatistical work. In both of the applications considered here, the raw semi-variogram is 'contaminated' by error, but the error variance varies widely between data points, so that the spatial autocorrelation structure of the underlying error-free variable is blurred. An adjusted semi-variogram, which removes the effect of the measurement error, is defined and incorporated into the kriging equations.

A number of measures have been proposed for kriging in the presence of trend, ranging from explicit modelling of a deterministic trend function to 'moving window' kriging, which assumes local stationarity as an approximation. The former approach is often inappropriate over large non-homogenous regions, while the latter approach tends to underestimate the kriging variance. As an alternative strategy it is proposed here that the trend function be considered as another random variable, with a long-range spatial autocorrelation. This approach is simple to implement, and can also be used as a basis for filtering the data to separate trend from local or high-frequency variation.

The daily rainfall model is based on a Fourier series representation giving rise to amplitude and phase parameters; the latter are circular in nature, and not amenable to analysis by standard techniques. This thesis describes a method of interpolation and smoothing, analogous to kriging, which is appropriate for unit vector data available at a number of spatial locations.

The cumulated values of species counts in the SABAP are essentially binomially distributed and thus again specialised techniques are required for interpolation. New geosta-

tistical methods which cater for both binomial and Poisson data are presented.

Another problem arises from the need to improve interpolated values of the rainfall model parameters by incorporating information on altitude. Although a number of approaches are possible, for example, using co-kriging or kriging with external drift, difficulties are caused by the complexity of the relationship between the rainfall at a point and the surrounding topography. This problem is overcome by the use of orthogonal functions of altitude to model the patterns of topography.

Acknowledgements

A number of people contributed to the successful completion of this thesis, and to all of them I extend my grateful appreciation.

I am especially indebted to my supervisor, Walter Zucchini who suggested the major topic for the thesis and without whose unfailing enthusiasm the thesis would never have been completed.

The Water Research Commission provided financial support for the rainfall modelling project. I would also like to acknowledge the help and support of my co-authors of the WRC report: Alison Joubert, who obtained all the data from the Computing Centre for Water Research, did much of the data checking and organised many of the computer runs, and Anabela Brandão who did the programming for the bootstrap estimates.

Les Underhill introduced me to the SABAP project, and also provided much-needed encouragement at various stages. James Harrison provided access to the SABAP data.

During the course of my thesis work I supervised the projects of a number of UCT honours students in topics related to the thesis: Mark Harrison tested the use of the maximum likelihood approach for covariance modelling, Silas Sedupane investigated the modelling of cross-covariances of rain and altitude in a number of regions, and Neville Verlander assisted with model validation for the SABAP project.

I would also like to thank the Computing Centre for Water Research for assistance in obtaining rainfall and altitude data and for the use of their computer facilities for the production runs.

All computer programs used for the data analysis and mapping were written in FORTRAN by myself.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mapping of environmental data has become increasingly important in recent years in application areas as diverse as mining exploration, meteorology and epidemiology. In some cases only small data sets are available, in others, particularly where satellite imagery is used, the data sets may be very large; in either case there is inevitably a demand to make maximum use of the available data to provide detailed and comprehensive coverage of the region of interest.

Features which are common to many problems involving spatial data include irregular spacing of data points and auto-correlation of neighbouring values. Several techniques have been developed in recent years to handle the estimation and smoothing of spatial variables with such characteristics, kriging and smoothing splines being perhaps the most notable of these. Although these tools are flexible and powerful, many applications have special features which pose new challenges, and this thesis addresses two such applications.

The first of the problems is to extend the estimation of the parameters of a daily rainfall model, originally developed by Zucchini and Adamson (1984a), to sites in southern Africa where insufficient or no rainfall data are available, so that the model parameters must be estimated using values from nearby sites. Once calibrated, the model can be used to generate long

artificial rainfall sequences (typically 1000-2000 years) which preserve all the statistical properties of daily rainfall; not merely the means and variances, but also the frequency of occurrence of any sequence of values. By averaging over such a generated sequence one can estimate derived statistics relating to any specific characteristic of daily rainfall, for example, the probability of receiving a minimum of x mm of rain during a specified period, or the percentage of annual rainfall attributable to storms. In addition, the artificial rainfall sequences may be used as input to other models such as crop yield, soil moisture or rainfall-runoff models.

In order to be of general use, the model must somehow be calibrated throughout the country; however, to estimate the model parameters directly from the historical rainfall record at a given site Zucchini and Adamson (1984a) found that a minimum of 30 years of daily rainfall data was needed. In some parts of southern Africa locations with the required amount of data are few and far between, as can be seen from Figure 1.1. A research project, funded by the Water Research Commission, was therefore undertaken to estimate the model parameters on a grid of 1 minute of a degree of latitude and longitude throughout the region by using spatial interpolation. Apart from allowing the use of the model at some half a million individual grid points, such an extension also allows the production of maps of derived parameters such as that given in Figure 1.2.

One special feature of this problem, from a point of view of the selection of an appropriate interpolation method, is that the model parameters at the available stations are not known accurately, but are estimated from the daily rainfall data, and the accuracy of the resultant parameters is thus largely dependent on the length of the rainfall record, which in turn varies markedly between stations. Thus what is really required here is in fact not a method of interpolation but a method of smoothing which can take into account the varying accuracy of the data points. This problem is addressed in Chapter 6.

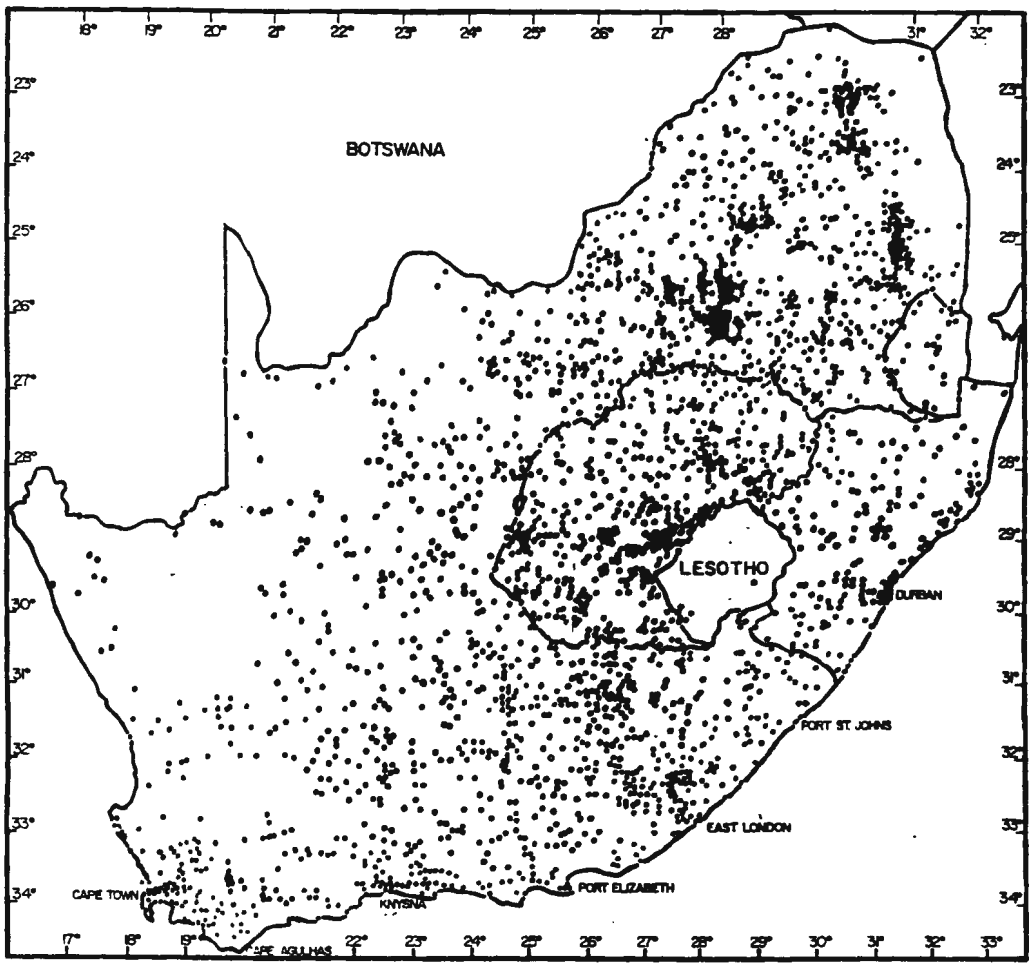


Figure 1.1: Rainfall stations used by Zucchini and Adamson.

The daily rainfall model, which is described in detail in Chapter 2, makes use of a truncated Fourier series representation, so that the resultant parameters to be estimated consist of both amplitude and phase parameters. The phase parameters of the model are circular in nature, and standard techniques for spatial interpolation cannot be applied to such data. Chapter 8 presents a new technique for the interpolation and smoothing of circular data.

Topography plays an important part in the local variation in rainfall,

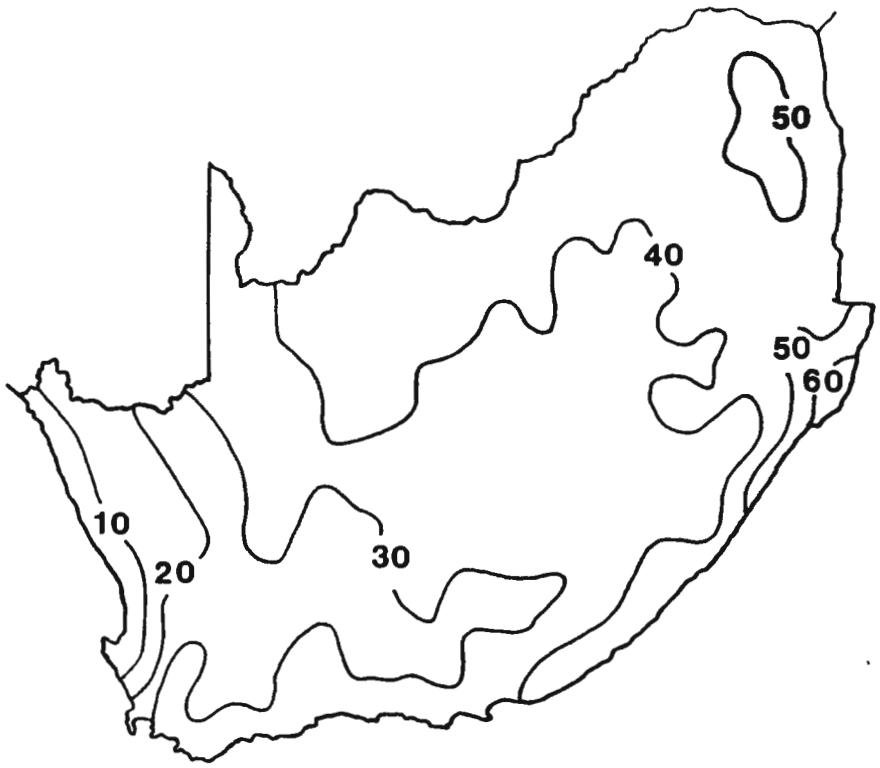


Figure 1.2: Mean annual percentage of rainfall attributable to storms.

and thus estimation of the parameters of the daily rainfall model should be improved by the incorporation of topographic information. The exact relationship between rain and topography is complex and localised, and previous attempts to study the relationship have been limited to very small areas. Chapter 7 considers approaches to this problem and attempts to find a methodology which can realistically be applied on a large scale.

The second problem which is considered in this thesis is that of producing smoothed maps of the spatial distribution of individual bird species for the Southern African Bird Atlas Project. This major atlassing project (Harrison, 1987 and 1992) has involved some 7500 data collectors, mainly volunteers,

over a five-year data collection period ending in 1992; and during this period over seven million individual sightings have been entered into the databank. The major aim of the project is to produce detailed distribution maps for each of the 900 or so species of birds occurring in southern Africa, with additional information on factors such as seasonal variation of distribution and breeding periods. Apart from its value to ornithologists, the resulting atlas will provide a basis for monitoring long-term changes in abundance, for the planning of conservation strategies based on species richness and diversity, and for the study of the relationship of avian distribution patterns with other environmental variables such as rainfall and human land use.

The data for the project are in the form of presence/absence records for individual species in each month and each quarter degree grid square, so that the resulting cumulated abundance measure for each species is in the form of a reporting rate which may reasonably be modelled by a binomial distribution. One result of the use of volunteers for data collection is that most of the data are concentrated around urban areas, so that, as with the daily rainfall model discussed above, the data are clustered in certain areas and very sparse in others. Thus the underlying parameters of the binomial model are accurately estimated in some grid squares and very poorly estimated in others, so that there is a need to improve these estimates by some form of spatial smoothing which is adapted to the binomial nature of the data. Chapter 9 of this thesis presents extensions to standard kriging techniques to cater for binomial and Poisson data.

There is a considerable degree of overlap, from a statistical point of view, in the two problems described above. Both deal with irregularly spaced auto-correlated spatial data subject to measurement error, the variance of which varies widely from point to point. Thus after describing the daily rainfall model and the Southern African Bird Atlas Project in some detail in Chapters 2 and 3 respectively, those aspects which are common to both

problems are considered, deferring special features of the individual problems to later chapters. In Chapter 4 methods of interpolation and smoothing for spatial data are reviewed and the advantages and disadvantages of each are discussed briefly. The special requirements of the two projects described above are discussed, and reasons are given as to why a geostatistical approach was selected for their solution.

An overview of the geostatistical technique of kriging, as used for spatial interpolation, is given in Chapter 5, together with a discussion of the methods of co-kriging and kriging with external drift. There is no difficulty in principle in extending kriging to the case where the data contain measurement error, although in practice there are very few published examples of kriging used for smoothing. In Chapter 6 the appropriate kriging equations for this situation are given and methods of estimation of the relevant components of the spatial covariance or semi-variogram are discussed, with specific reference to the case where the error variance is not constant across the data locations. An alternative approach to the problem of kriging in the presence of large-scale fluctuations is proposed. This approach is simpler to implement than other methods presently in use for modelling long-range trend. It is also demonstrated that kriging can provide an appropriate technique for non-parametric trend estimation for irregularly-spaced data with autocorrelated residuals. These ideas are illustrated on the amplitude parameters of the daily rainfall model.

The incorporation of topographical data into the estimation of the parameters of the daily rainfall model is considered in Chapter 7. Although either the method of co-kriging or kriging with external drift should in theory be appropriate for this task, difficulties arise due to the complex nature of the relationship between topography and rainfall, which shows significant directional and regional affects. Computational difficulties are also encountered as a result of the number of altitude values needed to describe the relevant

patterns of topography in sufficient detail. These problems are overcome by the use of orthogonal functions of altitude to summarise the topographical patterns in the neighbourhood of any given point.

Chapter 8 addresses the problem of estimation of the phase parameters of the daily rainfall model, which are circular in nature. Existing methods of interpolation and smoothing for such data are reviewed, and it is shown that none of the existing methods is entirely suitable for our purpose. An extension of the kriging methodology is therefore derived to cater for such data and is shown by means of a test data set to be more effective than a simple distance-weighted average method for this data set.

Chapter 9 considers the extension of the semi-variogram and kriging equations to binomial and Poisson data, and illustrates the use of this methodology for the smoothing of the bird atlas reporting rates.

Although the techniques described in Chapters 8 and 9 for smoothing circular, binomial and Poisson data are introduced here in the specific contexts of the daily rainfall model and the Southern African Bird Atlas Project, they are clearly of wider applicability, and further applications are discussed briefly in Chapter 10.

Chapter 2

The Daily Rainfall Model

Zucchini and Adamson (1984a) fitted a model of daily rainfall to some 2550 sites throughout South Africa, Lesotho and Swaziland. Briefly, the model provides conditional probabilities of rain (conditioned on the previous day's rain) for each day of the year, together with the mean and variance of rainfall amounts on wet days, again for each day of the year. Such a model is particularly versatile in that it allows one to generate very long sequences of simulated daily rainfall amounts, which in turn enables one to calculate any statistics of interest for daily, weekly, monthly or annual time periods. The probabilities of relatively complex events are easily calculated, for example, the probability that, between 15 October and 31 December, there will be at least 200 mm of rain, and that there will be no 10-day run having less than 5 mm. Derived values which can be defined in terms of daily rainfall, such as indices of probability of drought, can also be generated. The model thus provides a research tool which has found wide usage among water resource planners, agriculturalists and other research workers. Several examples of the use of the model are given in Zucchini, Adamson and McNeill (1991 and 1992).

2.1 Description of the Model

In the Zucchini and Adamson model the process of daily rainfall is described by means of two main components; the first component describes the occurrence of wet and dry days and the second component describes the distribution of rainfall amounts on wet days. Such a model has been used by other researchers in other parts of the world, for example, Gabriel and Neumann (1962). Woolhiser (1992) gives a recent review.

It is known that the probability of rainfall on a particular day varies with the season in a smooth fashion, and, in addition, wet days tend to occur in clusters owing to the movement of weather-generating systems (Caskey, 1963; Woolhiser and Pegram, 1979). Thus the occurrence of rain can be modelled using a seasonal Markov chain, and Zucchini and Adamson found that a first-order Markov chain, that is a process with a memory of one day, provided a good fit to data at sites throughout southern Africa. The first step in fitting the model at a given site is therefore to use all the available daily rainfall records to estimate, for each day of the year, t, the probability of a wet day on day t, given that day $t-1$ was wet, and the probability of a wet day on day t, given that day $t-1$ was dry; these are denoted as $\pi_{W|W}(t)$ and $\pi_{W|D}(t)$ respectively. The maximum likelihood estimates of these probabilities are simply the relevant (conditional) frequencies. Thus, for example,

$$\hat{\pi}_{W|W}(t) = N_{W|W}(t)/N_W(t-1) \qquad t = 1, 2, ...365$$

where

$\pi_{W|W}(t)$ = probability that day t is wet, given day $t-1$ is wet

$N_{W|W}(t)$ = number of years when it rained on day $t-1$ and day t

$N_W(t-1)$ = number of years when it rained on day $t-1$

Unless a very long historical record is available, of hundreds, perhaps

thousands, of years, these estimators provide very poor estimates of the required probabilities. The estimates of $\pi_{W|W}$ are particularly difficult to estimate accurately in a relatively dry climate like South Africa's since the denominator in the equation above will generally be very much smaller than the number of years of data. Thus the sampling variance is large and the seasonal cycle of probabilities, although fairly marked at most sites (Figure 2.1), is far from smooth, whereas it is reasonable to assume that the actual probabilities for any two consecutive days should be very similar. Furthermore, probability estimates of zero can occur for certain days in the year, which makes no physical sense. Finally, since $\pi_{W|W}(t)$ and $\pi_{W|D}(t)$ need to be estimated for each value of t, a model based on these empirical probabilities would require $2 \times 365 = 730$ parameters at each rainfall station. Parsimony and the required smooth estimates of the transition probabilities were achieved by using a truncated Fourier series representation for the logits of the daily probability estimates. The logit transformation was used in order to avoid the possibility of the model predicting probabilities outside the permissible [0, 1] range. The number of parameters required in the Fourier series approximation was determined to be five, using model selection criteria of Linhart and Zucchini (1986). Thus we have

$$\lambda_{W|W}(t) = \alpha_0 + \sum_{i=1}^{2} \alpha_i cos(2\pi i(t - 1 - \beta_i)/365) \qquad (2.1)$$

where

$$\lambda_{W|W}(t) = ln(\pi_{W|W}(t)/(1 - \pi_{W|W}(t)) \qquad t = 1, 2, ...365$$

and a similar Fourier series approximation holds for $\lambda_{W|D}(t)$, the logit of the probability of a wet day following a dry day. Thus each of the 365 daily probabilities is replaced by the three amplitude parameters α_0, α_1, α_2, and the two phase parameters β_1, β_2.

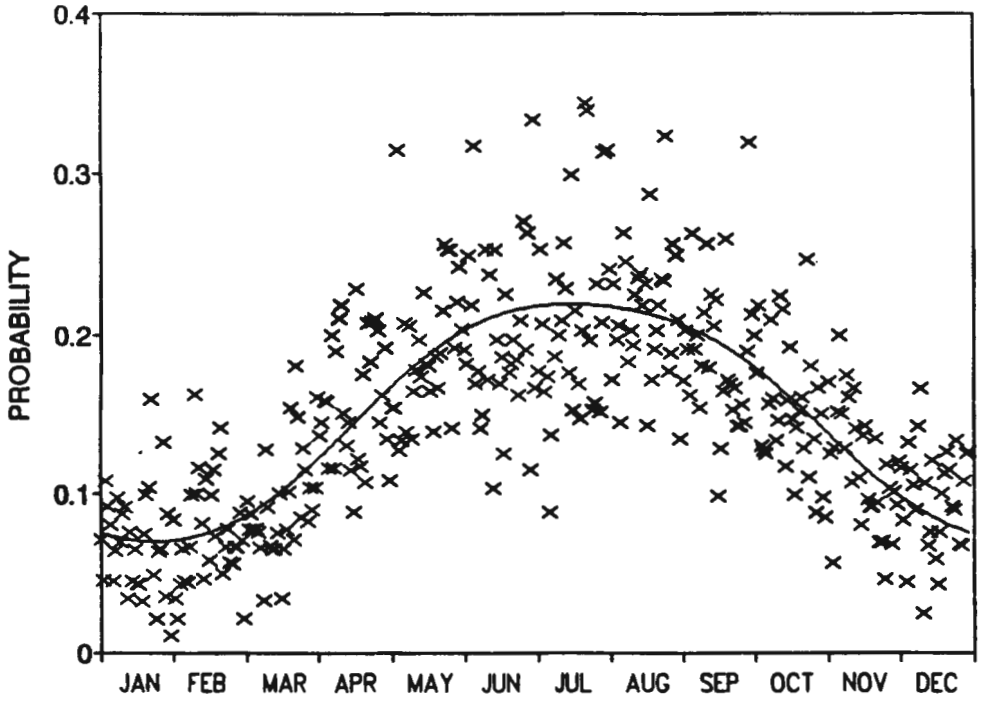


Figure 2.1: Empirical probabilities and Fourier series model for $\pi_{W|D}(t)$ at Stellenbosch.

The distribution of rainfall depths on rainy days also exhibits a distinct seasonal variation: both the average depth and the variance of the depth vary with the time of year. However the coefficient of variation was found by Zucchini and Adamson to be approximately constant over the year at any site. They therefore used a truncated Fourier series, also with two harmonics (five parameters), to approximate the seasonal variation in mean rainfall depths, and this, together with the constant coefficient of variation of the depths, allows one to fit any two-parameter family of univariate distributions to the rainfall depths at any site, with a mean and standard deviation which change in a smooth way throughout the year. The Weibull distribution was found to be the best choice for a sample of locations which were selected to be

representative of the full range of rainfall regimes in southern Africa, although the flexibility of the model allows the user to fit different distributions, for example the log-normal or gamma, at different sites, or at different times of year at one site where this may be appropriate.

Thus the model for rainfall depths on wet days is given by

$$\mu(t) = \alpha_0 + \sum_{i=1}^{2} \alpha_i cos(2\pi i(t - 1 - \beta_i)/365) \qquad (2.2)$$

$$CV(t) = CV \qquad t = 1, 2, ...365$$

where

$\mu(t)$ = mean depth on day t, given that day t was wet

$CV(t)$ = coefficient of variation of depth.

There are thus a total of sixteen model parameters for each site: three amplitudes and two phases for the seasonal mean depth and for the logits of each of the two probabilities, $\pi_{W|W}(t)$ and $\pi_{W|D}(t)$, plus the coefficient of variation of the depths. Methods and explicit algorithms for estimating the parameters, together with the estimates for 2550 rainfall stations across southern Africa, are given in Zucchini and Adamson (1984a and 1984b). Table 2.1 lists the parameters, together with a mnemonic for each parameter; these mnemonics are used for convenience in later chapters of this thesis.

2.2 Interpretation of the Parameters

Figure 2.1 shows the daily estimates and fitted Fourier series for the probability of a wet day following a dry day at Stellenbosch near Cape Town, based on a 104-year rainfall record. The amplitude parameter α_0 of equation 2.1 can be interpreted as a measure of the annual average of the logit of the probability of rain, while α_1 and α_2 relate to the amplitude of the seasonal variation; the phase parameters are indicators of the time of year of

WWA0	Zero'th amplitude:	Prob$(W_t \mid W_{t-1})$
WWA1	First amplitude:	Prob$(W_t \mid W_{t-1})$
WWA2	Second amplitude:	Prob$(W_t \mid W_{t-1})$
WWP1	First phase:	Prob$(W_t \mid W_{t-1})$
WWP2	Second phase:	Prob$(W_t \mid W_{t-1})$
DWA0	Zero'th amplitude:	Prob$(W_t \mid D_{t-1})$
DWA1	First amplitude:	Prob$(W_t \mid D_{t-1})$
DWA2	Second amplitude:	Prob$(W_t \mid D_{t-1})$
DWP1	First phase:	Prob$(W_t \mid D_{t-1})$
DWP2	Second phase:	Prob$(W_t \mid D_{t-1})$
DEPA0	Zero'th amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPA1	First amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPA2	Second amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPP1	First phase:	Mean depth on wet days $(mm \times 10^{-1})$
DEPP2	Second phase:	Mean depth on wet days $(mm \times 10^{-1})$
CV	Coefficient of Variation:	Depth on wet days

Table 2.1: List of symbols: daily rainfall model.

peaks in the rain probabilities. Figure 2.2 shows the individual harmonics of the curve shown in Figure 2.1 which clarifies this interpretation. A similar interpretation holds for the parameters of the Fourier series for the rainfall depths given in equation 2.2. For most areas of southern Africa, the seasonal patterns are reasonably well represented by a Fourier series with a single harmonic, so that α_2 is small and β_2 is not very well defined.

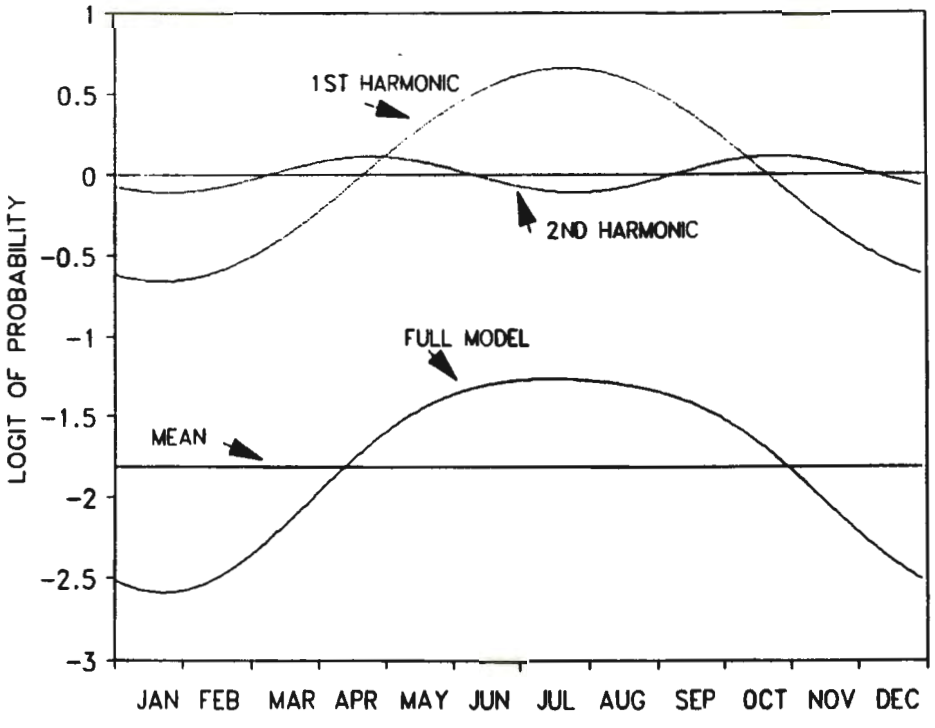


Figure 2.2: Individual harmonics of Fourier series model for $\pi_{W|D}(t)$ at Stellenbosch.

It is possible to use an alternative form of the Fourier series model based on sine and cosine terms instead of the amplitude and phase representation described above; the amplitude-phase representation was selected for this project because these parameters have a clearer physical interpretation

which should make it easier to relate the parameters to other factors such as topography.

2.3 Spatial Aspects of the Model

Zucchini and Adamson (1984a) originally fitted their model to some 2550 rainfall stations (Figure 1.1), being all stations at which a sufficiently long record of daily rainfall (30 years) was available. It can be seen from Figure 1.1 that the stations are clustered around major urban centres, and are very sparse in some rural areas, notably in Lesotho and in the north-west. A consequence of this distribution is that the locations are thus biased towards lower lying areas, rather than mountainous areas, which in turn could lead to a downward bias in areal rainfall estimates. In order to extend the applicability of the model, there is a need to provide estimated model parameters at a fairly dense grid of locations giving an even coverage throughout southern Africa.

Figure 2.3 shows averages of the model parameters calculated for each S.A.Weather Bureau block (half-degree grid square) . These maps indicate that there are fairly smooth large-scale trends, but also some significant local variation. In general one would not expect sharp changes in the phases, which reflect the time of year of highest rainfall, but it is likely that the amplitudes will be influenced by topography and could thus change rapidly over short distances in mountainous areas.

2.4 Estimation Error

The daily rainfall data used by Zucchini and Adamson were taken from the data base of the Computing Centre for Water Research (CCWR). These data originate from several sources, including the South African Weather

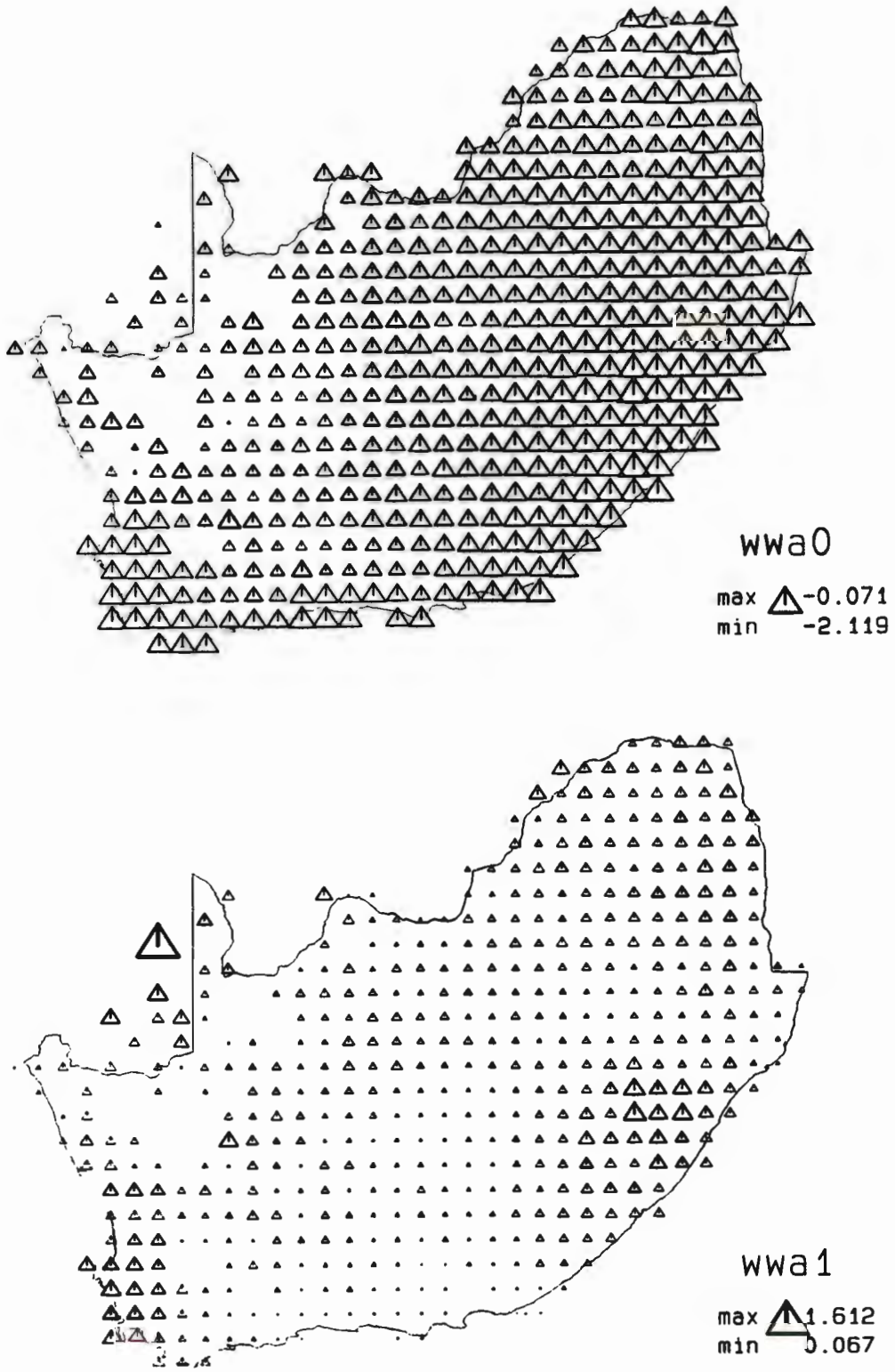


Figure 2.3: SA Weather Bureau block means.

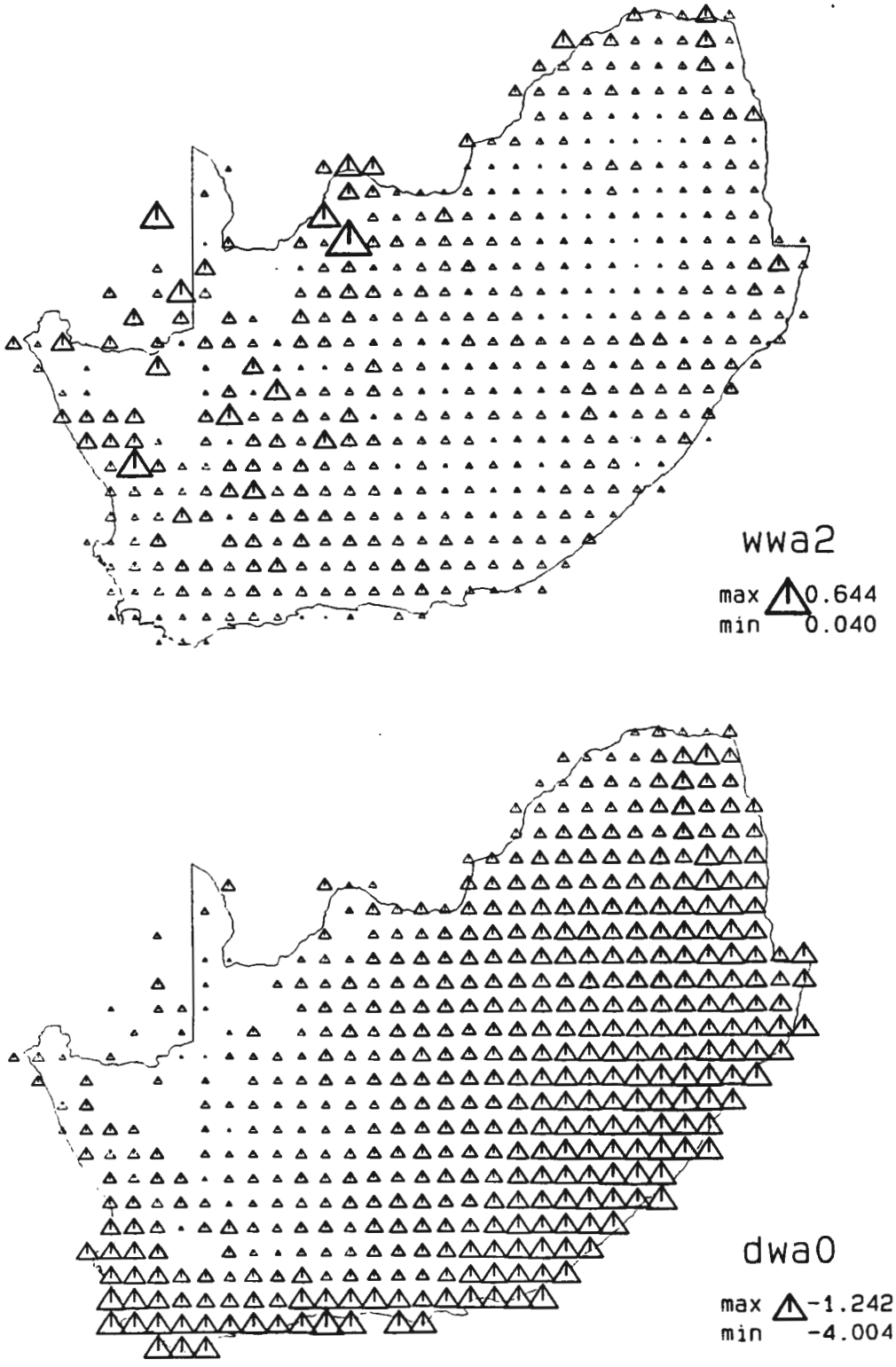


Figure 2.3: SA Weather Bureau block means (contd.).

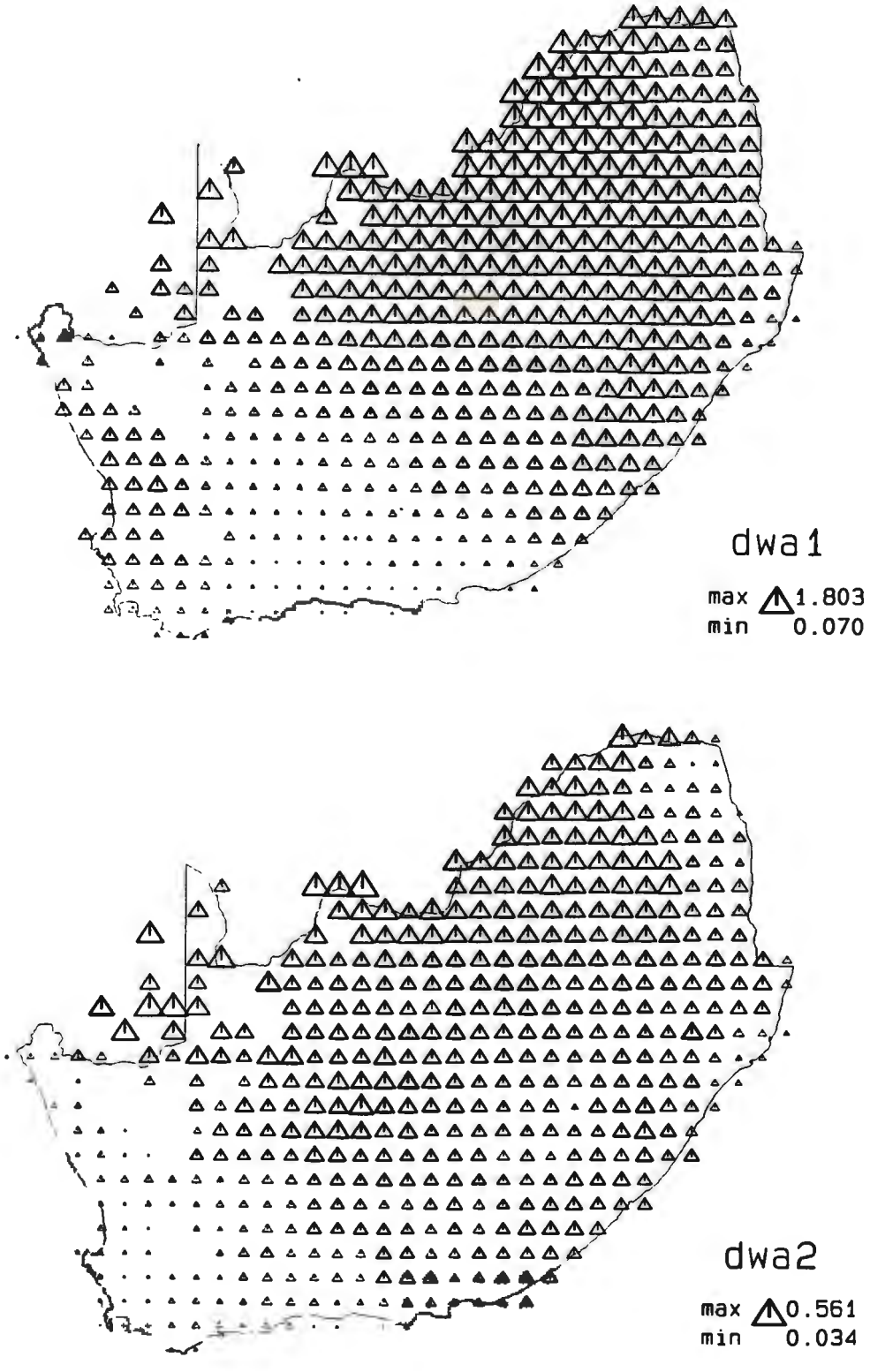


Figure 2.3: SA Weather Bureau block means (contd.).

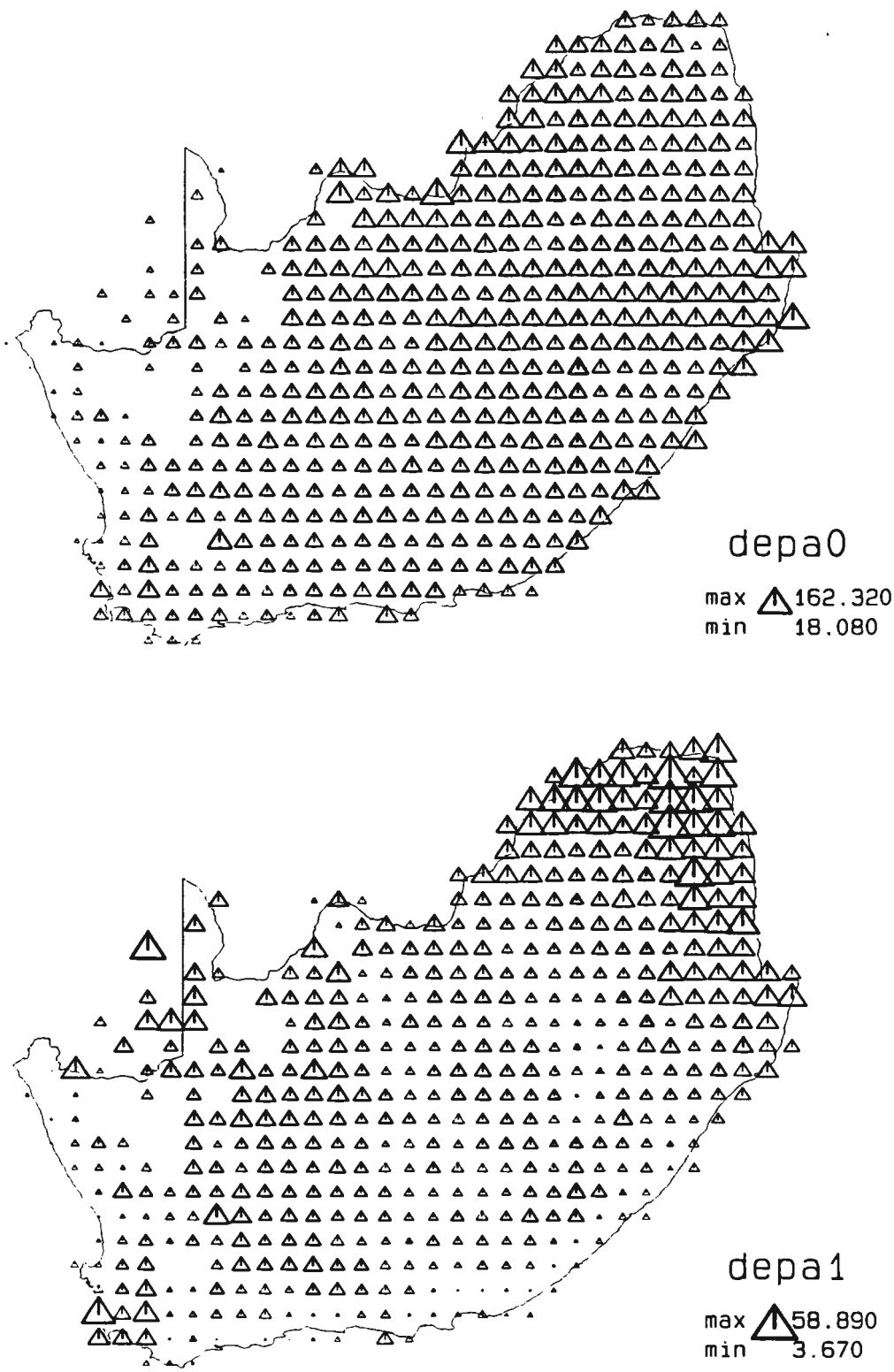


Figure 2.3: SA Weather Bureau block means (contd.).

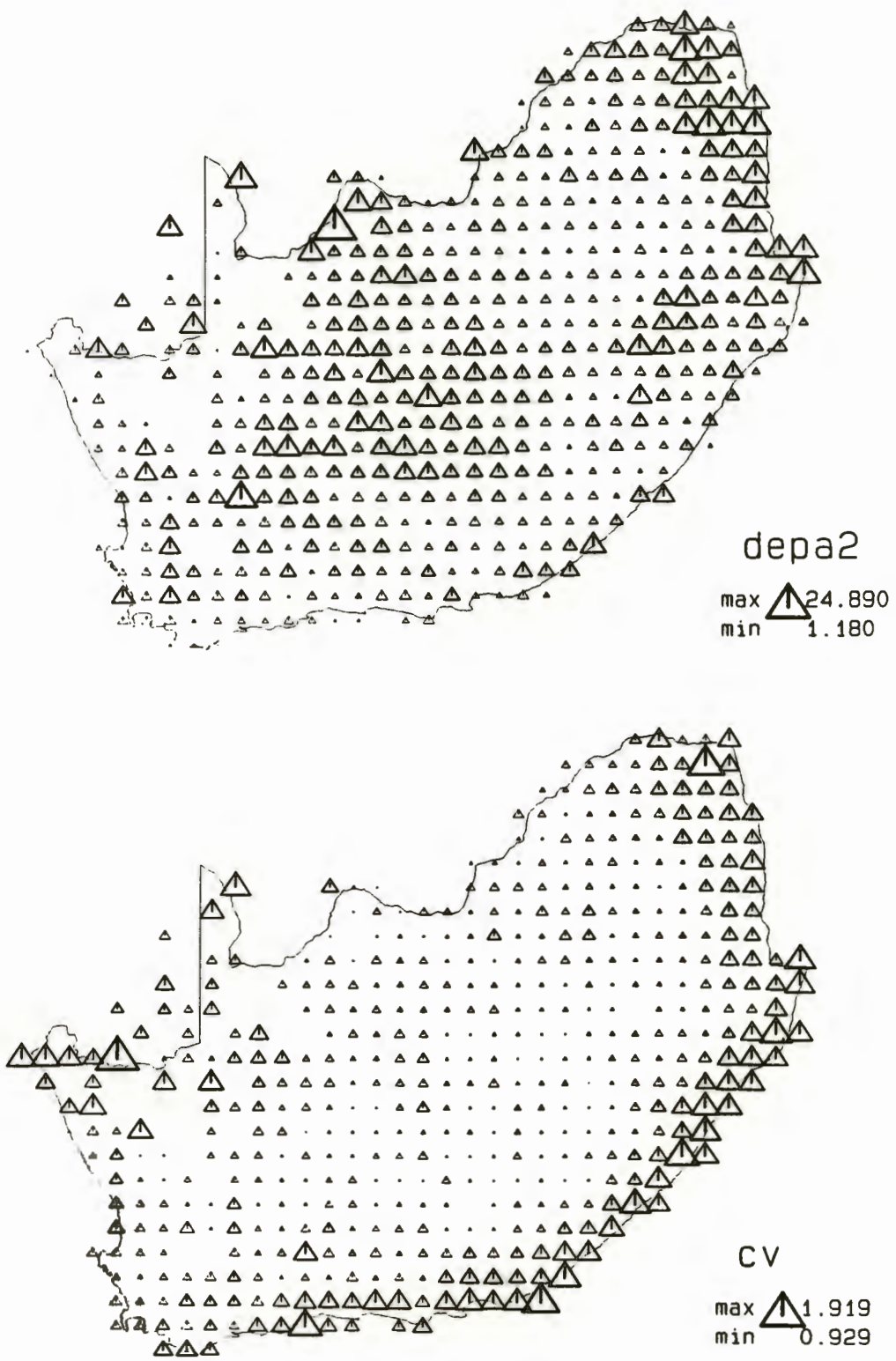


Figure 2.3: SA Weather Bureau block means (contd.).

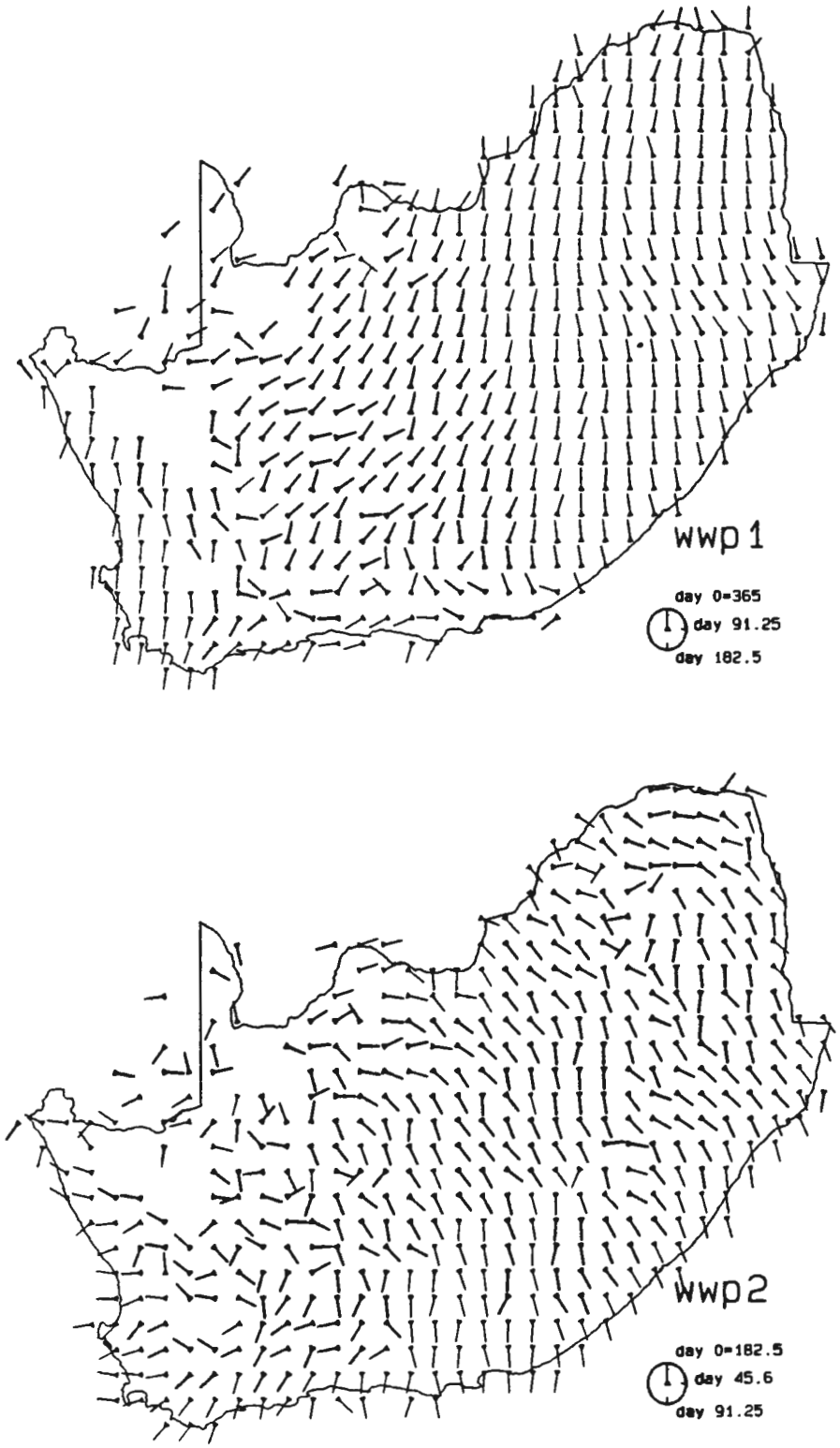


Figure 2.3: SA Weather Bureau block means (contd.).

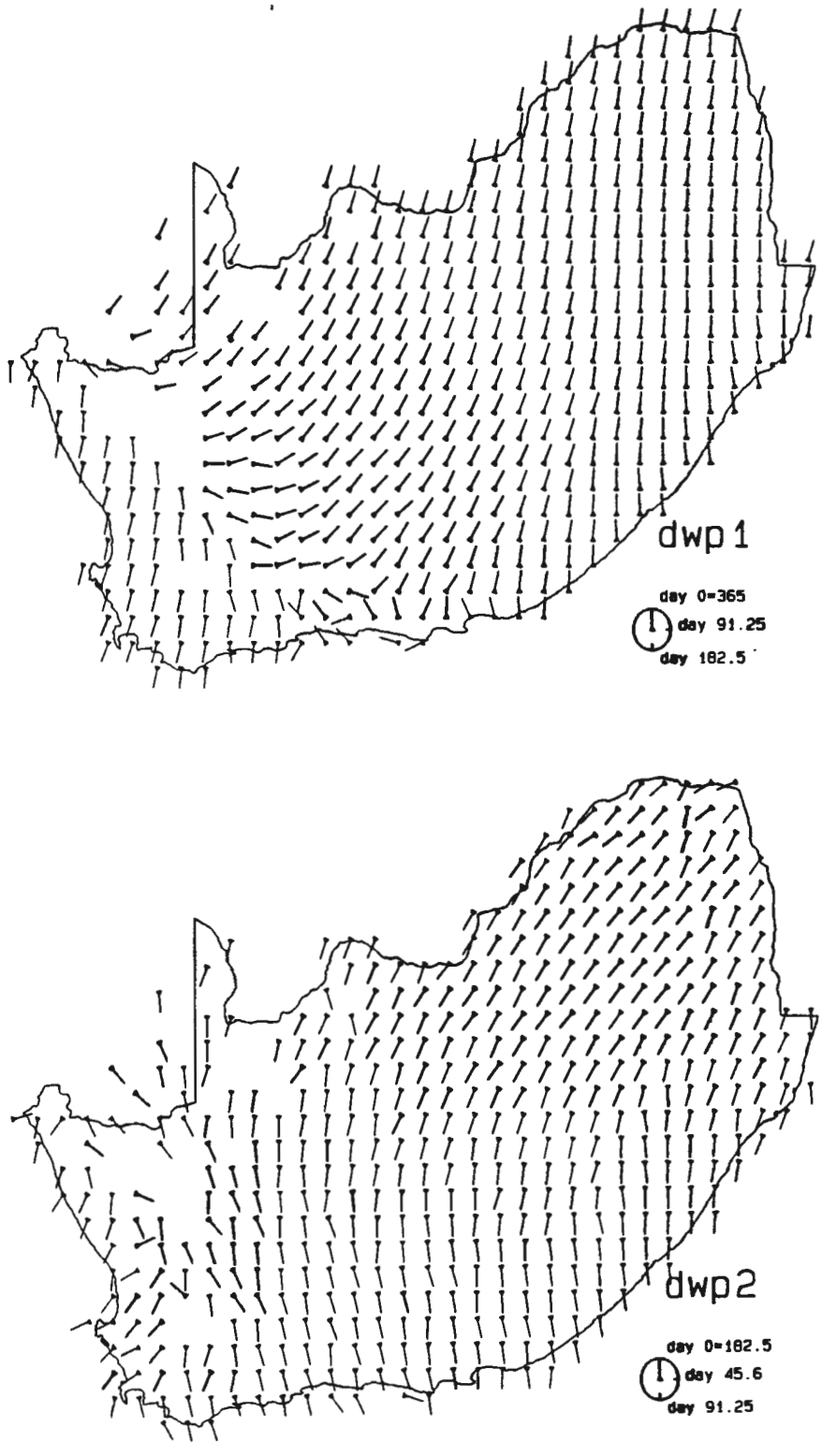


Figure 2.3: SA Weather Bureau block means (contd.).

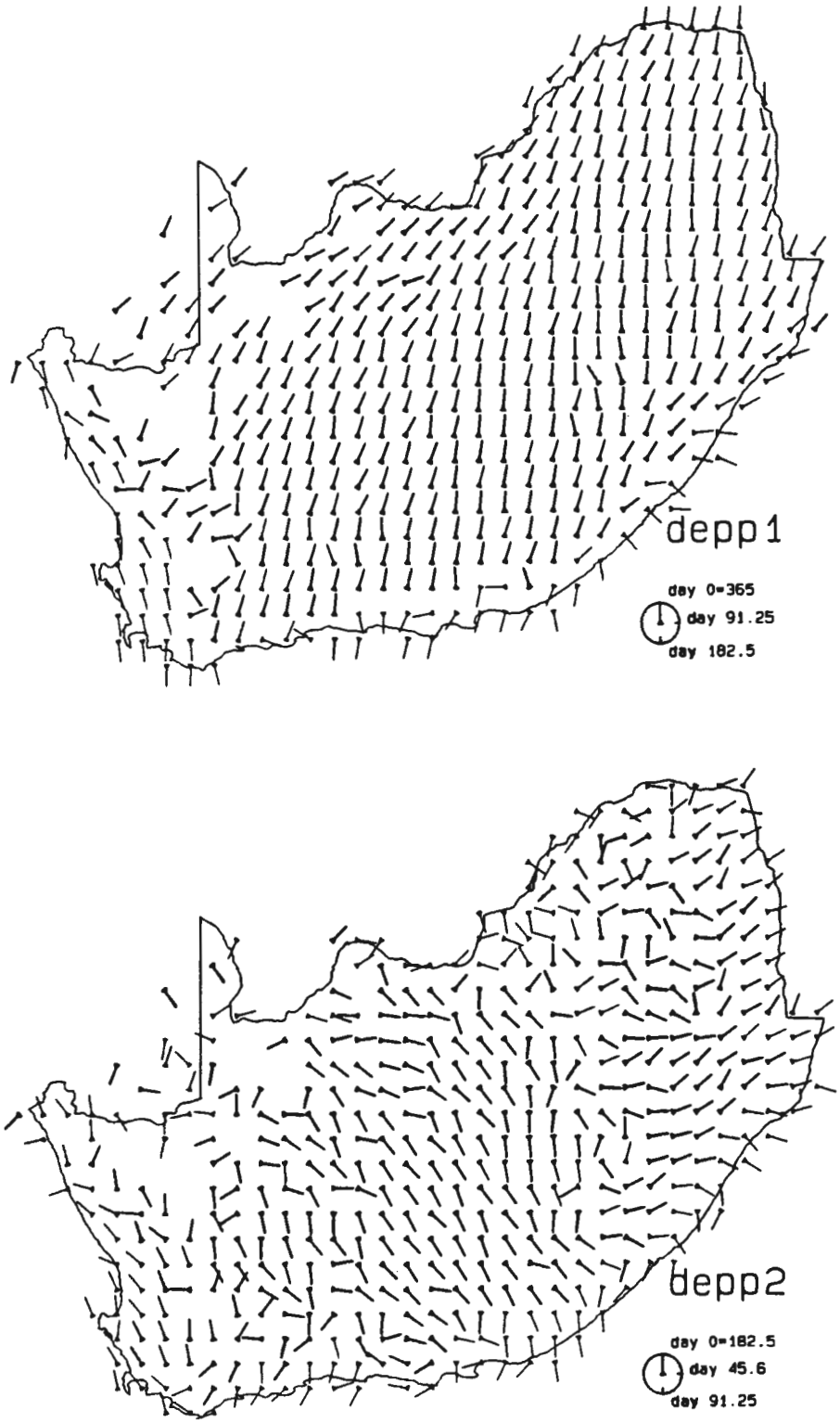


Figure 2.3: SA Weather Bureau block means (contd.).

Bureau, the Department of Forestry, the Department of Agriculture, the South African Sugar Association, as well as data collected by farmers and other members of the public. Dent *et al.* (1989) describe the data base in more detail and also discuss the quality of the data. While the data have been screened by CCWR for human recording and coding errors, and missing values are appropriately flagged, it is almost certain that there are still some errors in the data base. Apart from such errors, some of the apparent spatial variability in the parameters may be due to model fitting error, resulting from the estimation of probabilities and means from a finite record length. This error could be expected to vary between sites depending on the length of the rainfall record.

As the model fitting process was based on maximum likelihood it should in theory be possible to estimate the model fitting error directly using the inverse of the matrix of second derivatives of the likelihood, but this approach led to difficulties in that, for some stations, the parameters for the mean rainfall depth on wet days are such as to violate the regularity conditions required for the usual asymptotic distribution. The problem is described in more detail in McNeill *et al.* (1994). It was therefore decided to estimate the model fitting error at each site by means of a bootstrap procedure. Sampling directly from the original daily data would have been inappropriate because of the necessity to preserve the one day memory of the process and thus a parametric bootstrap was used. That is, at each site, 100 samples of n years of data, (where n was the length in years of the rainfall record at the given site), were simulated using the original maximum likelihood parameter estimates, and the model parameters were re-estimated from each of the generated data sets; the variability of these 100 estimates provides a measure of the model parameter estimation error at that site. This error is not constant across the sites; in particular, it tends to be greater for sites at which the number of years of rainfall record, n, is small, and also larger in areas where the

rainfall is less frequent or more variable from year to year. In Figure 2.4 the bootstrap variance is plotted against the number of years of data for each of the parameters.

Another limitation on the accuracy of any spatial interpolation of the model parameters arises from the station locations. These are recorded in the CCWR database to the nearest minute of a degree of latitude and longitude. This means that locations are accurate to within approximately 1 to 2 km. In most parts of the country the pattern of daily rainfall will change very little over such a distance, however in coastal and mountainous areas the changes can be quite significant. As an example, Table 2.2 lists the fitted model parameters at three stations on the slopes of Table Mountain in Cape Town which all have the same *recorded* location. It can be seen that for some parameters, notably WWA0 and DEPA0, the differences are quite considerable. This variability must be viewed as a limitation imposed by the resolution of the data; it cannot be removed but must be taken into account in the estimation process.

2.5 Extending the Data Set

In order to make use of rainfall data collected in recent years since the original analysis of Zucchini and Adamson (1984a), the model parameters, together with bootstrap estimates of the error variances, were re-calculated for all the original stations, and a number of additional stations, which at the end of February 1992 had 20 years of data, were included (Figure 2.5). This still resulted in a very sparse spatial coverage in certain parts of the country, thus in these areas stations with five or more years of data were also included. While models fitted at such sites might not be very accurate in themselves, they would contribute useful information to the interpolation process described in this thesis; the accuracy of the fitted model was incorporated into the final

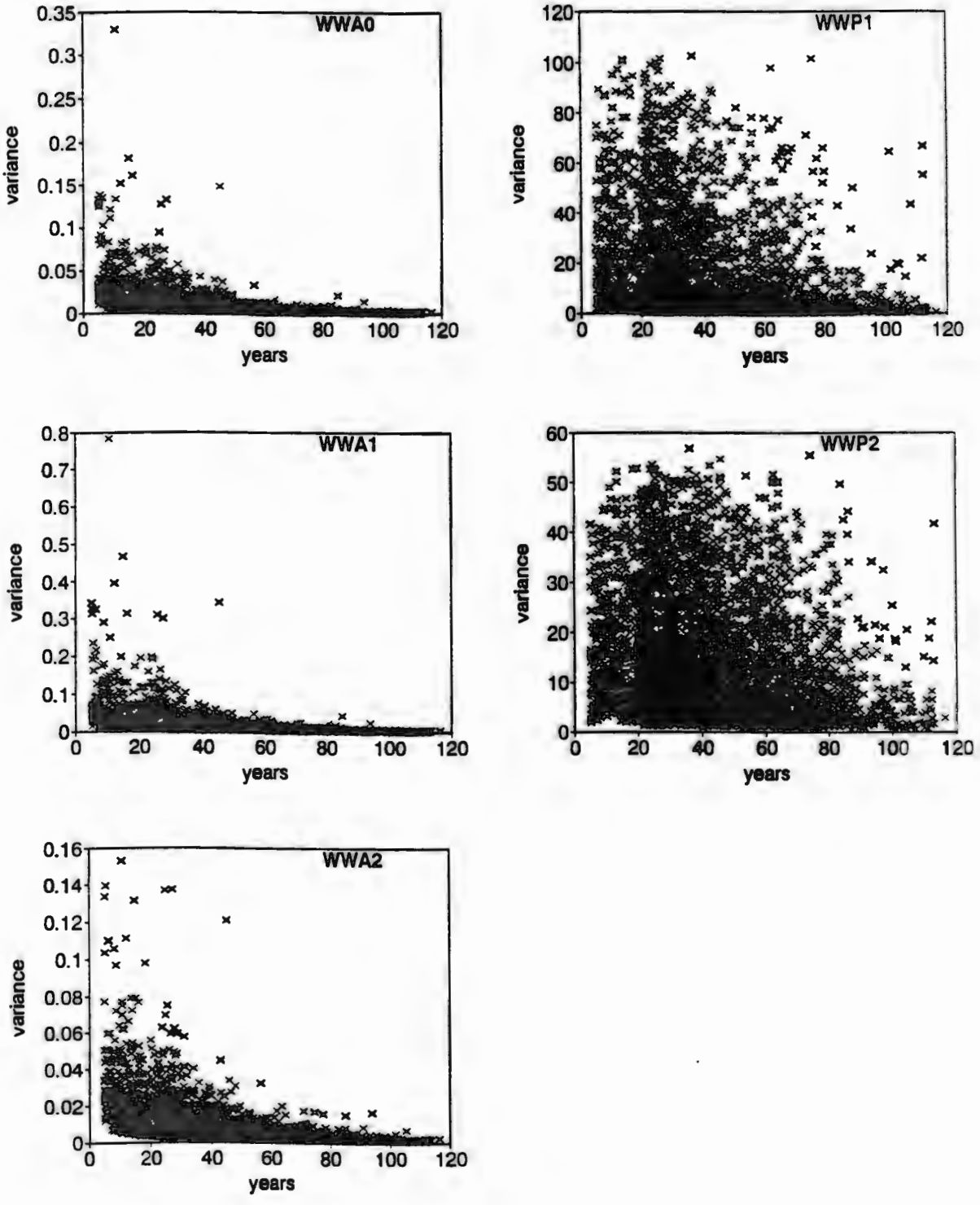


Figure 2.4: Bootstrap variance versus number of years of data.

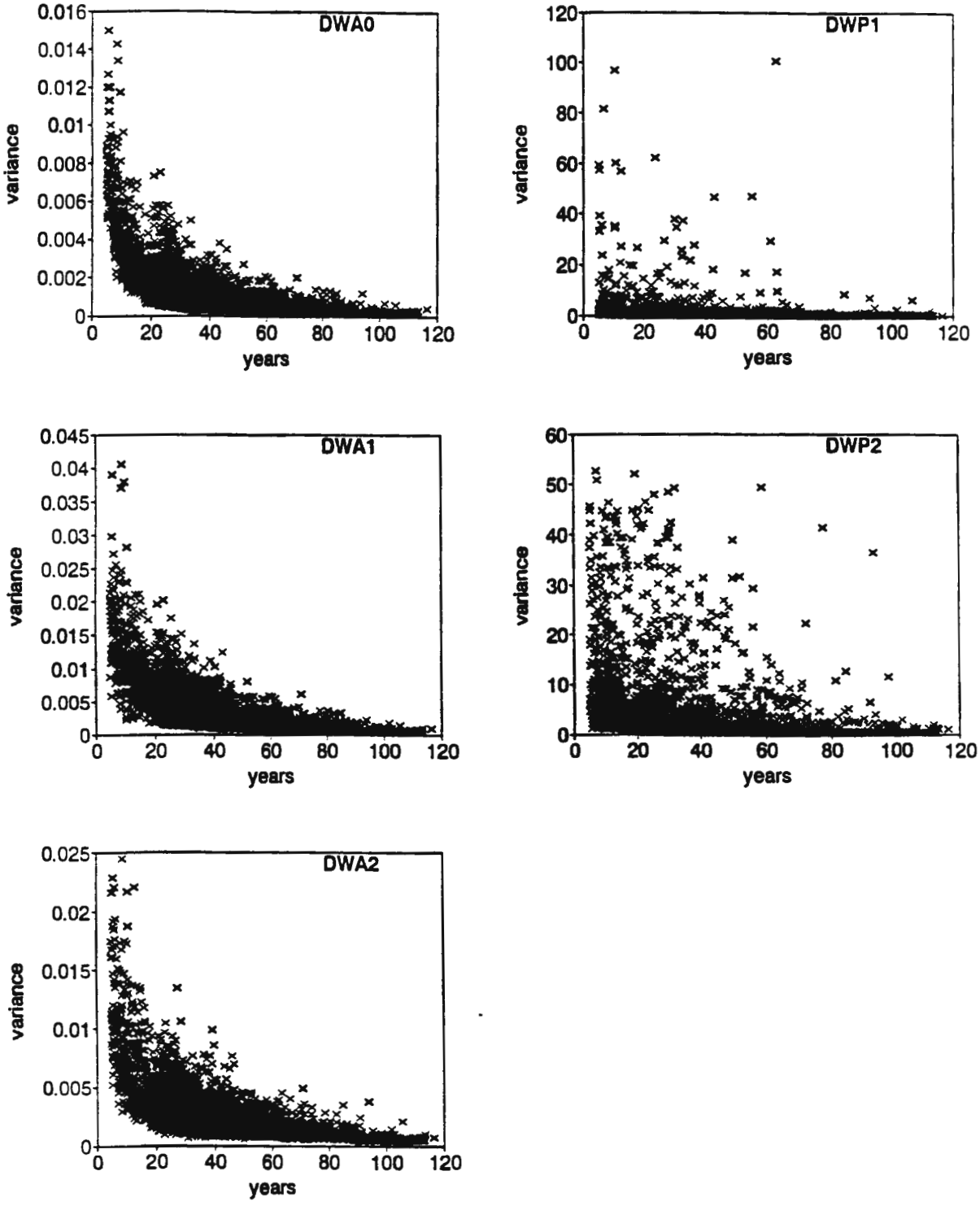


Figure 2.4: Bootstrap variance versus number of years of data (contd.).

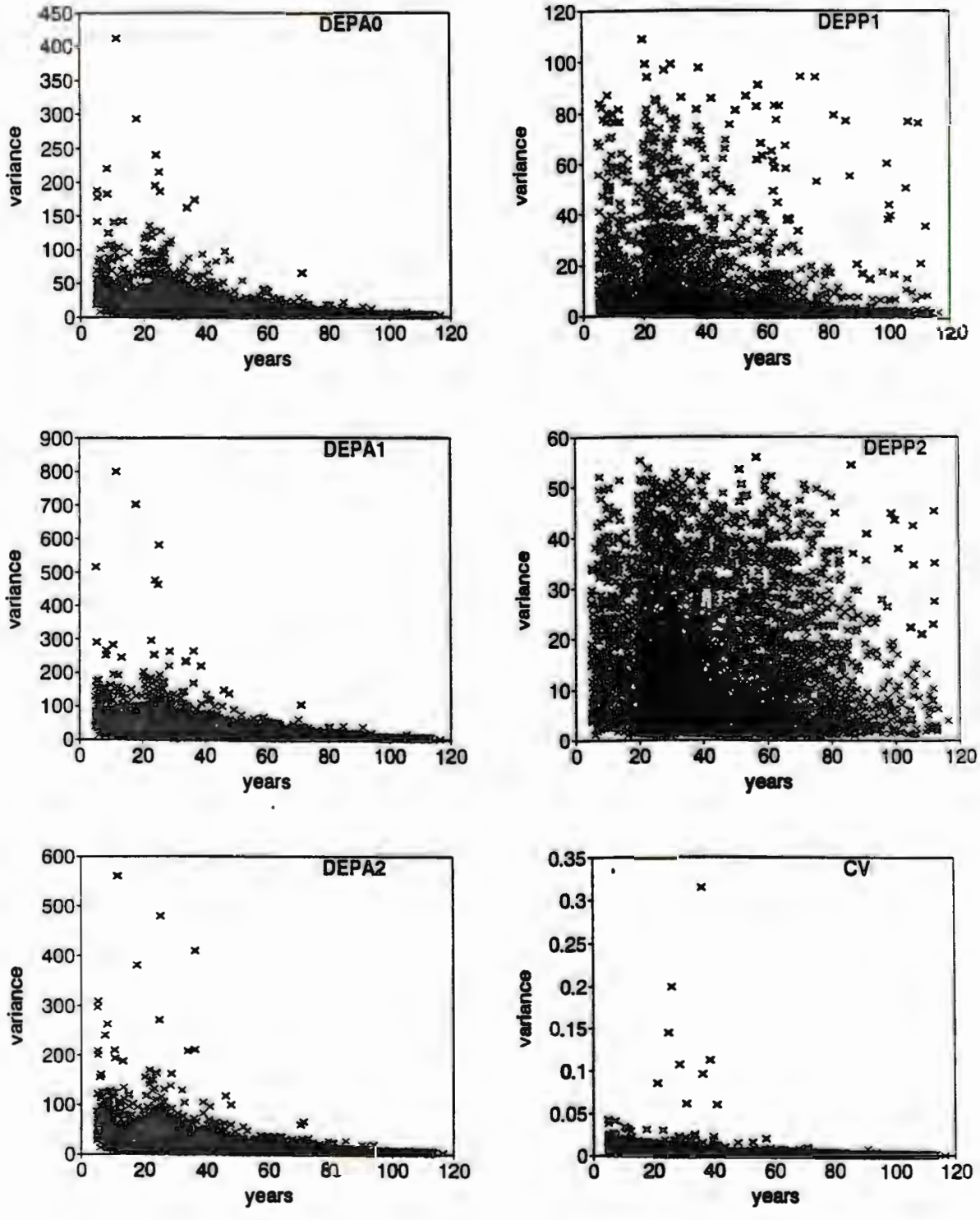


Figure 2.4: Bootstrap variance versus number of years of data (contd.).

	Station Code		
	20719 W	20719AW	20719BW
WWA0	-0.847	-1.018	0.183
WWA1	0.622	0.576	0.634
WWA2	0.140	0.183	0.085
WWP1	195.40	204.56	191.87
WWP2	131.09	127.22	132.51
DWA0	-1.614	-1.646	-1.175
DWA1	0.292	0.258	0.395
DWA2	0.051	0.067	0.033
DWP1	216.67	216.34	211.50
DWP2	49.42	53.02	97.74
DEPA0	203.40	192.38	114.20
DEPA1	88.63	91.82	37.90
DEPA2	25.23	27.12	11.77
DEPP1	173.44	173.40	176.81
DEPP2	164.90	163.71	175.26
CV	1.265	1.278	1.233

Table 2.2: Fitted parameters: stations on Table Mountain.

estimation process in such a way that stations where the fitted model had low accuracy were appropriately down-weighted.

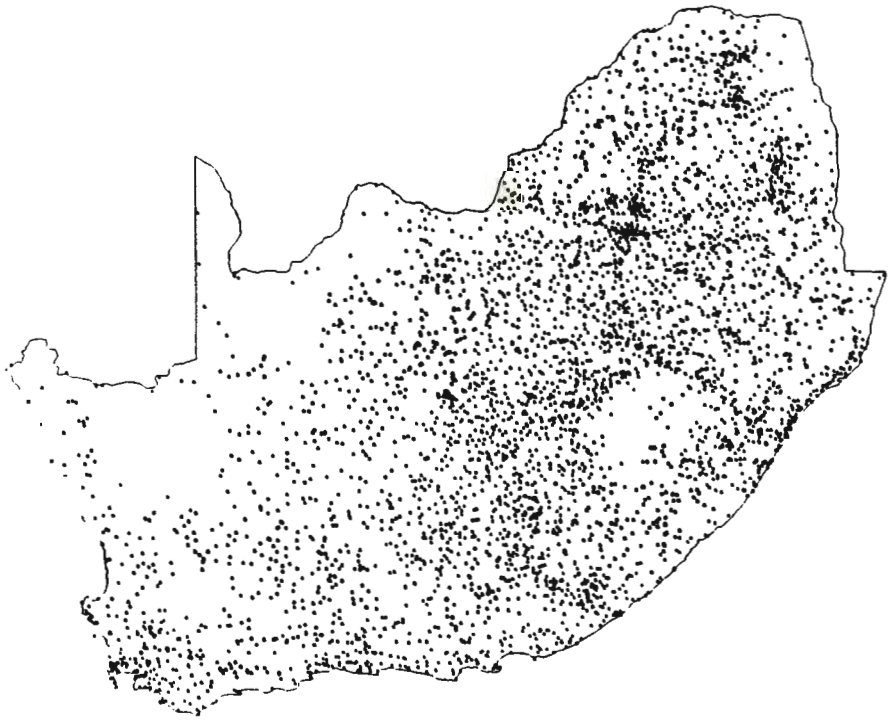


Figure 2.5: Rainfall stations with at least 20 years of data.

Simple outlier and consistency checks were done on the parameter estimates at each site, for example the maximum likelihood parameter estimates were compared with the mean of the bootstrap estimates to check for discrepancies which might suggest instability or bias in the model, and as a result a few sites were deleted from the data set. In all, there were 5070 stations finally selected; their locations are shown in Figure 2.6. As can be seen by comparing Figure 2.6 with Figure 2.5, however, the additional sites still do little to fill in the major gaps on the original map, for example in the north-west region and, to a lesser extent, in Lesotho.

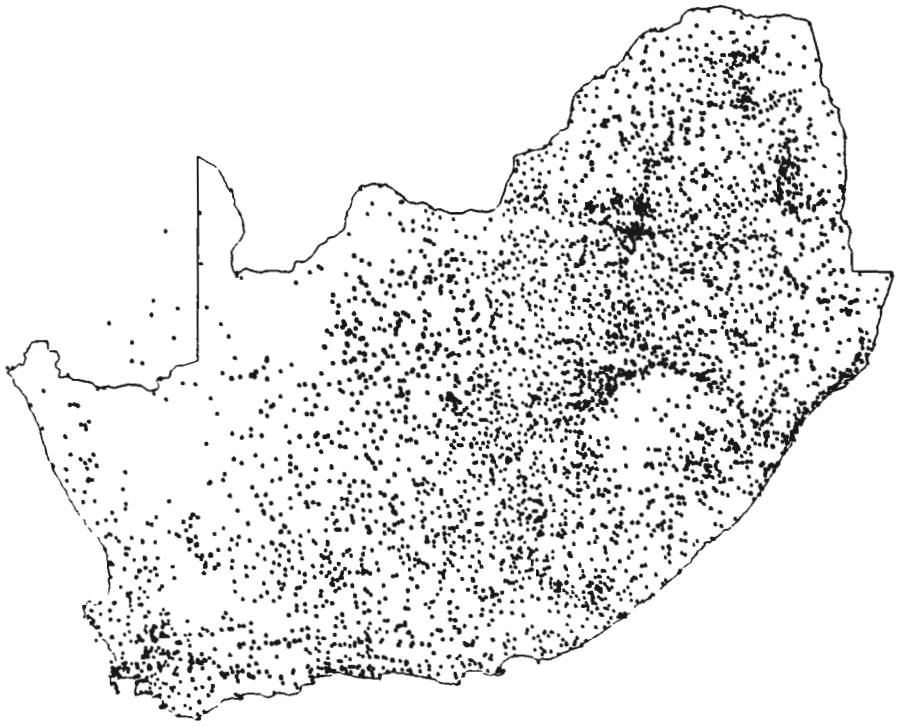


Figure 2.6: Selected rainfall stations.

Using this data set as a basis, the objective was then to estimate the model parameters on a grid of 1 minute of a degree of latitude by 1 minute of a degree of longitude (or approximately 1,9 km by 1,6 km) across southern Africa. Thus in total some 500000 sites are to be estimated in all.

Many methods for interpolation and smoothing of spatial data are available and these are reviewed in Chapter 4, but in choosing a method for this particular application the following features of the problem should be borne in mind:

- The data locations are irregularly spaced (Figure 2.6).
- The data locations are biased in the sense that there are more rainfall

gauges in places close to human habitation and thus fewer in more mountainous areas which in turn tend to be areas of increased rainfall.

- On a local scale, rainfall is known to be highly influenced by topography with orographic rain being induced on the windward sides of mountain slopes and rain shadow areas occurring on the leeward slopes, and thus (and also because of the bias mentioned above) it is essential to allow for the incorporation of topographical information in the estimation process. The effects of topography must be modelled on a local basis, partly because large scale trends may tend to obscure the relationship, and also because the effects will be different in different areas due, for example, to the changing direction of the prevailing rain-bearing wind.

- In order to study the local effect of topography it may be necessary to first remove the large scale trends which could otherwise obscure the relationship. Since the maps in Figure 2.3 show that the large scale trends do not follow any simple functional form, a non-parametric from of trend removal, appropriate for irregularly spaced data with autocorrelated residuals, would be required.

- The data (fitted model parameters) are known to be subject to error, and the error variance varies considerably from one site to another, depending on the number of years of data and the geographical location. In particular, sites with less than 20 years of data generally have very large error variances (Figure 2.4). The error variance for each site can be estimated by the bootstrap procedure.

- The phase parameters are circular variables (see Chapter 8) and ordinary interpolation and smoothing techniques are inappropriate for such data.

- There are sixteen rainfall model parameters, each estimated at some

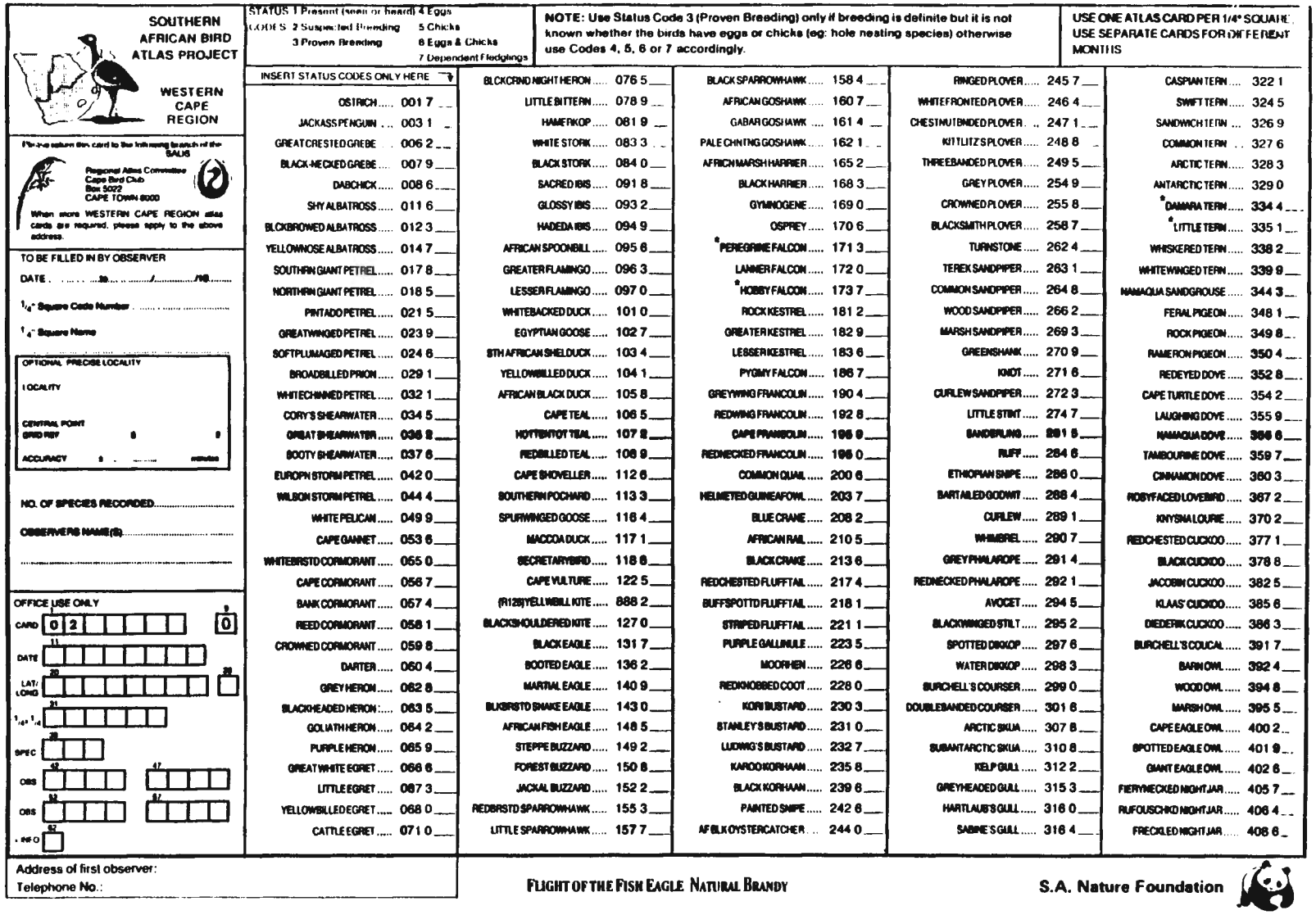
5000 sites. In addition there are some 500000 altitude values (on a grid of one minute of a degree of latitude by one minute of a degree of longitude). We wish to interpolate the sixteen rainfall model parameters to some 500000 sites. Thus the methodology chosen must be computationally feasible for very large data sets.

Chapter 3

The Southern African Bird Atlas Project

The Southern African Bird Atlas Project (Harrison, 1987 and 1992) has as its main objective to map the spatial distribution of each of the individual bird species occurring in southern Africa. The project, which has the Department of Environment Affairs, the Mazda Wildlife Fund, the Southern African Ornithological Society and the University of Cape Town as major sponsors, commenced in 1987. Between 1987 and 1992 some 7500 people, mainly volunteers, have participated in data collection, and over seven million individual records have been entered into the databank. By recording reporting rate frequencies on a grid square basis for individual months the resultant data base provides a wealth of information on distribution, seasonality and migration patterns, and allows for the study of the relationship between these patterns and other environmental variables such as vegetation, rainfall, topography and human land use. The atlas will also provide a basis for the planning of conservation strategies based on species richness and diversity and provide a basis for monitoring long-term trends in distribution and abundance.

The spatial unit for data collection and mapping is the quarter degree



SOUTHERN AFRICAN BIRD ATLAS PROJECT

WESTERN CAPE REGION

Regional Atlas Committee
Cape Bird Club
Box 5022
CAPE TOWN 8000

When more WESTERN CAPE REGION atlas cards are required, please apply to the above address.

TO BE FILLED IN BY OBSERVER

DATE ..... /...../.....

1/4° Square Code Number .....

1/4° Square Name .....

OPTIONAL PRECISE LOCALITY

LOCALITY

CENTRAL POINT GRID REF ° °

ACCURACY ° . ..... minutes

NO. OF SPECIES RECORDED.....

OBSERVERS NAME(S).....

.....

OFFICE USE ONLY

CARD 0 2 0

DATE

LAT/LONG

1/4° 1/4

SPEC

OBS

OBS

INFO

STATUS 1 Present (seen or heard) 4 Eggs
CODES 2 Suspected Breeding 5 Chicks
3 Proven Breeding 6 Eggs & Chicks
7 Dependent Fledglings

NOTE: Use Status Code 3 (Proven Breeding) only if breeding is definite but it is not known whether the birds have eggs or chicks (eg: hole nesting species) otherwise use Codes 4, 5, 6 or 7 accordingly.

USE ONE ATLAS CARD PER 1/4° SQUARE. USE SEPARATE CARDS FOR DIFFERENT MONTHS

INSERT STATUS CODES ONLY HERE

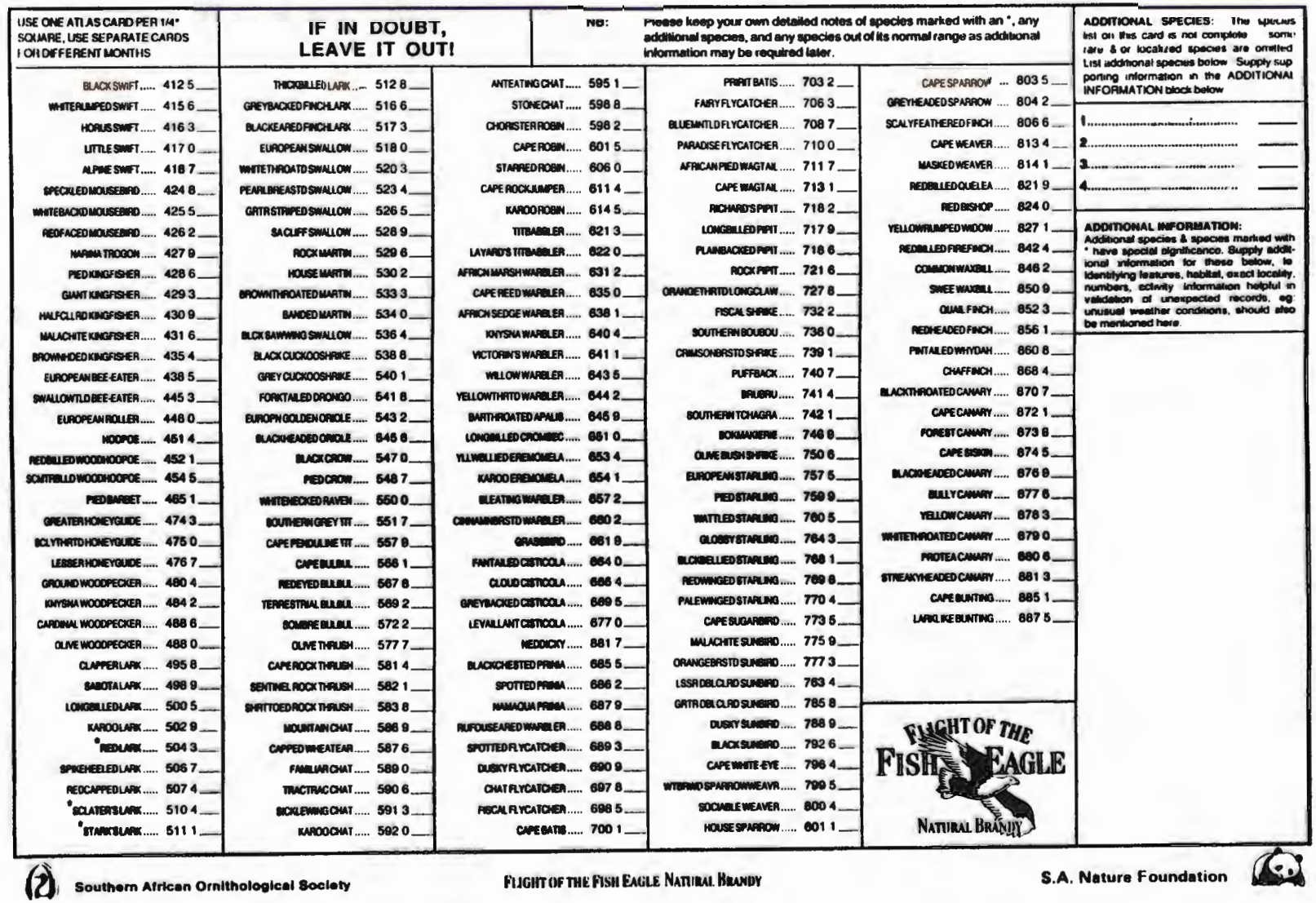
OSTRICH	001 7	BLCKCRND NIGHT HERON	076 5	BLACK SPARROWHAWK	158 4	RINGED PLOVER	245 7	CASPIAN TERN	322 1
JACKASS PENGUIN	003 1	LITTLE BITTERN	078 9	AFRICAN GOSHAWK	160 7	WHITEFRONTED PLOVER	246 4	SWIFT TERN	324 5
GREAT CRESTED GREBE	006 2	HAMERKOP	081 9	GABAR GOSHAWK	161 4	CHESTNUTBNDED PLOVER	247 1	SANDWICH TERN	326 9
BLACK-NECKED GREBE	007 9	WHITE STORK	083 3	PALE CHNTNG GOSHAWK	162 1	KITTLITZ'S PLOVER	248 8	COMMON TERN	327 6
DABCHICK	008 6	BLACK STORK	084 0	AFRICN MARSH HARRIER	165 2	THREEBANDED PLOVER	249 5	ARCTIC TERN	328 3
SHY ALBATROSS	011 6	SACRED IBIS	091 8	BLACK HARRIER	168 3	GREY PLOVER	254 9	ANTARCTIC TERN	329 0
BLCKBROWED ALBATROSS	012 3	GLOSSY IBIS	093 2	GYMNOGENE	169 0	CROWNED PLOVER	255 8	*DAMARA TERN	334 4
YELLOWNOSE ALBATROSS	014 7	HADEDA IBIS	094 9	OSPREY	170 6	BLACKSMITH PLOVER	258 7	*LITTLE TERN	335 1
SOUTHRN GIANT PETREL	017 8	AFRICAN SPOONBILL	095 6	*PEREGRINE FALCON	171 3	TURNSTONE	262 4	WHISKERED TERN	338 2
NORTHRN GIANT PETREL	018 5	GREATER FLAMINGO	096 3	LANNER FALCON	172 0	TEREK SANDPIPER	263 1	WHITEWINGED TERN	339 9
PINTADO PETREL	021 5	LESSER FLAMINGO	097 0	*HOBBY FALCON	173 7	COMMON SANDPIPER	264 8	NAMAQUA SANDGROUSE	344 3
GREATWINGED PETREL	023 9	WHITEBACKED DUCK	101 0	ROCK KESTREL	181 2	WOOD SANDPIPER	266 2	FERAL PIGEON	348 1
SOFTPLUMAGED PETREL	024 6	EGYPTIAN GOOSE	102 7	GREATER KESTREL	182 9	MARSH SANDPIPER	269 3	ROCK PIGEON	349 8
BROADBILLED PRION	029 1	STH AFRICAN SHELDUCK	103 4	LESSER KESTREL	183 6	GREENSHANK	270 9	RAMERON PIGEON	350 4
WHITECHINNED PETREL	032 1	YELLOWBILLED DUCK	104 1	PYGMY FALCON	186 7	KNOT	271 6	REDEYED DOVE	352 8
CORY'S SHEARWATER	034 5	AFRICAN BLACK DUCK	105 8	GREYWING FRANCOLIN	190 4	CURLEW SANDPIPER	272 3	CAPE TURTLE DOVE	354 2
GREAT SHEARWATER	036 9	CAPE TEAL	106 5	REDWING FRANCOLIN	192 8	LITTLE STINT	274 7	LAUGHING DOVE	355 9
SOOTY SHEARWATER	037 6	HOTTENTOT TEAL	107 8	CAPE FRANCOLIN	196 9	SANDERLING	281 5	NAMAQUA DOVE	356 6
EUROPN STORM PETREL	042 0	REDBILLED TEAL	108 9	REDNECKED FRANCOLIN	196 0	RUFF	284 6	TAMBOURINE DOVE	359 7
WILSON STORM PETREL	044 4	CAPE SHOVELLER	112 6	COMMON QUAIL	200 6	ETHIOPIAN SNIPE	286 0	CINNAMON DOVE	360 3
WHITE PELICAN	049 9	SOUTHERN POCHARD	113 3	HELMETED GUINEAFOWL	203 7	BARTAILED GODWIT	288 4	ROBYFACED LOVEBIRD	367 2
CAPE GANNET	053 6	SPURWINGED GOOSE	116 4	BLUE CRANE	208 2	CURLEW	289 1	KNYSNA LOURIE	370 2
WHITEBRSTD CORMORANT	055 0	MACCOA DUCK	117 1	AFRICAN RAIL	210 5	WHIMBREL	290 7	REDCHESTED CUCKOO	377 1
CAPE CORMORANT	056 7	SECRETARYBIRD	118 6	BLACK CRAKE	213 6	GREY PHALAROPE	291 4	BLACK CUCKOO	378 8
BANK CORMORANT	057 4	CAPE VULTURE	122 5	REDCHESTED FLUFFTAIL	217 4	REDNECKED PHALAROPE	292 1	JACOBIN CUCKOO	382 5
REED CORMORANT	058 1	(R128)YELLWBILL KITE	888 2	BUFFSPOTTD FLUFFTAIL	218 1	AVOCET	294 5	KLAAS' CUCKOO	385 6
CROWNED CORMORANT	059 8	BLACKSHOULDERED KITE	127 0	STRIPED FLUFFTAIL	221 1	BLACKWINGED STILT	295 2	DIEDERIK CUCKOO	386 3
DARTER	060 4	BLACK EAGLE	131 7	PURPLE GALLINULE	223 5	SPOTTED DIKKOP	297 6	BURCHELL'S COUCAL	391 7
GREY HERON	062 8	BOOTED EAGLE	136 2	MOORHEN	226 6	WATER DIKKOP	298 3	BARN OWL	392 4
BLACKHEADED HERON	063 5	MARTIAL EAGLE	140 9	REDKNOBBED COOT	228 0	BURCHELL'S COURSER	299 0	WOOD OWL	394 8
GOLIATH HERON	064 2	BLKBRSTD SNAKE EAGLE	143 0	KORI BUSTARD	230 3	DOUBLEBANDED COURSER	301 6	MARSH OWL	395 5
PURPLE HERON	065 9	AFRICAN FISH EAGLE	148 5	STANLEY'S BUSTARD	231 0	ARCTIC SKUA	307 8	CAPE EAGLE OWL	400 2
GREAT WHITE EGRET	066 6	STEPPE BUZZARD	149 2	LUDWIG'S BUSTARD	232 7	SUBANTARCTIC SKUA	310 8	SPOTTED EAGLE OWL	401 9
LITTLE EGRET	067 3	FOREST BUZZARD	150 8	KAROO KORHAAN	235 8	KELP GULL	312 2	GIANT EAGLE OWL	402 6
YELLOWBILLED EGRET	068 0	JACKAL BUZZARD	152 2	BLACK KORHAAN	239 6	GREYHEADED GULL	315 3	FIERYNECKED NIGHTJAR	405 7
CATTLE EGRET	071 0	REDBRSTD SPARROWHAWK	155 3	PAINTED SNIPE	242 6	HARTLAUB'S GULL	316 0	RUFOUSCHKD NIGHTJAR	406 4
		LITTLE SPARROWHAWK	157 7	AFBLK OYSTERCATCHER	244 0	SABINE'S GULL	316 4	FRECKLED NIGHTJAR	408 6

Address of first observer:
Telephone No.:

FLIGHT OF THE FISH EAGLE NATURAL BRANDY

S.A. Nature Foundation

Figure 3.1: SABAP field card.



USE ONE ATLAS CARD PER 1/4° SQUARE, USE SEPARATE CARDS FOR DIFFERENT MONTHS

IF IN DOUBT, LEAVE IT OUT!

NB: Please keep your own detailed notes of species marked with an *, any additional species, and any species out of its normal range as additional information may be required later.

BLACK SWIFT	412 5	THICKBILLED LARK	512 8	ANTEATING CHAT	595 1	PRIRIT BATIS	703 2	CAPE SPARROW	803 5
WHITE-RUMPED SWIFT	415 6	GREYBACKED FINCH-LARK	516 6	STONECHAT	598 8	FAIRY FLYCATCHER	706 3	GREYHEADED SPARROW	804 2
HORUS SWIFT	416 3	BLACKEARED FINCH-LARK	517 3	CHORISTER ROBIN	598 2	BLUE-MNTLD FLYCATCHER	708 7	SCALY-FEATHERED FINCH	806 6
LITTLE SWIFT	417 0	EUROPEAN SWALLOW	518 0	CAPE ROBIN	601 5	PARADISE FLYCATCHER	710 0	CAPE WEAVER	813 4
ALPINE SWIFT	418 7	WHITE-THROATD SWALLOW	520 3	STARRED ROBIN	606 0	AFRICAN PIED WAGTAIL	711 7	MASKED WEAVER	814 1
SPECKLED MOUSEBIRD	424 8	PEARL-BREASTD SWALLOW	523 4	CAPE ROCK-JUMPER	611 4	CAPE WAGTAIL	713 1	RED-BILLED QUELEA	821 9
WHITE-BACKD MOUSEBIRD	425 5	GRTR STRIPED SWALLOW	526 5	KAROO ROBIN	614 5	RICHARD'S PIPIT	716 2	RED BISHOP	824 0
RED-FACED MOUSEBIRD	426 2	SA CLIFF SWALLOW	526 9	TITBABBLER	621 3	LONG-BILLED PIPIT	717 9	YELLOW-RUMPED WIDOW	827 1
NARINA TROGON	427 9	ROCK MARTIN	529 6	LAYARD'S TITBABBLER	622 0	PLAIN-BACKED PIPIT	716 6	RED-BILLED FIREFINCH	842 4
PIED KINGFISHER	428 6	HOUSE MARTIN	530 2	AFRICN MARSH WARBLER	631 2	ROCK PIPIT	721 6	COMMON WAXBILL	846 2
GIANT KINGFISHER	429 3	BROWN-THROATED MARTIN	533 3	CAPE REED WARBLER	635 0	ORANGE-THRTD LONGCLAW	727 8	SWEE WAXBILL	850 9
HALF-CLLRD KINGFISHER	430 9	BANDED MARTIN	534 0	AFRICN SEDGE WARBLER	638 1	FISCAL SHRIKE	732 2	QUAIL FINCH	852 3
MALACHITE KINGFISHER	431 6	BLCK-SAWWNG SWALLOW	536 4	KNYSNA WARBLER	640 4	SOUTHERN BOUBOU	736 0	RED-HEADED FINCH	856 1
BROWN-HDED KINGFISHER	435 4	BLACK CUCKOOSHRIKE	538 8	VICTORIN'S WARBLER	641 1	CRIMSON-BRSTD SHRIKE	739 1	PINTAILED WHYDAH	860 8
EUROPEAN BEE-EATER	438 5	GREY CUCKOOSHRIKE	540 1	WILLOW WARBLER	643 5	PUFFBACK	740 7	CHAFFINCH	868 4
SWALLOW-TLD BEE-EATER	445 3	FORK-TAILED DRONGO	541 8	YELLOW-THRTD WARBLER	644 2	BRUBRU	741 4	BLACK-THROATED CANARY	870 7
EUROPEAN ROLLER	446 0	EUROPN GOLDEN ORIOLE	543 2	BARTHROATED APALIS	645 9	SOUTHERN TCHAGRA	742 1	CAPE CANARY	872 1
HOOPOE	451 4	BLACK-HEADED ORIOLE	545 6	LONG-BILLED CROMBEC	651 0	BOKMAKIERIE	746 9	FOREST CANARY	873 8
RED-BILLED WOOD-HOOPOE	452 1	BLACK CROW	547 0	YLLW-BILLIED EREMOMELA	653 4	OLIVE BUSH-SHRIKE	750 6	CAPE SISKIN	874 5
SCMTR-BILLD WOOD-HOOPOE	454 5	PIED CROW	548 7	KAROO EREMOMELA	654 1	EUROPEAN STARLING	757 5	BLACK-HEADED CANARY	876 9
PIED BARBET	465 1	WHITE-NECKED RAVEN	550 0	BLEATING WARBLER	657 2	PIED STARLING	759 9	BULLY CANARY	877 6
GREATER HONEYGUIDE	474 3	SOUTHERN GREY TIT	551 7	CINNAMN-BRSTD WARBLER	660 2	WATTLED STARLING	760 5	YELLOW CANARY	878 3
SCLYTHRTD HONEYGUIDE	475 0	CAPE PENDULINE TIT	557 9	GRASSBIRD	661 9	GLOSSY STARLING	764 3	WHITE-THROATED CANARY	879 0
LESSER HONEYGUIDE	476 7	CAPE BULBUL	566 1	FANTAILED CISTICOLA	664 0	BLCK-BELLIED STARLING	768 1	PROTEA CANARY	880 6
GROUND WOODPECKER	480 4	RED-EYED BULBUL	567 8	CLOUD CISTICOLA	666 4	RED-WINGED STARLING	769 8	STREAKY-HEADED CANARY	881 3
KNYSNA WOODPECKER	484 2	TERRESTRIAL BULBUL	569 2	GREY-BACKED CISTICOLA	669 5	PALE-WINGED STARLING	770 4	CAPE BUNTING	885 1
CARDINAL WOODPECKER	488 6	BOMBRE BULBUL	572 2	LEVAILLANT CISTICOLA	677 0	CAPE SUGARBIRD	773 5	LARK-LIKE BUNTING	887 5
OLIVE WOODPECKER	488 0	OLIVE THRUSH	577 7	NEDDICKY	681 7	MALACHITE SUNBIRD	775 9		
CLAPPER LARK	495 8	CAPE ROCK THRUSH	581 4	BLACK-CHESTED PRINIA	685 5	ORANGE-BRSTD SUNBIRD	777 3		
SABOTA LARK	498 9	SENTINEL ROCK THRUSH	582 1	SPOTTED PRINIA	686 2	LSSR DBL-CLRD SUNBIRD	763 4		
LONG-BILLED LARK	500 5	SHORT-TOED ROCK THRUSH	583 8	NAMAQUA PRINIA	687 9	GRTR DBL-CLRD SUNBIRD	785 8		
KAROO LARK	502 9	MOUNTAIN CHAT	586 9	RUFOUS-EARED WARBLER	688 6	DUSKY SUNBIRD	788 9		
*RED LARK	504 3	CAPPED WHEATEAR	587 6	SPOTTED FLYCATCHER	689 3	BLACK SUNBIRD	792 6		
SPIKE-HEELED LARK	506 7	FAMILIAR CHAT	589 0	DUSKY FLYCATCHER	690 9	CAPE WHITE-EYE	796 4		
RED-CAPPED LARK	507 4	TRACTRAC CHAT	590 6	CHAT FLYCATCHER	697 8	WTBRWD SPARROWWEAVR	799 5		
*SCLATER'S LARK	510 4	SICKLE-WING CHAT	591 3	FISCAL FLYCATCHER	698 5	SOCIABLE WEAVER	800 4		
*STARK'S LARK	511 1	KAROO CHAT	592 0	CAPE BATIS	700 1	HOUSE SPARROW	801 1		

ADDITIONAL SPECIES: The species list on this card is not complete - some rare & or localized species are omitted. List additional species below. Supply supporting information in the ADDITIONAL INFORMATION block below

1.....

2.....

3.....

4.....

ADDITIONAL INFORMATION:
Additional species & species marked with * have special significance. Supply additional information for these below, ie identifying features, habitat, exact locality, numbers, activity. Information helpful in validation of unexpected records, eg: unusual weather conditions, should also be mentioned here.

FLIGHT OF THE FISH EAGLE NATURAL BRANDY

Southern African Ornithological Society

FLIGHT OF THE FISH EAGLE NATURAL BRANDY

S.A. Nature Foundation

Figure 3.1: SABAP field card (contd.).

grid square (QDGS) of 15 minutes of a degree latitude by 15 minutes of a degree longitude, or approximately 28 × 24 km in the centre of South Africa. Observers complete monthly field cards similar to that shown in Figure 3.1 for individual QDGS, on which the presence or absence of each individual species is recorded. That is, if a species was sighted at least once during the month in a given QDGS, it is recorded as present.

The field cards are checked and collated by the atlas co-ordinator, with the help of regional assistants. The *reporting rate*, that is, the ratio of the number of cards indicating the presence of a given species to the total number of cards completed, is calculated for each QDGS and each month. These reporting rates form the basis of the maps of each species. In preparing the detailed map for each species to be published in the atlas, the aggregated data for the entire period will be used; that is, the ratio of the *total* number of monthly field cards showing the species as present to the *total* number of field cards completed for that QDGS; the small number of cards available in some areas does not allow for more detailed maps for individual months, although histograms showing the average reporting rate for each month, taken over all squares, will be included in the published atlas to give an indication of the seasonal variation of abundance. Figure 3.2 shows a map of the reporting rates for the Pied Crow *Corvus albus*, using data cumulated over all months of the year.

3.1 The Reporting Rate as a Measure of Abundance

The reporting rate is commonly used as a measure of abundance of a species. Clearly there are problems with this in that the relationship between abundance and reporting rate will depend on various factors such as the conspic-

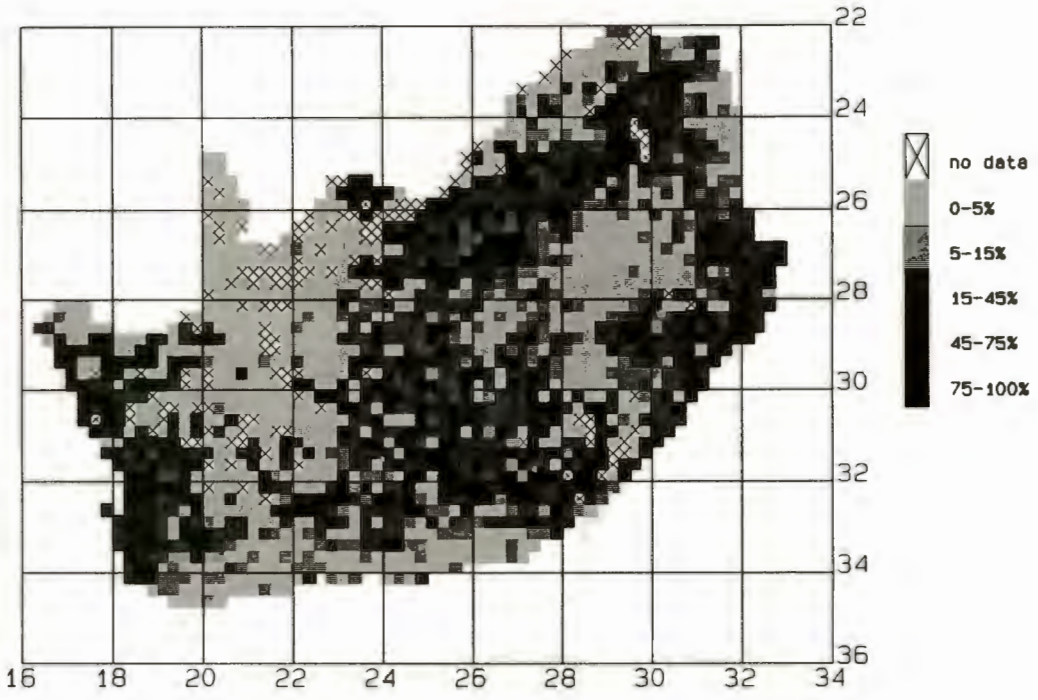


Figure 3.2: Unadjusted reporting rates: Pied Crow.

uousness and ease of identifiability of the species and the degree of clustering of the birds, and also on the search effort expended by the observer, both in terms of time and spatial coverage of the area. To a large extent, however, these other factors will be fairly constant across all QDGS for a given species, and previous studies show that the relationship between reporting rate and abundance is at least monotonic (Temple and Temple, 1986, Hockey *et al.*, 1989). Thus the maps of individual species' reporting rates do provide useful information on the spatial distribution of a given species at any point in time and could, for example, serve as a basis for monitoring changes over a period of some years.

3.2 Spatial Coverage of the Data

The use of volunteers for data collection has made it possible to collect an enormous amount of data in a fairly short time and with relatively little expense but a drawback is that the geographical coverage is very uneven, with observers' field efforts being concentrated mainly in the vicinity of the large urban centres. In this thesis I have made use of all the data that had been processed in April 1990 for the 1974 QDGS in South Africa, Lesotho and Swaziland. For some squares the total number of field cards completed and checked up to this time is quite large, while in more remote areas it is very small and, in some cases, zero (Table 3.1). More recently efforts have been made to obtain additional data in the more remote areas, but it is clear that there will always be squares for which there is very little available information; for example, at the end of February 1992 some 30 percent of all QDGS in the northern Cape still had fewer than 5 field cards.

No. of cards	No. of QDGS
0	139
1 - 5	592
6 - 10	277
11 - 20	273
21 - 50	339
51 and over	354
Total	1974

Table 3.1: Frequency distribution of field cards as at April 1990.

3.3 Smoothing the Data

The reporting rate may be interpreted as an estimate of the probability that a randomly chosen observer will see a given species (in a given area and time period). Clearly the observed reporting rate is subject to sampling variation, with the variability being larger for those QDGS where the number of completed field cards is small. Thus reporting rates based on a small number of field cards provide a poor estimate of the underlying probability. If, as is likely, the observed reporting rates exhibit spatial correlation, then a better estimate can be obtained by incorporating the observed reporting rates from neighbouring QDGS, that is, by some type of spatial smoothing of the data. Another advantage of smoothing the data is that this would reduce the rather discontinuous appearance of the species maps; while some discontinuity is inevitable if one maps rates for discrete subregions using a scale of grey shades, the sampling variation exacerbates this effect, especially in those areas where the number of field cards is low.

There are several aspects of this problem which are, in a statistical sense, very similar to those of the problem discussed in the previous chapter.

- Data are available at a large number of spatial locations (1974 in all).
- There is error (sampling variability in this case) in the observed data and the variance of this error depends on the sample size which varies markedly across the QDGS (from zero to over 1500).
- Estimation may be improved by the incorporation of covariate information such as climate, vegetation or topography, since many species are observed to occur only in specific types of habitat.
- There are a large number of values to be estimated; in the Southern African Bird Atlas Project the objective is to map reporting rates across some 2000 QDGS for each of 900 species or more.

The bird atlas data differs however from the daily rainfall model data in that the reporting rates are essentially binomial in nature, so that the variance of the error depends on the reporting rate. Thus, what is needed here is a method of spatial interpolation and smoothing for binomially distributed data. This problem is taken up in Chapter 9.

Chapter 4

Methods of Interpolation and Smoothing of Spatial Data

In this chapter it is assumed that values of some variable are available at a number of spatial locations and that the objective is to predict values at one or more additional spatial locations. Discussion is restricted to the case where the data locations are in a two-dimensional space, although the results generalise readily to three (or more) dimensions. In general the predicted values are required to be estimated at all points or at a regular grid of points within a given boundary so that one is essentially fitting a surface to the original data points. In the *interpolation* problem the fitted surface is required to coincide with the original values at the data points. The interpolation problem can be viewed as a limiting case of the *smoothing* problem, in which the fitted surface need not coincide with the original values. The smoothing problem arises in various situations:

- In mapping spatial data there is often a necessity to smooth in order that the major features of the data can be readily visualised without excessive detail. This is often used as a preliminary step in contouring the data.

- In cases when the data contain measurement error or other 'noise' one may wish to predict underlying 'measurement-error-free' values at all locations, that is, to separate 'signal' and 'noise'. This is the case for both the applications described in this thesis.

- Smoothing is sometimes part of the process of decomposition of data which is undertaken so that individual components may be studied separately; this is analogous to the decomposition of time-series into trend and other components. This approach is used in this thesis to filter out large-scale trends from the rainfall parameter values so that the local relationship of rain and topography may be studied in more detail.

These differing objectives may require different smoothing techniques; furthermore, different aspects of the data set, such as irregular spacing of the data, or the need to incorporate information on covariates, may influence the choice of methodology.

Once a method of smoothing has been selected, there still remains the problem of determining the degree of smoothing required. Once again there are several options. Many smoothing methods have a built-in smoothing parameter which may be set at a level chosen by the user; for example, the 'bandwidth' in kernel smoothing or the weighting parameter for the roughness penalty in smoothing splines. This parameter may be selected to optimise some user-selected criterion of goodness of fit, typically by means of cross-validation. Other methods proceed directly to minimise some model-based criterion such as the least squares fit (as in trend surface analysis) or the prediction variance (as in kriging); in order to do this some assumption must be made about the underlying model and, in particular, some knowledge of the covariance function must be assumed.

It is perhaps of interest to point out some fundamental differences between the estimation problem for spatial data as compared with time series data. It is unusual in the spatial context to try to predict beyond the geographical boundaries of the observed data so that one does not have the problem of extrapolation, and thus methods based on simple local averaging are often sufficient. This advantage is offset by some disadvantages, particularly the lack of a natural ordering in more than one dimension, and also the fact that spatial data are typically collected at *irregularly* spaced locations. These differences render much of the standard time series methodology inapplicable in the spatial context.

4.1 Review of Smoothing Methods for Spatial Data

We consider the more commonly used methods of interpolation and smoothing for spatial data which have been applied in such areas as soil mapping, mining, rainfall modelling and hydrology.

Three of the methods described below are model based, the remainder are non-parametric and appear to make fewer assumptions about the data; however this flexibility is often illusory, as several of the non-parametric methods have been shown to be equivalent to some form of kriging with a pre-specified covariance function. For comparison, at the end of this section we provide a common model framework for those methods which are linear smoothers.

Throughout this chapter we use the notation that the variable v_i is measured at locations $\mathbf{z}_i = (x_i, y_i)'$ where $i = 1, 2, \ldots n$, and x and y represent appropriate co-ordinates such as longitude and latitude. The location of the point to be estimated is given by $\mathbf{z}_0 = (x_0, y_0)'$, and d_{ij} represents the dis-

tance between $\mathbf{z}_i$ and $\mathbf{z}_j$, while d_{i0} represents the distance between $\mathbf{z}_i$ and $\mathbf{z}_0$.

4.1.1 Trend Surface Analysis

In trend surface analysis a simple polynomial function such as a plane or quadratic surface is fitted to the data using ordinary least squares (Grant, 1957; Krumbein, 1959; Watson, 1971 and 1972). For example, if the fitted function is quadratic in the x and y coordinates of the data locations, then the fitted surface has the form:

$$f(x,y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$$

While this method may be appropriate when the trend has a simple functional form, this is rarely the case in practice in the earth and atmospheric sciences, except perhaps over fairly small areas. The degree of the polynomial must be selected by the user, and in fact this is the only way in which the user can control the degree of smoothing; interpolation is possible for most data sets only by allowing the number of terms in the model to equal the number of data points. When the residuals from the trend, or local 'anomalies', are spatially correlated, as they generally are in spatial applications, use of the usual F-tests will often lead to the fitting of a surface of too high an order which is perceived by the user as 'too wavy'. Ripley (1981, Chapter 4) illustrates this effect. In addition, clustering of the data points tends to give excessive weight to the fit of the surface in the vicinity of the clusters. In the presence of spatial correlation of the residuals it would be more appropriate to use generalised (weighted) least squares (Draper and Smith, 1981).

The fitting of polynomial models lends itself to the inclusion of information on covariates, and thus the method of trend surface analysis has been popular for interpolating rainfall values, incorporating information on continentality, altitude, and other topographic features (Wolfson (1975), Hutchin-

son (1968), Hughes (1982), Dent *et al.* (1989)). However, a problem often encountered is that the relationship with the covariates may change across the study area and this may necessitate partitioning the area, which is in itself a major problem in that homogeneous areas must be delineated, and also leads to the subsequent problem of patching together the various fitted equations in a smooth way as described for example by Dent *et al.* (1989).

4.1.2 Harmonic Trend Analysis

This method is similar to trend surface analysis except that double Fourier series surfaces are used in place of polynomials. Thus we have

$$\begin{aligned} f(x,y) \;=\; & \sum_{m=0}^{M}\sum_{n=0}^{N} \lambda_{mn}(a_{mn}\cos(m\pi x/L)\cos(n\pi y/H) \\ & + b_{mn}\sin(m\pi x/L)\cos(n\pi y/H) + c_{mn}\cos(m\pi x/L)\sin(n\pi y/H) \\ & + d_{mn}\sin(m\pi x/L)\sin(n\pi y/H)) \end{aligned}$$

where the λ_{mn} are constants depending only on m and n and L and H are equal to one half of the distance covered by the data locations in the x and y directions respectively. The number of harmonics, M and N, are chosen by the user, while the values of a_{mn}, b_{mn}, c_{mn} and d_{mn} are determined by a least squares fit criterion. Harbaugh and Merriam (1968, Chapter 6) describe the methodology and give a comparison with polynomial trend surface fitting. James (1966) gives a FORTRAN program for fitting double Fourier series to irregularly spaced data.

Most of the criticisms made above for trend surface analysis apply also to harmonic trend analysis and the method is probably only appropriate when there is reason to expect some underlying periodicity in the trend.

4.1.3 Smoothing Splines

The idea of fitting local polynomial functions leads naturally to the concept of smoothing splines. There are various generalisations of spline smoothing to two dimensions, but the most commonly used is the thin-plate smoothing spline (two-dimensional Laplacian smoothing spline) which can be viewed as the function f which minimises the penalised least-squares expression

$$n^{-1}\sum_{i=1}^{n}[v_i - f(x_i, y_i)]^2 + \lambda J_2(f)$$

where

$$J_2(f) = \int\int [f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2]\, dx\, dy$$

(Duchon, 1976; Wahba and Wendelberger, 1980). The degree of smoothing is controlled by the smoothing parameter λ; if λ is set to zero, the solution will interpolate the original data. If there is measurement error in the data the smoothing parameter is usually selected by generalised cross-validation; software for this is available in GCVPACK (Bates *et al.*, 1987).

Unequal error variance in the data could be accommodated into the spline smoothing method by weighting the fit differently at individual points. Information on covariates, such as altitude, could possibly be included by using 'partial spline' models (Wahba, 1990a) so that

$$\hat{v}_i = f(x_i, y_i) + \sum_{j=1}^{k}\beta_j\psi_{ij}$$

where the β_j are parameters to be fitted and the ψ_{ij} are the covariate values.

4.1.4 Kriging and Optimal Interpolation

The technique of kriging was developed by Matheron (1963); the almost identical but less well known method of *optimal interpolation* was developed at about the same time by Gandin (1963). In these methods the data are

modelled as a realisation of a stochastic process with a covariance function which is assumed stationary, that is, dependent only on distance, at least locally, and the kriging predictor is derived as the minimum variance unbiased linear predictor. By explicitly modelling the covariance of the data points, the method is especially suited to clustered data exhibiting spatial autocorrelation.

If we use the model

$$v_i = \tau_i + \eta_i \tag{4.1}$$

where τ represents large-scale trend and η represents the local spatially correlated component, then kriging provides an estimator of v_0 of the form

$$\sum_{i=1}^{n} w_i v_i$$

where the weights w_i are chosen to minimise the expected squared error of estimation, that is, to minimise

$$\mathrm{E}\left[(\sum_{i=1}^{n} w_i v_i - v_0)^2\right]$$

In the case of so-called *simple* kriging, the data are assumed to be detrended so that the τ terms may be assumed to be zero. In this case the solution is given (Section 5.1.1) by

$$\mathbf{Cw} = \mathbf{c} \tag{4.2}$$

so that

$$\hat{v}_0 = \mathbf{c}'\mathbf{C}^{-1}\mathbf{v} \tag{4.3}$$

where the matrix $\mathbf{C}$ has elements $c_{ij} = \mathrm{cov}(v_i, v_j)$ and $c_i = \mathrm{cov}(v_i, v_0)$.

More generally, the trend term is either assumed to be a constant ('ordinary kriging'), or else modelled as a polynomial in x and y as in trend surface analysis ('universal kriging'). Full details are given in the next chapter.

Although kriging is almost invariably used in applications as an interpolator, it can also be utilised as a smoother; this aspect is discussed in Chapter 6.

The method of kriging can be extended to the situation where data on covariates are also available, using either *co-kriging*, in which the estimate is given by $\hat{v}_0 = \sum w_i v_i + \sum \tilde{w}_j u_j$ where the u_j are the covariate values, or alternatively by incorporating the covariates as part of the trend function. These two options are discussed further in the next chapter.

4.1.5 Kernel Smoothing

These methods use a simple weighted average of the neighbouring data points, with the weights being chosen as some (inverse) function of distance or *kernel* function. Thus to estimate the value v_0 at the location $\mathbf{z}_0$ based on values v_i at locations $\mathbf{z}_i$ we have

$$\hat{v}_0 = \sum_{i=1}^{n} w_i v_i$$

where the weights w_i are given by

$$w_i = \frac{c_0}{\lambda} K(d_{i0}/\lambda)$$

where K is the chosen kernel function (a decreasing function of distance) and λ is known as the *bandwidth*. The constant c_0 is usually included to ensure that the weights sum to unity (Hastie and Tibshirani, 1990). Interpolation of the data is achieved by choosing a weighting function which tends to infinity as the distance tends to zero. In general, the degree of smoothing is determined by the bandwidth and the rate of decay of the kernel function.

Kernel smoothing methods are computationally simple and do not require the assumption of any functional form for the underlying trend. They are however inappropriate for clustered data which exhibit short-scale spatial

correlation, since in this case the clusters tend to dominate the smooth in their vicinity, leading to bias. For one-dimensional data Gasser and Muller (1979) proposed a form of kernel smoothing which to some extent overcomes this problem by using as the weights, not the value of the kernel function itself, but rather the integral of this function over an interval around each data point, so that the weight w_i is given by

$$w_i = (c_0/\lambda) \int_{s_{i-1}}^{s_i} K((s - x_0)/\lambda)\, ds$$

where the s_i are the midpoints between the ordered data points so that $s_i = \frac{1}{2}(x_i + x_{i+1})$. This would however seem to give undue weight to the points on the edge of each cluster which would have a relatively large range of integration compared to the interior points. Furthermore, extension of their method to two dimensions would destroy the attractive computational simplicity of kernel smoothing as it would necessitate the integration of the kernel function over an area or 'tile' around each data point, where a tile consists of the set of points of the plane which is closer to that data point than to any other; this in turn requires a Dirichlet tessellation of the plane (Green and Sibson, 1978).

If the purpose of the smoothing is to remove large-scale trends only, then we need to allow for the fact that the residuals will be auto-correlated; this leads to difficulties in the choice of the optimum smoothing parameter. As with trend surface analysis, ignoring the auto-correlation leads to a smooth which follows the data too closely; Hart (1991) discusses this problem in the one-dimensional context.

Another disadvantage of kernel-based methods for estimating the parameters of the rainfall model or for smoothing the SABAP data is that it is not clear how one might incorporate covariate information. It is also not obvious how one would incorporate information on heterogeneity of the measurement error into these methods, although a possible approach might be to multiply

the weight of each data point by some inverse function of the variance.

4.1.6 Median Polish

Cressie (1986) proposed a method of smoothing of gridded data using the median polish technique of Tukey (1977). Trend is estimated as the sum of row and column effects which in turn are estimated by successively extracting the row and column medians until convergence is obtained, that is, the row and column medians of the residuals are zero. This technique is intended to be used to remove trend so as to produce a mean-stationary set of residuals which can then be used in the kriging process. While this does eliminate some of the difficulties involved in kriging non-stationary data, and is suitable for arbitrary trend functions, its use as a general-purpose smoother is somewhat limited in that there is no control over the degree of smoothing, and the method is most appropriate for data which lie, at least approximately, on a regular grid.

4.1.7 Multiquadric Surface Interpolation

A predictor for two-dimensional data based on the fitting of multi-quadric surfaces was proposed by Hardy (1971). The surface to be interpolated is represented by the summation of the heights of a series of n quadric surfaces, where the ith surface has its vertex at the ith data point. The parameters of the individual surfaces, which may be circular hyperboloids of two sheets, paraboloids or cones, are determined in such a way as to ensure that the final surface interpolates the original data. Lee *et al.* (1974) tested several types of quadric surface for the estimation of areal rainfall and concluded that the cone was the most appropriate choice of surface, giving good estimates and being simple to compute.

For the circular cone with vertex at (x_i, y_i) given by

$$v^2 = c^2((x - x_i)^2 + (y - y_i)^2)$$

the height at a point with coordinates (x_j, y_j) is given by

$$v_j = c_j d_{ij}$$

where d_{ij} is the distance between (x_i, y_i) and (x_j, y_j). Thus if the sum of the heights of the cones is to interpolate the original data, the constants c_i must be determined by the equations

$$\mathbf{v} = \mathbf{D}\mathbf{c}$$

where the elements of the $n \times n$ matrix $\mathbf{D}$ are the inter-point distances, and the elements of the vector $\mathbf{v}$ are the data values. From this we see that $\mathbf{c} = \mathbf{D}^{-1}\mathbf{v}$ and hence

$$\hat{v}_0 = \mathbf{d}'\mathbf{D}^{-1}\mathbf{v} \tag{4.4}$$

where $\mathbf{d}$ is the vector of distances d_{i0} between (x_i, y_i) and (x_0, y_0). The method is specifically designed to interpolate the original data and is thus not a general purpose smoother.

4.1.8 Natural Neighbour Interpolation

Sibson (1981) proposed a method of interpolation which he called natural neighbour interpolation. The method results in a fitted surface which is continuously differentiable and reproduces exactly an underlying spherical quadratic surface. In one dimension it reduces to a Hermite cubic interpolant.

As a first step in the calculation for two dimensional data, it is necessary to calculate what Sibson calls the *local coordinates* for each of the n data points, which are based on the areas of the polygons or 'tiles' resulting from a number of Dirichlet tessellations. Specifically, if we define T_i to be the tile

around the ith data point in a Dirichlet tessellation of the data locations, and T_{ij} to be that part of T_i that is closest to the jth location if the ith data point is excluded from the tessellation, and similarly define T_0 and T_{0j} when the location $\mathbf{z}_0$ (of the point at which an estimate is required) is included with the data points. Then if we let κ_{ij} represent the area of the tile T_{ij} and similarly for κ_i, κ_0 and κ_{0j}, then the local coordinates for the jth data point are given by $\lambda_{0j} = \kappa_{0j}/\kappa_0$ and $\lambda_{ij} = \kappa_{ij}/\kappa_i$.

These local coordinate values form the basis for the calculation of the estimate at the location $\mathbf{z}_0$; full details are given by Sibson (1981). Unfortunately no explicit representation for the λ values is readily available; this makes comparison with other methods difficult. Sibson (1980) mentioned the possibility of extending the technique to provide a method of smoothing, but as yet no such extension appears to have been published.

4.2 Comparison of Smoothing Methods

Discussions of most of these methods as well as some other simpler methods, together with further references, can be found in Ripley (1981) and in Cressie (1991).

Several researchers have compared some or all of these methods on real and simulated data. Creutin and Obled (1982) tested splines, optimal interpolation and kriging, amongst other methods, to estimate rainfall amounts in southern France, while Tabios and Salas (1985) used trend surface analysis, kriging, optimal interpolation, multi-quadric interpolation and inverse-distance averaging to estimate annual precipitation in Nebraska and Kansas. Both studies used a group of known sites as test sites to evaluate the various methods being tested. Generally the more sophisticated methods gave better results. In their conclusions, Creutin and Obled recommended optimal interpolation while Tabios and Salas recommended optimal interpolation or

kriging.

Despite the apparent differences between the various methods described above, there are numerous connections between them; some of these have been mentioned in the discussion above. With the exception of natural neighbour interpolation and median polish, all the methods can be expressed as linear functions of the data which makes detailed comparison feasible. To facilitate such comparison first specify a general model as

$$v_z = \tau_z + \eta_z + \epsilon_z$$

where τ is a large-scale trend function, η is an autocorrelated spatial process and ϵ is white noise.

In trend surface analysis, the form of τ is taken to be a polynomial in z, and the objective is to predict τ with η either taken to be zero (ordinary least squares), or to have known covariance (generalised least squares). Goldberger (1962) considered the prediction problem, that is, the estimation of $\tau+\eta$, and his formulation is identical to so-called *universal kriging* (see Section 5.1.3). He did not, however, discuss the problem of estimation of the covariance functions of the model components.

In kriging, the objective is generally to predict $\tau + \eta$, although, as we shall see in Chapter 6, the method may also be used to filter out trends, that is, to estimate τ alone. In most practical applications ϵ, the measurement error or 'noise', has been assumed to be zero, which seems rather unrealistic in many cases. Various assumptions may be made about the form of the trend: in *simple* kriging, τ is assumed zero, in *ordinary* kriging, it is assumed constant, while in *universal* kriging and *IRF-k* kriging it is generally assumed to be a low order polynomial, at least locally. The covariance of η is assumed to be some function of distance which is estimated from the data. In Gandin's optimal interpolation method the mean is separately estimated so that subsequent estimation is equivalent to simple kriging.

Although the thin-plate smoothing spline is not formulated as a model based method, it has been shown that it is equivalent to kriging with a particular covariance function and (local) trend model. (Kimeldorf and Wahba, 1970; Watson, 1984). The relationship between the two methods is also discussed in some detail by Cressie (1990) and Wahba (1990). Thus the smoothing spline can be seen as an estimator of $\tau + \eta$ with the covariance of η assumed to have a specific functional form. If it is required to estimate τ alone, then the spatial autocorrelation of η presents a problem for the generalised cross-validation (GCV) approach as discussed by Diggle and Hutchinson (1989). They show that, in the presence of autocorrelation of the errors, the GCV-based smoothing spline is inconsistent and suggest a modification using penalised maximum likelihood; this works reasonably well when the trend function has a relatively low frequency compared with the 'noise' component, and provided that the autocorrelation function is known.

Kernel smoothers can be considered as estimators of a non-parametric trend function, that is, no explicit assumption is made about the form of τ, except that it is reasonably smooth. Similarly, no explicit assumption is made about the error and thus in general kernel based methods will work best on irregularly spaced data only when η is zero, otherwise, as mentioned previously, undue weight is given to spatially clustered, and thus highly correlated, observations. Silverman (1985) discusses the relationship between splines and kernel smoothers and shows that the one-dimensional cubic smoothing spline is (approximately) equivalent to a kernel smoother with a bandwidth which is varied according to the local density at each data point used in the estimation, so that more clustered points are thus down-weighted.

The multi-quadric interpolator can also be expresses as a linear estimator. By comparing the solution given in equation 4.4 with that in equation 4.3 we see that the multi-quadric interpolator is equivalent to simple kriging with a linear covariance function. Thus multi-quadric surface interpolation can be

viewed as a special case of kriging, with a pre-specified covariance structure.

4.3 Selection of a Smoothing Method

Some the simpler methods described in the first section of this chapter are inappropriate for irregularly spaced data, or for data exhibiting autocorrelation, and are thus totally unsuitable for the problems being described in this thesis. Thus the choice would appear to be essentially between a spline-based approach or a geostatistical approach. The geostatistical approach was selected for the following reasons:

- Kriging was specifically developed for the prediction of random variables in the presence of spatial autocorrelation and irregular spacing of data points; the appropriate degree of smoothing is based on the spatial covariance function which is estimated directly from the data.

- The model based formulation of kriging makes it suitable for the accommodation of a varying error variance and for extension to the interpolation of binomial data, such as the reporting rates of the SABAP.

- Several techniques for including information on covariates (via co-kriging or kriging with external drift) already exist; in particular the co-kriging approach makes it possible to relate the rainfall at one point to the altitude at neighbouring points, thus largely obviating the need to try to define functions such as 'exposure'.

- The factorial kriging approach (to be discussed in Chapters 6) allows one to use kriging as a filter to extract trend or other components from the overall data 'signal'. In the rainfall model this allows us to separate the large scale effects from the more local effects due to topography.

- Kriging allows for the calculation of a local measure of precision associated with each estimated value.

Several problems still remain in adapting kriging to our specific needs, in particular to cater for binomial data and for circular data, and to allow for the widely varying error variances. These topics are the subject of chapters 6 to 9.

Chapter 5

Kriging

The stochastic modelling approach to the interpolation of spatial data was developed by Matheron (1963) and independently by Gandin (1963) who drew on earlier work of Kolmogorov (1941), and the branch of statistics known as *geostatistics* has grown out of Matheron's work. The books by Clark (1979), Cressie (1991) and Isaaks and Srivastava (1989) describe current theory and practice in some detail.

The geostatistical model for spatial data assumes that observations are available for a *regionalised variable*, that is, a random variable observed at a number of spatial locations. In general, the best predictor of the value v_0 at some location $\mathbf{z}_0$ based on values $v_1, ..v_n$, in the sense of minimising the expected squared error of prediction, is given by the conditional expectation of v_0, given $v_1, ..v_n$. If v is a Gaussian random process, then the required conditional expectation is linear in $v_1, ..v_n$, and thus the usual starting point is to look for a best *linear* predictor, that is, to find weights $w_1, ..w_n$ to minimise $E[\hat{v}_0 - v_0]^2$, where $\hat{v}_0 = \sum w_i v_i$. Matheron coined the name ***kriging*** for this process. The main aspects of the theory and practice of kriging are outlined in the remainder of this chapter.

It has been common practice amongst users of geostatistics to assume that the data are free of measurement error; we will make this assumption

throughout this chapter, and defer the discussion of measurement error until the following chapter. Thus the usual basic model for the geostatistical approach can be written as

$$v_{\mathbf{z}} = \tau_{\mathbf{z}} + \eta_{\mathbf{z}}$$

where z is a location in (usually) two or three dimensional space, τ is a large-scale trend function while η is an autocorrelated spatial process with mean zero.

5.1 The Kriging Equations

Several variants of the kriging method have developed, differing mainly in their approach to the modelling of the trend and the estimation of the covariance of the process $\eta_{\mathbf{z}}$ in the presence of trend, and these are discussed below. A further extension of the kriging method allows for the inclusion of information on a covariate or covariates and this is discussed in Section 5.3.1.

5.1.1 Simple Kriging

In the case of *simple* kriging, it is assumed that the trend term is zero, so that

$$E[v_{\mathbf{z}}] = 0$$

and thus

$$\begin{aligned} E(\hat{v}_0 - v_0)^2 &= E(\sum_{i=1}^{n} w_i v_i - v_0)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{ij} - 2\sum_{i=1}^{n} w_i c_{i0} + c_{00} \end{aligned}$$

where $c_{ij} = \mathrm{cov}(v_i, v_j) = \mathrm{cov}(\eta_i, \eta_j)$.

By taking derivatives with respect to the w_i and setting these equal to zero, we obtain the equations

$$2[\sum_{j=1}^{n} w_j c_{ij} - c_{i0}] = 0 \qquad i = 1, \ldots, n$$

so that the solution for the w_i is given by

$$\mathbf{w} = \mathbf{C}^{-1}\mathbf{c_0} \tag{5.1}$$

where the matrix C has elements c_{ij} and the vector $\mathbf{c}_0$ has elements c_{i0}.

This solution assumes that the covariance function is known; more realistically it must be estimated from the data. This is discussed in Section 5.2.

It is easy to see that the solution given above interpolates the data, since if $\mathbf{z}_0 = \mathbf{z}_j$ for some j then the vector $\mathbf{c}_0$ will be equal to the jth row and column of the matrix C so that the weight vector $\mathbf{w}$ will have a 1 in the jth position and zeros elsewhere and thus $\hat{v}_0 = v_j$.

5.1.2 Ordinary Kriging

In simple kriging one assumes that $E[v_{\mathbf{z}}] = 0$. A slightly more realistic model in practice would be

$$E[v_{\mathbf{z}}] = \mu$$

where μ is some constant. In order to ensure that the estimator will be unbiased, we now introduce the constraint $\sum_{i=1}^{n} w_i = 1$.Thus

$$\begin{aligned} E(\hat{v}_0 - v_0)^2 &= E(\sum_{i=1}^{n} w_i v_i - v_0)^2 \\ &= E(\sum_{i=1}^{n} w_i(v_i - \mu) - (v_0 - \mu))^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{ij} - 2\sum_{i=1}^{n} w_i c_{i0} + c_{00} \end{aligned}$$

as before. Using the Lagrange multiplier technique, we then set

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{ij} - 2\sum_{i=1}^{n} w_i c_{i0} + c_{00} + \lambda(\sum_{i=1}^{n} w_i - 1)$$

where λ is the Lagrangian parameter. Taking the derivative of G with respect to each of the w_i and λ and equating the derivatives to zero gives the equations:

$$\begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{c}_0 \\ 1 \end{pmatrix} \tag{5.2}$$

where the matrix $\mathbf{C}$ has elements c_{ij} and the vector $\mathbf{c}_0$ has elements c_{i0}, as in the simple kriging solution.

In practice it is usual to model the covariance indirectly via the *semi-variogram*, which is defined in Section 5.2. However, although the kriging equations can be written in terms of the semi-variogram, the covariance is generally used in algorithms for solving the kriging equations since the largest elements are then located on the diagonal, leading to greater numerical stability for systems based on Gaussian elimination.

5.1.3 Kriging in the Presence of Trend

A further relaxation of the assumptions made in the previous section is obtained by allowing the mean to vary across the study area. Thus the model becomes

$$v_{\mathbf{z}} = \mu_{\mathbf{z}} + \eta_{\mathbf{z}}$$

where

$$E[\eta_{\mathbf{z}}] = 0$$

Various approaches to the estimation of v_0 are possible. One common practice is to assume a simple functional form for the trend, for example, a quadratic

function of the x and y coordinates. Thus we have, at any location $\mathbf{z}$,

$$v_{\mathbf{z}} = \sum_{k=1}^{p} f_k(\mathbf{z})\beta_k + \eta_{\mathbf{z}}$$

where the $f_k(\mathbf{z})$ are functions of x and y and the β_k are coefficients to be determined. The problem is then equivalent to a generalised least squares prediction problem, and we can write, in the usual regression notation,

$$\mathbf{v} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$$

By using the Lagrange multiplier technique as above, and introducing the unbiasedness constraint $\mathbf{w}'\mathbf{X} = \mathbf{x_0}$, where $\mathbf{x_0}$ is the vector of values $f_k(\mathbf{z}_0)$ at the location $\mathbf{z}_0$ for which an estimate is sought, we obtain the solution for the $\mathbf{w}$ as:

$$\begin{pmatrix} \mathbf{C} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ -\boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{x}_0 \end{pmatrix} \tag{5.3}$$

where $\boldsymbol{\lambda}$ is a $p \times 1$ vector of Lagrange multipliers, $\mathbf{0}$ is a $p \times p$ matrix of zeros, $c_{ij} = \mathrm{cov}(\eta_i, \eta_j)$ and $\mathbf{c}_0$ has elements $\mathrm{cov}(\eta_i, \eta_0)$.

By partitioning the left hand matrix in the equation above the solution for $\hat{v}_0 = \sum w_i v_i = \mathbf{w}'\mathbf{v}$ can also be written (Goldberger, 1962 or Stein and Corsten, 1991) as

$$\hat{v}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} + \mathbf{c}_0'\mathbf{C}^{-1}(\mathbf{v} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5.4}$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$. Comparing this with equation 5.1 we can see that the procedure is equivalent to generalised least squares regression combined with simple kriging of the residuals.

There are several difficulties with this approach, named *universal kriging* by Matheron. Firstly, an appropriate form for the trend function may not be known and a simple polynomial may not suffice throughout the study area, and secondly, the covariances required in the solution are those of the regionalised variable η, but η itself is not directly observable, since only v is

actually observed. The usual solution to this latter problem is to use an iterative approach, starting with an ordinary least squares analysis of the data (trend surface analysis), followed by modelling the covariance of the residuals, and then using the generalised least squares analysis as given above, re-modelling the residuals, and so on until convergence is reached. An example of an application to the estimation of regional groundwater levels is described by Neuman and Jacobson (1984). However this procedure leads to a slight bias in the estimation of the semi-variogram function (Cressie, 1991). Alternatively, one could attempt to model the trend and covariance simultaneously using a maximum likelihood approach as suggested by Vecchia (1988). This method was implemented and tested for some of the daily rainfall model data described in Chapter 2, using fairly simple parametric models for both the trend and the residual covariance structure, but computational difficulties were encountered which seemed to be due to the likelihood function being rather flat and having numerous local maxima (Harrison[1], 1989). This problem has been noted and discussed by other authors (Warnes and Ripley, 1987; Mardia and Watkins, 1989).

An alternative approach, suggested by Cressie (1986), is to use median polish to remove the trend and then to use ordinary kriging on the residuals. This corresponds to a non-linear estimate of the trend function. As mentioned in the previous chapter, median polish is most appropriate for gridded or approximately gridded data.

Matheron (1973) takes another approach to the problem by defining *generalised increments* of order k, which are linear combinations of the data which are chosen to filter out polynomial trends of order k. This allows for the definition of *generalised covariances*, such that the covariance of two generalised increments can be written as a linear combination of the generalised

[1] An honours student project supervised by the author of this thesis.

covariances and the kriging is carried out on these generalised functions. The procedure has been largely automated in the software package BLUEPACK, (Delfiner *et al.*, 1978) although the order k must be specified by the user. However, some authors have suggested that the parametric covariance modelling procedure used in BLUEPACK is not entirely satisfactory (Starks and Fang, 1982; Zimmerman and Zimmerman, 1991). In addition, unless the data locations lie on a grid, to obtain the generalised increments of the data one must first group adjacent data points and solve a set of linear equations to obtain each increment (Delfiner, 1975).

In practice, many statisticians use *local* or *moving-window* kriging, that is, only data points within a certain distance of the point to be estimated are included in the kriging estimation. The assumption is then that $\mu_{\mathbf{z}}$ will not vary much within this neighbourhood, so that *local* stationarity may be assumed, and the ordinary kriging model described in the previous section may thus be applied. Local kriging is common practice in any case even when there is no trend in the mean, for purely computational reasons, that is, to avoid having to invert a very large covariance matrix. Journel and Rossi (1989) show that local kriging in the presence of apparent trend produces estimates which are virtually identical to those produced by kriging with a simple polynomial trend model, except in those situations where one attempts to extrapolate beyond the spatial boundaries of the data.

A further possibility for kriging in the presence of trend, based on treating the trend as another, large scale, random process, and partitioning the overall semi-variogram in a commensurate way, is suggested in the next chapter.

5.2 Estimation of the Covariance Function

In the kriging equations as given above it has been assumed that the covariance is known. In practice it is estimated from the data, and, since only one

value of the regionalised variable v is usually available at each data location it is usual to make some assumption of second-order stationarity which implies that the covariance between two points will be dependent, not on their specific locations, but only on the vector distance between them.

Thus the spatial covariance is defined as:

$$\sigma(\mathbf{h}) = E[(v_{\mathbf{z}} - \mu_{\mathbf{z}})(v_{\mathbf{z}+\mathbf{h}} - \mu_{\mathbf{z}+\mathbf{h}})]$$

where μ denotes the mean value at a given location.

Matheron (1963) develops the theory of geostatistics on a slightly less restrictive set of assumptions known as the ***intrinsic hypothesis*** which does not actually require the existence of a covariance function, but only a semi-variogram function, defined as

$$\gamma(\mathbf{h}) = \frac{1}{2}E[(v_{\mathbf{z}} - v_{\mathbf{z}+\mathbf{h}})^2] \tag{5.5}$$

The term 'semi-variogram' is due to Matheron although its use had been recommended earlier in a time series context by Jowett (1955). Even when the covariance exists there are several advantages in working with the semi-variogram, using the estimator

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N_{\mathbf{h}}}\sum(v_{\mathbf{z}_i} - v_{\mathbf{z}_j})^2$$

where the summation is over all $N_{\mathbf{h}}$ pairs which are a vector distance h apart. In practice, for non-gridded data, the summation is calculated over all pairs belonging to specified distance *intervals*, for example, [0-1) km, [1-2) km, [2-3) km etc. If the spatial continuity is more marked in some directions than others, then it is necessary to calculate separate semi-variograms for each direction, but often there is no directional effect so that one need only consider the distance $h = \| \mathbf{h} \|$.

One of the advantages of working with the semi-variogram is that its estimation does not require any prior estimate of the mean or trend function.

The effect of trend on the semi-variogram is dependent on the exact form of the trend. If the trend is constant then it is clear from equation 5.5 that it will cancel out of $\gamma(h)$. In general the trend will vary little over small distances, so that the effect of the trend will be relatively minor for small lag distances, these being the more critical values for the kriging process, since in practice it is usual to obtain the kriging estimate at a given location using only the data points which are in the vicinity of that location. If the trend is not constant, for example if a linear deterministic trend is present, then the variance of the process may not show an upper bound, as illustrated in Figure 5.1. Additional reasons for preferring to use the semi-variogram rather than the covariance function are given by Srivastava (1988) and Cressie and Grondona (1992).

As defined by Matheron's *intrinsic hypothesis* the semi-variogram exists for a wider class of processes than the covariance function; however, for a second-order stationary process where both exist there is a simple relationship between them, given by

$$\gamma(h) = \sigma(0) - \sigma(h)$$

so that, having estimated the semi-variogram, one may readily obtain the corresponding covariance to be used in the solution of the kriging equations.

In order to ensure that the fitted semi-variogram corresponds to a positive-definite covariance function, it is usual to select a parametric model from one of several standard models which are known to have this property. Christakos (1984) gives details of necessary and sufficient conditions to ensure a valid variogram model. Four of the most commonly used models are the *nugget*, *linear*, *spherical*, *exponential* and *Gaussian* models:

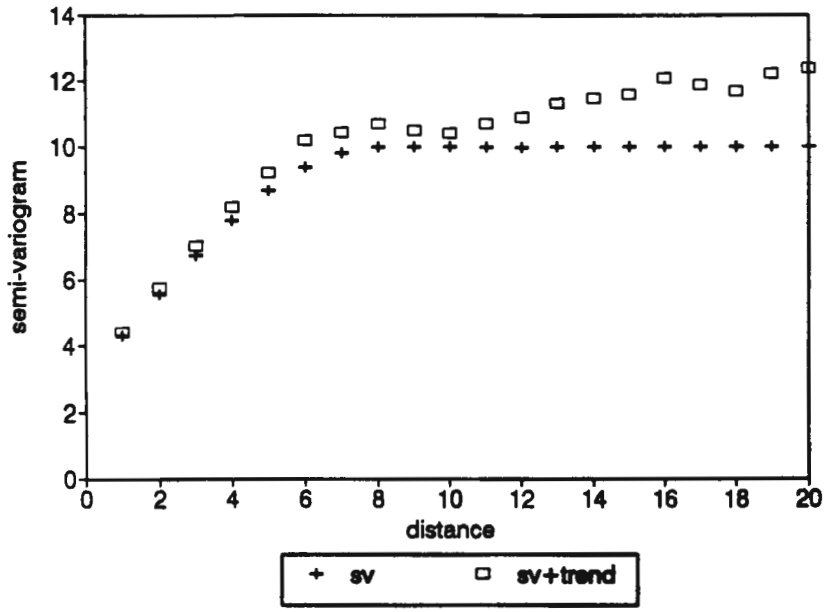


Figure 5.1: Effect of trend on the semi-variogram.

Nugget:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c & h > 0 \end{cases}$$

Linear:

$$\gamma(h) = ah \qquad h \geq 0$$

Spherical:

$$\gamma(h) = \begin{cases} s[(3/2)(h/r) - (1/2)(h/r)^3] & 0 \leq h \leq r \\ s & h > r \end{cases}$$

where s, the asymptote, is commonly known as the *sill*, while r, the *range*, indicates the maximum extent of the spatial correlation.

Exponential:

$$\gamma(h) = s(1 - \exp(-h/r)) \qquad h \geq 0$$

where s is the sill and r is a range parameter. The *effective range*, that is, the range at which the value of γ reaches 95% of the sill, is $3r$.

Gaussian:

$$\gamma(h) = s(1 - \exp(-h^2/r^2)) \qquad h \geq 0$$

where s is the sill and r is the range parameter, the effective range for this model being $\sqrt{3}r$.

These models are illustrated in Figure 5.2. In practice, a combination of models is frequently used, for example, a nugget model plus a spherical model.

Zimmerman and Zimmerman (1991) review and compare a number of methods of fitting semi-variogram models. The method used for applications described in this thesis was the weighted least squares method of Cressie (1985), which Zimmerman and Zimmerman found to perform consistently well, and which weights points according to the number of data pairs on which they are based, and also gives more weight in the fitting process to points at smaller lag distances. Thus the parameters of the semi-variogram model are chosen to minimise the expression

$$\sum_{i=1}^{k} N_i \left(\frac{\hat{\gamma}_i - \gamma_i}{\gamma_i} \right)^2$$

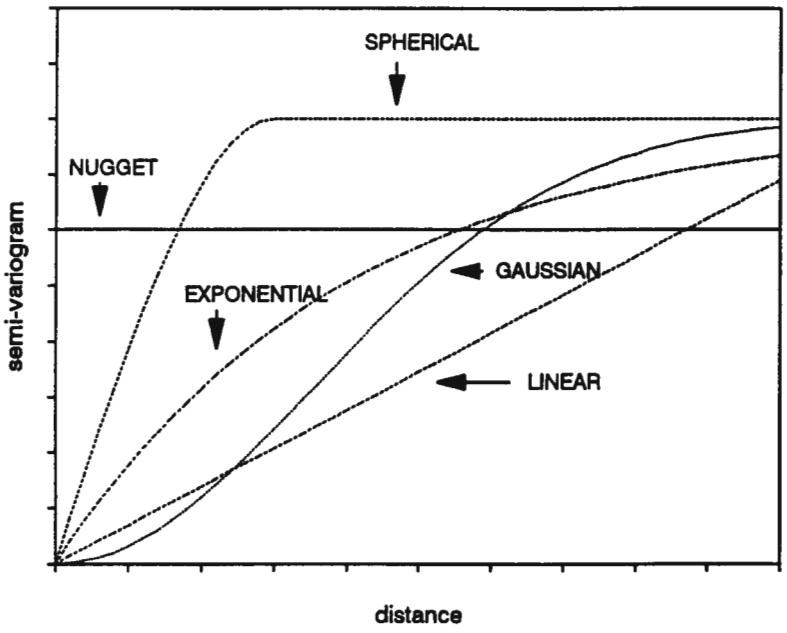


Figure 5.2: Semi-variogram models.

summed over the k lag distances, where γ_i is the value given by the model and $\hat{\gamma}_i$ is the value estimated from the data.

Although the kriging equations may be re-written in terms of the semi-variogram function it is common practice to estimate the semi-variogram from the data and then convert this to a covariance function prior to solving the kriging equations, as this has computational advantages, as mentioned earlier.

Clearly the optimality claims made for the kriging estimator rest on the assumption of correctly specified models. The effects of errors in the estimation of the semi-variogram are not normally included in the calculation of the kriging variance. Some research has been done on the effects of mis-specification, for example Stein (1988) and Stein and Handcock (1989).

Cressie (1991) notes that while model mis-specification will lead to a negative bias in the mean square prediction error, in practice other factors such as strong spatial correlation and a tendency to underfit the trend may act as compensatory factors.

5.3 Kriging with Covariate Information

In some situations we may have values of more than one regionalised variable; if the variables are correlated we may then be able to incorporate the information on all of the variables to improve prediction accuracy. The different variables need not all be measured at the same sites. For example, in estimating rainfall parameters one might wish to include information on the altitude, which is available on a grid of 1 minute of latitude by 1 minute of longitude. There is no difficulty in theory in incorporating such additional information, and several approaches are possible.

5.3.1 Co-Kriging

One possibility is to use co-kriging (Matheron, 1971) in which the covariates are essentially treated as an extension of the data vector so that the solution (for a single covariate) is a weighted sum of values of the variate to be interpolated and the values of the covariate. Thus we have

$$\hat{v}_0 = \sum_{i=1}^{n} w_i v_i + \sum_{j=1}^{m} \tilde{w}_j u_j$$

with unbiasedness constraints $\sum w_i = 1$ and $\sum \tilde{w}_j = 0$.

Assuming a no-trend model for the data, the co-kriging weights are given

by

$$\begin{pmatrix} \mathbf{C_{vv}} & \mathbf{C_{vu}} & 1 & 0 \\ \mathbf{C_{uv}} & \mathbf{C_{uu}} & 0 & 1 \\ 1' & 0' & 0 & 0 \\ 0' & 1' & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \tilde{\mathbf{w}} \\ -\lambda_v \\ -\lambda_u \end{pmatrix} = \begin{pmatrix} \mathbf{c_{v0}} \\ \mathbf{c_{u0}} \\ 1 \\ 0 \end{pmatrix}$$

where the matrix $\mathbf{C_{vu}}$ is now the $n \times m$ matrix of cross-covariances, $\mathbf{C_{uv}} = \mathbf{C_{vu}}'$ and the vector $\mathbf{c_{u0}}$ also contains cross-covariances (of the u_i with v_0), while λ_v and λ_u are Lagrangian parameters. More generally, a polynomial trend model may be included, as in universal kriging. Stein and Corsten (1991) show how co-kriging with a polynomial trend function may be expressed as a generalised least squares predictor.

In using co-kriging, the covariates need not be available at the same points as the variate of interest, nor at the sites to be estimated, although the locations of the covariate information do affect the method of estimation of the cross-covariance function, as explained below. Co-kriging is generally most valuable when the covariates are sampled more intensely than the predictand. An application of co-kriging to the estimation of rainfall data is described by Krajewski (1987).

In order to use the co-kriging approach it is necessary to model the *cross-covariance* of u and v in addition to their respective covariances. The spatial cross-covariance of two variates v and u is defined as

$$\sigma_{vu}(\mathbf{h}) = E[(v_{\mathbf{z}} - \mu^v_{\mathbf{z}})(u_{\mathbf{z}+\mathbf{h}} - \mu^u_{\mathbf{z}+\mathbf{h}})]$$

where μ^v and μ^u denote the mean values, of v and u respectively, at the relevant locations.

While the cross-covariance function may be estimated directly it is natural, in view of the advantages of the semi-variogram over the ordinary covariance function, to look for an appropriate way of defining a *cross semi-variogram*. The traditional definition (see for example Journel and Hui-

jbregts, 1978) is

$$\gamma_{vu}(\mathbf{h}) = \frac{1}{2}E[(v_{\mathbf{z}} - v_{\mathbf{z}+\mathbf{h}})(u_{\mathbf{z}} - u_{\mathbf{z}+\mathbf{h}})]$$

In order to estimate this function the variables v and u must be available at a number of common locations. The function is also symmetric in v and u. This implies that $\gamma_{vu}(\mathbf{h}) = \gamma_{vu}(-\mathbf{h})$ and this is not always appropriate; for example, in studying the relationship between rain and altitude, it is generally the windward slopes of mountains which receive more rain, so that the direction of a mountain in relation to a point at which rainfall is to be estimated cannot be ignored.

Myers (1982) suggested making use of the fact that the cross-covariance of the sum of the values $v+u$ can be expressed as the sum of the covariances of the individual components plus twice the cross-covariance. Thus if one fits models for the semi-variograms of v, u and $v+u$, and uses these to estimate the corresponding covariances, then the cross-covariance is readily obtained. However, to obtain the semi-variogram of $v+u$, one again needs to have a number of common data locations. This method is also symmetric in v and u, and thus suffers from the same problem as the previous definition in that it does not cater for directional effects. Armstrong (personal communication) comments that the use of this approach occasionally leads to inadmissible models, for example models with negative variances.

Clark *et al.* (1989) have suggested that a better definition of the cross semi-variogram is

$$\gamma_{vu}(\mathbf{h}) = \frac{1}{2}E[(v_{\mathbf{z}} - u_{\mathbf{z}+\mathbf{h}})^2]$$

Use of this definition does not require common data locations; furthermore the definition is *not* symmetric in v and u. They suggest that the two variables first be standardised to zero mean and unit variance so that values are commensurate, since gross differences in scale could adversely affect the precision of computations. A disadvantage with this definition is that it is

not possible to establish a relationship between the cross semi-variogram and the cross-covariance without estimating the mean value for each of the two variables.

5.3.2 Kriging with External Drift

An alternative to the co-kriging approach is to include the covariates as part of a trend function, which is essentially similar to the kriging formulation given in Section 5.1.3 so that, in the equation

$$v_{\mathbf{z}} = \sum_{l=1}^{p} f_l(\mathbf{z})\beta_l + \eta_{\mathbf{z}} \tag{5.6}$$

some of the functions f_l may be functions of the covariates. Thus, for example, the functions f_l might be functions of altitude as well as latitude and longitude. This approach requires firstly that the form of the relationship between predictand (e.g. rainfall parameter) and covariate (e.g. altitude or function of altitude) is known or can be approximated by a simple function such as a polynomial, and also that the covariate information is available at all the sites at which the predictand is known and also at all those points at which an estimate of the predictand is required. This method is known as *kriging with external drift* (Galli and Meunier, 1987; Ahmed and de Marsily, 1987 and Renard and Nai-Hsien, 1988). It has recently been used by Armstrong (1992) to estimate monthly rainfall in Lesotho, using (sometimes estimated) annual rainfall as the covariate or 'drift', and by Hudson (1992) to estimate monthly temperatures in Scotland using elevation as the covariate.

The use of this approach means, in effect, that we would be using a generalised least squares multiple regression of the variable to be predicted (rainfall parameter) on some function of the covariate (altitude), together with ordinary kriging of the residuals. This would appear to bring us back to the problems previously mentioned for the multiple regression approach, namely

the need for restricting the regression calculations to homogenous sub-regions in which the β_l of equation 5.6 could be treated as constant, and also the need to pre-define the appropriate functions of the surrounding topography to be included in the regression model. The use of a moving-window (local) kriging approach as discussed in Section 5.1.3 avoids the first of these problems by re-calculating the estimates at each point using only data points within a limited neighbourhood, thus effectively re-fitting the regression at each point. While computationally intensive, the process is computationally stable, and does not produce the sharp discontinuities in the output that can occur at regional boundaries when regional regression models are used. By contrast, in co-kriging, if the relationship cannot be assumed to be constant throughout the area under study either the area must be partitioned or the cross-covariance function will have to be re-modelled locally prior to kriging each point; this latter option will increase the computational load very considerably, besides making it impractical for the user to interact fully with the modelling process, which is generally considered essential for the successful application of kriging.

The second problem mentioned above, that is the need to define suitable functions of altitude to use as the external drift, was overcome by using as the covariates orthogonal functions of the gridded altitude values in the vicinity of each point, thus effectively including all possible polynomial functions up to a given degree. This is described further in Chapter 7.

5.4 Variance of the Kriging Estimator

Using a general notation which encompasses ordinary kriging, universal kriging, kriging with external drift and co-kriging, we may write our model as

$$\begin{pmatrix} \mathbf{v} \\ \mathbf{u} \\ v_0 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \\ \mathbf{x}_0' & 0' \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\eta}$$

where

$$E[\boldsymbol{\eta}] = 0$$

and

$$\text{cov}[\boldsymbol{\eta}] = \begin{pmatrix} \mathbf{C} & \mathbf{c}_0 \\ \mathbf{c}_0' & c_{00} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{vv} & \mathbf{C}_{vu} & \mathbf{c}_{v0} \\ \mathbf{C}_{uv} & \mathbf{C}_{uu} & \mathbf{c}_{u0} \\ \mathbf{c}_{v0}' & \mathbf{c}_{u0}' & c_{00} \end{pmatrix}$$

where v_0 is the value to be predicted, the vector $\mathbf{v}$ is the vector of data values and $\mathbf{u}$ is the vector of covariates (included only in co-kriging), the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ contain trend function values (polynomials in latitude and longitude) or external drift values, or simply a vector of 1's for ordinary kriging, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of unknown parameters. Then it can be shown (Stein and Corsten, 1991) that the variance of the prediction error (for v_0) is given by

$$c_{00} - \mathbf{c}_0'\mathbf{C}^{-1}\mathbf{c}_0 + \mathbf{x}_a'(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{x}_a \tag{5.7}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{pmatrix}$$

and

$$\mathbf{x}_a = \mathbf{x}_0 - \mathbf{X}'\mathbf{C}^{-1}\mathbf{c}_0$$

In ordinary kriging, when the matrix $\mathbf{X}$ reduces to a vector of 1's, we can see from the equation above that the error variance depends only on the

covariance values, which in turn reflect the covariance model and the density of the data points in the vicinity of the point to be estimated However, when a trend or external drift function is used, equation 5.7 shows that the estimation variance will also depend on the values of the trend or drift function.

Chapter 6

Kriging as a Smoother

Some of the assumptions made in the previous chapters are clearly not appropriate for the data sets described in Chapters 2 and 3. In particular, both sets of data show measurement error, and the variance of this measurement error is not constant, being heavily dependent on the number of years of data in the case of the rainfall model or on the number of field cards in the bird atlas project. Although kriging in the presence of measurement error has been discussed in the literature (Cressie, 1988) such discussion does not deal with the case of non-constant variance. A further problem arises from the need to de-trend the rainfall data in order to study the effect of topography on local rainfall variability. The maps of the parameters (Figure 2.3) suggest that a simple polynomial will not be suitable as a model of trend, while the irregular spacing of the data points makes non-parametric smoothing methods such as kernel smoothing or median polish more difficult to implement. In this chapter some possible solutions to these problems are considered.

We now extend our discussion of kriging to the model

$$v_{\mathbf{z}} = \mu + \tau_{\mathbf{z}} + \eta_{\mathbf{z}} + \epsilon_{\mathbf{z}} \tag{6.1}$$

where

μ is a constant (the mean of the process over the whole study area)

τ_z is a large-scale stochastic process which is assumed to have zero mean and covariance some function σ_τ of distance, and which takes the place of a non-constant deterministic 'trend' in the usual geostatistical models

η_z is a small-scale or 'local' process with zero mean and covariance function σ_η which is also a function of distance and typically will have a much shorter range than σ_τ

ϵ_z represents spatially uncorrelated zero mean noise, typically measurement error, and which is independent of both τ and η.

Thus the large-scale variation, or 'trend', commonly modelled as a deterministic process, is here treated as another random variable, but with a spatial correlation extending over a longer range than that of the 'local' process.

The decomposition into trend (low frequency component) and local variation (high frequency component) is not uniquely defined, but is generally dependent on the scale at which the data are measured, and also on underlying processes which generate the data; for example, with rainfall data the trend might reflect the effect of large-scale weather generating systems in the atmosphere (global circulation patterns) while the local effects might result from local topographic variability. Some further guidance as to a possible separation of these components may be obtained from a study of the spatial covariance function or spectrum, or simply from a map of the data. For certain applications it may be appropriate to decompose the observations into more than the three components shown in equation 6.1, and the methods discussed in this chapter are readily extended to such situations. We assume that the large-scale process τ and the local process η are independent of one another. This will be a reasonable assumption in many applications where two separate underlying processes occur, as suggested above for the rainfall model. This independence assumption is not in fact necessary unless we

wish to use kriging as a filter to separate the components, as described in Section 6.4.

The measurement error term ϵ can be quite considerable in certain practical situations, often as a result of some type of sub-sampling. For example, in mining applications, chemical determinations of borehole core samples are typically done on sub-samples of the individual cores which are themselves not homogeneous. In the case of the rainfall model parameters introduced in Chapter 2 there is measurement error due to the fact that the model is fitted to a finite daily rainfall record. The variance of this error can be estimated by the bootstrap procedure described in Section 2.4. In the Southern African Bird Atlas Project discussed in Chapter 3 the species counts are binomially distributed and thus subject to the usual binomial sampling error which depends on the underlying probability and the sample size.

In many applications it will be reasonable to assume that the measurement error term ϵ has zero expectation and shows no spatial correlation. In practice it is quite common to ignore the error altogether, that is, the data are assumed to be error-free, so that interpolation methods are used and thus predictions at existing data locations 'honour' the data. However, when there is known to be error in the data it makes more sense to try and estimate values of $\tau + \eta$ rather than v itself, so that the resulting predicted grid or surface would *smooth* the original data.

This chapter begins by showing the effect of measurement error on the semi-variogram and considers the situation where the variance of the measurement error is not constant but is known or can be estimated for each data point. We then look at how the kriging equations need to be adjusted to estimate values of $\tau + \eta$, that is, estimation of the underlying error-free values. In the third section we look at kriging in the presence of large-scale variation in the case where a simple parametric model of the trend is inappropriate. In the fourth section we discuss how kriging may be used as a

filter to extract only the long-range or the local component.

6.1 Estimating the Components of the Semi-variogram

In the case where $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma_\epsilon^2$ and the ϵ_i are uncorrelated with one another and with the τ and η then it is easy to show that

$$E[(\tau_i + \eta_i + \epsilon_i) - (\tau_j + \eta_j + \epsilon_j)]^2 = E[(\tau_i + \eta_i) - (\tau_j + \eta_j)]^2 + 2\sigma_\epsilon^2 \quad (6.2)$$

so that we have

$$\gamma_{\tau+\eta+\epsilon}(\mathbf{h}) = \gamma_{\tau+\eta}(\mathbf{h}) + \sigma_\epsilon^2$$

so that the error term increases the semi-variogram by a constant amount equal to σ_ϵ^2 (Figure 6.1).

If, in addition, the long-range and local components are uncorrelated then we have also

$$\gamma_{\tau+\eta+\epsilon}(\mathbf{h}) = \gamma_\tau(\mathbf{h}) + \gamma_\eta(\mathbf{h}) + \sigma_\epsilon^2$$

Figure 6.1 shows a typical semi-variogram composed of spherical models for τ and η together with a measurement error component. Modelling such a composite semi-variogram presents no new problems, since in any case many geostatisticians commonly use nested models for semi-variogram estimation, as described in Section 5.2.

Unless there are repeated measurements at some locations, or some other way of independently estimating σ_ϵ^2, the error variance can only be estimated from the empirical semi-variogram by extrapolating the fitted model to the point $\mathbf{h} = \mathbf{0}$. As it is quite possible that there is also significant short-scale variation in η, commonly referred to as the *nugget effect*, such extrapolation may be quite inaccurate (Figure 6.2).

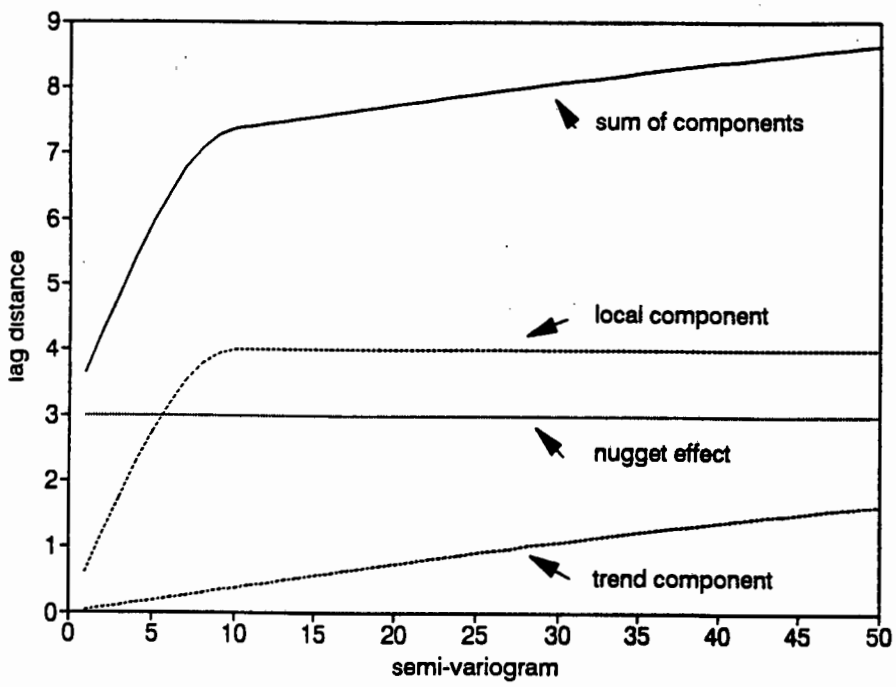


Figure 6.1: Components of the semi-variogram.

In the case of non-constant error variance, as is the case in both the applications described in this thesis, then the effect of error on the semi-variogram is more complex. We have then

$$E[(v_i - v_j)^2] = E[(\tau_i - \tau_j)^2] + E[(\eta_i - \eta_j)^2] + \sigma^2_{\epsilon_i} + \sigma^2_{\epsilon_j}$$

so that the increase in the semi-variogram will differ for each pair of data locations. If the individual error variances are known or can be estimated, one can estimate the semi-variogram of the error-free values $\tau + \eta$ by using

$$\hat{\gamma}_{\tau+\eta}(\mathbf{h}) = \frac{1}{2N_{\mathbf{h}}} \left\{ \sum (v_i - v_j)^2 - \hat{\sigma}^2_{\epsilon_i} - \hat{\sigma}^2_{\epsilon_j} \right\}$$

where the summation is over all $N_{\mathbf{h}}$ pairs which are a vector distance $\mathbf{h}$ apart.

The use of the adjusted semi-variogram values also relieves us of the necessity of trying to estimate separately the measurement error variance and the nugget effect from the empirical semi-variogram, that is, any residual

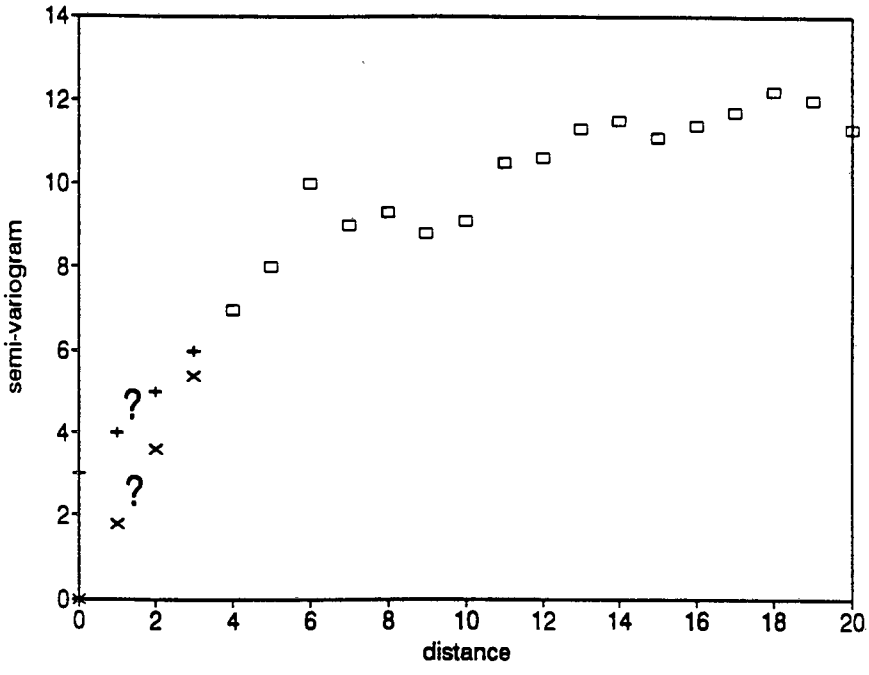


Figure 6.2: Estimating the nugget effect.

nugget effect observed in the adjusted semi-variogram can presumably be attributed entirely to short-scale variation rather than measurement error.

6.1.1 The Daily Rainfall Model

It is clear from the maps of the parameter values (Figure 2.3) that there are large-scale trends in the values of all the parameters; more detailed study in various regions shows that there are also more local topographical effects. There is also 'measurement error' in the data resulting from the estimation of probabilities and mean depths from a finite record length. As discussed in Chapter 2, the assumption of constant σ_ϵ^2 would be unreasonable for this data, given that, firstly, the number of years of rainfall data on which the parameter estimation was based varied from as few as five years to as many as 115 years, and, secondly, the year-to-year variation in rainfall differs markedly

in different areas of the country. Therefore the bootstrap procedure described in Section 2.4 was used to estimate σ_ϵ^2 for each of the sixteen parameters at each site so that the adjusted semi-variogram, as described above, could be calculated.

Figure 6.3 shows the unadjusted and adjusted semi-variograms for the nine amplitude parameters and the coefficient of variation of the daily rainfall model. For most of the parameters, the unadjusted values suggest a definite nugget effect, while in many of the graphs the adjusted values appear to pass approximately through the origin, that is, $\gamma_{\tau+\eta}(0) = 0$. Those which still show an apparent nugget effect after adjustment (notably DEPA0, DEPA1, CV, WWA0, and DWA0) suggest that the corresponding parameters are sensitive to local topographical changes, or possibly other sources of small-scale variation. Another reason for an apparent nugget effect could be that, since the rainfall station locations are recorded only to the nearest minute of latitude and longitude, two or more distinct stations may have the same recorded location, but differing rainfall parameters, as illustrated in table 2.2.

For one parameter, WWA2, the adjustment seems to have over-corrected, resulting in negative values throughout the empirical semi-variogram. This is probably due to the somewhat skew distribution of this parameter, and also the fact that the error variance of this parameter is relatively large compared to the actual parameter values, which are typically quite small. Models were fitted to each of the adjusted semi-variograms shown in Figure 6.3; in each case the model fitted was the sum of a constant (nugget effect) together with a spherical model for the local component η, plus a linear model for the long-range component τ.

One problem arises as a result of the negative values in the adjusted semi-variogram; the weighted least squares method of fitting variogram models described in Section 5.2 does not operate correctly when negative values are present (of course such values are theoretically impossible!). This is

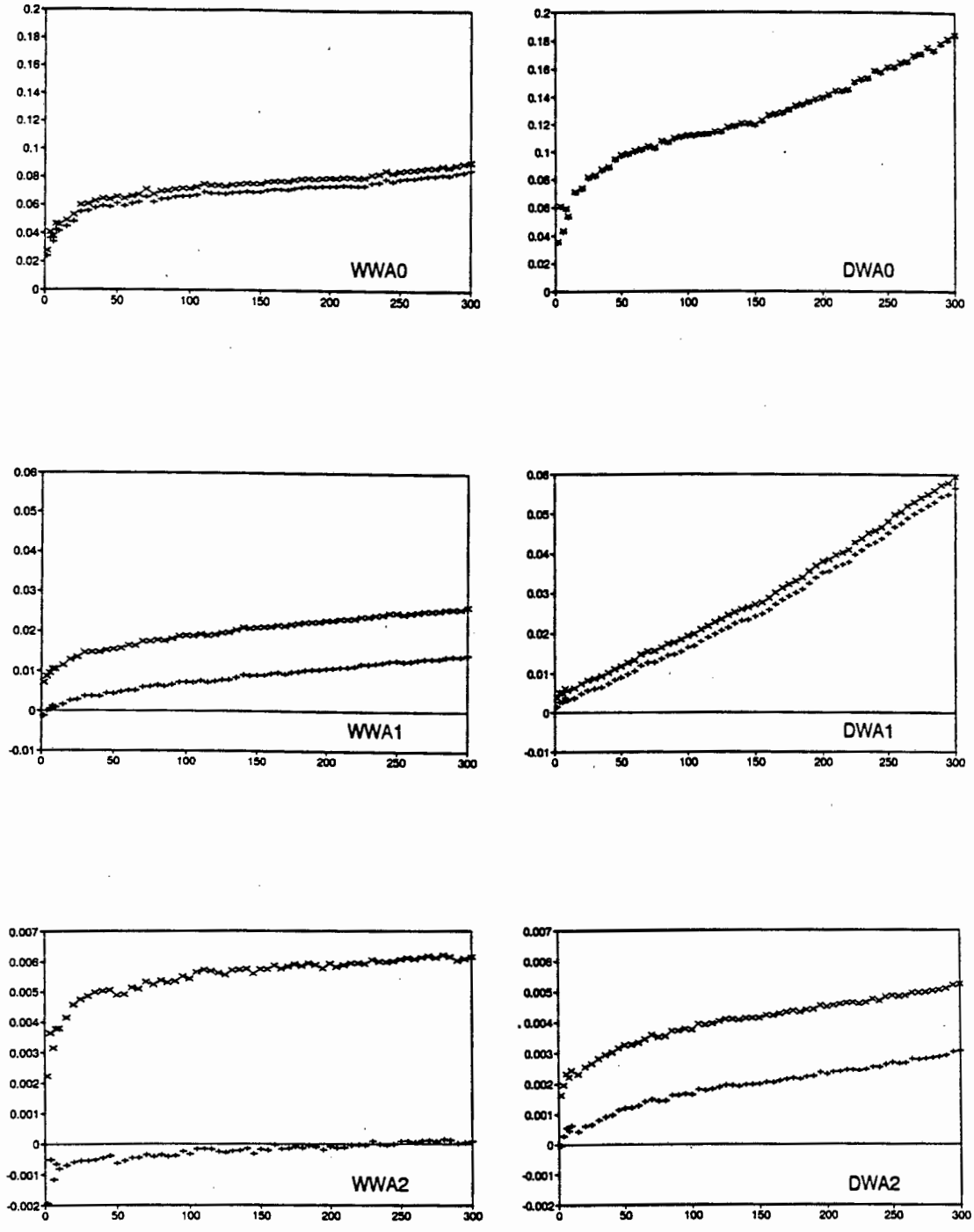


Figure 6.3: Semi-variograms: amplitude parameters.

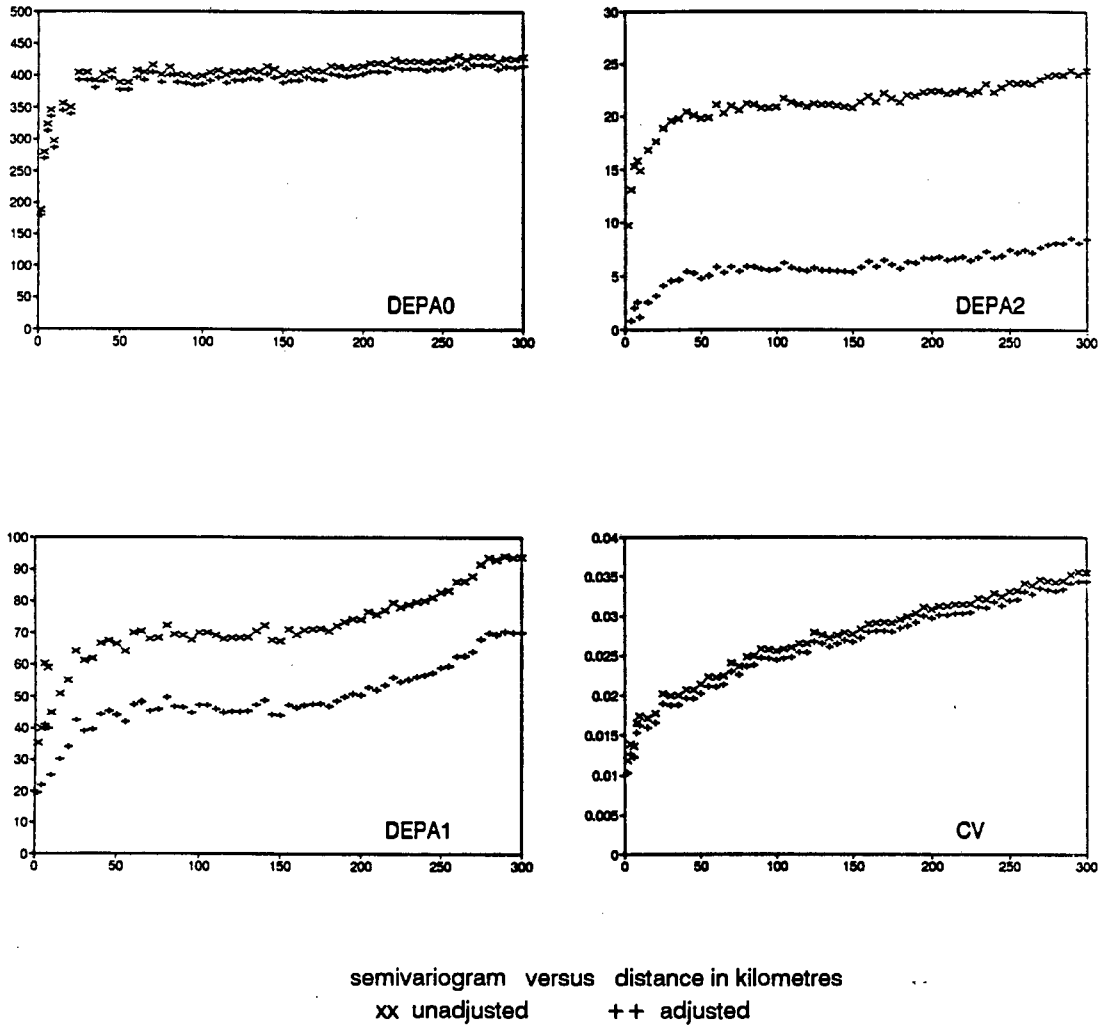


Figure 6.3: Semi-variograms: amplitude parameters (contd.).

overcome by adding a suitably chosen constant to both the observed and predicted values during the fitting process, so that all values are positive. For WWA2 the semi-variogram was fitted to the unadjusted data, incorporating a separate estimate of the (average) measurement error variance.

The fitted values of the sill and range of the spherical semi-variogram, the slope of the linear semi-variogram, and the nugget value for each rainfall model parameter, are listed in Table 6.1. The ranges of the local component vary between 10 and 40 kilometres. All the models fit well up to a distance of at least 240 km, which corresponds to the maximum distance which was

later used in the local kriging calculations. These fitted models are used to provide the covariances required for the kriging calculations.

parameter	sill	range	slope	nugget
WWA0	0.0350	24	0.000110	0.020
WWA1	0.0035	36	0.000036	0.000
WWA2	0.0027	19	0.000007	0.000
DWA0	0.0450	35	0.000300	0.035
DWA1	0.0030	10	0.000140	0.000
DWA2	0.0007	12	0.000010	0.000
DEPA0	210	26	0.0400	180
DEPA1	22	40	0.0330	20
DEPA2	5	40	0.0020	0
CV	0.0060	12	0.000064	0.010

Table 6.1: Fitted semi-variogram models: amplitude parameters.

A similar adjustment was made for the phase parameters of the daily rainfall model, but as the semi-variogram for these circular variables needs to be defined somewhat differently discussion of these is deferred to Chapter 8.

6.1.2 The Southern African Bird Atlas Project

In the Southern African Bird Atlas Project we observe the reporting rate in each QDGS which is effectively a frequency and thus reflects the underlying probability of observation of the species in question, but is subject to the usual binomial sampling variance. This variance in turn is dependent on the number of field cards available for each QDGS which varies greatly (from zero to 1698). The semi-variogram of the underlying probabilities can again be estimated by adjusting the observed semi-variogram for the estimated binomial variances; full details are given in Chapter 9.

6.2 Kriging in the Presence of Measurement Error

In the case where the data contain measurement error it is appropriate to estimate the underlying measurement error-free values, either at the existing data locations or at new locations. That is, we assume that

$$v_{\mathbf{z}} = \mu + \tau_{\mathbf{z}} + \eta_{\mathbf{z}} + \epsilon_{\mathbf{z}}$$

and the objective is then to estimate the value $\mu + \tau_0 + \eta_0$ at some location $\mathbf{z}_0$. Thus we seek w_i such that

$$E[\sum_{i=1}^{n} w_i v_i - (\mu + \tau_0 + \eta_0)]^2$$

is a minimum, subject to $\sum_{i=1}^{n} w_i = 1$.

We have:

$$\begin{aligned} E[\sum_{i=1}^{n} w_i v_i - (\mu + \tau_0 + \eta_0)]^2 &= E[\sum_{i=1}^{n} w_i(\tau_i + \eta_i) - (\tau_0 + \eta_0) + \sum_{i=1}^{n} w_i \epsilon_i]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{ij} - 2\sum_{i=1}^{n} w_i c_{i0} + c_{00} + \sum_{i=1}^{n} w_i^2 \sigma_{\epsilon_i}^2 \end{aligned}$$

where $c_{ij} = \mathrm{cov}(\tau_i + \eta_i, \tau_j + \eta_j)$ and $c_{i0} = \mathrm{cov}(\tau_i + \eta_i, \tau_0 + \eta_0)$.

If we set

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j c_{ij} - 2\sum_{i=1}^{n} w_i c_{i0} + c_{00} + \sum_{i=1}^{n} w_i^2 \sigma_{\epsilon_i}^2 - 2\lambda(\sum_{i=1}^{n} w_i - 1)$$

where λ is a Lagrange multiplier, then, taking the derivative of G with respect to each of the w_i and λ leads to the set of equations

$$\frac{\partial G}{\partial w_i} = 2[\sum_{j=1}^{n} w_j c_{ij} - c_{i0} + w_i \sigma_{\epsilon_i}^2 - \lambda] = 0 \qquad i = 1, \ldots n$$

and

$$\sum_{i=1}^{n} w_i = 1$$

or, in matrix notation,

$$\begin{bmatrix} c_{11}+\sigma^2_{\epsilon_1} & c_{12} & c_{13} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22}+\sigma^2_{\epsilon_2} & c_{23} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn}+\sigma^2_{\epsilon_n} & 1 \\ 1 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ -\lambda \end{bmatrix} = \begin{bmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{n0} \\ 1 \end{bmatrix}$$

which gives the required values of the w_i and hence the estimated value at $\mathbf{z}_0$, assuming that the relevant covariances are known or can be modelled as described in the previous section..

Since $\mathrm{cov}(v_i, v_j) = \mathrm{cov}(\tau_i + \eta_i, \tau_j + \eta_j)$ when $i \neq j$, it might at first appear that the equations above are no different to those of ordinary kriging (equation 5.2). However, there are two important differences. Firstly, if we seek an estimate at a point which is also a data location, i.e. $\mathbf{z}_0 = \mathbf{z}_i$ for some i, then the ordinary kriging solution would lead to an exact interpolation, since the ith row of the data covariance matrix would then be identical to the right hand vector. However, in the equations above this is no longer so, since the term $\sigma^2_{\epsilon_i}$ appears in the left hand matrix but not in the right hand vector, so we no longer have exact interpolation. Thus we avoid the 'spikes' in the estimated surface at the data points which are characteristic of ordinary kriging when there is measurement error present. A second difference is that, if there are duplicated data, that is, two data values with the same position vector $\mathbf{z}$ then the ordinary kriging equations will be singular while the equations above are not. In practice, the position vectors are only measured to a limited degree of accuracy in most situations, so that duplicate values may occur, this is certainly the case for the daily rainfall data, where rain gauge positions are recorded to the nearest minute of a degree and several apparently duplicated locations thus exist.

6.3 Kriging in the Presence of Trend

In the previous chapter current methods of kriging in the presence of trend were reviewed. In all the methods discussed the assumption was made that the trend could be approximated by a low order polynomial. For two dimensional data a simple quadratic model contains six parameters while a cubic model contains ten and yet such a model is often a poor representation of the trend over the entire region of interest. A more realistic option is thus to allow for a non-parametric modelling of the trend. The model described at the beginning of this chapter provides such a framework by considering the 'trend' to be another random process with a longer range of spatial correlation than the 'local' process.

The discussion of the previous section shows that such a trend is readily incorporated into the kriging process. The solution to the kriging equations derived there may be written as

$$\begin{pmatrix} \mathbf{w} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{c}_0 \\ 1 \end{pmatrix} \tag{6.3}$$

where $c_{ij} = \text{cov}(v_i, v_j)$ and $\mathbf{c}_0$ has elements $\text{cov}(v_i, (\mu + \tau_0 + \eta_0))$. Thus the covariances in the matrix $\mathbf{C}$ are the covariances of the observed data points (including the component due to the trend) while the covariances in the vector $\mathbf{c}_0$ are the *adjusted* covariances described in Section 6.1, that is, excluding the error variance component.

In practice, if a local window is used for kriging then it is only the covariance for values of $\mathbf{h}$ not exceeding the window width which is actually of interest. Essentially then, kriging using the above model is similar to the local kriging method recommended by Journel and Rossi (1989) which was mentioned in Chapter 5. Note however that the equations above show that the semi-variogram should be modelled so as to include effects which are due

to trend; such effects are often ignored in semi-variogram modelling when a local kriging approach is used. In practice the difference will be small since, by definition, the effect of trend on the semi-variogram will be negligible for small lags. However, by not modelling the trend one is effectively underestimating the variance and thus kriging estimation errors will probably also be underestimated. This may explain results such as that of Dagbert *et al.* (1993) who noticed in the results of cross-validation exercises with hydrographic data that

> *'any de-trending generates variograms which are too low i.e. predicted interpolation errors from those variograms are optimistic compared to the actual errors. In fact, the overall variogram which we calculated ... without any "de-trending" was all right in the prediction of the average interpolation error'.*

The method proposed above has advantages over universal kriging and IRF-kriging in that, firstly, there is no need to define a deterministic model for the trend, and, secondly, it is not necessary to try and separate out the effect of trend in modelling the covariance. Also, the kriging equations are simpler (equivalent to those used in ordinary kriging). In practice, the results of the various methods may not differ markedly, although if we attempt to extrapolate to points outside the region of the data points then estimates based on a deterministic trend model will be dominated by the extrapolated trend while the method described above will give estimates tending towards the mean as the distance from the nearest data points increases.

Compared with the usual kriging model with a deterministic trend function as discussed in Chapter 4, we may now have some additional parameters to estimate in the structural analysis, such as the slope of the linear component of the semi-variogram described in Section 6.1.1, but this is compensated for by the fact that the parameters of the trend function are no longer re-

quired; for spatial data there are generally at least three parameters in this function, and thus there will usually be a net saving in the total number of parameters to be estimated, besides the extra flexibility of not having to determine an appropriate functional form of the trend function.

6.4 Kriging as a Filter for Trend

In most practical situations involving spatial data the available data are not regularly spaced, so that removal of trend by simple moving averages such as are commonly used in time-series analysis, or by the median polish approach of Cressie (1986), is often not possible.

By a simple modification of the kriging procedure described in the previous section one can use the kriging process as a filter to separate the trend component from the short-scale component. We now wish to obtain the w_i which minimise

$$\mathrm{E}\left[(\sum_{i=1}^{n} w_i v_i - (\mu + \tau_0))^2\right]$$

subject again to the unbiasedness constraint $\sum_{i=1}^{n} w_i = 1$.

Using the Lagrange multiplier technique as before we find that the solution is now given by the following equation:

$$\begin{pmatrix} \mathbf{w} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{c}_0 \\ 1 \end{pmatrix} \tag{6.4}$$

where $c_{ij} = \mathrm{cov}(v_i, v_j)$ and $c_{0i} = \mathrm{cov}(v_i, (\mu + \tau_0)) = \mathrm{cov}((\mu + \tau_i), (\mu + \tau_0))$ and λ is the Lagrangian multiplier. Thus the values c_{0i} required here differ from those of equation 6.3 in that they exclude the component due to the small-scale process η.

In order to estimate the values of $\mathbf{c}_0$ we need to separately estimate the covariances σ_τ and σ_η, or, equivalently, to estimate the semi-variograms of

these components. Provided that the components τ and η are independent, then their combined covariance will be the sum of the individual covariances, and similarly for the semi-variograms. Thus we have

$$\gamma_{\tau+\eta}(\mathbf{h}) = \gamma_{\tau}(\mathbf{h}) + \gamma_{\eta}(\mathbf{h}) \tag{6.5}$$

Clearly the success of the method depends on being able to correctly identify and estimate the individual components of the covariance. It should be noted that such a decomposition is not unique; while the empirical semi-variogram may suggest a particular decomposition there is no reason why the user should not use additional information to decide on an appropriate split into trend and local variation. Any decomposition of the covariance function will lead to some smoothing of the data; the amount of smoothing will depend on precisely how the total semi-variogram is partitioned in equation 6.5. A study of the semi-variograms in Figure 6.3 indicates that most of them show a levelling off at between 10km and 40km which probably corresponds to the range of spatial correlation of the 'local' component; this was used as the basis of the fitted models described in Section 6.1.1.

Using the fitted semi-variogram given in table 6.1 for the parameter DEPA0 of the daily rainfall model, with the linear semi-variogram attributed to the trend component τ and the spherical model attributed to the local component η, equation 6.4 was applied to the data for DEPA0 to estimate the underlying trend values. Figure 6.4 shows the original values of DEPA0 at stations in the south-west Cape while Figure 6.5 shows the estimated trend values at the same sites; these figures show clearly the smoothing effect of the procedure. In many applications one would want to estimate trend values at a regular grid of points and display the results by means of a contour map; I have deliberately chosen not to do this here since the process of gridding and contouring generally involves additional smoothing which would obscure the effect which I wish to illustrate; however, the procedure can equally well be

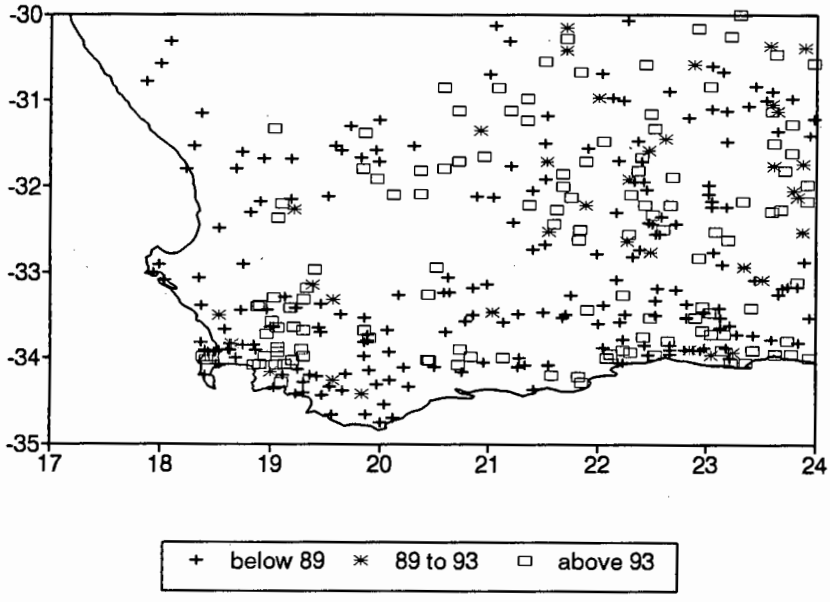


Figure 6.4: SW Cape: DEPA0 – Raw data.

used to estimate trend values at a grid of points.

The decomposition is analogous to the method known as *factorial kriging* (Matheron, 1982), in which the stochastic part of the model is split in to a number of components which are separately estimated, but the method of factorial kriging does not appear to have been used to estimate trend. The process is also analogous to Wiener filtering which is usually applied in the frequency domain, using Fourier transforms, to estimate the trend or signal in the time-series or signal-processing context (Press *et al.*, 1989).

Figure 6.6 shows the semi-variograms of the residuals, after the trend has been removed, of the four major parameters of the daily rainfall model. It is clear from the graphs that the trend has effectively been removed from the data. The graphs also show some evidence of a spike at small lags, which may indicate spurious negative correlations introduced at small lags by the smoothing process. Such an effect is well known in the time series field; for

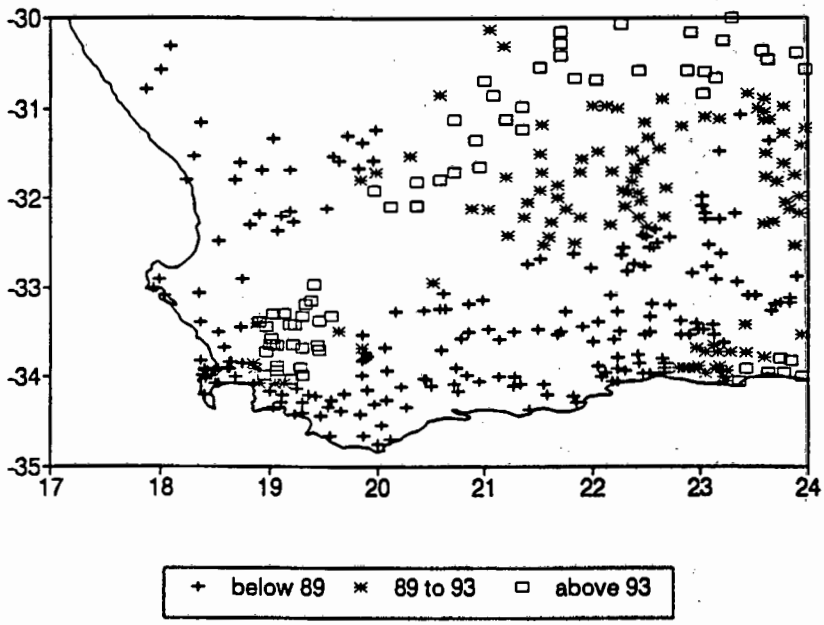


Figure 6.5: SW Cape: DEPA0 – Trend.

example, Diggle (1990, Section 2.6) gives an example of a simple 3-point moving average applied to an equally-spaced data series with linear trend and uncorrelated residuals, giving rise to a negative correlation at lag 1 in the residuals after the smoothing.

The procedure described here was used to separate the trend from the local component for all the amplitude parameters of the daily rainfall model as a pre-cursor to the study of the effects of topography on the 'local' component; this is discussed further in the next chapter.

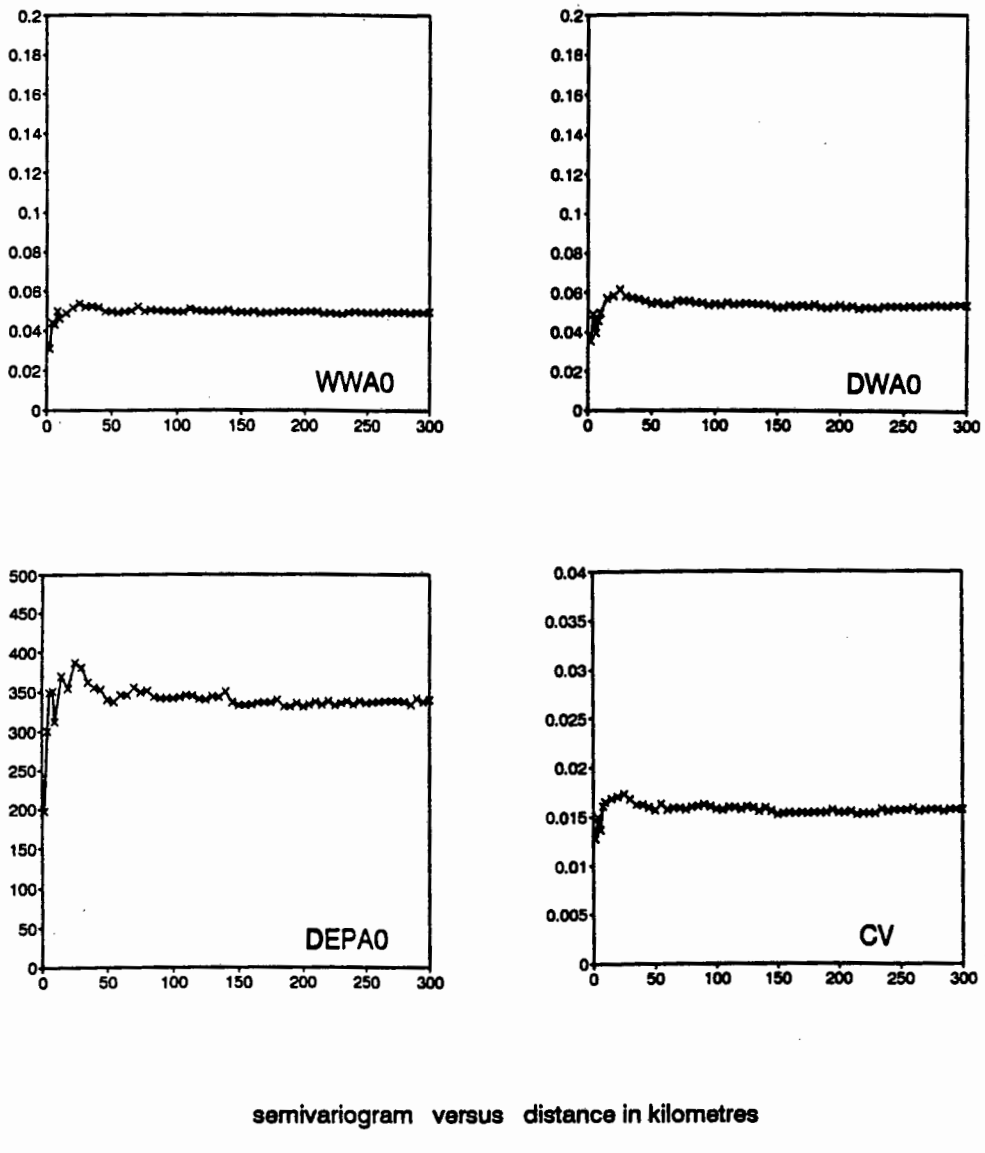


Figure 6.6: Semi-variograms: de-trended amplitude parameters.

Chapter 7

Rainfall Model: Interpolation of the Amplitude Parameters

In this chapter[1] we consider the problem of estimating the amplitude parameters of the daily rainfall model at sites where insufficient rainfall data are available. Specifically, our objective is to estimate values of all the parameters on a grid of 1 minute of a degree of latitude and longitude throughout South Africa, Lesotho and Swaziland. For various reasons, the treatment of the amplitude and phase parameters is somewhat different, and thus this chapter looks only at the amplitude parameters (and the coefficient of variation); the phase parameters are considered in the next chapter.

The approach taken is univariate; that is, each of the parameters is estimated independently of the others. Although there will probably be some spatial correlation between the parameters, which might suggest some advantage to be gained from a multivariate approach, possibly using co-kriging, it has been found in practice that co-kriging is generally only beneficial when the covariates are sampled more densely than the variable of interest. For the rainfall model, the data locations are the same for all parameters, so

[1]The work described in this chapter has previously been published in McNeill *et al.* (1994).

that a multivariate kriging is unlikely to give much advantage. In addition, a multivariate approach would be considerably more complex, necessitating the modelling of all pairwise cross-covariances amongst the 16 parameters, as well as greatly increasing the time required for the kriging computations of the roughly 500000 points to be estimated. Most users of the model will not be interested in the individual parameters but rather in descriptive measures of rainfall which can be deduced from the model, for example, mean rainfall over a given time period, or the probability of exceedance of a minimum rainfall criterion, as described in Chapter 2. While it might be optimal to interpolate such measures directly, as has been done by Dent *et al.* (1989) for mean annual and mean monthly precipitation throughout southern Africa, such a procedure is very time-consuming, while the model may be used to obtain a wide range of such derived measures in a matter of minutes on a personal computer. At the end of Chapter 8 the performance of the estimated model parameters in deriving mean annual precipitation is compared with values obtained from two other sources.

The map of selected rainfall stations (Figure 2.6) shows that in some areas of the country there is a high density of stations while in others, notably the north-western Cape, the data are very sparse. Available data tend to be clustered around areas of human habitation. One consequence of this is that, in mountainous regions of the country, the higher-lying areas tend to be less well covered by rain gauges, so that to ignore this in the analysis would tend to give rise to under-estimation of rainfall.

Large-scale spatial patterns are clearly observable in most of the model parameters (Figure 2.3). These large scale trends may be attributed to general circulation patterns affecting the climate of southern Africa and involving the movements of large masses of air, giving rise to *frontal* rainfall. On a smaller scale rainfall patterns are affected by the local topography and other physical features; in particular *orographic* rainfall is induced by the forced

ascent of air on the windward side of mountain barriers, while *convectional* rainfall is due to updraughts caused by localised heating and can thus be affected by ground cover and land use. In all types of rainfall, rising air is cooled so that it approaches saturation; a further factor in the formation of actual rain droplets is the presence of suitable nuclei; these may be provided by ice crystals in the clouds or by other particles such as occur in dust or man-made air pollution so that, for example, large cities may have more rainfall than the surrounding rural areas. It is clear that local anomalies can be accurately estimated only if the rainfall data are sufficiently dense in a given locality or if information on local explanatory variables is incorporated into the estimation process.

While an appropriately detailed data base on ground cover and land use is not available at this stage, elevation data are available on a grid of 1 minute of a degree of latitude by 1 minute of a degree of longitude throughout South Africa, and one would expect that local estimation of model parameters could be improved by incorporating this information. In addition, by making use of elevation data we would hope to overcome the bias in the station locations towards the lower-lying parts of each region.

One might expect that the amplitude parameters, which relate to rainfall *amounts*, would be more susceptible to topographic effects than the phase parameters which relate to *seasonality* of rainfall. This is exemplified by a comparison between the models for Tamboerskloof in Cape Town (station code 20716 W, elevation 100m) and the station at Woodhead Dam on the slopes of Table Mountain (station code 20719BW, elevation 747m) as shown in Figure 7.1. These two sites are only about five kilometres apart but show a large difference in the parameter DEPA0 (as can be seen by comparing the curves for mean rainfall depth) and a somewhat smaller difference in DEPA1 (the amplitude of the seasonal variation of mean depth), but show very little difference in the phase parameters which indicate the time of year of maxi-

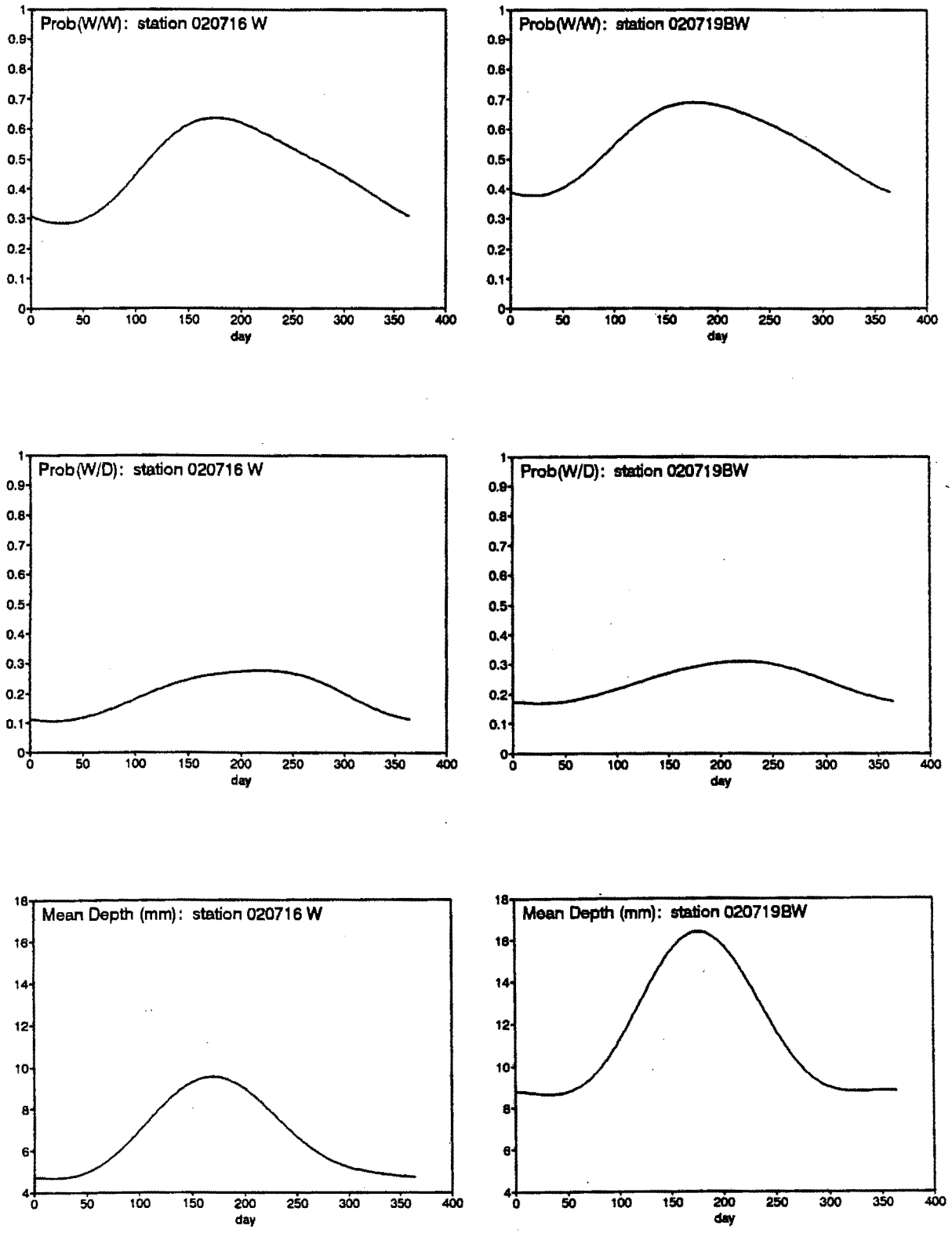


Figure 7.1: Comparison of two stations on Table Mountain.

mum probabilities or mean depth. It was therefore decided to make use of the altitude information only in the estimation of the amplitude parameters.

In the following section we review some of the approaches taken by previous researchers in the field of rainfall modelling to the incorporation of the effects of topography into the modelling process.

7.1 Rainfall and Topography: A Review

As mentioned in the previous section, orographic rain results when air rises over mountains, so that one may expect the most rain to occur on the windward slopes; for narrow mountain ranges the tops of the mountains and leeward slopes may also experience relatively heavy rainfall, however for more extensive mountain ranges the leeward slopes may be in a rain shadow area. This suggests that using only the altitude at a given point to predict the rain anomaly at that point will in general not be very successful, and this has been found to be the case by several researchers, for example, Armstrong (1993) and Creutin and Obled (1982). Thus a considerable body of research has been directed at deriving functions of the altitude at surrounding points which will be more suitable for predicting local rainfall patterns.

An early study is that of Spreen (1947) who investigated the relationship between elevation, slope, orientation (or aspect) and exposure (defined below). Using a graphical regression technique Spreen found that 88% of the variation in mean winter precipitation in western Colorado could be explained by these four variables, compared with 30% for elevation alone. Other studies, using the same measures, plus a number of others such as 'roughness', have been carried out in other parts of the world, for example in New Zealand by Hutchinson (1968), in Israel by Wolfson (1975), and in South Africa by Whitmore (1968), Schulze (1976), Hughes (1982) and Dent *et al.* (1989).

Most of these authors have used multiple regression techniques to incorpo-

rate the topographic variables into the rainfall modelling process; a difficulty with this approach is that if the area under study is large it may first need to be segmented into homogenous sub-regions within each of which the relationship between rainfall and the topographic variables is approximately constant. Dent *et al.* (1989) initially delineated some 712 regions in their study of mean annual and monthly rainfall in southern Africa, but experienced considerable difficulty in patching together the resultant estimates at the sub-region boundaries.

All the topographic variables used by these authors are based on gridded altitude data, using a local grid centred on a given point to calculate the relevant variates at that point. Definitions of the most commonly used measures are given below.

Gradient and Aspect Given a tangent plane to the surface at any point, the gradient is the maximum rate of change in altitude on this plane and the aspect is the compass direction of this maximum (decreasing) rate of change (Skidmore, 1989). The estimate of these values will depend on the grid size and limits used, as well as the algorithm used; Skidmore compares six possible algorithms. Some authors (Spreen, 1947; Hutchinson, 1968) define aspect as the direction in which the exposure (defined below) is a maximum. Aspect is a circular variable, and thus cannot be used directly in a standard regression model.

Roughness In view of the fact that the roughness of the terrain may cause updraughts and turbulence which may in turn influence the occurrence and longevity of storms (London and Emmitt, 1986) some researchers have included a measure of roughness. Hobson (1972) gives three methods for the calculation of roughness: one based on 'bump frequency', one based on comparing estimated surface area with the corresponding planar area and a third based on the variation in the direction of nor-

mals to planar surfaces defined by adjacent groups of three elevation readings.

Exposure Several authors have attempted to define directly a function of topography which encapsulates the fact that windward slopes tend to get more rain due to their greater 'exposure' to the rain bearing winds. Dent *et al.* (1989) used the definition of exposure suggested by Seed (1987) which involves counting the number of points in a 5 minute by 5 minute mask which have a lower elevation than the point at the centre. Spreen (1947) used as his definition of exposure the number of one-degree sectors of a 20 mile radius circle centred on the station in which there is no land higher than 1000 feet above the station. Hutchinson (1968) used a similar definition but with a five mile radius. Hughes (1982) used an index based on the (weighted) sum of areas of grid squares with elevation higher than the gauge, taken over all squares of area 0,25 km^2 lying in a 45 degree sector oriented south-west and of radius 10 km. The weighting used was the logarithm of the excess elevation while the south-west orientation was chosen to coincide with the main rain-bearing wind direction in the area; the other aspects of the measure were chosen after a number of trials with exposure indices of varying complexity, and Hughes comments that *'the choice of a measure of exposure proved to be very difficult'*. It is clear from these different definitions that, apart from the difficulty of finding a satisfactory definition of exposure, there is almost certainly a need to 'customise' the measure for different geographical regions.

In practice all these measures are calculated as a function of the a_i where the a_i are the local values of altitude, usually available at a grid of points, and are thus influenced by the grid spacing and also by the extent of the local area or mask used in the calculation. Many of the measures can be expressed

in the form $\sum w_i a_i$, that is, a linear function of the local altitudes. In view of the fact that researchers have noted the difficulty in finding an appropriate measure of 'exposure' based on a priori considerations, it is appropriate to ask whether it may not be possible to use the data themselves to determine, on a local basis, that function of the a_i which best explains the rainfall anomalies, and let this function provide a local definition of 'exposure' which can then be calculated at ungauged locations to predict the anomalies there. By defining a single measure in this way one would also avoid the difficulty that arises when a number of correlated measures are used as the explanatory variables in a multiple regression and also the need to consider the possible interacting effects of such variables. This approach is discussed further in Section 7.2.2.

7.2 Modelling Topography via Kriging

In Chapter 5 we described two possible approaches to the problem of incorporating covariate information into the kriging process, namely co-kriging and kriging with external drift. We now consider each of these in turn as applied to the amplitude parameters of the daily rainfall model, with the altitudes or functions of altitude as covariates.

7.2.1 Co-Kriging

In order to use the co-kriging approach, as outlined in Section 5.3.1, it is necessary first to model not only the spatial covariance functions of each parameter, but also the cross-covariances of each parameter with altitude.

For our application, where the cross-covariance will certainly show directional effects, it was decided to try modelling the cross-covariance directly, rather than using one of the cross semi-variogram measures introduced in Section 5.3.1. In exploring this approach the cross-covariance of the param-

eter DEPA0 with altitude was studied for the south western Cape region (approximately west of Mossel Bay (22°E) and south of Calvinia (32°S)). Calculations were done for eight directions, and the results are shown in Figure 7.2 (in which the distance units are minutes of a degree of latitude or longitude). Figure 7.3 displays the same information in the form of a contour map. It is clear from the figures that the cross-covariance is generally positive and decreases with distance, but there is a distinct group of negative values at a lag distance of approximately 20 units (about 35 km) in a north-west direction. This corresponds with the knowledge that the main rain-bearing wind direction in this area is approximately north-west, so that it is likely that rain gauges which are sited so as to have points of high elevation to the north-west will be in the rain shadow of that higher ground and thus have reduced rain.

It is clear, however, that in order to take account, for example, of locally varying directional effects, the cross-covariance models would have to be re-calculated and fitted regionally, or perhaps, to avoid discontinuities at regional boundaries, re-computed in a neighbourhood of each point being estimated as suggested by Haas (1990). This would necessitate an enormous amount of computation. Further, there is a then a need to parameterise appropriate cross-covariance models which could be used as part of an automatic fitting procedure, since it would be impractical for the user to interactively model cross-covariances at the half a million or so locations being estimated in this project. Some further research was done to investigate the feasibility of such automatic modelling, but the results were generally disappointing (Sedupane[2], 1992).

It was therefore decided not to continue with the co-kriging approach and to look at the alternative possibility of kriging with an 'external drift'

[2]An honours student project supervised by the author of this thesis

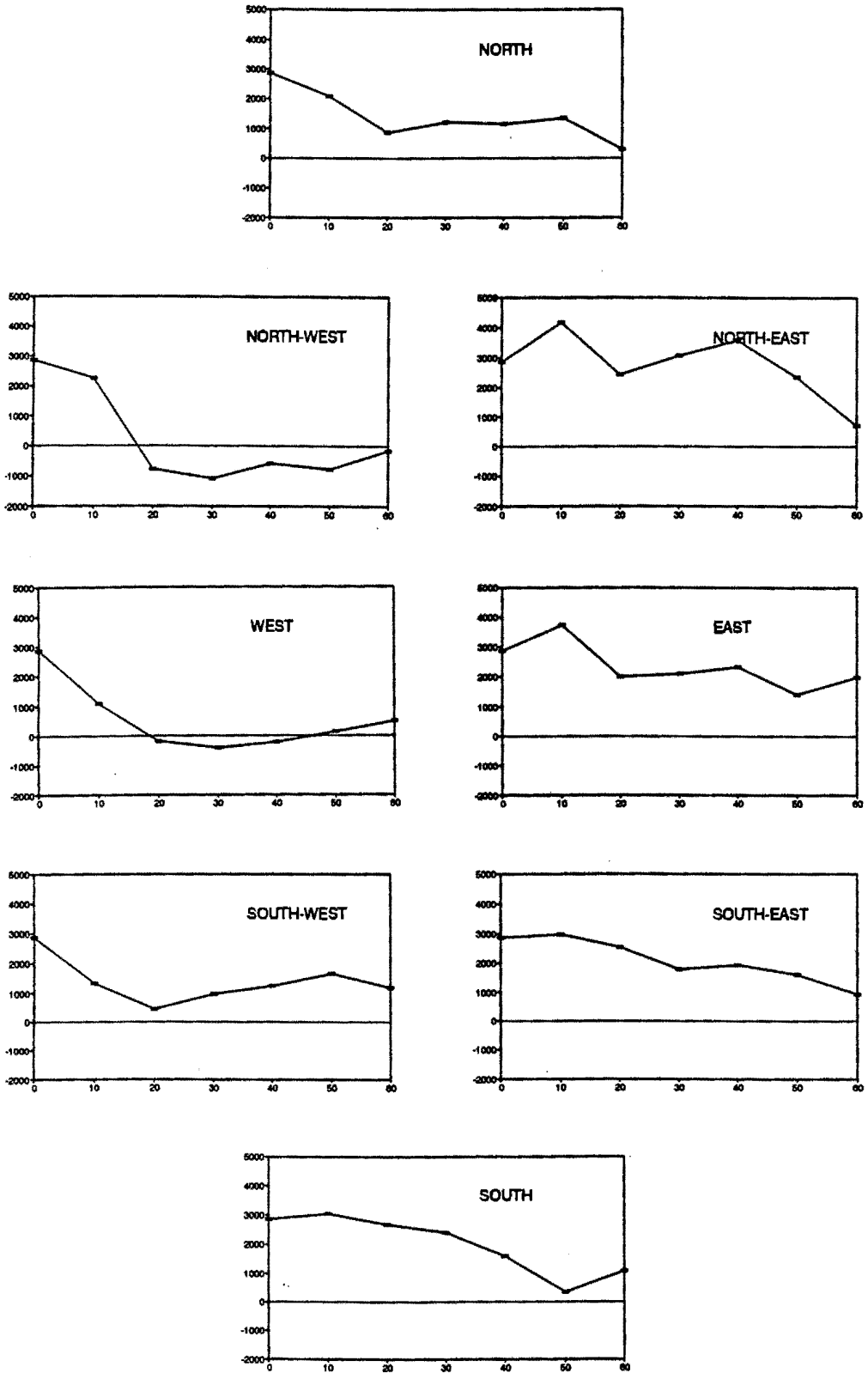


Figure 7.2: Cross-covariance of rain and altitude: SW Cape.

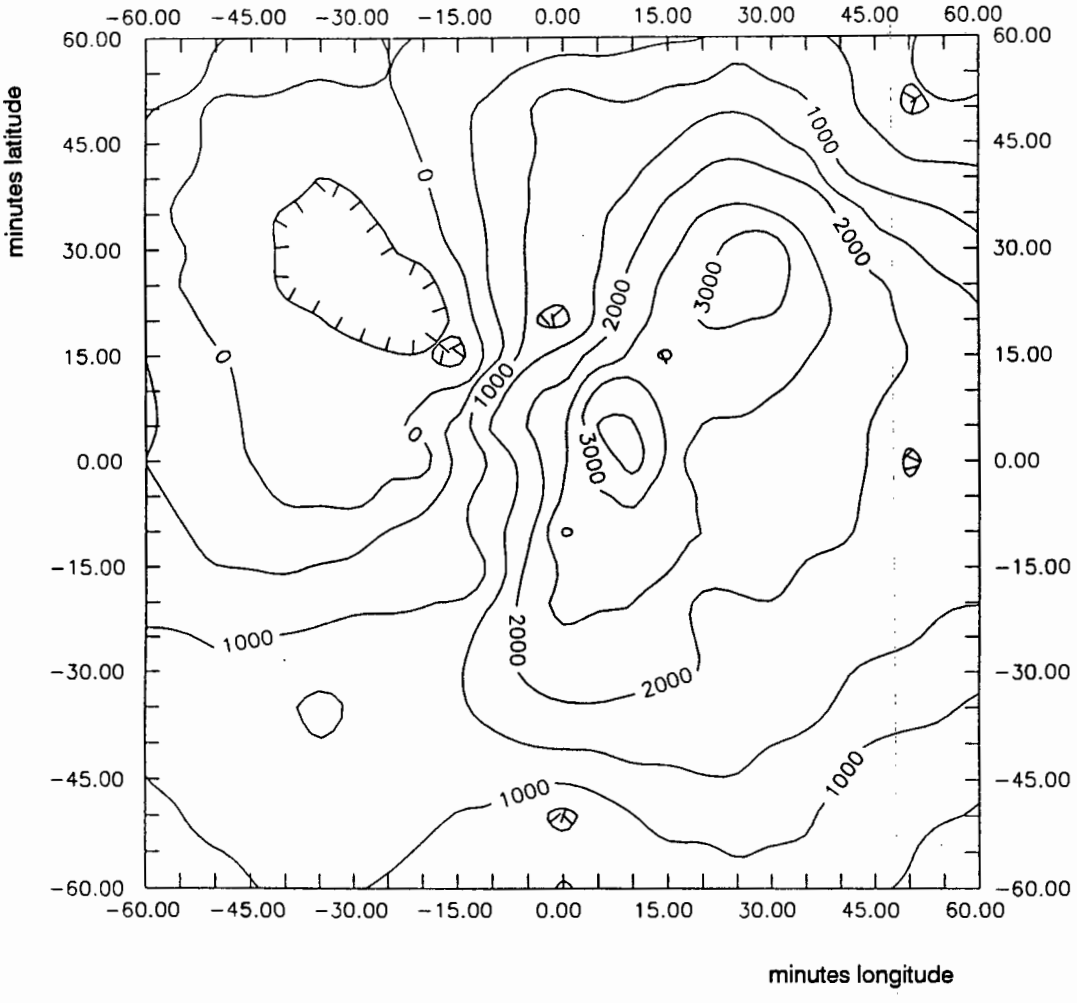


Figure 7.3: Contoured cross-covariance: SW Cape.

function.

7.2.2 Kriging with External Drift

As outlined in Section 5.3.2, kriging with external drift uses a model similar to that of universal kriging, that is,

$$v_{\mathbf{z}} = \sum_{l=1}^{p} f_l(\mathbf{z})\beta_l + \eta_{\mathbf{z}} \tag{7.1}$$

However the functions f_l are now taken to be functions of the covariates. Thus we require firstly that the form of the relationship between predictand (rainfall parameter) and covariate (altitude or function of altitude) is known or can be approximated by a simple function such as a polynomial, and also that the covariate information is available at all the sites at which the predictand is known and also at all those points at which an estimate of the predictand is required. As discussed above, the relationship between altitude and rainfall is complex, and certainly the rainfall at a point is influenced not only by the altitude at that point but by the pattern of topography in the vicinity of the point. Thus one possibility would be to define a function such as 'exposure' and use this as the covariate; however, as mentioned previously, it is unlikely that a single definition would suffice throughout the country, and the alternative of attempting to find appropriate local measures of 'exposure' is prohibitively computer intensive, since such a measure will necessarily incorporate altitude values on a fairly large grid surrounding each point.

A solution to this difficulty was obtained by first calculating orthogonal functions of the surrounding elevations at each gridded altitude point, which together would account for all possible (local) patterns up to a third degree surface. This would effectively incorporate a number of the functions defined in Section 7.1; for example, both slope and aspect can be measured in terms of first degree functions, while some of the definitions of roughness and exposure

could also be expressed as low-order polynomials of the gridded altitude values. It should be emphasised, however, that what is proposed here is more general than the use of a pre-defined function of altitude such as slope, in that no particular polynomial is chosen *a priori*, but rather, a set of functions is used which effectively encompasses all possible patterns that can be described by third degree functions; the kriging process then estimates appropriate weightings to give to the component functions in the neighbourhood of each point being estimated, and the use of moving-window kriging ensures that the weightings are updated for each point to be kriged.

An advantage of defining orthogonal functions is that they are by definition uncorrelated and thus we avoid the multicollinearity problems commonly associated with multiple regression.

For gridded data the calculation of the orthogonal functions is a simple matter. If we write the altitude values at a grid of points in an $q \times q$ matrix, D, and then calculate the matrix $M'DM$ where M is the $q \times 4$ matrix whose columns give the coefficients for orthogonal polynomials of degree 0,1,2,3 respectively, then the resultant matrix has as its elements the required orthogonal functions.

We illustrate the procedure using a 5×5 grid of altitude points. Let the altitude values at a grid of points be given by:

$$D = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

Then the matrix M of orthogonal polynomial coefficients is given by:

$$M = \begin{bmatrix} 1 & -2 & 2 & -1 \\ 1 & -1 & -1 & 2 \\ 1 & 0 & -2 & 0 \\ 1 & 1 & -1 & -2 \\ 1 & 2 & 2 & 1 \end{bmatrix}$$

and $M'DM$ can be written as

$$M'DM = \begin{bmatrix} \xi_{00} & \xi_{01} & \xi_{02} & \xi_{03} \\ \xi_{10} & \xi_{11} & \xi_{12} & \xi_{13} \\ \xi_{20} & \xi_{21} & \xi_{22} & \xi_{23} \\ \xi_{30} & \xi_{31} & \xi_{32} & \xi_{33} \end{bmatrix}$$

where ξ_{00} is simply the sum of the elements of D, while ξ_{01} is the linear contrast of the columns of D, which corresponds to a plane with E-W slope, and ξ_{10} corresponds to a plane sloping N-S. By including both ξ_{10} and ξ_{01} in the external drift function we allow for a plane of any inclination to form the 'drift' function. By including also ξ_{02}, ξ_{11} and ξ_{20} we allow for an arbitrary second degree surface and so on. I decided to include all terms up to third degree, thus allowing for a cubic surface, and using 10 orthogonal functions in all. This allows for reasonably complex topographical patterns.

Tables of orthogonal polynomial coefficients for small values of q, together with formulae for their computation in the general case, are given in various books of statistical tables, for example, Pearson and Hartley (1962).

Thus the full external drift function has the form:

$$\beta_{00}\xi_{00}+\beta_{10}\xi_{10}+\beta_{01}\xi_{01}+\beta_{20}\xi_{20}+\beta_{11}\xi_{11}+\beta_{02}\xi_{02}+\beta_{30}\xi_{30}+\beta_{21}\xi_{21}+\beta_{12}\xi_{12}+\beta_{03}\xi_{03}$$

where the β coefficients will be selected optimally by the kriging program to model the relationship between the rainfall model parameter and the components of the pattern of topography in the neighbourhood of the point being estimated.

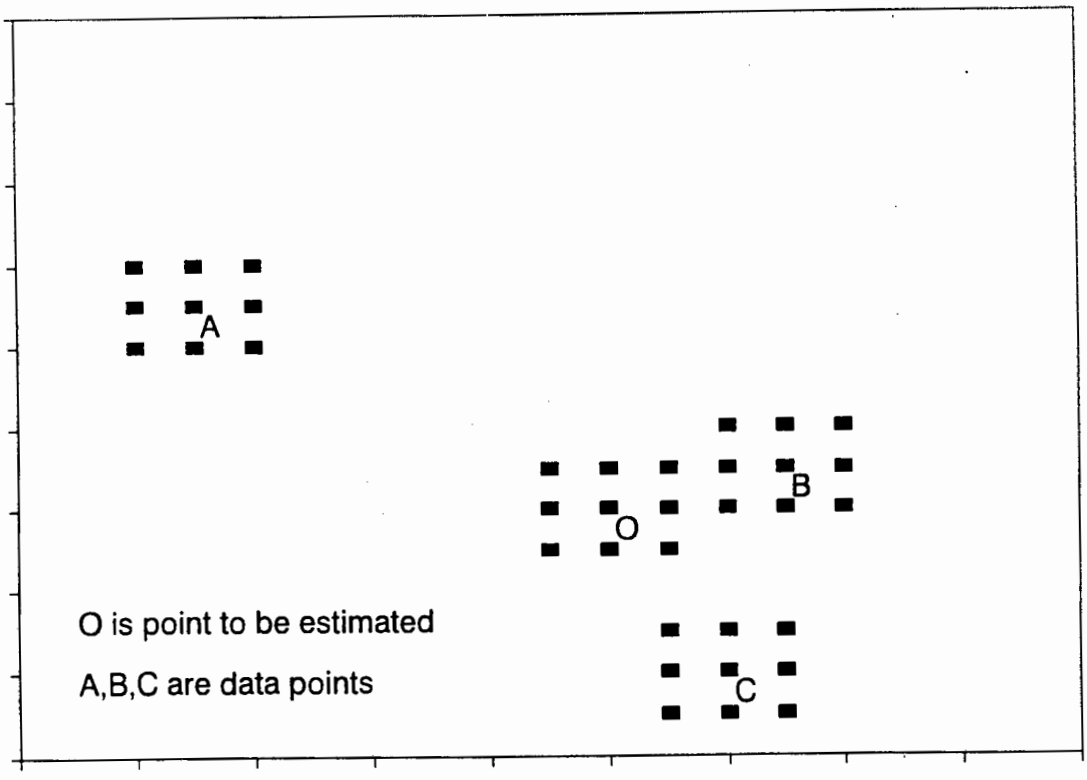


Figure 7.4: Calculating the functions of topography.

The values $\xi_{00}, \ldots \xi_{03}$ are calculated at each data point and at each point to be estimated. Figure 7.4 illustrates this, using a 3×3 grid. In practice, a 3×3 grid is too small to allow the calculation of independent functions up to third degree; a grid of at least 5×5 is required, but it is not immediately obvious how to choose the optimal size of the grid. In addition, the optimal choice may depend on whether the data have been de-trended. If one is working with de-trended residuals, then they will probably only reflect local topographic effects, while the effects of larger mountain features will have been absorbed by the trend. On the other hand if no prior smoothing has been used then the data will possibly reflect the effects of mountain features some distance away, so that a larger mask should probably be used for the altitude function calculations. A disadvantage of using a larger mask would

be that as the number of points in the mask increases, so does the potential complexity of possible topographic patterns, so that it might be necessary to use orthogonal functions of higher degree to obtain a realistic approximation to the surface. Another problem with using a large mask is that the gridded altitude data are not currently available for some of the areas north of the South African border, so that as the grid size increases, there are an increasing number of data sites for which we cannot calculate the necessary functions without some additional estimation.

7.2.3 Cross-Validation

In order to decide on the optimal grid size and also on the optimal degree of the ξ functions to be used in the external drift kriging procedure, a set of test sites was selected as described below, and the values at these sites were estimated from the remaining sites using a range of grid sizes and functions. The sum of the squared estimation errors at the test sites was then used to compare the various options.

Thus, for a given grid size, the corresponding orthogonal altitude functions were first calculated and stored for each point within southern Africa. Then the kriging estimation process was carried out firstly using no topographical information, then using only the term ξ_{00}, that is, the average altitude, then using only first degree functions, that is, including ξ_{10} and ξ_{01}, then using first and second degree functions, and finally using the full set of third degree functions. For any sites where the external drift matrix consisting of the values of the ξ was singular, (this could happen for example if all the altitude values in the mask around that site were identical) a drift of lower order was substituted until a non-singular matrix was obtained.

Spherical models fitted to the adjusted semi-variograms described in Section 6.1.1 were used throughout.

The whole process was repeated twice; once using the de-trended parameters, (the trend estimation procedure was described in Section 6.4), and once with the original parameters. In all cases, a moving window version of kriging was used, such that the closest 33 points (within a maximum search distance of 120 km) were selected. In those parts of the country where the stations are fairly dense the closest 33 points were generally all within a radius of not more than 60 km. If the number of points found within the maximum search distance of 120 km was insufficient for estimation with the chosen degree of orthogonal functions then a drift of lower order was used at that site. For example, if a cubic drift was selected, but less than 30 data points were found within 120 km of a given location to be estimated, then a quadratic drift was used, if less than 20 data points were found, then linear drift was used, while if less than 10 data points were available, then only ξ_{00} was used. In practice only three of the test sites were affected in this way.

In selecting a set of data points as test sites to cross-validate the various options it must be remembered that the data, that is, the rain parameter values, are themselves estimated values subject to error, so that we do not have 'true' values with which to compare our estimates. In order to minimise the effect of this unknown error in the data I decided to select from each Weather Bureau block (half degree grid square) the rainfall station at which the bootstrap variance of the estimate of DEPA0 was a minimum, and to use these points as the test points. No test point was selected from blocks having less than five data points, as this would mean that the resultant data set was rather sparse around that point. Apart from these omitted blocks the 373 test points are thus roughly on a grid across the country, with one in each Weather Bureau block. The decision to use the variance of the parameter DEPA0 as the selection criterion for the test sites was based on the fact that this parameter is probably the one most sensitive to topography.

The levels of the various factors used in the cross-validation exercise were:

- grid size: (5×5, 15×15, 25×25, 35×35 minutes of a degree)
- degree of orthogonal functions: (0,1,2,3)
- prior de-trending / no prior de-trending

For each rain model parameter the mean squared estimation error, averaged over the 373 test points, was calculated for various possible combinations of the factors shown above. The optimal factor combination could vary depending on the specific rain parameter under consideration. In practice, results were very similar for all parameters and thus only results for the parameter DEPA0 are given here (Table 7.1).

	degree of orthogonal function							
	with no de-trending				with prior de-trending			
grid size	0	1	2	3	0	1	2	3
5×5	363	378	437	568	373	387	443	558
15×15	366	373	410	466	377	383	417	466
25×25	374	385	391	437	not calculated			
35×35	378	400	413	465	not calculated			
no altitude	381	not applicable			386	not applicable		

Table 7.1: Mean squared estimation error: DEPA0.

It is clear from the results that prior de-trending of the data gives no improvement. The spurious correlations at short lags which were induced by the smoothing process (see Section 6.4) may be responsible for this. Also, the fact that local kriging was used throughout means that the effects of trend are likely to be small, and thus the de-trending only introduces an extra level of complexity into the kriging process, apparently without any compensating gain in accuracy.

A rather surprising feature of the results is that fitting the more complex topographical models produces poorer results, although the deterioration

is less marked for the larger grid sizes. The estimation using the average altitude ξ_{00} does however give a small improvement over estimation ignoring altitude.

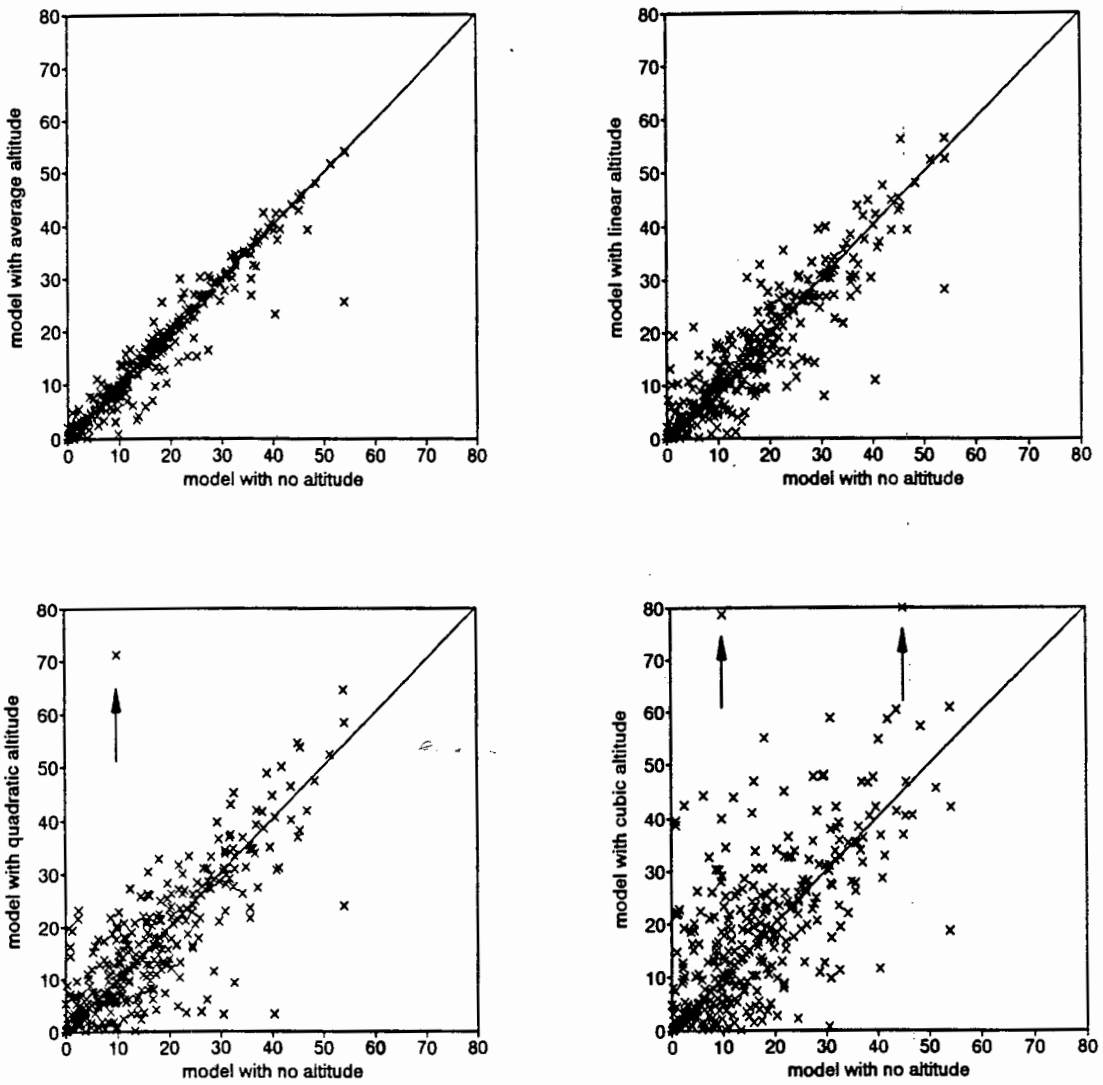


Figure 7.5: Estimation errors at individual test sites.

These results are made clearer if we graph the absolute values of the errors at individual test sites, as in Figure 7.5. Points below the diagonal line in each diagram represent test sites for which the inclusion of the relevant

functions of altitude leads to a reduction in the estimation error, and *vice versa*. We see from these graphs, which are based on a grid size of 9×9, that, although the models using average, linear or quadratic altitude generally give rise to smaller absolute error than the model ignoring altitude, the quadratic model gives rise to one very poor estimate, and the cubic model gives two. Bearing in mind that we may be extrapolating the altitude functions; that is, the values of the ξ at the point being estimated could be outside the range of the ξ values at the neighbouring data locations, it is not so surprising that we occasionally get rather poor estimates. Thus it seems that the more complex models are less robust. The fact that the models which do not include altitude at all do almost as well as the models with altitude is probably due to the fact that the test points, which were chosen for their low variance, are typically stations with many years of data and are not necessarily at high altitude, so that a simple interpolation from the neighbouring data points gives fairly accurate results.

In order to show more clearly the effect of including altitude in the models, parameter estimates were calculated at one minute intervals along two transects; Figure 7.6 shows the estimates of DEPA0 along the two transects, together with the altitude values. In the first, running west to east in the Jonkershoek mountains near Stellenbosch (latitude 33°59′S), there are a number of rainfall stations within the mountains, so that the ordinary kriging model without any altitude functions follows the shape of the mountains quite well. In the second transect, running west to east across the mountain ranges just north of Porterville (latitude 32°50′S), the model without altitude does not pick up the individual mountain peaks at all as there are few rainfall stations in the area, while the model including ξ_{00} shows a small rise in DEPA0 as each peak is crossed. By contrast, the values of DWA0, which is a measure of the *probability* of rain, taken along the same transect (Figure 7.7), show almost no response to altitude; the values decrease steadily as

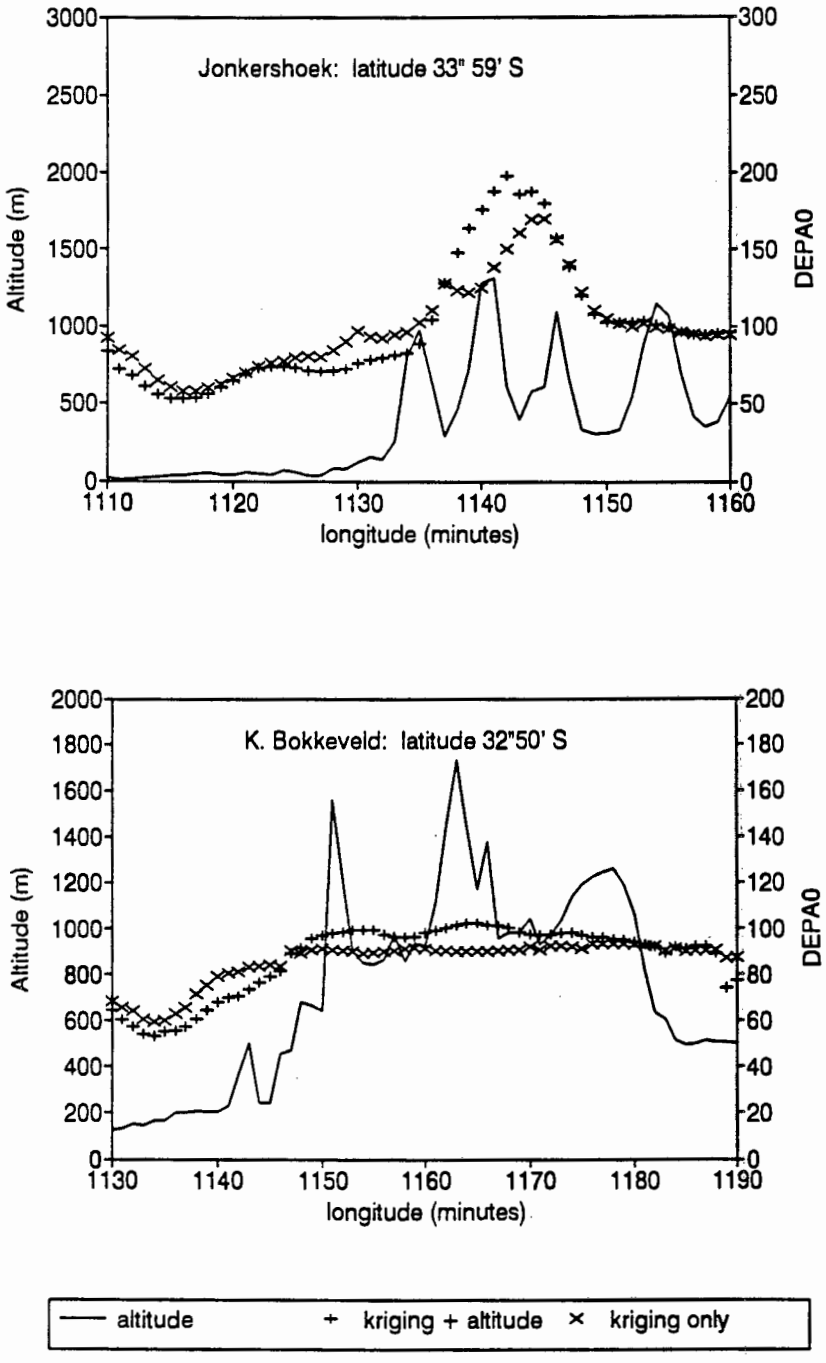


Figure 7.6: Predicted DEPA0 and altitude.

one moves west, away from the coast.

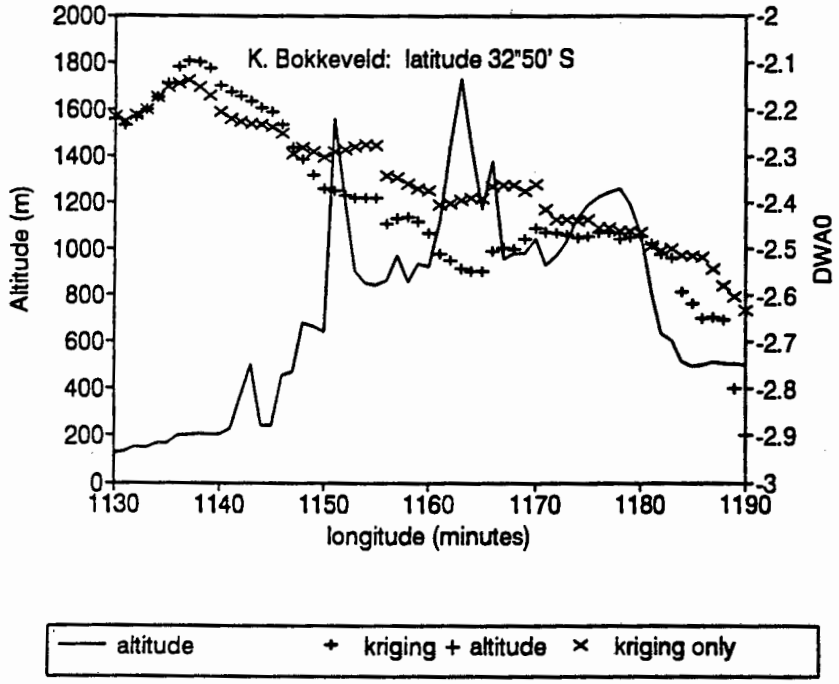


Figure 7.7: Predicted DWA0 and altitude.

On the basis of these results it was decided that the final estimates should be done without prior de-trending and including only the term ξ_{00} in the drift function. Further checking of grid sizes suggested that a grid of 9×9 would be optimal, and this was used for the final estimation at a grid of sites covering the country at 1 minute by 1 minute intervals. Maps of the estimates at intervals of 30 minutes by 30 minutes, that is, at the centre of each Weather Bureau block, are shown in Figure 7.8.

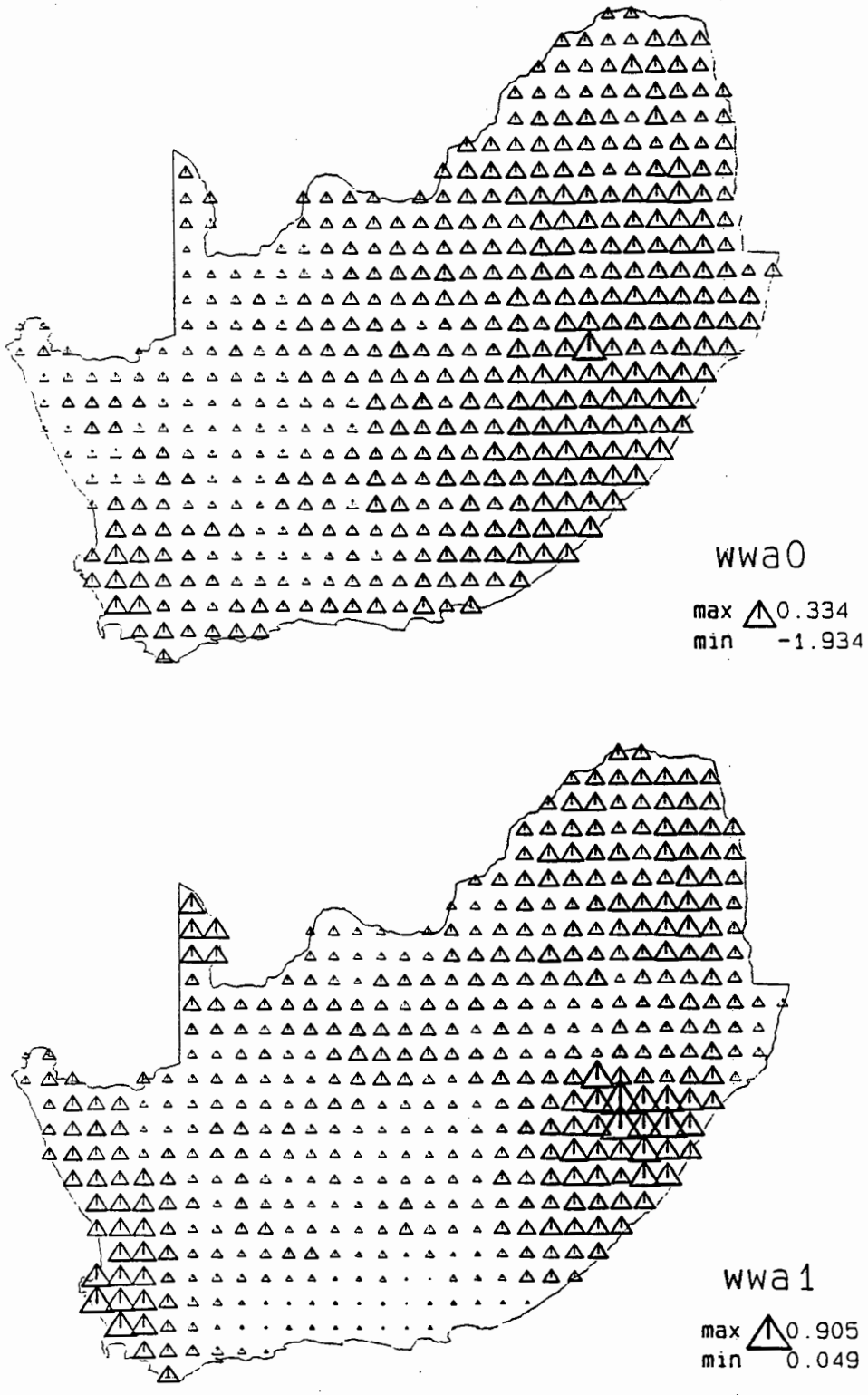


Figure 7.8: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters).

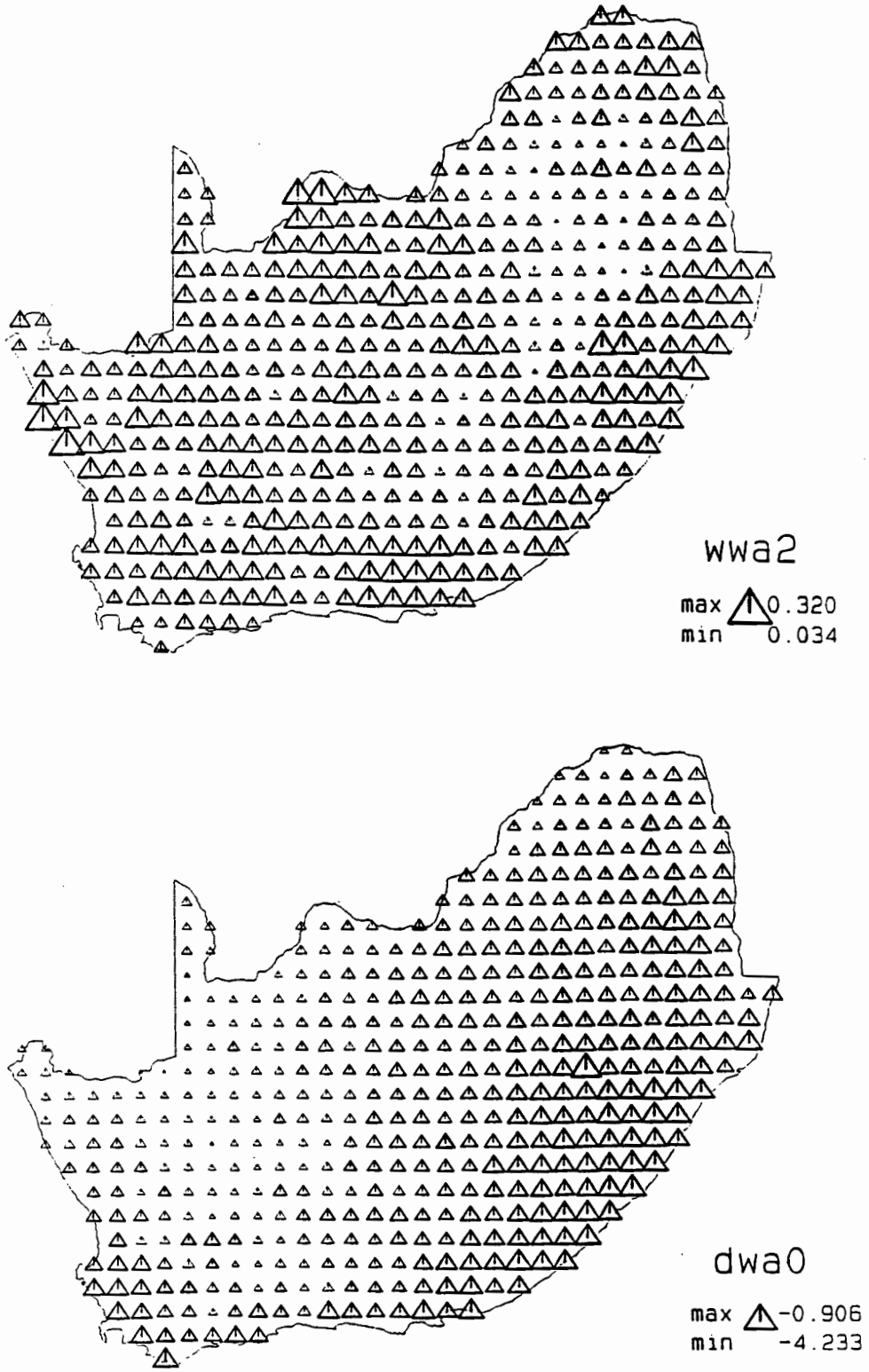


Figure 7.8: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

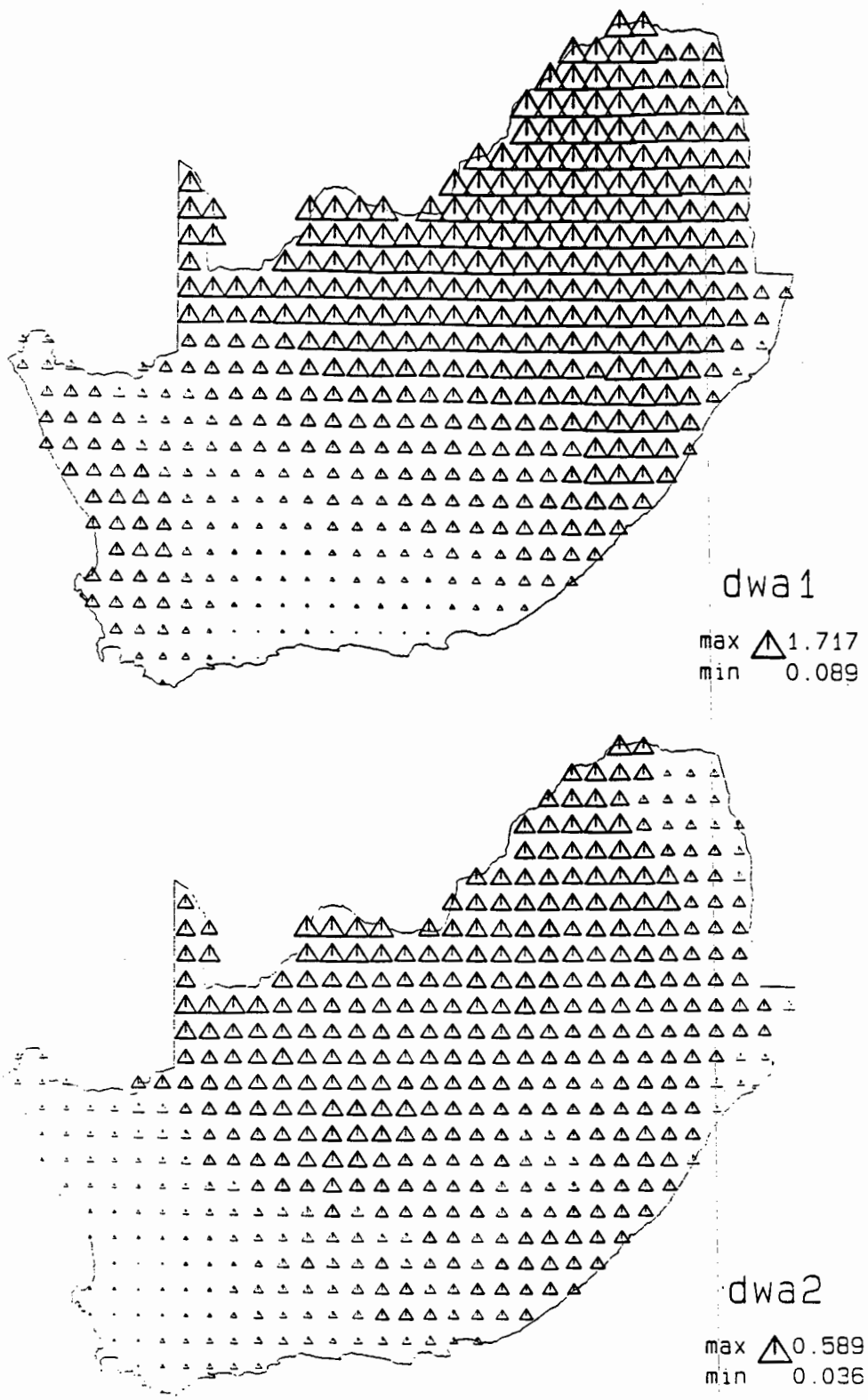


Figure 7.8: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

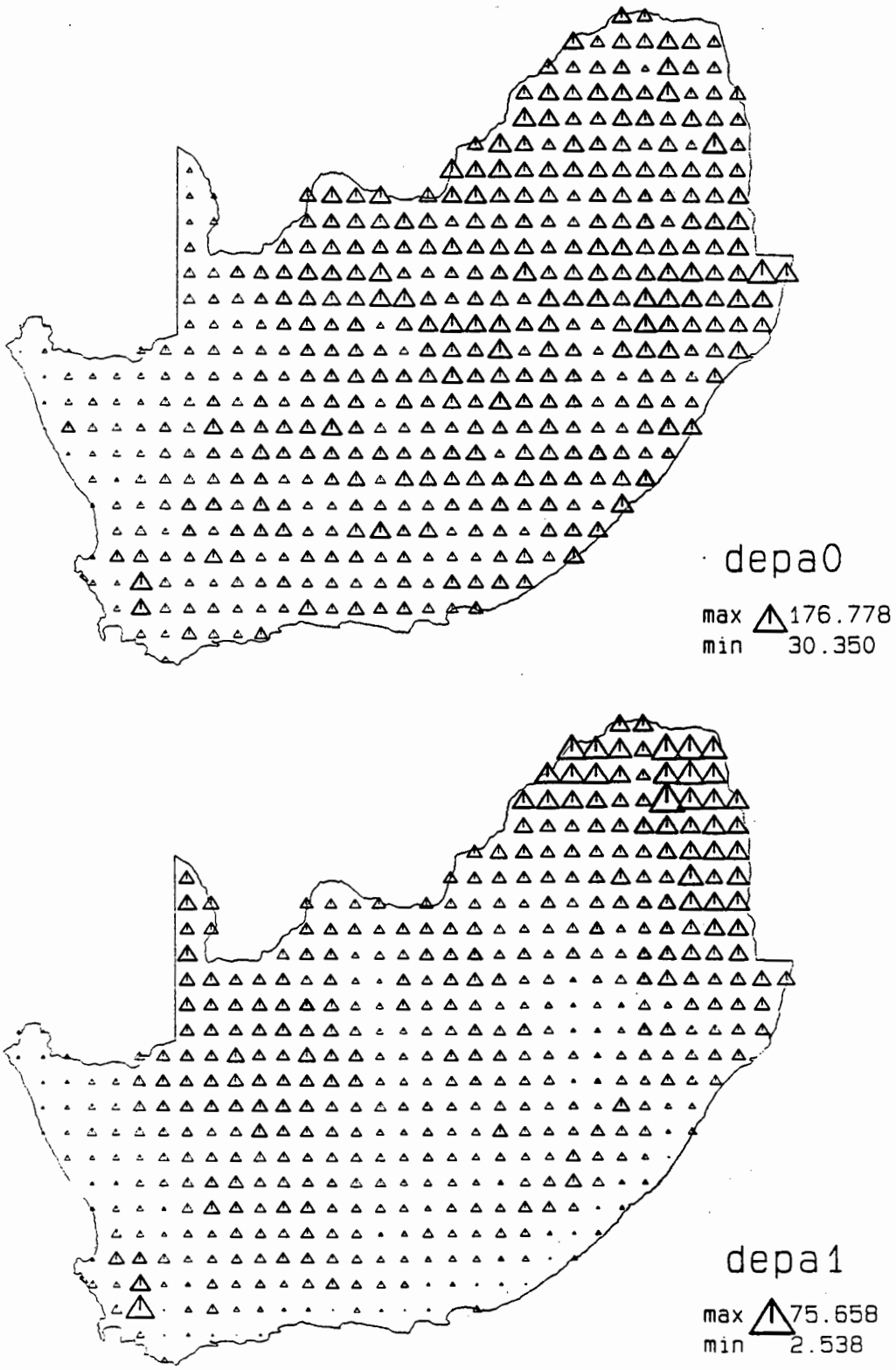


Figure 7.8: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

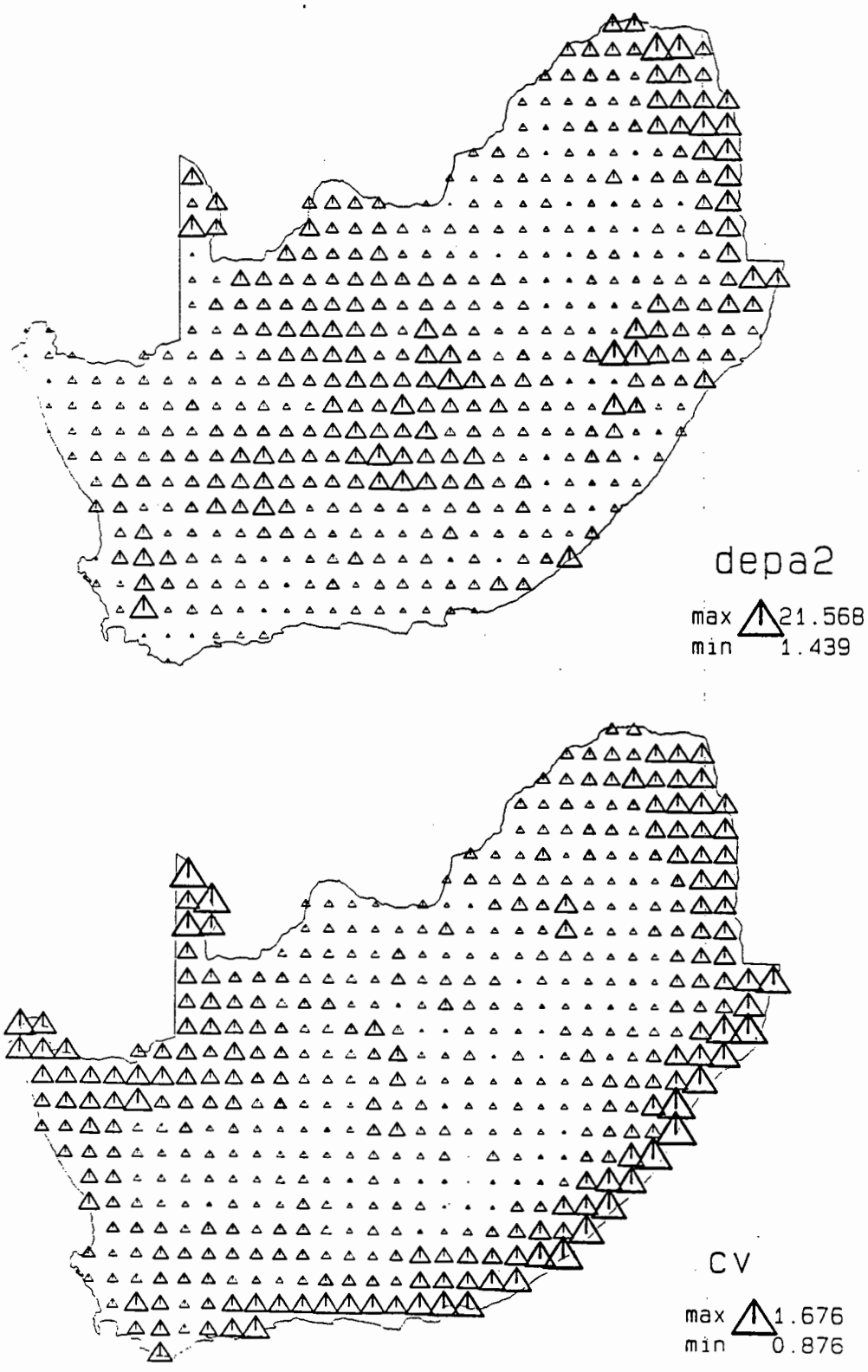


Figure 7.8: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

Chapter 8

Rainfall Model: Interpolation of the Phase Parameters

The phase parameters of the daily rainfall model [1] are *circular* in nature. In particular the first phase parameters take values between 0 and 365 while the second phase parameters take values between 0 and 365/2. Given two sites, one with $\phi_1 = 364$ (December 30) and the other with $\phi_1 = 3$ (January 3), say, an obvious estimate of the value of this phase parameter at a site situated midway between the two sites would be given, *not* by the arithmetic average $(364 + 3)/2 = 183.5$, but by the value $\phi_1 = 1$ (January 1). From this simple example it is clear that normal methods of calculation are inappropriate for circular data. Such data arise in various fields. The most common examples arise either from directional data in two-dimensional space, such as in studies of wind direction or direction of magnetisation of rock specimens, or else from periodic phenomena, such as the time of day or the time of year of the occurrence of certain events. The phase parameters of the rainfall model are of the latter type. Many other examples are given in the texts by Mardia (1972), Batschelet (1981) and Upton and Fingleton (1989).

[1] The work described in this chapter has previously been published in McNeill (1993) and McNeill *et al.* (1994).

Statistical techniques for circular data tend to be more computationally complex than their counterparts for data taking values on the real line. Notions of correlation, regression and bias are still the subject of discussion (Mardia, 1975; Jupp and Mardia, 1989). In particular, the development of smoothing techniques for spatially distributed circular data is very much in its infancy.

8.1 Smoothing Methods for Circular Data

Almost all of the published work on interpolation and smoothing of circular or spherical data has been directed towards the problem of data indexed in one dimension. Many of these methods are not truly vectorial in nature. For example, Parker and Denham (1979), studying the problem of plotting the position of the magnetic north pole at various geological times, suggested splining the individual x and y coordinates of the unit vector independently, and then re-normalizing to get unit vectors; this method could be extended to data available at locations in two dimensions using thin plate splines, but the method has been criticised by several authors, including Watson (1985) and Jupp and Kent (1987), as it ignores the geometry of the sphere. Other methods, such as that of Clark and Thompson (1984) which involves projecting the data on to a tangent plane, do not work well unless the data are restricted to a portion of the sphere or circle, and produce fitted paths which are not invariant to change of origin. Since the phase parameters of the rainfall model take values throughout the full range of angles on the circle, such methods are not suitable for our application.

The method proposed by Jupp and Kent (1987) does not readily extend to spatial data as they require that the circular data be 'unwrapped' on to the real line; it is not clear that this can be done consistently for data indexed in two dimensions.

Another feature of almost all the methods suggested for smoothing circular or spherical data is that they are spline based, and Watson (1985) suggests that, even in the relatively simple case when the data locations are on a line, rather than a plane, there is a need to penalise excessive spiralling of the estimated values around the unit circle, in addition to the usual roughness penalty used for splines, and this leads to a considerably more complex problem.

Watson (1985) is perhaps the first to discuss the problems of interpolating and smoothing circular data available at a number of spatial locations, and he outlines a couple of possible approaches. For the interpolation problem he suggests the use of a weighted average of the data points, with the weights chosen proportional to the inverse of the square of the distance from the point to be estimated. He goes on to outline a possible method for smoothing spatial directional data, represented by angles $\theta_1, ..\theta_n$, based on calculating estimated values $f(\mathbf{z}_1)$ to $f(\mathbf{z}_n)$ at the given spatial locations $\mathbf{z}_i$, $(i = 1, 2...n)$ to maximise

$$\sum_{i=1}^{n} k_i \cos(\theta_i - f(\mathbf{z}_i)) + \tau \sum_{1 \leq i < j \leq n} \cos(f(\mathbf{z}_i) - f(\mathbf{z}_j))/d_{ij}$$

where k_i is some inverse function of the measurement error variance of the ith data point, d_{ij} is a some monotonic increasing function of distance between the ith and jth data points and τ is a smoothing parameter. Watson does not give full details of the method; in particular, the discussion does not explain how the method generates estimates at points other than the original data locations $\mathbf{z}_i$. Also, if the original data locations are clustered in space, then excessive weight will be given to 'high density' areas. This criticism applies also to the simple weighted average method.

Mendoza (1986) has subsequently implemented a method of smoothing circular data available at locations on a plane and illustrates its use to smooth data on the cross-bedding directions of sandstone. His method, which is

rather similar to Watson's except that it uses a spline-based measure of smoothness, finds $f(\mathbf{z})$ to minimize

$$\sum_{i=1}^{n} w_i(1 - \cos(\theta_i - f(\mathbf{z}_i)) + \lambda R_2(f)$$

where θ_i is the observed angle and $f(\mathbf{z}_i)$ the smoothed angular value at the ith location, w_i is a weighting factor for the ith data point and $R_2(f)$ is a measure of roughness of the function f. The roughness criterion is similar to that used in spline-smoothing and is given by

$$R_2(f) = \int\int (\frac{\partial^2}{\partial x^2} f)^2 + 2(\frac{\partial^2}{\partial x \partial y} f)^2 + (\frac{\partial^2}{\partial y^2} f)^2 \, dx \, dy$$

Whereas for most calculations involving directional data only trigonometric functions of the data are used, so that θ_i and $\theta_i + 360$ (in degrees) are equivalent, in the calculation of $R_2(f)$ such values are *not* equivalent, so that it is first necessary to choose the value θ_i to represent each data point. Mendoza suggests that these values should be chosen '*so that observations are not rougher than they should be*'. Thus, for example, a sequence of adjacent values (in degrees) of 240 300 350 30 is re-expressed as 240 300 350 390. This may not be easy to achieve consistently for spatial data. For example, Figure 8.1 shows some data in the south-western Cape, taken from the values of WWP1 (converted to degrees), in which smoothing the data along the path indicated by the dotted line from the point A (192 degrees) to the point B leads to a labelling of 381 degrees for point B whereas smoothing the data along the path indicated by the dashed line leads to a value of 21 degrees for the same point; the lack of a natural ordering of points in two dimensional space means that appropriate values may not be uniquely defined.

Young (1987) extends the technique of kriging to *vector* data in a natural way, using the Euclidean norm as a measure of distance, and suggests that the resulting technique will be appropriate for directional or circular data. This is not the case however, since there is no guarantee that the vector estimate

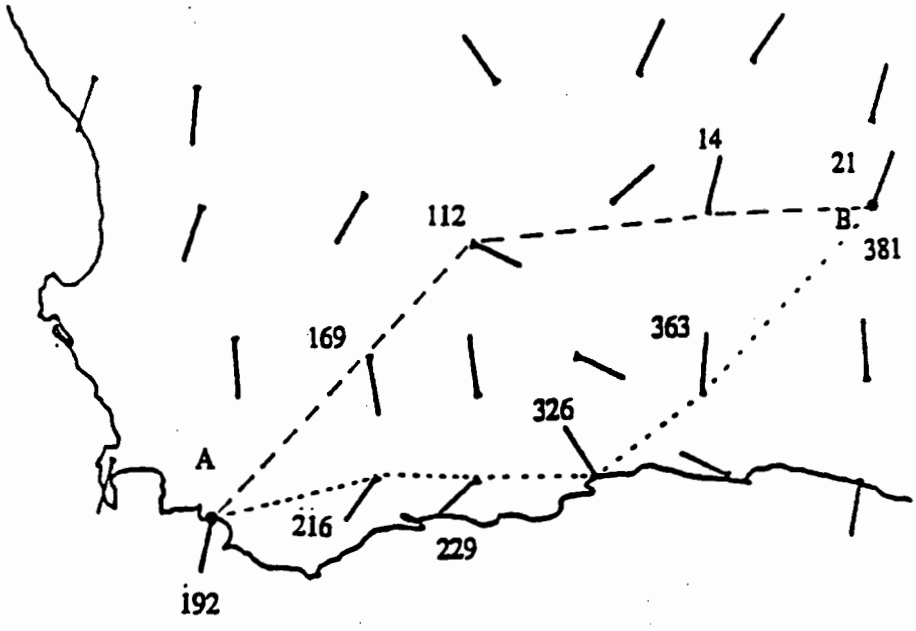


Figure 8.1: Re-labelling of circular values.

resulting from the vector kriging process will in fact be of unit length, and thus it may minimise the vector distance but not the angular distance. Thus, for example, in Figure 8.2 the vector B is closer (as measured by the Euclidean norm distance) to vector O than is vector A. However, in terms of angular distance, A is closer to O than B is. To get an estimate which minimises the angular distance it is necessary either to constrain the solution to lie on the unit circle, or else to try to minimise angular distance directly.

In the next section we explore the feasibility of extending kriging in this way.

8.2 Kriging for Circular Data

We start by reviewing some basic notation and summary statistics for circular data. This leads us to define a weighted average for circular data and

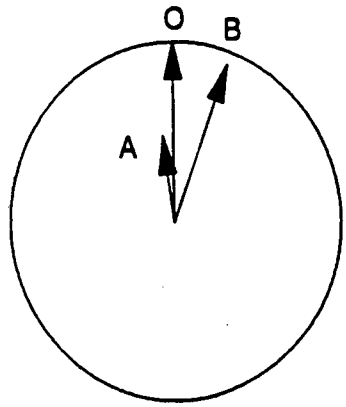


Figure 8.2: Vector distance and angular distance.

appropriate measures of angular distance and covariance.

8.2.1 Means and Variances for Circular Data

Circular data can be represented by points on a circle of unit radius. Where the data are not directions, as in the present application, the range of values can easily be mapped on to the circle; for example, in the case of the first phase parameters of the rainfall model , which take values between 0 and 365, we can multiply the values by $2\pi/365$ to get an equivalent value in radians. The mean of the data is then defined to be the direction of the resultant vector. That is, if we represent the data points by the unit vectors $\mathbf{e}_1$,$\mathbf{e}_2$, $\ldots \mathbf{e}_n$, with polar coordinates $(1,\theta_1),(1,\theta_2),\ldots(1,\theta_n)$ then the mean vector of the sample is given by

$$\mathbf{m} = \sum_{i=1}^{n} \mathbf{e}_i/n$$

If we assign a unit mass to each data point in Figure 8.3, then $\mathbf{m}$ represents the centre of gravity of the data. The cartesian coordinates of $\mathbf{m}$

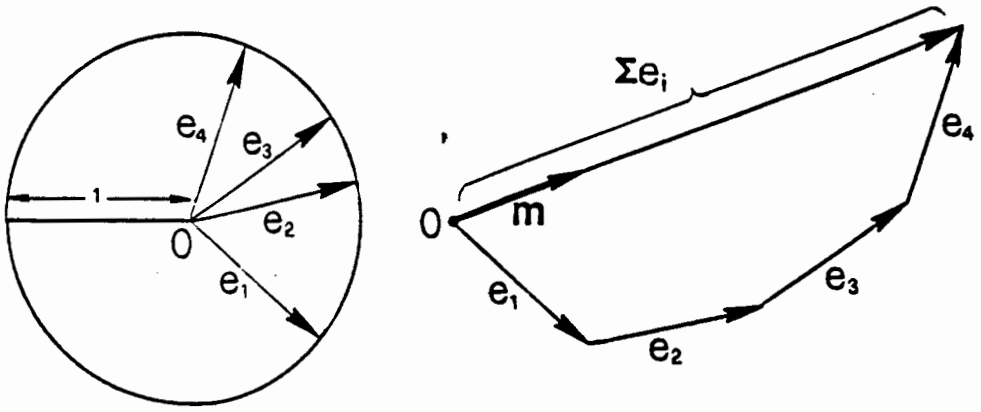


Figure 8.3: Mean of circular data.

are $\overline{x} = \sum \cos\theta_i/n$ and $\overline{y} = \sum \sin\theta_i/n$ and the polar coordinates are $(\overline{R}, \overline{\theta})$ where

$$\overline{R}^2 = \overline{x}^2 + \overline{y}^2$$

and

$$\tan\overline{\theta} = \sum \sin\theta_i / \sum \cos\theta_i$$

This has a singularity if $\sum \sin\theta_i = \sum \cos\theta_i = 0$, so that the centre of gravity is at the origin, and thus the resultant direction is not uniquely defined.

It can be shown that

$$\overline{R} = \sum_{i=1}^{n} \cos(\theta_i - \overline{\theta})/n$$

and thus the measure $[1 - \overline{R}]$ provides a measure of the variance, with properties which are in many ways analogous to those of the variance for non-circular data (Mardia, 1972)

To obtain a weighted mean with weights $w_1, \ldots w_n$ it is natural to define this as the direction corresponding to the point with coordinates $\overline{x} = \sum w_i \cos\theta_i/n$ and $\overline{y} = \sum w_i \sin\theta_i/n$, which is equivalent to assigning the

weights as masses to the data points and finding the centre of gravity as before. Note that multiplying all the weights by a non-zero constant, l, does not affect the direction of the weighted mean, but changes the length of the mean vector by a factor l.

8.2.2 The Kriging Equations

We consider a model equivalent to that used in *ordinary* kriging, that is, we assume the data are a realisation of a stochastic process with common mean and variance, and that the covariance, to be defined below, is a function of distance only.

Given unit vectors $\mathbf{e}_1,\mathbf{e}_2,...\mathbf{e}_n$ at spatial locations $\mathbf{z}_1,\mathbf{z}_2,...\mathbf{z}_n$, we seek an estimator, based on the weighted sum $\sum w_i\mathbf{e}_i$, of the value $\mathbf{e}_0$ at some location $\mathbf{z}_0$. Specifically, if $\sum w_i\mathbf{e}_i$ is written in polar form as $(R_0, \hat{\theta}_0)$ then we seek $\mathbf{w}$ to minimise

$$E[1 - \cos(\hat{\theta}_0 - \theta_0)] \tag{8.1}$$

where θ_0 is the unknown angular value at location $\mathbf{z}_0$. The use of the function $1-\cos(\hat{\theta}-\theta)$ as a measure of estimation error is common in circular statistics, and is analogous to the usual least squares criterion (Fisher and Lewis, 1985).

If we write $\mathbf{e}_i = (x_i, y_i)'$ so that $\theta_i = \arccos(x_i) = \arcsin(y_i)$ and $\mathbf{e}_0 = (x_0, y_0)'$ with $\theta_0 = \arccos(x_0) = \arcsin(y_0)$, then we have

$$\sin\hat{\theta}_0 = \sum_{i=1}^{n} w_i y_i / R_0 = \sum_{i=1}^{n} w_i \sin\theta_i / R_0$$

and

$$\cos\hat{\theta}_0 = \sum_{i=1}^{n} w_i x_i / R_0 = \sum_{i=1}^{n} w_i \cos\theta_i / R_0$$

where R_0 is the length of the vector $\sum w_i\mathbf{e}_i$ so that

$$R_0^2 = (\sum_{i=1}^{n} w_i x_i)^2 + (\sum_{i=1}^{n} w_i y_i)^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \cos\theta_i \cos\theta_j + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sin\theta_i \sin\theta_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \cos(\theta_i - \theta_j) \\
&= \mathbf{w}'Q\mathbf{w}
\end{aligned}
$$

where $q_{ij} = \cos(\theta_i - \theta_j)$.

Now

$$
\begin{aligned}
\cos(\hat{\theta}_0 - \theta_0) &= \cos\hat{\theta}_0 \cos\theta_0 + \sin\hat{\theta}_0 \sin\theta_0 \\
&= 1/R_0 \left(\sum_{i=1}^{n} w_i(\cos\theta_i \cos\theta_0 + \sin\theta_i \sin\theta_0) \right) \\
&= \mathbf{w}'\mathbf{c}/\sqrt{(\mathbf{w}'Q\mathbf{w})}
\end{aligned}
$$

where $c_i = \cos(\theta_i - \theta_0)$

Thus in order to minimise $E[1 - \cos(\hat{\theta}_0 - \theta_0)]$ we need to find $w_1, ...w_n$ to maximise

$$E[\mathbf{w}'\mathbf{c}/\sqrt{\mathbf{w}'Q\mathbf{w}}] \tag{8.2}$$

It is clear from the formula above that the solution will be unique only up to a constant multiplier; that is, if $\mathbf{w}$ is a solution, then so is $l\mathbf{w}$ for any non-zero constant l. In order to make the solution unique, we might introduce a constraint, such as $\sum w_i = 1$, or $E[R_0^2] = E[\mathbf{w}'Q\mathbf{w}] = 1$. Note that the constraint $\mathbf{w}'Q\mathbf{w} = 1$ is *not* an appropriate choice, since then the w_i will depend on the data values, so that, for example, we would not have $E[\mathbf{w}'\mathbf{c}] = \sum w_i E[c_i]$.

The expected values of the elements of Q and $\mathbf{c}$ in equation 8.2 can be estimated by modelling an appropriate spatial covariance function as described in Section 8.2.3 below; however, this is not of immediate help in finding a solution, since the expression to be maximised cannot be expressed directly as a linear or quadratic function of these, so that the usual method of maximisation used in kriging does not apply directly in this case.

If we use a first order Taylor series expansion of $E[\mathbf{w}'\mathbf{c}/\sqrt{\mathbf{w}'Q\mathbf{w}}]$, so that we approximate it by $\mathbf{w}'\mathbf{s}/\sqrt{\mathbf{w}'K\mathbf{w}}$, where $k_{ij} = E[q_{ij}]$ and $s_i = E[c_i]$, and use the uniqueness constraint $E[\mathbf{w}'Q\mathbf{w}] = \mathbf{w}'K\mathbf{w} = 1$ then we can use the Lagrange multiplier approach to find the optimal values of the w_i. With the chosen uniqueness constraint, the function to be maximised becomes simply $\mathbf{w}'\mathbf{s}$, so if we set

$$G = \mathbf{w}'\mathbf{s} - \lambda(\mathbf{w}'K\mathbf{w} - 1)$$

then

$$\frac{\partial G}{\partial \mathbf{w}} = \mathbf{s} - 2\lambda K\mathbf{w}$$

and

$$\frac{\partial G}{\partial \lambda} = \mathbf{w}'K\mathbf{w} - 1$$

which leads to the solution

$$\mathbf{w} = \frac{K^{-1}\mathbf{s}}{\sqrt{\mathbf{s}'K^{-1}\mathbf{s}}} = \frac{K^{-1}\mathbf{s}}{r} \tag{8.3}$$

where $r = \sqrt{\mathbf{s}'K^{-1}\mathbf{s}}$ is a scalar normalising constant.

For the case where the data include measurement error, as is the case with the rainfall model parameters, so that we observe $\vartheta_i = \theta_i + \epsilon_i$ instead of θ_i, a similar argument shows that the approximate solution to obtaining an optimal estimate of θ_0 is given by the same expression but with $k_{ij} = E[\cos(\vartheta_i - \vartheta_j)]$ and $s_i = E[\cos(\vartheta_i - \theta_0)]$.

The form of the solution given by equation 8.3 is, apart from the normalising constant r, exactly analogous to the solution of the usual (non-circular) problem of *simple* kriging. The form of this solution is intuitively appealing, in that it gives more weight to those data points which are close (in space) to the point to be estimated (via $\mathbf{s}$) and gives less weight to points which are clustered with other data points (via K^{-1}). Thus, although equation 8.3 gives only an approximate solution to the minimisation of equation 8.1, it can be justified in its own right as a form of weighted average which caters

specifically for clustered data, and can also cater for varying error variances. It is thus of interest to see how well such a method performs in practice.

The performance of the method was therefore compared with the simple weighted average method described in Section 8.1, using a set of test sites. Before carrying out the estimation it is first necessary to model the spatial covariance so as to have values for $E[\cos(\vartheta_i - \vartheta_j)]$ and $E[\cos(\vartheta_i - \theta_0)]$.

8.2.3 Modelling the Spatial Covariance

Many measures of association have been proposed for circular data, and reviews can be found in Jupp and Mardia (1989) and Breckling (1989). For our purposes here it suffices to find a measure of the relationship between two circular variables with the same distribution, and, in particular, with the same mean, since, in using local kriging, trends can be ignored. In view of the form of the kriging equations an obvious choice is

$$\sigma_{ij} = E[\cos(\theta_i - \theta_j)]$$

This is in fact equivalent to the measure proposed by Breckling (1989) in the case where the means of θ_i and θ_j are the same.

If we assume that the covariance is a function of distance only then we can estimate the covariance function from the data by plotting $\cos(\theta_i - \theta_j)$ as a function of d_{ij}, the distance between the two data points. Alternatively we may prefer to define a circular semi-variogram as

$$\gamma(h) = E_{(d_{ij}=h)}[\frac{1}{2}(1 - \cos(\theta_i - \theta_j))] \tag{8.4}$$

as a circular analogue of the usual semi-variogram, where the expectation is over all locations i and j with separation distance h. The circular semi-variogram as thus defined takes values between 0 and 1. In order to study the empirical circular semi-variogram as a function of separation distance we

can plot the average of the values $(1 - \cos(\theta_i - \theta_j))/2$ for all pairs of points with a given separation distance as a function of the separation distance.

When the data are measured with error, this empirical semi-variogram will be increased by an amount depending on the error variance. Specifically, suppose that we observe angular values ϑ_i such that

$$\vartheta_i = \theta_i + \epsilon_i$$

where the angles ϵ_i represent measurement error, so that we may assume that the values of ϵ_i at different sites are independent of one another and also of the values θ_i. We also assume that the distribution of ϵ_i is symmetric with mean zero so that $E[\sin \epsilon_i] = 0$. Then we have:

$$\begin{aligned}
E[\cos(\vartheta_i - \vartheta_j)] &= E[\cos(\theta_i + \epsilon_i - \theta_j - \epsilon_j)] \\
&= E[\cos(\theta_i - \theta_j)\cos(\epsilon_i - \epsilon_j) - \sin(\theta_i - \theta_j)\sin(\epsilon_i - \epsilon_j)] \\
&= E[\cos(\theta_i - \theta_j)(\cos\epsilon_i \cos\epsilon_j + \sin\epsilon_i \sin\epsilon_j) \\
&\quad - \sin(\theta_i - \theta_j)(\sin\epsilon_i \cos\epsilon_j - \cos\epsilon_i \sin\epsilon_j) \\
&= E[\cos(\theta_i - \theta_j)\cos\epsilon_i \cos\epsilon_j] \\
&= E[\cos(\theta_i - \theta_j)]E[\cos\epsilon_i]E[\cos\epsilon_j]
\end{aligned}$$

Similarly, we can show that

$$E[\cos(\vartheta_i - \theta_0)] = E[\cos(\theta_i - \theta_0)]E[\cos\epsilon_i]$$

Estimates of the terms $E[\cos\epsilon_i]$ were calculated for each parameter at each rainfall station as part of the bootstrap variance calculations described in Chapter 2, and thus it is possible to estimate the covariance of the underlying θ values using the adjusted covariance estimator

$$\hat{\sigma}_\theta(h) = 1/N_h \left\{ \sum \left(\frac{\cos(\vartheta_i - \vartheta_j)}{E[\cos(\epsilon_i)]E[\cos(\epsilon_j)]} \right) \right\}$$

where the summation is over all N_h pairs of points a distance h apart. If one prefers to work with the circular semi-variogram defined above then a

similar adjustment may be made to the semi-variogram calculated from the observed ϑ values to get an estimate of the semi-variogram of the θ values.

Figure 8.4 shows the unadjusted and adjusted semi-variogram for each of the phase parameters of the daily rainfall model. The effect of the adjustment is less marked for the first phase parameters than for the second phase parameters, indicating that the former are subject to less estimation error. After adjustment, the graphs show little evidence of any residual nugget effect, which means that the phase parameters do not change significantly over short distances. This confirms the assumption that while the amplitude parameters would be sensitive to local topography, the phase parameters would not. The few negative values in some of the adjusted graphs result from the measurement error adjustment; clearly, in fitting a model to such data we require all the fitted values to be positive.

8.2.4 Validation and Discussion

To test the circular kriging method proposed in the previous section, the method was compared with the simpler method of inverse distance weighting using a test set of 101 rainfall stations and a data set of 325 rainfall stations selected from the full data set. The test sites selected lie approximately on a regular grid, with one station having been selected at random from every alternate Weather Bureau block, while the 'data' sites were obtained by selecting one station at random from every sequence of ten stations remaining in the data file after removal of the test sites. Since the stations are ordered in the data file by Weather Bureau block, this method of selection ensures that the 'data' sites have a similar spatial distribution to the full data set, but are much more sparse (Figure 8.5).

While the sparseness of the selected data points can be expected to result in rather poor estimates, it should also help to highlight the difference

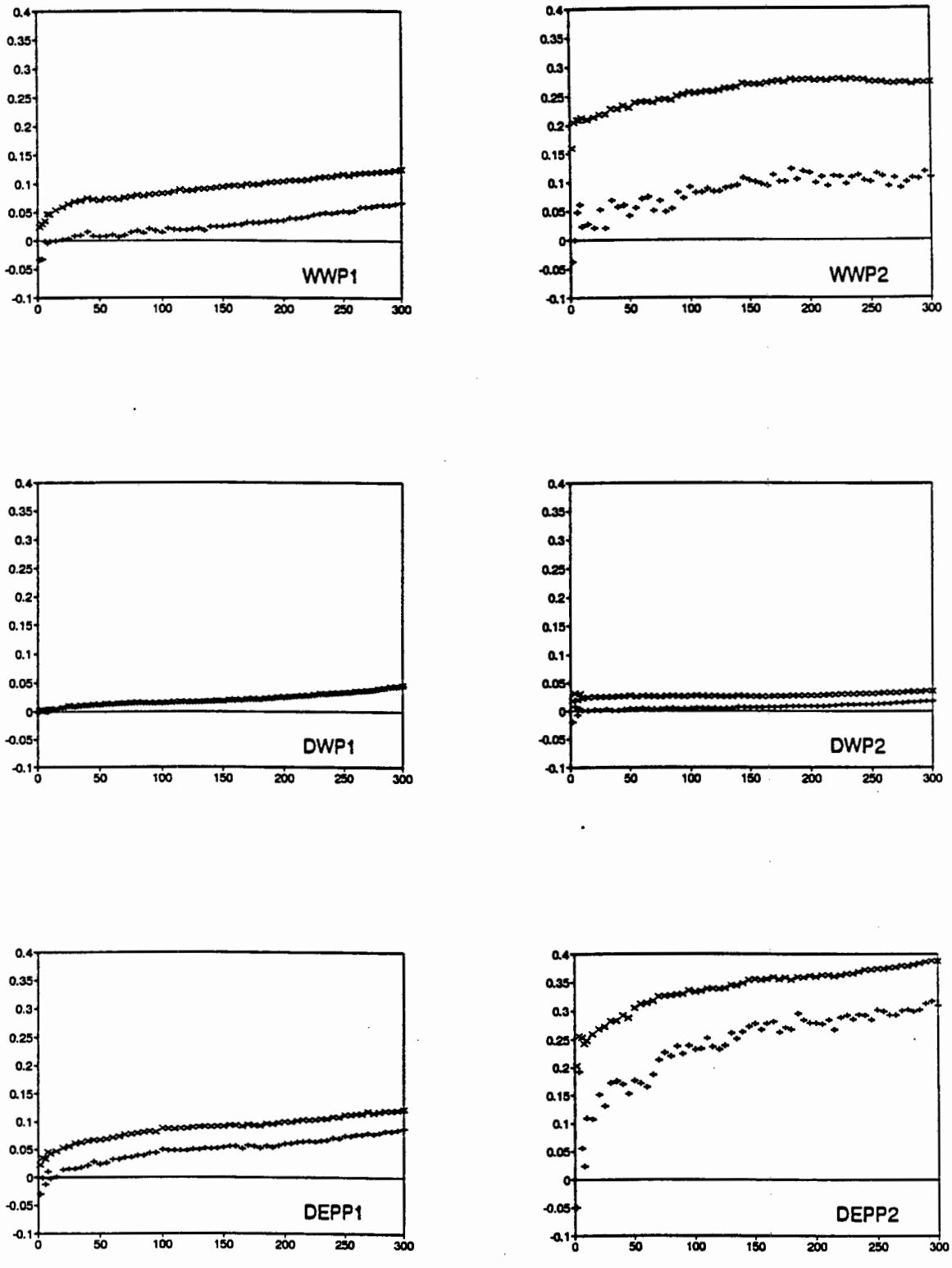


Figure 8.4: Semi-variograms: phase parameters.

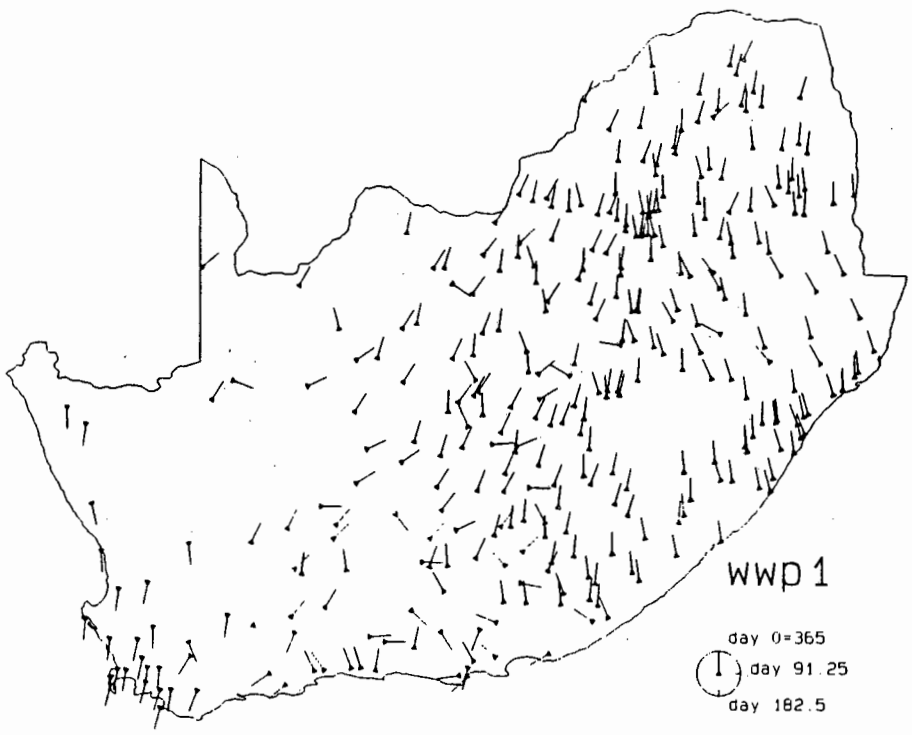


Figure 8.5: Data set used in circular kriging validation.

between the two methods being tested, since, if too dense a data set is used, almost all smoothing methods will give good results. In both methods a search radius of 300 kilometres was used, that is, only points within 300 kilometres of the point to be estimated were included in the calculation.

The unadjusted and adjusted semi-variogram was calculated using the selected data sites for the parameter WWP1, and a Gaussian model, given by

$$\gamma(h) = s(1 - \exp(-h^2/r^2)) \tag{8.5}$$

was fitted, with sill and range parameters 0.0929 and 233 respectively for the adjusted semi-variogram. The average of the error terms, $[1 - \cos(\vartheta_i - \hat{\theta}_i)]$ (averaged over the 101 test data points), was compared for the two methods,

and the results are shown Table 8.1 below, from which it is clear that the kriging method has resulted in considerably lower errors on average.

Method	Average Error
Inverse distance	0.1318
Kriging	0.0988

Table 8.1: Comparison of prediction errors: WWP1.

There are several reasons why the kriging method may give better results than the inverse distance method. One is, of course, that in using the inverse distance method no attempt has been made to optimise the particular inverse distance function used; it would be possible to use a parametric function of distance, with the parameter value controlling the effective bandwidth selected, for example, by cross-validation, but this to some extent reduces the main advantage of the inverse distance method, namely its simplicity. Another possible reason for the superiority of the kriging method is that the inverse distance method does not take account of clustering in the data; however, a study of Figure 8.5 suggests that this is probably not of great consequence for this particular data set, as the clustered data points generally have similar values to the more isolated points around them. A third reason for the superiority of the kriging method is its explicit use of the error variance of the data; the inverse distance method will give relatively high weight to the few points which are closest to the point to be estimated even if those data points have high measurement error, whereas the kriging method will adjust for this; this is quite important in the present application where the error variance, as measured via the bootstrap procedure, was rather high at some sites.

In comparing individual estimated values with the original values in the test data set one must bear in mind that even the values in the test data set are not entirely accurate but are subject to the parameter estimation errors

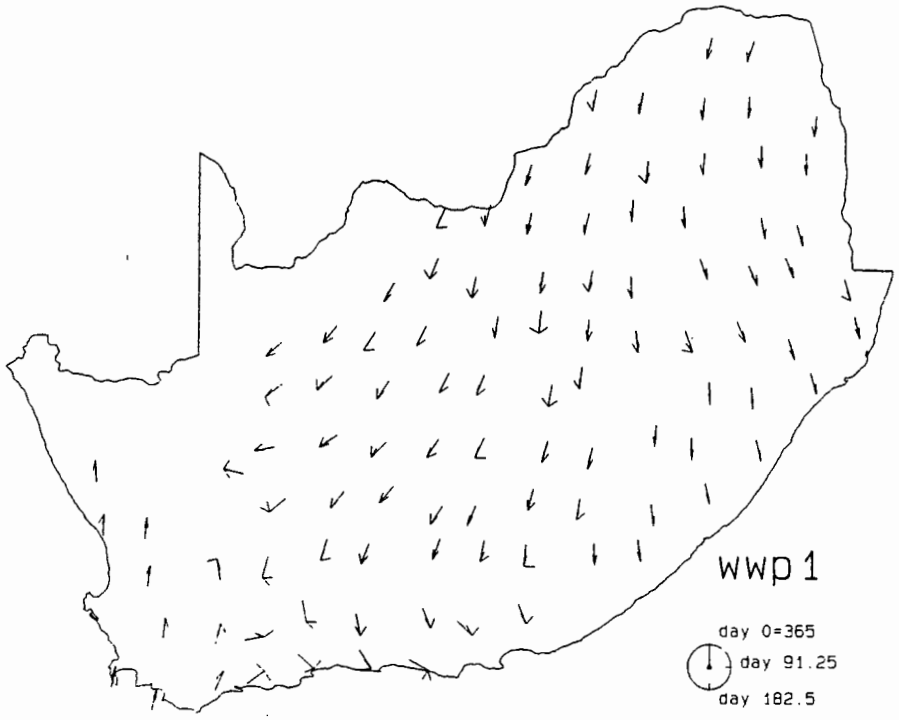


Figure 8.6: Kriging estimates at test sites.

as estimated by the bootstrap procedure. Therefore, in plotting a map of the values estimated by the kriging method (Figure 8.6) I have included for comparison, not the original data values, but a *range* of values (indicated by the two short lines emanating from each site in the figure) which represent a range of $\vartheta_i \pm \varepsilon_i$, where ϑ_i is the original data value at that site and $\varepsilon_i = \arccos(\sum_{j=1}^{100} \cos(\vartheta_{ij} - \overline{\vartheta_{ij}}))$, where the ϑ_{ij} are the individual bootstrap estimates described in Chapter 2 and $\overline{\vartheta_{ij}}$ is their mean. For our data set this range corresponds roughly to an approximate 95% confidence interval for data having a von Mises distribution, based on the formulae given in Section 9.6 of Upton and Fingleton (1989).

It can be seen in Figure 8.6 that the fit of the estimated values is generally

good, except for five sites which lie in the area of change-over between the winter rainfall area in the south west and the summer rainfall area further to the north and east. The test data set is relatively sparse in this area; clearly more data points are needed for accurate estimation in this region. In practice, of course, the full data set has over 5000 points compared with the 325 used here, which should give much more accurate results throughout the country.

In the comparison described above the semi-variogram parameters were estimated directly from the empirical semi-variogram. However, since the solution given by equation 8.3 is only approximate as it is based on the linear Taylor series expansion there is no reason why these parameters should be optimal and it is likely that better results would be obtained if the sill and range parameters were estimated via cross-validation. However, the cross-validation approach is more computationally intensive and it is thus of interest to test the sensitivity of the kriging method to the semi-variogram parameters. The estimation process was therefore repeated with a range of values of these parameters, and some of the results are shown table 8.2, where the average error is given as a function of the sill and range parameters. It is apparent from the results that the method is not particularly sensitive to the choice of parameters, as the error surface is fairly flat in the region of the parameter values used, although the global optimum (0.0942) was found at values of roughly 0.20 for the sill and 350 for the range. In fact, a detailed search suggests that the error surface may well have several local minima, for example in the region of $s = 0.09$ and $r = 300$, a fact that has been noted by other researchers in connection with maximum likelihood estimation of spatial covariance functions (Warnes and Ripley, 1987). Thus it would seem that, for this data set at least, using cross-validation to estimate optimal parameters is not likely to give much improvement over the computationally quicker method used here, based on modelling the empirical semi-variogram.

	Sill				
Range	0.07	0.08	0.09	0.10	0.11
100	0.1083	0.1081	0.1080	0.1078	0.1077
150	0.1069	0.1072	0.1075	0.1076	0.1077
200	0.1013	0.1019	0.1024	0.1029	0.1034
250	0.0964	0.0967	0.0971	0.0975	0.0978
300	0.0948	0.0947	0.0947	0.0947	0.0948
350	0.0961	0.0956	0.0953	0.0950	0.0948
400	0.0973	0.0969	0.0966	0.0963	0.0961
450	0.0973	0.0970	0.0968	0.0967	0.0965
500	0.0975	0.0970	0.0966	0.0963	0.0962

Table 8.2: Average error as a function of sill and range parameters.

Another possible method of improving the accuracy of estimation would be by re-estimating the semi-variogram locally, as suggested by Haas (1990). For example, it is likely that the effective range of the spatial covariance would be smaller in the change-over region between the winter and summer rainfall area than it is in the middle of the summer rainfall area. However, as mentioned in Section 7.2.1, such a moving-window approach is excessively computationally intensive and thus effectively impractical in a project such as this. In addition, the advantage of a locally-calibrated semi-variogram model must be offset against the fact that relatively fewer data points will be used to estimate each local model and thus the model-fitting procedure will be less robust.

For the final estimation of the phase parameters throughout South Africa, the semi-variograms for all parameters were estimated using the full data set (Figure 8.4), and the kriging weights were calculated for each grid point using a search radius of 120 km. Figure 8.7 shows maps of the resultant estimates at the centre of each Weather Bureau block.

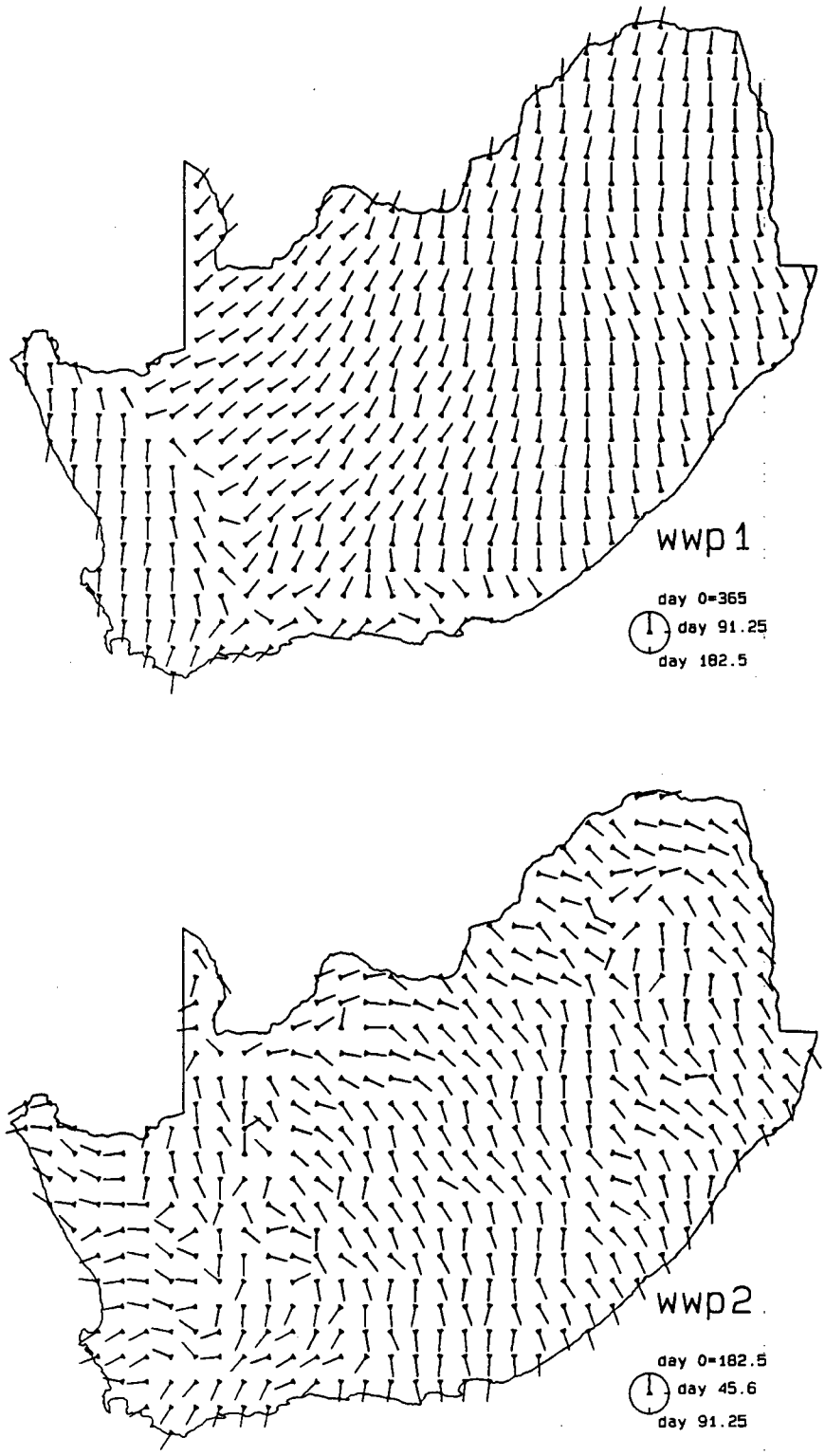


Figure 8.7: Estimated parameter values at centres of Weather Bureau blocks (phase parameters).

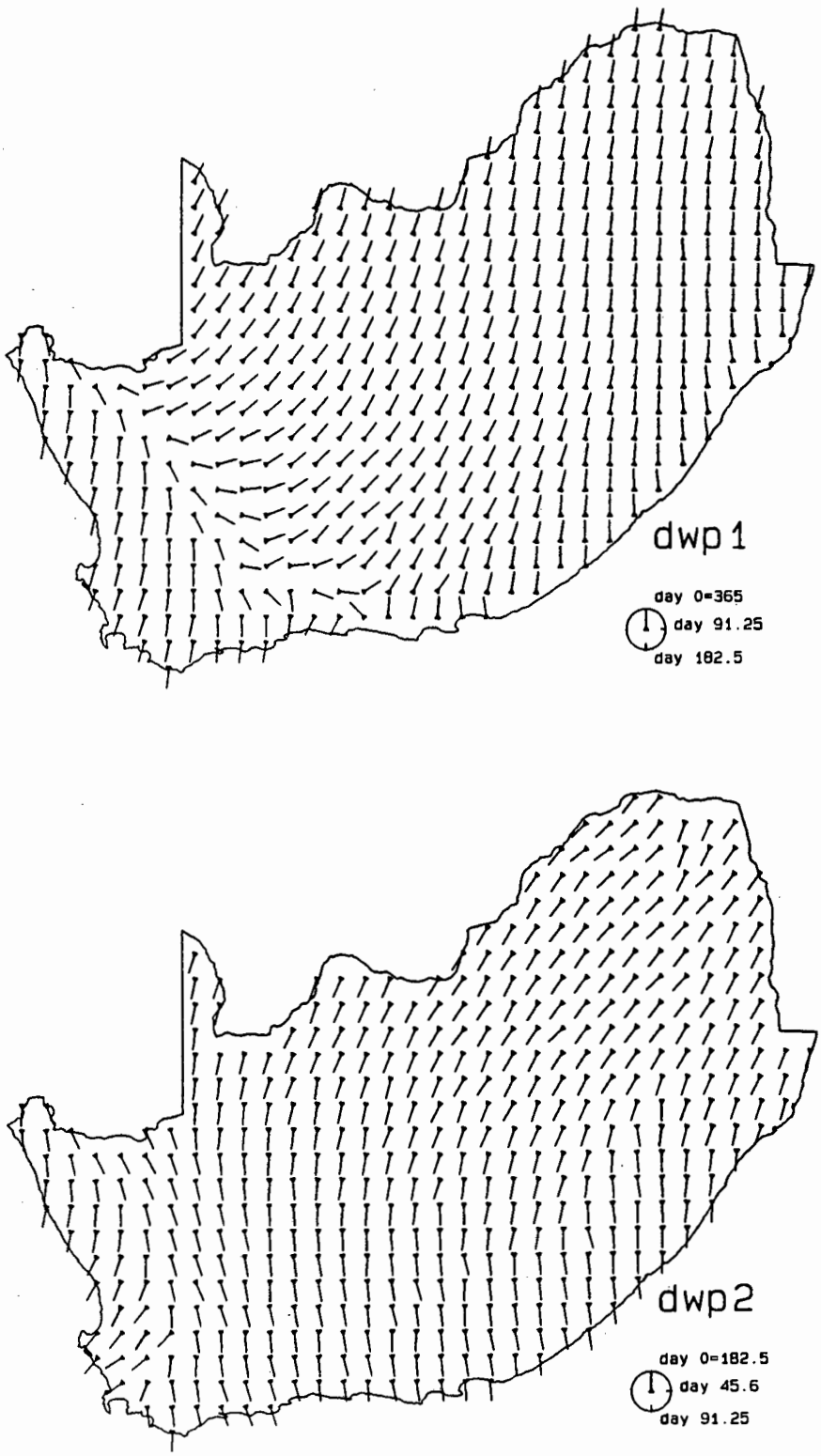


Figure 8.7: Estimated parameter values at centres of Weather Bureau blocks (phase parameters) (contd.).

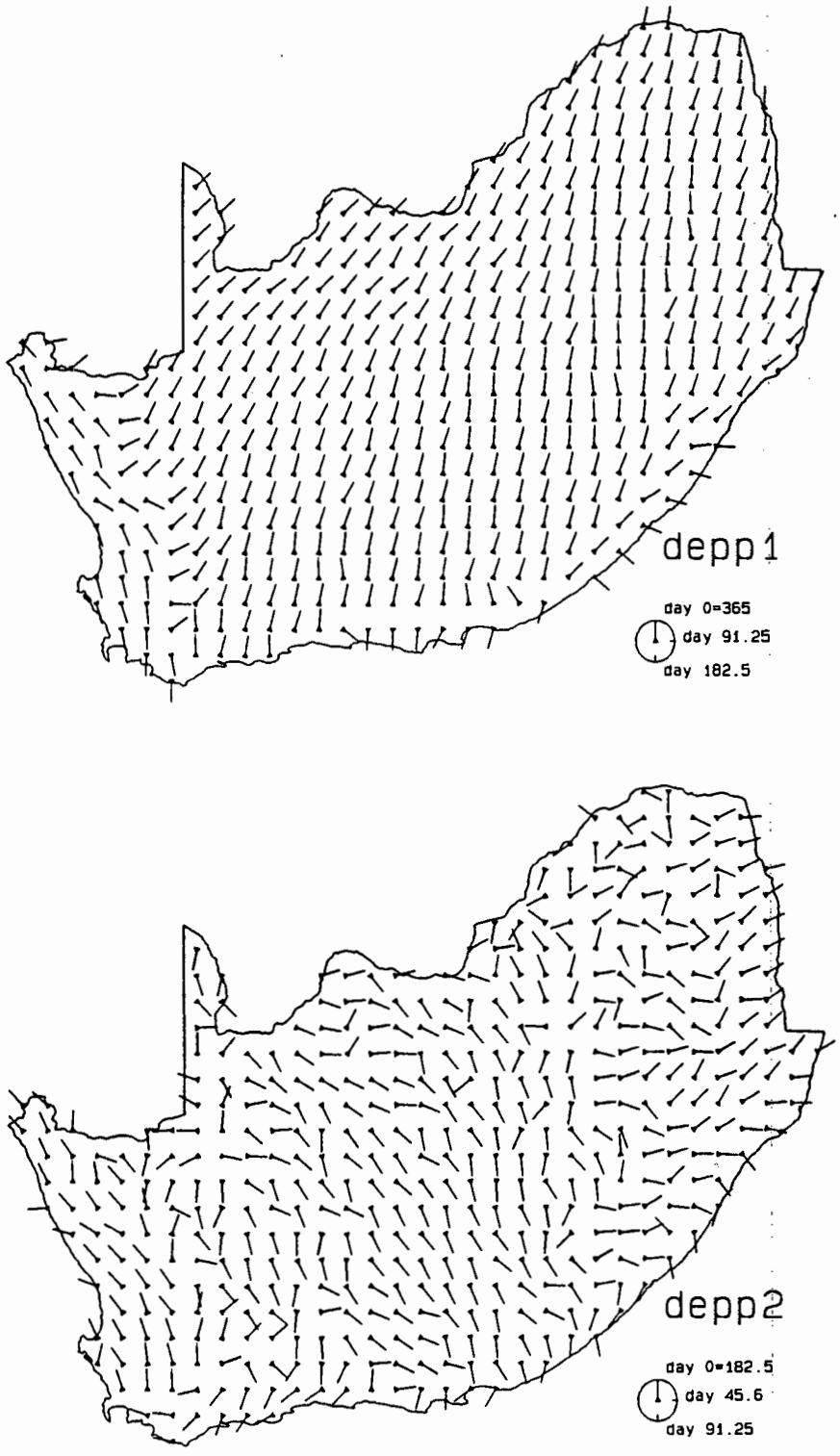


Figure 8.7: Estimated parameter values at centres of Weather Bureau blocks (phase parameters) (contd.).

8.2.5 Measure of Estimation Error

One of the advantages of the kriging method is that it provides a measure of the error of estimation, although, as discussed in Section 5.2, the kriging variance may under-estimate the error, since it does not take account of the fact that the covariance model is estimated.

For the estimator given by equation 8.3 we have

$$\begin{aligned} E[\cos(\hat{\theta}_0 - \theta_0)] &= E[\mathbf{w}'\mathbf{c}/\sqrt{\mathbf{w}'Q\mathbf{w}}] \\ &= E[\frac{\mathbf{s}'K^{-1}\mathbf{c}/r}{\sqrt{(\mathbf{s}'K^{-1}/r)Q(K^{-1}\mathbf{s}/r)}}] \\ &= E[\frac{\mathbf{s}'K^{-1}\mathbf{c}}{\sqrt{\mathbf{s}'K^{-1}QK^{-1}\mathbf{s}}}] \\ &\approx \sqrt{\mathbf{s}'K^{-1}\mathbf{s}} \end{aligned} \tag{8.6}$$

using again a first order Taylor series approximation. A similar expression holds for the case where the data are measured with error provided that k_{ij} is defined as $E[\cos(\vartheta_i - \vartheta_j)]$ and s_i as $E[\cos(\vartheta_i - \theta_0)]$.

In the analysis described in this thesis I have assumed that the covariance model is constant throughout the entire area, and thus the kriging estimation error at any point will reflect only the configuration of data points in that locality and the inherent measurement error of those data values.

It is difficult with our data set to check the estimation errors, since we do not know the exact values of θ_0 at the test points but only $\vartheta_0 = \theta_0 + \epsilon_0$. However, an examination of the estimation errors as given by $\cos(\hat{\theta}_0 - \vartheta_0)$ for the 101 test points showed that none of the errors exceeded the approximate expectation as given in equation 8.6 by more than a very small amount, with the notable exception of the five points in the change-over area between the summer and winter rainfall regions, where the observed error was much greater than predicted. This again suggests that it might be more realistic to use locally computed semi-variogram models, although this

will not necessarily result in more accurate estimation, as discussed in the previous section.

8.3 Rainfall Model Validation

Zucchini and Adamson (1984a) validated the daily rainfall model by comparing the relevant characteristics of simulated daily rainfall data generated by the model at individual rainfall stations with the corresponding values deduced directly from the original data. The kriging process described in this thesis extends the original 5000 stations at which the model is available to some 500000 points throughout southern Africa. Since most of these points do not coincide with the location of rain gauges, it is not possible to validate them in the same way. In addition, at those grid points which do coincide with rainfall stations we do not expect the estimated model parameters for the grid point to be equal to the values fitted to the original data at the station, since the kriging process takes into account the estimation error in the fitting of the original parameters and also the error introduced by the limited accuracy of the station locations. However, a comparison of grid point and station values will give some indication of the validity of the kriging process.

Rather than comparing individual model parameter values, it is more meaningful to compare derived characteristics, such as the mean annual precipitation, based on simulated data generated by the model; this enables us to test the model as a whole in the form in which it will be used in practice, and also allows comparison with the same statistics derived from other sources. The mean annual precipitation (MAP) was therefore calculated at the location of each of the 373 test sites described in Section 7.2.3 using four different methods:

- Using a 100 year simulation based on the daily rainfall model parameters estimated for that station.

- Using a 100 year simulation based on the daily rainfall model parameters estimated by the kriging procedure at the grid point with the same latitude and longitude as the station.

- Using the MAP calculated directly from the daily rainfall data for that station held by Computing Centre for Water Research (CCWR).

- Taking the value of MAP from the CCWR data base of gridded MAP values, as estimated by Dent *et al.* (1987).

The results are shown in Table 8.3. There are several reasons for the differences between the four values; in particular, sampling variability introduced by the simulation process, uncertainty in the exact station location relative to the grid point, estimation error in the daily rainfall model parameters, estimation error in the kriging procedure, estimation error in the CCWR gridded MAP value calculations, outliers in the daily rainfall data, and also the use of data for a different time period (the grid values estimated by Dent *et al.* (1987) include data up to May 1987 and thus exclude the more recent rainfall data up to the end of February 1992 which were incorporated in the other three values). In general, however, the agreement between the four sets of figures is quite close.

The MAP estimates based on the kriged values are also compared graphically with the other three sets of values in Figure 8.8. As might be expected, the agreement with the values based on the daily rainfall model fitted to the station data (diagram A) is the closest; any discrepancy is due to the allowance for model-fitting error and the uncertainty of the exact station locations in the kriging process and also to the sampling variation of the simulation process. In diagram B, where the kriged values are compared with those calculated directly from the CCWR rainfall data at the station, the discrepancies incorporate also any inherent 'lack of fit' of the seasonal

Markov chain model described in Chapter 2. In diagram C, the discrepancies are generally greater, as they now include also the effects of estimation errors inherent in the regression procedure used by Dent *et al.* (1987) and also the effect of using a slightly different historical set of daily rainfall data, as mentioned above.

Thus while Sections 8.2.4 and 7.2.3 gave some confirmation of the validity of the kriging methodology applied to individual rainfall model parameters, the results shown here indicate that the process of separately estimating the 16 parameters and then combining these estimates to provide a model of rainfall at any location is also valid. The kriging estimation procedures have thus allowed us to extend the applicability of the original Zucchini and Adamson model to sites throughout southern Africa.

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
2885 W	-34	45	20	0	77	464	468	471	483
3032 W	-34	32	20	2	112	473	478	467	466
4891 W	-34	21	18	30	30	370	373	368	363
5605 A	-34	5	18	51	94	654	648	650	642
6733 W	-34	13	19	25	112	534	538	529	473
7698 A	-34	8	19	54	53	438	436	431	430
8136 A	-34	16	20	5	60	404	399	402	393
9815 W	-34	5	20	58	90	408	407	412	401
11065 W	-34	5	21	33	42	450	452	451	428
12215 W	-34	5	22	8	24	451	454	466	391
17582 A	-34	12	24	50	94	671	657	657	657
21055 W	-33	55	18	32	85	481	483	472	484
22038 W	-33	38	19	2	87	766	763	751	767
23678 W	-33	48	19	53	106	333	317	322	320
24197 W	-33	47	20	7	79	315	312	321	269
25599 W	-33	59	20	50	93	1026	1016	1019	1015
26510 W	-34	0	21	17	55	654	644	642	645
27302 W	-33	32	21	41	112	200	196	194	198
28838 W	-33	58	22	28	107	884	894	878	778
29805 W	-33	55	22	57	99	838	839	828	830
30090 W	-34	0	23	3	80	916	920	900	953
31237 W	-33	57	23	38	105	1011	1008	994	1003
32209 W	-33	59	24	7	98	1145	1145	1140	1144
33384 W	-33	54	24	43	43	566	564	550	642
34767 W	-33	47	25	26	28	435	425	449	400
35179 A	-33	59	25	36	39	615	634	604	611
36729 W	-33	39	26	25	105	649	653	641	637
37696 W	-33	36	26	54	89	666	667	660	672
40653 W	-33	23	18	22	98	479	479	475	461
41417 W	-33	27	18	44	112	462	456	453	452
42227 W	-33	17	19	6	113	487	485	473	415
43109 W	-33	19	19	34	33	589	584	605	577
44050 W	-33	20	20	2	85	227	228	222	223
45134 W	-33	14	20	35	98	167	166	158	165
46479 W	-33	29	21	16	112	320	321	318	368
47716 W	-33	26	21	54	63	415	414	424	415
48043 W	-33	13	22	2	98	166	166	169	169
49060 W	-33	30	22	32	76	321	321	333	324
50887 W	-33	17	23	30	78	269	272	262	233
51430 W	-33	10	23	45	63	261	260	263	255
52590 W	-33	20	24	20	98	232	233	236	206
53432 W	-33	12	24	45	60	238	238	247	231
54805 W	-33	25	25	27	29	385	385	384	355
55300 W	-33	30	25	40	31	336	334	332	328
56709 W	-33	19	26	24	101	607	610	605	693
57048AW	-33	18	26	32	110	708	695	704	696
58192 W	-33	12	27	7	110	527	527	510	523
59722 W	-33	2	27	55	72	844	826	832	757
60620 W	-32	50	17	51	30	248	242	236	235
61298 W	-32	58	18	10	17	284	290	278	263
62444 W	-32	54	18	45	111	483	467	464	410
63538 A	-32	58	19	18	38	622	623	614	529
68857 W	-32	47	21	59	106	124	118	120	117
69559 W	-32	49	22	19	67	163	164	167	169
70033 W	-32	33	22	32	64	191	189	195	193
71264 W	-32	54	23	9	72	180	161	186	184
72712 W	-32	52	23	54	47	193	199	202	214
73871 W	-32	31	24	30	98	281	270	276	284
74296 W	-32	56	24	40	80	273	273	271	268
75215 W	-32	35	25	8	90	320	324	323	327
76884 W	-32	44	26	0	86	477	480	465	469
77522 W	-32	42	26	18	97	439	452	428	439

Table 8.3: Comparison of MAP values (in mm).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
78227 A	-32	47	26	38	112	522	524	514	513
79632 W	-32	32	27	22	104	977	970	952	956
80694 W	-32	34	27	54	105	738	738	724	647
81007 W	-32	37	28	1	72	821	819	799	693
83572 A	-32	2	18	20	14	232	221	227	223
84159 W	-32	9	18	36	55	240	234	230	240
85112 W	-32	22	19	4	82	704	694	692	457
87186 W	-32	6	20	7	52	366	355	371	369
88293 W	-32	23	20	40	61	246	276	240	339
89385 W	-32	25	21	13	57	295	294	296	293
90196 W	-32	16	21	37	71	173	173	172	176
91835 W	-32	25	22	28	67	189	188	192	194
92141 W	-32	21	22	35	96	237	238	239	238
93314 W	-32	14	23	11	101	213	213	218	219
94730 W	-32	10	23	55	59	406	416	392	406
95119 W	-32	29	24	4	100	286	276	282	285
96101 W	-32	11	24	34	104	284	285	279	281
97239 W	-32	29	25	8	68	362	363	346	348
98190 W	-32	10	25	37	86	321	323	315	312
99811 W	-32	1	26	28	84	446	447	438	431
100329 W	-32	29	26	41	106	1033	1018	997	1026
101804 W	-32	24	27	27	102	786	779	765	733
102762 W	-32	12	27	56	101	716	720	694	702
103516 W	-32	6	28	18	60	613	615	597	605
104762 W	-32	12	28	56	89	1145	1146	1133	1116
106850 W	-31	40	18	29	23	147	143	130	136
107396 W	-31	36	18	44	97	146	146	149	147
109215 W	-31	35	19	38	55	215	214	214	208
110385 W	-31	55	20	13	76	143	143	144	165
111373 W	-31	43	20	43	65	123	127	126	131
112346 W	-31	46	21	12	53	167	167	168	169
113025 W	-31	55	21	31	83	181	178	176	181
114747 W	-31	57	22	25	76	205	204	207	212
116083 W	-31	53	23	3	95	225	224	229	229
117447 W	-31	57	23	45	87	280	281	282	257
118395 W	-31	35	24	14	81	327	331	318	323
119209 W	-31	59	24	37	95	336	335	330	345
120338 W	-31	38	25	12	97	350	349	348	323
121518 W	-31	38	25	48	78	358	359	361	332
122480 W	-32	0	26	16	111	448	443	444	445
123304 W	-31	34	26	41	90	545	537	526	496
125150 W	-32	0	27	35	93	668	664	653	649
127485 A	-31	35	28	47	61	620	625	604	595
128032 W	-31	32	29	2	72	784	817	775	812
134478 A	-31	28	19	46	100	222	213	218	210
137337 W	-31	7	21	12	61	171	169	168	165
138041 W	-31	11	21	32	77	162	162	157	160
139658 W	-31	28	22	22	67	233	235	237	234
140616 W	-31	16	22	51	76	245	245	243	243
141329 W	-31	29	23	11	76	226	228	226	224
142805 W	-31	25	23	57	113	325	326	323	323
143579 W	-31	9	24	20	78	296	299	291	285
144900 W	-31	30	25	0	87	325	321	322	322
145029 A	-31	29	25	1	75	368	371	361	356
146588 W	-31	18	25	50	110	433	434	426	422
147409 W	-31	19	26	14	69	502	500	505	483
148352 A	-31	22	26	42	101	561	561	550	542
149082 A	-31	22	27	3	99	622	620	598	592
150085 W	-31	25	27	33	100	722	725	695	710
151604 W	-31	4	28	21	93	772	785	746	757
152190 W	-31	10	28	37	67	1115	1099	1091	1091
157035 W	-30	35	17	32	30	110	103	145	133

Table 8.3: Comparison of MAP values (contd.).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
165898 A	-30	58	22	0	60	204	205	209	204
166238 W	-30	58	22	8	45	198	195	188	201
167665 W	-30	35	22	53	87	216	216	221	216
168250 W	-30	40	23	9	61	230	238	234	221
169090 W	-31	0	23	33	71	289	292	302	286
170009 A	-30	39	24	1	94	298	294	287	303
171756 W	-30	36	24	56	81	348	351	344	324
172163 W	-30	43	25	6	112	400	400	393	388
173497 W	-30	47	25	47	84	435	438	419	408
174550 W	-30	40	26	19	106	483	486	473	451
175371 W	-30	41	26	43	72	522	509	507	524
176631AW	-30	31	27	22	73	678	695	640	622
177178 A	-30	58	27	36	106	621	618	590	588
178689 W	-30	59	28	23	29	820	808	804	813
179790 W	-30	40	28	57	67	912	911	897	912
180032 W	-30	32	29	2	64	817	820	818	773
181073 W	-30	43	29	33	77	954	951	923	832
182379 A	-30	49	30	13	62	1231	1214	1195	1206
185023 W	-30	23	17	31	28	113	99	105	102
193561 A	-30	21	21	49	65	171	177	179	175
196375 W	-30	15	23	13	72	213	203	215	210
198836 W	-30	26	24	28	112	332	332	327	335
199107 W	-30	17	24	34	68	315	316	313	307
200466 W	-30	16	25	16	97	401	406	390	395
201361 W	-30	1	25	43	76	437	436	421	428
202575 W	-30	5	26	20	33	523	531	489	487
203657 W	-30	27	26	52	80	636	641	623	587
204138 W	-30	18	27	5	81	715	714	686	685
205385 W	-30	25	27	43	42	830	824	834	707
206843 W	-30	3	28	29	43	617	621	609	585
207560 W	-30	20	28	49	76	707	711	673	682
208406 W	-30	16	29	14	75	750	754	733	743
209039 W	-30	9	29	32	88	1156	1133	1130	1108
210002 W	-30	2	30	1	75	881	873	810	925
211661 A	-30	1	30	53	39	1042	1047	1019	1019
214670 W	-29	40	17	53	114	219	210	216	216
223344 W	-29	44	22	12	24	235	238	235	202
224430 W	-29	40	22	45	57	250	243	250	228
225679 W	-29	49	23	23	79	267	269	264	271
226327 W	-29	57	23	41	85	244	248	248	243
227127 W	-29	37	24	5	86	321	324	311	302
228567 W	-29	57	24	49	82	383	390	367	369
229556 A	-29	46	25	19	41	438	458	428	422
230816 W	-29	36	25	58	46	515	508	509	489
231279 W	-29	39	26	10	85	500	496	483	479
232823 W	-29	43	26	58	77	618	619	585	582
233044 W	-29	44	27	2	70	537	559	430	503
236677 W	-29	47	28	53	18	619	654	586	582
237471 W	-29	51	29	16	53	1207	1199	1192	1184
238837 A	-29	57	29	58	49	881	866	864	862
239482 A	-29	32	30	17	73	900	889	875	876
240891 W	-29	51	31	0	111	1032	1025	1020	1020
241019 W	-29	49	31	1	56	995	1000	966	967
251261 W	-29	21	21	9	78	138	144	144	141
252894 W	-29	24	22	0	63	181	184	191	196
253648 W	-29	18	22	22	69	216	222	227	223
255202 W	-29	22	23	7	89	223	221	229	235
256453 W	-29	3	23	46	103	340	334	331	330
257845 W	-29	5	24	29	77	380	386	364	366
258458 W	-29	8	24	46	97	395	400	376	381
259727 W	-29	7	25	25	83	443	441	426	420
260678 W	-29	18	25	53	61	495	500	467	478

Table 8.3: Comparison of MAP values (contd.).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
261722 W	-29	2	26	25	86	566	568	540	487
262129 W	-29	9	26	35	66	564	556	554	516
263859 A	-29	19	27	29	47	755	745	717	714
264022 W	-29	22	27	31	50	820	819	776	735
268640 A	-29	10	29	52	80	894	901	877	877
269532 A	-29	22	30	18	68	1221	1215	1184	1195
270544 W	-29	4	30	49	63	1085	1072	1092	1081
271099 W	-29	9	31	4	66	1089	1086	1083	1080
272121 W	-29	1	31	35	78	1077	1073	1025	1063
282823 W	-28	43	20	58	47	169	163	165	159
283098 W	-28	38	21	4	70	149	151	151	155
286824 W	-28	44	22	58	18	308	313	297	310
287441 W	-28	51	23	15	89	294	299	296	293
288528 W	-28	48	23	48	71	340	343	323	329
289102 W	-28	42	24	4	55	334	332	328	333
290463 W	-28	43	24	46	51	419	418	406	410
291899 A	-28	59	25	30	77	431	422	410	410
292461 W	-28	41	25	46	65	436	444	433	447
293597 A	-28	57	26	20	64	575	581	542	529
294847 W	-28	37	26	59	62	590	588	570	563
295408 W	-28	48	27	14	66	654	652	635	631
296379 W	-28	49	27	43	78	710	706	691	711
297694 W	-28	34	28	24	42	777	780	752	761
298301 W	-28	31	28	41	68	838	836	825	822
299419 W	-28	59	29	14	26	1292	1260	1356	1262
300567 A	-28	57	29	49	86	761	761	724	723
301692 A	-28	32	30	24	62	804	803	754	772
302687 W	-28	57	30	53	43	938	917	883	758
303633 W	-28	33	31	22	28	827	822	833	911
304822 W	-28	42	31	58	70	1145	1142	1102	1109
305037 W	-28	37	32	2	71	1026	1022	994	1002
317447 A	-28	27	21	15	86	163	162	151	174
320348 W	-28	18	22	42	92	339	347	326	334
321110 W	-28	20	23	4	74	332	336	327	328
322071 W	-28	11	23	33	70	386	390	384	371
323649 W	-28	19	24	22	95	412	414	407	383
324607 W	-28	7	24	51	79	443	448	424	426
325870 W	-28	30	25	29	39	367	376	368	409
326668 W	-28	8	25	53	60	522	510	496	470
327883 W	-28	13	26	30	73	503	504	491	482
328628 W	-28	28	26	51	45	602	598	580	545
329215 W	-28	5	27	8	80	549	550	545	546
330750 W	-28	30	27	55	65	666	663	656	643
331058 W	-28	28	28	2	61	810	805	790	781
332663 W	-28	3	28	53	53	685	680	663	660
333226 W	-28	16	29	8	76	646	643	618	618
334825 W	-28	15	29	58	72	922	915	909	907
335550 A	-28	10	30	19	60	806	796	778	778
336283 W	-28	13	30	40	61	804	798	782	853
337148 W	-28	28	31	5	53	848	839	826	994
339354 A	-28	24	32	12	66	935	924	886	884
356285 W	-27	45	22	40	70	365	356	368	339
358049 W	-27	49	23	32	40	509	516	490	463
359808 W	-27	58	24	27	88	466	459	450	415
360597 A	-27	57	24	50	51	462	462	441	430
361354 W	-27	54	25	12	65	447	449	414	436
362862 W	-27	52	25	59	53	551	544	548	532
363651 W	-27	51	26	22	66	476	473	450	451
364322 W	-27	52	26	41	83	512	514	502	497
365400 W	-27	40	27	14	65	607	606	596	580
366710 W	-27	50	27	54	47	659	653	626	628
367768 W	-27	48	28	26	86	715	718	692	698

Table 8.3: Comparison of MAP values (contd.).

Station Code	Latitude	Longitude	Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
368634 W	-27 34	28 52	80	758	750	721	760
369238 W	-27 58	29 8	84	717	717	691	691
370486 W	-27 36	29 47	67	808	804	785	863
371579 W	-27 39	30 20	73	762	766	736	743
372852 W	-27 42	30 59	75	799	801	781	764
373680 W	-27 50	31 23	63	1559	1544	1531	1547
374264 W	-27 54	31 39	71	933	929	883	887
375366 W	-27 36	32 13	41	680	680	642	622
392148 W	-27 28	22 35	65	327	320	332	314
393778 W	-27 28	23 26	34	475	443	436	480
394874 W	-27 4	24 0	14	359	362	346	400
395855 W	-27 15	24 29	36	329	352	325	332
396813 W	-27 3	24 58	74	474	474	466	427
397086 W	-27 26	25 3	21	419	418	413	460
398556 W	-27 16	25 49	84	526	526	511	512
399667 W	-27 7	26 23	42	575	559	568	557
400203 W	-27 23	26 37	80	558	554	526	541
401798 W	-27 18	27 27	75	608	609	586	575
402081 W	-27 21	27 33	74	644	654	613	625
403886 W	-27 16	28 30	51	658	672	623	647
404316 W	-27 16	28 41	79	566	570	546	536
405753 W	-27 3	29 26	56	727	728	731	705
406607 W	-27 7	29 51	84	776	768	768	760
407639 W	-27 9	30 22	63	795	804	770	796
408798 W	-27 18	30 57	66	855	864	820	820
409375 W	-27 15	31 13	84	805	808	768	783
410133 W	-27 13	31 35	40	934	914	894	896
411175 W	-27 25	32 6	45	645	633	603	603
430354 W	-26 54	23 42	37	349	349	345	343
431896 W	-26 56	24 30	42	462	473	452	463
432237 A	-26 57	24 38	71	466	465	445	437
433858 W	-26 48	25 29	67	535	532	509	504
434020 W	-26 50	25 31	65	541	535	529	526
435400 W	-26 40	26 14	61	627	633	622	614
436747 W	-26 57	26 55	69	607	598	600	597
437134 A	-26 44	27 5	77	644	631	622	618
438315 W	-26 45	27 41	75	699	691	694	657
439764 W	-26 44	28 26	76	720	718	695	684
440157 W	-26 37	28 36	78	737	727	718	721
441777 W	-26 57	29 26	68	729	723	697	692
442781 A	-26 31	29 57	49	755	764	724	723
443451 W	-26 31	30 16	63	813	808	797	786
444746 W	-26 56	30 55	49	782	776	742	825
446741 S	-26 51	31 55	67	580	580	553	553
468318 W	-26 18	24 11	78	449	451	437	407
471490 W	-26 10	25 47	54	573	573	571	565
472175 W	-26 25	26 6	62	607	611	595	590
473686 W	-26 26	26 53	69	620	608	609	574
474198 W	-26 18	27 7	78	659	662	638	603
475881 W	-26 11	28 0	91	817	807	614	807
476072 W	-26 12	28 3	96	860	798	844	839
477309 W	-26 9	28 41	83	702	706	692	678
478360 W	-26 30	29 12	62	750	739	740	731
479545 W	-26 5	29 49	68	680	686	645	863
480889 W	-26 19	30 30	38	838	831	839	837
481167 W	-26 17	30 36	78	900	882	885	881
482357 W	-26 27	31 12	65	1137	1129	1126	1121
483053 W	-26 23	31 32	52	687	684	658	703
504838 W	-25 58	23 58	20	407	390	394	421
505834 W	-25 54	24 28	34	407	397	397	368
508649 W	-25 49	25 52	74	580	585	589	547
509759 W	-25 39	26 26	79	599	605	596	599

Table 8.3: Comparison of MAP values (contd.).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
510712 W	-25	52	26	54	77	667	672	667	648
511469 W	-25	49	27	16	70	692	684	668	673
512613 W	-25	43	27	51	69	696	689	667	675
513382 W	-25	52	28	13	74	706	702	685	688
514618 W	-25	48	28	51	83	724	723	694	684
515826 W	-25	46	29	28	52	709	695	686	689
516285 A	-25	45	29	40	62	784	777	764	757
517430 W	-25	40	30	15	83	798	797	781	790
518859 W	-25	49	30	59	81	819	828	786	872
519017 W	-25	47	31	1	28	715	713	744	682
520450 W	-26	0	31	45	35	764	768	739	594
545626 W	-25	26	25	51	70	616	605	583	588
546082 W	-25	22	26	3	66	605	610	596	587
549354 W	-25	24	27	42	62	594	594	579	604
550567 W	-25	27	28	19	52	601	615	570	612
551120 W	-25	30	28	34	74	681	671	660	644
552610 W	-25	10	29	21	39	623	631	604	625
553651 W	-25	21	29	52	66	739	725	718	692
554786 W	-25	6	30	27	72	695	691	677	678
555487 W	-25	7	30	47	49	1138	1151	1098	1095
556110 W	-25	20	31	4	54	915	900	901	648
557029 W	-25	29	31	31	35	699	673	670	652
585528 W	-24	48	26	18	53	583	591	566	585
586441 W	-24	51	26	45	67	580	587	551	572
587350 W	-24	50	27	12	30	697	695	669	632
588406 W	-24	46	27	44	83	617	622	601	614
589594 W	-24	54	28	20	30	654	631	632	629
590307 W	-24	37	28	41	72	635	632	623	615
591538 W	-24	58	29	18	50	520	522	500	515
593015 W	-24	45	30	1	64	562	558	559	552
594141 W	-24	51	30	35	67	598	602	570	551
595202 W	-24	52	31	7	36	1053	1046	1022	1019
630511 W	-24	1	27	18	34	512	493	504	506
631011 W	-24	11	27	31	58	525	523	513	505
632465 W	-24	15	28	16	48	599	586	563	610
633503 W	-24	23	28	47	49	673	668	664	666
634011 W	-24	11	29	1	27	595	592	597	624
635076 W	-24	16	29	33	54	563	561	534	537
636308 W	-24	8	30	11	61	959	971	948	968
637720 W	-24	30	30	54	45	834	827	790	903
638748 W	-24	28	31	25	35	572	581	561	581
639504 W	-24	24	31	47	41	597	589	556	562
673284 W	-23	44	27	10	36	473	463	442	447
675117 W	-23	57	28	4	63	556	564	540	553
676523 W	-23	43	28	48	37	475	489	474	499
677834 W	-23	54	29	28	82	493	486	482	485
678725 W	-23	35	29	55	44	640	629	618	589
679268 W	-23	58	30	9	51	1320	1338	1268	1147
680354 W	-23	54	30	42	41	537	541	521	537
681069 W	-23	39	31	3	33	486	497	471	501
718874 W	-23	4	28	0	51	416	416	420	403
719370 A	-23	10	28	13	28	418	434	402	391
720727 W	-23	7	28	55	60	555	536	557	686
721772 W	-23	22	29	26	53	415	416	405	402
722497 W	-23	17	29	47	52	428	430	405	411
723080 W	-23	20	30	3	58	787	796	749	841
724790 W	-23	10	30	57	33	550	548	575	566
762532 W	-22	52	28	18	49	369	378	366	465
763313 W	-22	43	28	41	29	448	456	417	410
764161 W	-22	41	29	6	61	364	347	374	377
765869 W	-22	59	29	59	39	783	786	746	794
766863 A	-22	53	30	29	37	1072	1080	1026	1077

Table 8.3: Comparison of MAP values (contd.).

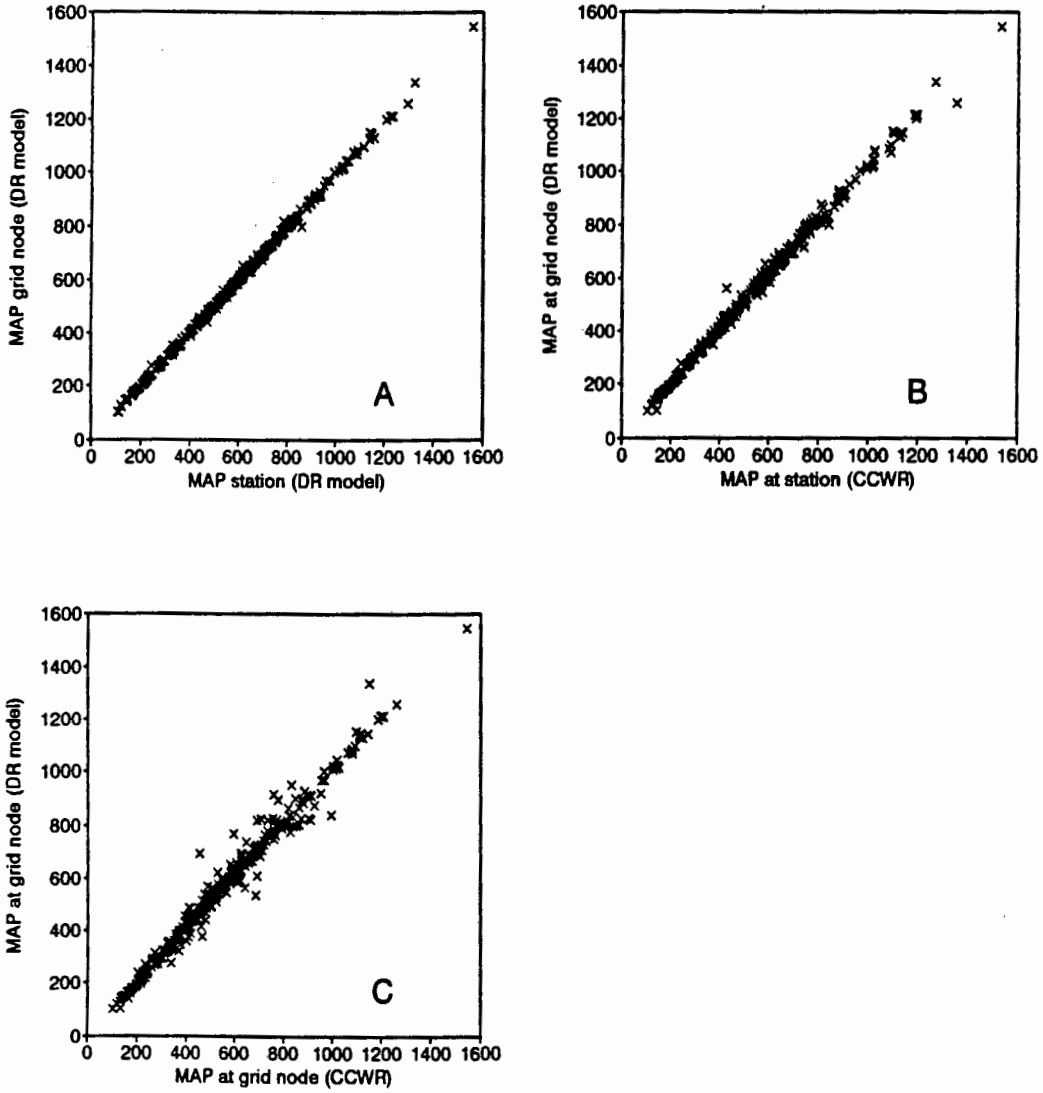


Figure 8.8: Comparison of MAP values (in mm).

Chapter 9

Kriging Binomial or Poisson Data

In this chapter we look at how kriging may be adapted to the situation where the data values are viewed as a sample from a binomial or Poisson distribution[1], and where the underlying distribution parameter may be expected to show spatial correlation. We may then wish to estimate that parameter at the given locations or to interpolate to obtain estimates at unsampled locations. This may be considered as a special case of the situation discussed in Chapter 6 in that there is error in the observed data (we observe $\hat{\pi} = X/n$ and we wish to estimate π); in this case however the 'measurement error' variance is dependent on the value of the (unknown) underlying parameter π. Examples of such data arise in many fields; for example the reporting rates of diseases or the occurrence of rare events such as rain storms. Our particular interest is to smooth the reporting rates of the data collected for the South African Bird Atlas Project described in Chapter 3 and the discussion in Section 9.3 will focus exclusively on that application; other applications are described briefly in the next chapter.

Previously published methods for smoothing binomial or Poisson data

[1]The work described in this chapter has previously been published in McNeill (1991).

have been suited to gridded data and involved the merging of adjoining grid cells (Musmeci and Vere-Jones, 1986; Byers, 1992); the methods described here are applicable to data collected at irregularly spaced locations and allow for a model in which the underlying parameter changes continuously.

9.1 Binomial Model for Spatial Data

For the ith location we define:

$$R_i = X_i/n_i = \pi_i + \varepsilon_i \tag{9.1}$$

where n_i is the number of samples for this location and X_i is the number of these in which the condition of interest (such as the presence of a species or the occurrence of a disease) is recorded; we refer to R_i as the 'observed reporting rate'. The distribution of X_i, conditional on π_i, is Binomial (n_i, π_i), and we assume that, if $i \neq j$, then X_i and X_j are conditionally independent, given π_i and π_j.

A simple model for the spatial structure of the π_i is obtained by assuming second-order stationarity, that is:

$$\mathrm{E}[\pi_i] = \mu \tag{9.2}$$

$$\mathrm{var}[\pi_i] = \sigma^2 = \rho(0) \tag{9.3}$$

$$\mathrm{cov}(\pi_i, \pi_j) = \rho(d_{ij}) \tag{9.4}$$

where $\rho(d_{ij})$ is some function of the distance d_{ij} between the ith and jth locations.

We also have the restriction that $\pi_i \in [0,1]$.

Clearly the model could be extended to cover the situation where large-scale trend is present as discussed in Chapter 6 , so that μ is not constant for

all i, but for simplicity that case will not be considered here. The assumption of constant variance might also need to be relaxed if the data locations differ greatly in size, this is the so-called 'change of support' problem in geostatistics.

9.1.1 The Kriging Equations

We seek a best unbiased linear estimator of each of the π_i, that is, for each i, we seek an estimator of the form $\sum_{j=1}^{N} w_j R_j$, where the summation is over all N of the locations for which data are available, and the w_j are chosen to minimise

$$\mathrm{E}\left[(\sum_{j=1}^{N} w_j R_j - \pi_i)^2\right]$$

subject to the constraint $\sum_{j=1}^{N} w_j = 1$, the latter condition guaranteeing unbiasedness. Using the Lagrange multiplier technique as in Section 5.1.2 we can show that, for each i, the solution is given by the following equation

$$\begin{pmatrix} \mathbf{w} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{K} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{c} \\ 1 \end{pmatrix} \tag{9.5}$$

where $k_{jk} = \mathrm{cov}(R_j, R_k)$ and $c_j = \mathrm{cov}(R_j, \pi_i)$ and λ is a Lagrangian multiplier.

9.1.2 Estimation of the Covariance

In order to apply the kriging equations given above we need to know or estimate the relevant covariance values.

We use the fact that

$$\mathrm{E}[X_i^k] = \int_0^1 \mathrm{E}[X_i^k/\pi_i] f(\pi_i)\, d\pi_i$$

and thus, in particular,

$$\mathrm{E}[X_i] = n_i \mu$$

$$\mathrm{E}[X_i^2] = n_i\mu + n_i(n_i - 1)(\sigma^2 + \mu^2)$$

so that

$$\mathrm{var}(X_i) = n_i\mu(1 - \mu) + n_i(n_i - 1)\sigma^2$$

$$\mathrm{var}(R_i) = \mu(1 - \mu)/n_i + \sigma^2(n_i - 1)/n_i$$

Also,

$$\begin{aligned}\mathrm{E}[X_i\pi_j] &= \int_0^1 \int_0^1 \mathrm{E}[X_i\pi_j/\pi_i\pi_j]f(\pi_i\pi_j)\,d\pi_i\,d\pi_j \\ &= n_i[\rho(d_{ij}) + \mu^2]\end{aligned}$$

so that

$$\mathrm{cov}(X_i\pi_j) = n_i\rho(d_{ij})$$

$$\mathrm{cov}(R_i\pi_j) = \rho(d_{ij})$$

and, if $i \neq j$,

$$\begin{aligned}\mathrm{E}[X_iX_j] &= \int_0^1 \int_0^1 \mathrm{E}[X_iX_j/\pi_i\pi_j]f(\pi_i\pi_j)\,d\pi_i\,d\pi_j \\ &= n_in_j[\rho(d_{ij}) + \mu^2]\end{aligned}$$

so that

$$\mathrm{cov}(X_iX_j) = n_in_j\rho(d_{ij})$$

$$\mathrm{cov}(R_iR_j) = \rho(d_{ij})$$

Summarising the results above, we have

$$\mathrm{cov}(R_j, R_k) = \begin{cases} \rho(d_{jk}) & j \neq k \\ 1/n_j[\mu(1 - \mu) + \sigma^2(n_j - 1)] & j = k \end{cases} \tag{9.6}$$

and

$$\mathrm{cov}(R_j, \pi_i) = \begin{cases} \rho(d_{ji}) & j \neq i \\ \rho(0) & j = i \end{cases} \tag{9.7}$$

Note that, since $\text{cov}(R_i, \pi_i) \neq \text{cov}(R_i, R_i)$, the vector $\mathbf{c}$ in equation 9.5 is not equal to the ith row and column of the matrix K and thus the solution is *not* given by $w_i = 1$ so that we do *not* get exact interpolation at the data points, as is the case when the data are error-free.

9.1.3 The Semi-Variogram

In order to find a suitable estimator of ρ we consider the use of the semi-variogram in the binomial situation. For the isotropic situation the semi-variogram of the random variable π is defined as

$$\gamma_\pi(h) = \frac{1}{2}E\left[(\pi_{\mathbf{z}} - \pi_{\mathbf{z+h}})^2\right]$$

where h is the length of the vector $\mathbf{h}$. It is easy to show that:

$$\gamma_\pi(h) = \rho(0) - \rho(h)$$

Clearly the empirical semi-variogram based on the observed R values will not be a good estimator of γ_π, but will be inflated due to the binomial sampling error. In this case, however, the measurement error variance, that is, the variance of ε_i in equation 9.1, is not constant across all locations , but depends on both n_i and π_i. We have, when $j \neq k$,

$$\begin{aligned}
\text{E}[(R_j - R_k)^2] &= [\text{var}(R_j) + \mu^2 + \text{var}(R_k) + \mu^2 - 2(\text{cov}(R_j R_k) + \mu^2)] \\
&= [\text{var}(R_j) + \text{var}(R_k) - 2\rho(d_{jk})] \\
&= \sum_{r=j,k} [\mu(1-\mu)/n_r + \sigma^2(n_r - 1)/n_r] - 2\rho(d_{jk}) \\
&= 2[\sigma^2 - \rho(d_{jk}) + 1/2 \sum_{r=j,k} [\mu(1-\mu) - \sigma^2]/n_r]
\end{aligned}$$

and thus

$$\frac{1}{2}E[(R_j - R_k)^2] = \rho(0) - \rho(d_{jk}) + \frac{1}{2}\sum_{r=j,k}\frac{\mu(1-\mu)}{n_r} - \frac{1}{2}\sum_{r=j,k}\frac{\sigma^2}{n_r} \qquad (9.8)$$

so that we may estimate $\gamma_\pi(h)$ by calculating an adjusted semi-variogram as

$$\frac{1}{2N_h}\sum\left[(R_j - R_k)^2 - \sum_{r=j,k}\frac{\hat{\mu}(1-\hat{\mu})}{n_r} + \sum_{r=j,k}\frac{\hat{\sigma}^2}{n_r}\right] \tag{9.9}$$

where the summation is over all N_h pairs of points which are a distance h apart. Thus the correction to the semi-variogram requires a prior estimate of both μ and σ^2, the mean and variance respectively of the π_i. These may be calculated as:

$$\hat{\mu} = \frac{\sum X_i}{\sum n_i} \tag{9.10}$$

and

$$\hat{\sigma}^2 = \frac{\sum(R_i - \hat{\mu})^2 - \hat{\mu}(1-\hat{\mu})\sum(1/n_i)}{N - \sum(1/n_i)} \tag{9.11}$$

where the summations are over all the N locations with $n_i > 0$. The variance estimator is based on the fact that

$$\begin{aligned} \mathrm{E}[\sum_{i=1}^{N}(R_i - \mu)^2] &= \sum_{i=1}^{N}\mathrm{var}(R_i) \\ &= \mu(1-\mu)\sum_{i=1}^{N}(1/n_i) + \sigma^2(N - \sum_{i=1}^{N}(1/n_i)) \end{aligned} \tag{9.12}$$

The estimator of the mean is unbiased, but the estimator of the variance is not; the bias is a function of the variance of $\hat{\mu}$ which depends on the unknown $\rho(d_{ij})$. However, the fitted semi-variogram provides another estimate of σ^2 in the sill parameter, and if the difference between the two estimates is large the new value may be substituted in equation 9.9 and the process repeated until convergence is obtained.

9.2 Poisson Model for Spatial Data

Suppose now that the observations X_i can be assumed to be samples from Poisson distributions with parameters λ_i, and that, if $i \neq j$, then X_i and X_j

are conditionally independent, given λ_i and λ_j. Then if we use a second-order stationary model as before, we can write

$$\mathrm{E}[\lambda_i] = \mu \tag{9.13}$$

$$\mathrm{var}[\lambda_i] = \sigma^2 = \rho(0) \tag{9.14}$$

$$\mathrm{cov}(\lambda_i, \lambda_j) = \rho(d_{ij}) \tag{9.15}$$

Then

$$\mathrm{E}[X_i^k] = \int_0^\infty \mathrm{E}[X_i^k/\lambda_i] f(\lambda_i)\, d\lambda_i$$

and thus, in particular,

$$\mathrm{E}[X_i] = \int_0^\infty \lambda_i f(\lambda_i)\, d\lambda_i = \mu$$

$$\mathrm{E}[X_i^2] = \int_0^\infty (\lambda_i - \lambda_i^2) f(\lambda_i)\, d\lambda_i = \sigma^2 + \mu + \mu^2$$

so that

$$\mathrm{var}(X_i) = \sigma^2 + \mu$$

Also,

$$\begin{aligned} \mathrm{E}[X_i\lambda_j] &= \int_0^\infty \int_0^\infty \mathrm{E}[X_i\lambda_j/\lambda_i\lambda_j] f(\lambda_i\lambda_j)\, d\lambda_i\, d\lambda_j \\ &= \rho(d_{ij}) + \mu^2 \end{aligned}$$

so that

$$\mathrm{cov}(X_i, \lambda_j) = \rho(d_{ij})$$

and, if $i \neq j$,

$$\begin{aligned} \mathrm{E}[X_iX_j] &= \int_0^\infty \int_0^\infty \mathrm{E}[X_iX_j/\lambda_i\lambda_j] f(\lambda_i\lambda_j)\, d\lambda_i\, d\lambda_j \\ &= \rho(d_{ij}) + \mu^2 \end{aligned}$$

so that

$$\mathrm{cov}(X_i, X_j) = \rho(d_{ij})$$

Summarising the results above, we have

$$\text{cov}(X_j, X_k) = \begin{cases} \rho(d_{jk}) & j \neq k \\ \sigma^2 + \mu & j = k \end{cases} \tag{9.16}$$

and

$$\text{cov}(X_j, \lambda_i) = \begin{cases} \rho(d_{ji}) & j \neq i \\ \sigma^2 & j = i \end{cases} \tag{9.17}$$

From this it follows that, if $j \neq k$,

$$\begin{aligned} \frac{1}{2}\text{E}[(X_j - X_k)^2] &= \frac{1}{2}\text{E}[X_j^2 + X_k^2 - 2X_jX_k] \\ &= \sigma^2 + \mu - \rho(d_{jk}) \end{aligned}$$

so that the Poisson case is simpler than the binomial case discussed above in that the semi-variogram of the observed X_i differs from that of the underlying parameter values λ_i by the constant amount μ, which may be estimated either directly from the data, or alternatively as the apparent nugget effect of the empirical semi-variogram of the X_i.

The kriging equations are then similar to those given for the binomial model; that is, to estimate λ_i we use $\sum w_j X_j$ where the vector $\mathbf{w}$ is obtained from:

$$\begin{pmatrix} \mathbf{w} \\ -\text{L} \end{pmatrix} = \begin{pmatrix} \mathbf{K} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{c} \\ 1 \end{pmatrix} \tag{9.18}$$

where $k_{jk} = \text{cov}(X_j, X_k)$ and $c_j = \text{cov}(X_j, \lambda_i)$ and L is a Lagrangian multiplier.

9.3 The Southern African Bird Atlas Project

An outline of the Southern African Bird Atlas Project (SABAP) was given in Chapter 3. The major objective of the SABAP is to obtain and publish definitive information on the spatial distribution and relative abundance of

individual bird species based on detailed field records. The raw data are in the form of field cards detailing presence/absence data on individual species based on the quarter degree grid square (QDGS) sampling units. The resultant maps show, for each QDGS, the 'reporting rate', that is, the ratio of the number of cards indicating the presence of a given species to the total number of cards completed. As discussed in Section 3.1, the relationship between reporting rate and relative abundance is dependent on a number of factors, so that a map of reporting rates can be seen at best as an indicator of abundance. There is also the problem of random sampling error in the reporting rates, particularly for those more remote areas of the region where sampling intensity is very low, in some QDGS as few as five field cards or less (table 3.1). In the most extreme cases, where only one field card is available the observed reporting rate must necessarily take one of the values zero or one. Clearly values based on small samples provide a very poor estimate of the underlying probability that a species will be observed by a randomly selected observer in a given QDGS, and hence a poor indicator of abundance.

Since the data are effectively on a regular grid one possibility for improving the estimates might be to smooth the data using a simple average or weighted average of neighbouring values, as was used, for example, by Byers (1992) for contour mapping of plant and animal densities. However, there is a need for an objective method to select appropriate weights which takes into account the extent of spatial correlation present in the data. In this particular application, there is also a need for a method which will allow for the large variation in sampling effort across QDGS, so that squares with more field cards should receive more weight. The map of the raw reporting rates of the Pied Crow given in Figure 3.2 shows clear evidence of spatial continuity in the data but also illustrates the effect of the lower sample sizes in the north-western and southern areas of the country which give rise to a somewhat less smooth map in those areas.

9.3.1 The Model

We assume that, for a given bird species, the reporting rate in the ith QDGS would tend to some value π_i as the number of completed field cards tends to infinity. Thus π_i can be loosely interpreted as the probability that a 'typical' observer in the ith QDGS will see the given species in any month. In practice the number of completed field cards depends on the number of observers participating in the project in a given region. Then, if n_i is the number of field cards for the ith QDGS and X_i is the number of these in which the given species is reported as present, the distribution of X_i may be modelled as Binomial (n_i, π_i), The equations described in Section 9.1 may then be used to model the spatial structure of the π_i.

The area of the QDGS decreases gradually as one moves to the south, from 28×25 km to 28×23 km, and in practice one might expect the variance σ^2 of the π to be larger for smaller areas; however the differences in area are small and have been ignored in this analysis.

In order to calculate the adjusted semi-variogram described in Section 9.1.3 we need to obtain preliminary estimates of μ and σ^2 using equations 9.10 and 9.11. For the Pied Crow the initial estimates of μ and σ^2 were 0,334 and 0,0716 respectively; the first iteration of the variogram fitting process gave σ^2 as 0,0712, and a second iteration gave 0,0711.

Figure 9.1 shows both the unadjusted semi-variogram of the reporting rates and the semi-variogram of the underlying probabilities estimated as in equation 9.9. In plotting the semi-variograms values are grouped into lag distance classes of 5 km and only the average of each class is shown on the graph. The distances between QDGS were calculated as the distances between their respective centres; since the centres of the QDGS are approximately on a square lattice the distances tend to cluster, thus the smallest observed distance is of the order of 26 km and the next group of distances

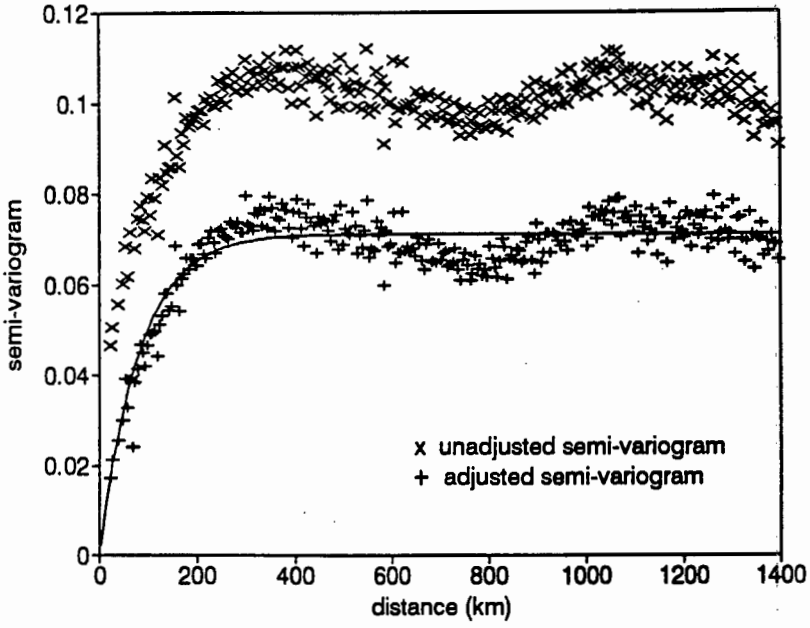


Figure 9.1: Semi-variogram: Pied Crow reporting rates.

is around $26 \times \sqrt{2}$ km. The 'waviness' of the graph at large lags is due to the existence of a small number of large areas of high density which can be seen in Figure 3.2. In practice, using local kriging, only the values of the variogram for small lags are required and thus one need not be especially concerned with the fit of the model for large lags.

The exponential variogram model, fitted by the weighted least squares method (Cressie, 1985), was found to provide a satisfactory fit. The value of the sill in the final fitted semi-variogram (Figure 9.1) was 0.0711 and the effective range was 240 km.

9.3.2 Solving the Kriging Equations

The kriging equations given in equation 9.5 were used to estimate the value of π for each QDGS. In solving the kriging equations a 7×7 neighbourhood of squares was used as the local window to compute each estimated value. Although the data points are (approximately) on a regular grid, the fact that the diagonal elements of $\mathbf{K}$ involve the n_i means that the covariance matrix has to be inverted separately for each point to be estimated. The use of local-window kriging effectively obviates the need to de-trend the data prior to kriging. Alternatively the data could have been de-trended using the median polish technique mentioned in Section 5.1.3 and then a global kriging approach could be used, requiring only a single matrix inversion of the full covariance matrix, (in this case a 1974×1974 matrix).

The estimation process does not explicitly constrain the values to lie in the range [0,1]. In practice, none of the estimates exceeded 1 and only 28 of the 1974 estimates were less than zero, with the smallest being -0.004; these 28 values were simply adjusted to zero. Barnes and You (1992) show that this is in fact the appropriate solution to the corresponding constrained kriging problem, and that the associated estimation variance is equal to the ordinary kriging variance plus a non-negative correction term.

The possibility of working with transformed data was also explored. The logit transformation is an obvious candidate, but as many of the reporting rates were zero it is necessary to use a modified logit. Various adjustments were considered, for example, Cox's empirical logistic transformation (Cox, 1970) given by:

$$Z_i = \log\left[\frac{X_i + \frac{1}{2}}{n_i - X_i + \frac{1}{2}}\right]$$

However, the transformed values are then dependent on the n_i; for example, if $X_i = 0$ then $Z_i = \log(\frac{1}{3})$ when $n_i = 1$ and Z_i approaches $-\infty$ as n_i gets large. This leads to instability in the variogram and the estimation process.

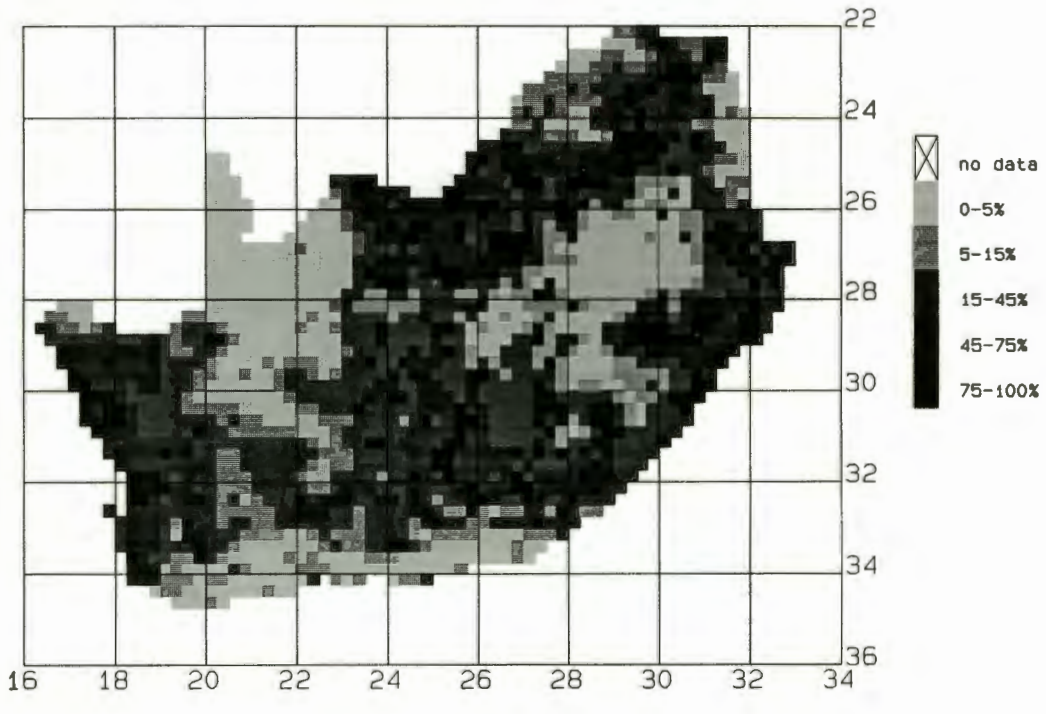


Figure 9.2: Pied Crow: smoothed reporting rates.

In the map of the kriged estimates (Figure 9.2) the sharp discontinuities evident in the original map (Figure 3.2) have been smoothed out. The degree of smoothing is greater in the more remote areas of the country, where the number of field cards tends to be smaller.

The estimated kriging variance, that is, the variance of the estimation error, is dependent on the values of $\mathbf{K}$ and $\mathbf{c}$ and hence on the fitted semi-variogram model. For gridded data the inter-point distances, and thus the off-diagonal values of $\mathbf{K}$, remain constant at each point being estimated, unless it is close to the boundary; however the diagonal values of $\mathbf{K}$ depend on the n_i and thus a plot of the standard errors should largely reflect the values of the n_i in the QDGS and its neighbours. Figure 9.3 shows that this

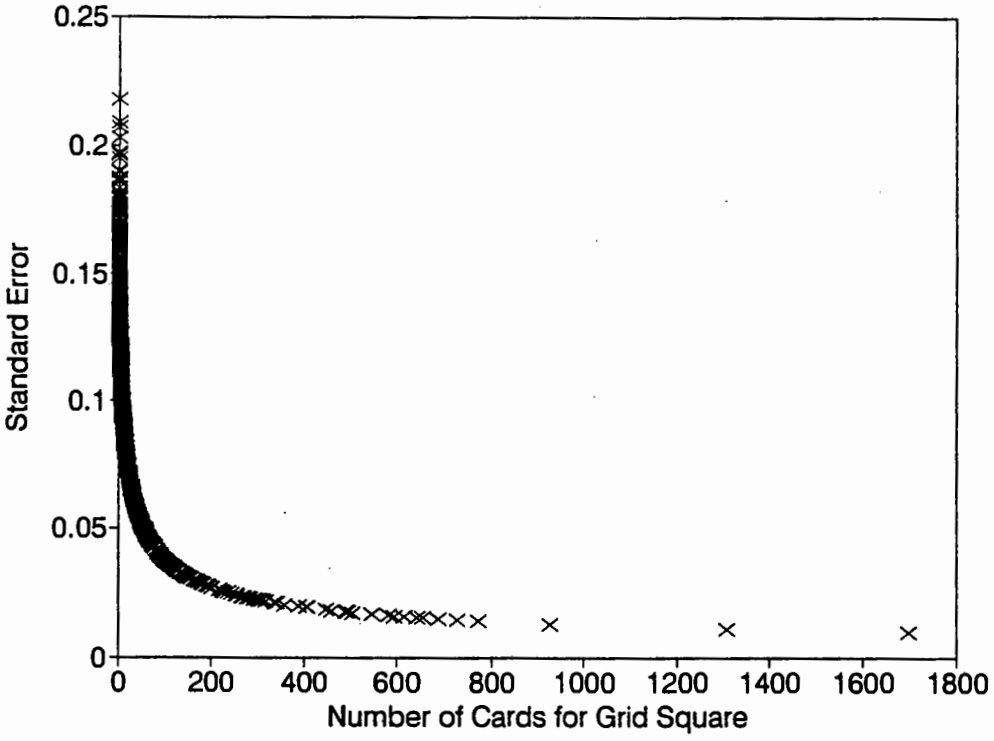


Figure 9.3: Pied Crow: kriging error plotted against number of field cards.

is so, and also shows that, when the number of cards in the QDGS is large, the standard error is determined almost entirely by this number, irrespective of the number of cards in adjacent QDGS. In fact, when n_i is large, one would expect that $\hat{\pi}_i \approx R_i$. This is confirmed by a plot of the absolute value of $(\hat{\pi}_i - R_i)$ against n_i (Figure 9.4).

9.3.3 Discussion

There are several possible criticisms of our model; these are discussed in this section and various refinements and improvements are suggested.

The π_i may vary with time, due to seasonal effects caused by migration or long term trends in species abundance. However, the five-year data collection

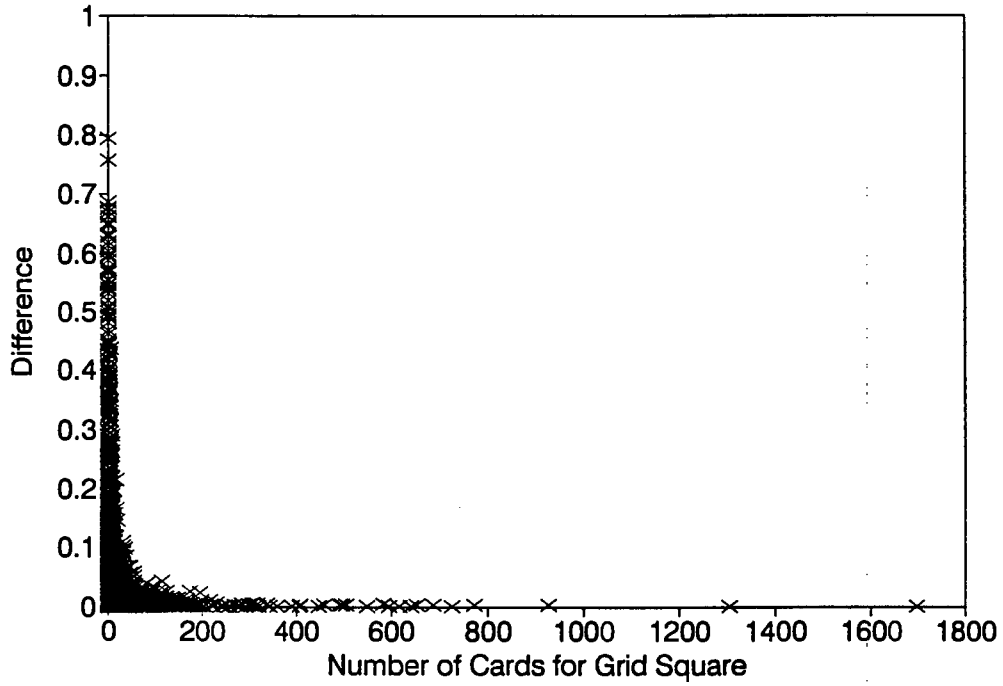


Figure 9.4: Pied Crow: amount of smoothing plotted against number of field cards.

period is considered sufficiently short for such long-term trends to have a fairly small effect. The number of field cards sent in for any QDGS does not vary much with time of year, so that the process of aggregating monthly data is not unreasonable, and serves the aim of providing a measure of relative spatial abundance, averaged over the period of the survey.

It is also likely that the π_i will vary between observers, depending on their ability to identify the species in question, on their exact location in the QDGS, and on the time and effort they expend in locating species. This could lead to a distribution with variance greater than the usual binomial model. In addition, the fact that each observer may send in a card each

month, and that the cards are aggregated over months, means that there will often be more than one card from a given observer in the raw reporting rate, which may violate the assumption of independent observations.

For species with very specific habitat or climatic requirements one might be able to improve the estimation by incorporating information on the habitat or climate using the method of co-kriging or kriging with external drift, as was done for the rainfall model parameters; this assumes of course that the relevant habitat information is actually available throughout the area of interest.

A problem that occurs for species with a more localised distribution is that the distribution of reporting rates is very skew with many zeros, leading to a very skew distribution of the observed semi-variogram values at any lag, since a large number of pairs of zeros will occur. For such species, a more robust estimator of the variogram would probably be advantageous, although Verlander[2] (1990) successfully applied the procedures described in this chapter to two other species which have rather localised distributions, namely the Greenshank *Tringa nebularia*, a common non-breeding Palaearctic migrant with a habitat largely restricted to dams, pans, seashores and estuaries, and the Hadeda Ibis *Bostrychia hagedash*, a common resident of grassland areas which is generally absent from the dry western part of the country.

In considering these various criticisms and suggestions one must bear in mind that the ultimate objective in this project is to provide maps for close to 900 individual species based in each case on almost 2000 individual reporting rates and thus it is important that the resultant methodology does not make excessive demands of computer time for negligible gain in accuracy.

---

[2]An honours student project supervised by the author of this thesis

9.4 Conclusion

The methods developed in this chapter allow for the technique of kriging to be used as a means of estimation of the underlying parameter of data with a binomial or Poisson distribution. In the binomial case, the sample sizes need not be constant across all spatial locations; the procedure automatically takes the varying accuracy of different sized samples into account both in the variogram estimation and in the kriging equations. The estimated spatial correlation function of the underlying binomial or Poisson parameter is used to determine the appropriate degree of weighting of neighbouring sample values.

While the data used in Section 9.3 were obtained at regularly spaced locations, this makes no difference to the calculations, and thus the methods have wide application to estimation and mapping in the biological and social sciences. For example, Oliver *et al.* (1993) recently used the methodology to derive regional estimates of risk of childhood cancer in the West Midlands of England based on local frequencies of the disease.

Chapter 10

Summary

In this thesis I have addressed a number of problems arising from the need to estimate the values of a spatially correlated variable at a grid of points covering an extensive geographical area. Although extensive data were available, the data locations were not evenly spaced throughout the region and the values were subject to some form of measurement error; the variance of such error could be estimated, either as a function of the mean or via bootstrap calculations, but varied considerably between data points. The calculation of an adjusted semi-variogram, free of the measurement error, was suggested as a means of overcoming this problem.

Further complications arose from the fact that some of the data showed large scale trends which could not be realistically modelled by simple parametric functions. This led to a consideration of modelling trend as another, large-scale, random function which allows a flexible and yet simple solution to the problem of kriging in the presence of trend. The use of kriging as a means to filter out the large-scale trend from the local variation was also demonstrated, thus allowing for the study of the dependence of the local variation on relevant covariates.

The study of the effect of topography on the local component of the parameters of the rainfall model was simplified by using orthogonal functions

of altitude as covariates in an external drift model; the orthogonal functions effectively encompass all patterns which can be described in terms of low-order polynomial functions, thus obviating the need to pre-define suitable functions of topography such as 'roughness' or 'exposure'.

A new method of smoothing of circular variables was presented, based on an analogy with ordinary kriging techniques. The technique was successfully applied to the phase parameters of the daily rainfall model, and shown to be more effective than a simple distance-weighted average smoothing method. The technique may be used either for interpolation or smoothing of any circular data in one or more dimensions. Typical applications would be in stratigraphic analysis or in studies of the direction of movement of pollutants in the ocean or atmosphere. Other examples arise from the phase values of periodic phenomena, such as the time of year of biological events such as spawning of fish, onset of plant growth or outbreaks of diseases in crops or animals, or of extreme events in atmospheric or oceanographic systems, as in the study of maximum sea levels in coastal flood prevention schemes.

The process of kriging was also extended to cater for variables having a binomial or Poisson distribution; this approach was used to smooth the reporting rates of individual bird species for the Southern African Bird Atlas Project. Apart from its use for mapping in environmental and land-use applications based on counts or presence/absence data, the technique has many other potential applications in the smoothing and mapping of socio-economic and epidemiological data such as crime rates or the incidence of rare diseases, where it is often of interest to study how incidence patterns correlate with other environmental variables.

Thus, while the focus of the thesis was to find practical means of solving two specific problems in the area of spatial estimation, the methodology developed here has wide-ranging and important uses, especially in view of the increasing demand for detailed geographical information systems based

on data of varying accuracy and uneven spatial coverage.

References

AHMED, S. and DE MARSILY, G. (1987). Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, **23**, 1717–1737.

ARMSTRONG, M., CHETBOUN, G. and HUBERT, P. (1993). Kriging the rainfall in Lesotho. In *Geostatistics Troia 92*, A. Soares, ed., Kluwer.

BARNES, R.J. and YOU, K. (1992). Adding bounds to kriging. *Math. Geol.*, **24**, 171–176.

BATES, D.M., LINDSTROM, M.J., WAHBA, G. and YANDELL, B.S. (1987). GCVPACK - routines for generalized cross-validation. *Commun. Statist. Simula.*, **16**, 263–297.

BATSCHELET, E. (1981). *Circular Statistics in Biology.* London, Academic Press.

BRECKLING, J. (1989). *The Analysis of Directional Time Series: Applications to Wind Speed and Direction.* Springer-Verlag, Berlin.

BYERS, J.A. (1992). Grid cell contour mapping of point densities: bark beetle attacks, fallen pine shoots, and infested trees. *Oikos*, **63**, 233–243.

CASKEY, J.E. (1963). A Markov chain model for the probability of precipitation occurrence in intervals of various length. *Monthly Weather Review*, **91**, 298–301.

CHRISTAKOS, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, **20**, 251–265.

CLARK, R.M. and THOMPSON, R. (1984). Statistical comparison of palaeomagnetic directional records from lake sediments. *Geophys. J. R. Astron. Soc.*, **76**, 337–368.

CLARK, I. (1979). *Practical Geostatistics.* Elsevier, London

CLARK, I., BASINGER, K.L. and HARPER, W.V. (1989). MUCK: A novel approach to co-kriging. In *Proceedings of the Conference on Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modelling*, B.E.Buxton, ed. p473–493. Battelle Press, Columbus, Ohio.

COX, D.R. (1970). *The Analysis of Binary Data.* Chapman and Hall, London.

CRESSIE, N. (1985). Fitting variogram models by weighted least squares. *Math. Geol.*, **17**, 563–586.

CRESSIE, N. (1986). Kriging nonstationary data. *J. Amer. Statist. Assoc.*, **81**, 625–634.

CRESSIE, N. (1988). Spatial prediction and ordinary kriging. *Math. Geol.* **20**, 405–421.

CRESSIE, N. (1990). Reply to Wahba's letter to the editor. *American Statistician,* **44**, 256–258.

CRESSIE, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

CRESSIE, N. and GRONDONA, M.O. (1992). A comparison of variogram estimation with covariogram estimation. In *The Art of Statistical Science,* K. V. Mardia, ed. Wiley, New York.

CREUTIN, J.D. and OBLED, C. (1982). Objective analyses and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research*, **18**, 413–431.

DAGBERT, M., KIELLAND, P., L'ESPERANCE, M. and COWAN, A. (1993) Geostatistics to assist hydrographic survey design. In *Geostatistics Troia 92*, A. Soares, ed., Kluwer.

DELFINER, P. (1975). Linear estimation of non stationary spatial phenomena. In *Advanced Geostatistics in the Mining Industry*, M. Guarascio, ed. p49–68. D. Reidel, Dordrecht.

DELFINER, P., RENARD, D., and CHILES, J.P. (1978). *BLUEPACK-3D Manual.* Centre de Geostatistique, Ecole des Mines, Fontainebleau, France.

DENT, M.C., LYNCH, S.D. and SCHULZE, R.E. (1989). Mapping mean annual and other rainfall statistics over southern Africa. *Water Research Commission Report* 109/1/89, Water Research Commission, Pretoria.

DIGGLE, P.J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

DIGGLE, P.J. and HUTCHINSON, M.F. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist.*, **31**, 166–182.

DRAPER, N.R.and SMITH, H. (1981). *Applied Regression Analysis*. Wiley, New York.

DUCHON, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique*, **10**, 5–12.

FISHER, N.I. and LEWIS, T. (1985). A note on spherical splines. *J. R. Statist. Soc. B,* **47**, 482–488.

GABRIEL, K.R. and NEUMANN, J. (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, **88**, 90–95.

GALLI, A. and MEUNIER, G. (1987). Study of a gas reservoir using the external drift method. In *Geostatistical Case Studies.* G. Matheron and M. Armstrong, eds. p105–119. D. Reidel, Dordrecht.

GANDIN, L.S. (1963). *Objective Analysis of Meteorological Fields.* Leningrad: GIMIZ. (translated from the Russian in 1965 by R. Hardin, Israel Program for Scientific Translations: Jerusalem).

GASSER, T. and MÜLLER, H.G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation.* T. Gasser and M. Rosenblatt, eds. p23–68. Springer, Heidelberg.

GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Amer. Statist. Assoc.*, **57**, 369–375.

GRANT, F. (1957). A problem in the analysis of geophysical data. *Geophysics*, **22**, 309–344.

GREEN, P.J. and SIBSON, R. (1978). Computing Dirichlet tessellations in the plane. *The Computer Journal*, **21**, 168–173.

HAAS, T.C. (1990). Lognormal and moving window methods of estimating acid deposition. *J. Amer. Statist. Assoc.*, **85**, 950–963.

HARBAUGH, J.W. and MERRIAM D.F. (1968). *Computer Applications in Stratigraphic Analysis.* Wiley, New York.

HARDY, R.L. (1971). Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research*, **76**, 1905–1915.

HARRISON, J. (1987). The Southern African Bird Atlas Project. *S. Afr. J. Sci.*, **83**, 400–401.

HARRISON, J. (1992). The Southern African Bird Atlas Project databank: five years of growth. *S. Afr. J. Sci.*, **88**, 410–413.

HARRISON, M. (1989). An investigation into the relationship between rainfall parameters and altitude. Unpublished Honours project report. Dept. Statistical Sciences, University of Cape Town.

HART, J.D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc. B*, **53**, 173–187.

HASTIE, T.J. and TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HOBSON, R.D. (1972). Surface roughness in topography: quantitative approach. In *Spatial Analysis in Geomorphology*, R.J.Chorley, ed. p221–245. Methuen, London.

HOCKEY, P.A.R., UNDERHILL, L.G., NEATHERWAY, M. and RYAN, P.G. (1989). *Atlas of the Birds of the Southwestern Cape.* Cape Bird Club, Cape Town.

HUDSON, G. (1993). Kriging temperature in Scotland using the external drift method. In *Geostatistics Troia 92*, A. Soares, ed., Kluwer.

HUGHES, D.A. (1982). The relationship between mean annual rainfall and physiographic variables applied to a coastal region of southern Africa. *S. Afr. Geog. Jour.*, **64**, 41–50.

HUTCHINSON, P. (1968) An analysis of the effect of topography on rainfall in the Taieri catchment area, Otago. *Earth Science Journal*, **2**, 51–68.

ISAAKS, E.H. and SRIVASTAVA, R.M. (1989). *Applied Geostatistics.* Oxford University Press, New York.

JAMES, W.R. (1966). FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data. *Kansas Geol. Survey Computer Contr.*, **5**.

JOURNEL, A.G. and HUIJBREGTS, C.J. (1978). *Mining Geostatistics.* Academic Press, London.

JOURNEL, A.G. and ROSSI, M.E. (1989). When do we need a trend model in kriging? *Math. Geol.*, **21**, 715–739.

JOWETT, G.H. (1955). Sampling properties of local statistics in stationary stochastic series. *Biometrika*, **42**, 160–169.

JUPP, P.E. and KENT, J.T. (1987). Fitting smooth paths to spherical data. *Appl. Stat.*, **36**, 34–46.

JUPP, P.E. and MARDIA, K.V. (1989). A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review*, **57**, 261–294.

KIMELDORF, G.and WAHBA, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

KOLMOGOROV, A.N. (1941). Interpolated and extrapolated stationary random sequences, *Isvestia AN SSSR, Seriya Matematicheskaya*, **5**.

KRAJEWSKI, W.F. (1987). Cokriging radar-rainfall and rain gage data. *Journal of Geophysical Research*, **92**, 9571–9580.

KRUMBEIN, W.C. (1959). Trend-surface analysis of contour-type maps with irregular control-point spacing *Journal of Geophysical Research*, **64**, 823–834.

LEE, P.S., LYNN, P.P. and SHAW, E.M. (1974). Comparison of multiquadric surfaces for the estimation of areal rainfall. *Hydrological Sciences Bulletin*, **19**, 303–317.

LINHART, H. and ZUCCHINI, W. (1986). *Model Selection.* Wiley, New York.

LONDON, W. and EMMITT, G.D. (1986). Topographical influences on radar echo properties – implications to weather modification projects in mountainous terrain. 2nd Conference on Planned and Inadvertent Weather Modification.

MARDIA, K.V. (1972). *Statistics of Directional Data.* Academic Press, London.

MARDIA, K.V. (1975). Statistics of directional data (with discussion). *J. R. Statist. Soc. B,* **37**, 349–393.

MARDIA, K.V. and WATKINS, A.J. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika*, **76**, 289–95.

MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.*, **58**, 1246–1266.

MATHERON, G. (1971). The theory of regionalized variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique, No.***5.** Fontainebleau, France.

MATHERON, G. (1973). The intrinsic random functions and their applications. *Adv. Appl. Prob.*, **5**, 439–468.

MATHERON, G. (1982). Pour une analyse krigeante de données régionalisées. Note interne, N-732, Centre de Géostatistique, Fontainebleau, France.

McNEILL, L. (1991). Interpolation and smoothing of binomial data for the Southern African Bird Atlas Project. *S. African Statistical Journal*, **25**, 129–136.

McNEILL, L. (1993). Interpolation and smoothing of mapped circular data. *S. African Statistical Journal* **27**, 23–49.

McNEILL, L., BRANDÃO, A., ZUCCHINI, W. and JOUBERT, A. (1994). *Interpolation of the Daily Rainfall Model.* Report to the Water Research Commission, Pretoria, (in press).

MENDOZA, C.E. (1986). Smoothing unit vector fields. *Math. Geol.*, **18**, 307–322.

MUSMECI, F. and VERE-JONES, D. (1986). A variable-grid algorithm for smoothing clustered data. *Biometrics*, **42**, 483–494.

MYERS, D.E. (1982). Matrix formulation of co-kriging. *Math. Geol.*, **14**, 249–257.

NEUMAN, S.P. and JACOBSON, E.A. (1984). Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *J. Int. Assoc. Math. Geol.*, **16**, 499–521.

OLIVER, M.A., LAJAUNIE, C., WEBSTER, R., MUIR, K.E. and MANN, J.R. (1993). Estimating the risk of childhood cancer. In *Geostatistics Troia 92*, A. Soares, ed., Kluwer.

PARKER, R.L. and DENHAM, C.R. (1979). Interpolation of unit vectors. *Geophys. J. R. Astron. Soc.*, **8**, 685–687.

PEARSON, E.S. and HARTLEY, H.O. (1962). *Biometrika Tables for Statisticians: vol. 1.* Cambridge University Press, Cambridge.

PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A. and VETTERLING, W.T. (1989). *Numerical Recipes.* Cambridge University Press, Cambridge.

RENARD, D. and NAI-HSIEN, M. (1988). Utilisation de dérives externes multiples. *Sciences de la Terre*, **28**, 281–301.

RIPLEY, B.D. (1981). *Spatial Statistics.* Wiley, New York.

SCHULZE, R.E. (1976). On the application of trend surfaces of precipitation to mountainous areas. *Water SA*, **2**, 110–118.

SEDUPANE, S.M. (1992). Modelling cross-covariance between rainfall and altitude in the western Cape, Transvaal and Natal. Unpublished Honours project report. Dept. Statistical Sciences, University of Cape Town.

SEED, A.W. (1987). Techniques for mapping rainfall. Unpublished M.Sc.Eng. thesis. Dept. Agricultural Engineering, University of Natal, Pietermaritzburg.

SIBSON, R. (1980). A vector identity for the Dirichlet tessellation. *Math. Proc. Camb. Phil. Soc.*, **87**, 151–155.

SIBSON, R. (1981). A brief description of natural neighbour interpolation. In *Interpreting Multivariate Data*, V. Barnett, ed. p21–36. Wiley, New York.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J.R. Statist. Soc. B*, **47**, 1–52.

SKIDMORE, A.K. (1989). A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *Int. J. Geographical Information Systems*, **3**, 323–334.

SPREEN, W.C. (1947). A determination of the effect of topography upon precipitation. *Trans. Amer. Geophys. Union*, **28**, 285–290.

SRIVASTAVA, R.M. (1988). A non-ergodic framework for variograms and covariance functions. Technical Report no. 114, Dept. Statistics and Dept. Applied Earth Sciences, Stanford University, California. 113p.

STARKS, T.H. and FANG, J.H. (1982). On the estimation of the generalized covariance function. *Math. Geol.*, **14**, 57–64.

STEIN, A. and CORSTEN, L.C.A. (1991). Universal kriging and cokriging as a regression procedure. *Biometrics*, **47**, 575–587.

STEIN, M.L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, **16**, 55–63.

STEIN, M.L. and HANDCOCK, M.S. (1989). Some asymptotic properties of kriging when the covariance function is misspecified. *Math. Geol.*, **21**, 171–190.

TABIOS, G.Q. and SALAS, J.D. (1985). A comparative analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin,* **21**, 365–380.

TEMPLE, S.A. and TEMPLE A.J. (1986). Geographic distribution and patterns of relative abundance of Wisconsin birds: a WSO research project. *Passenger Pigeon*, **48**, 58–68.

TUKEY, J.W. (1977). *Exploratory Data Analysis.* Addison Wesley, Reading, MA.

UPTON, G.J.G. and FINGLETON, B. (1989). *Spatial Data Analysis by Example, vol.2: Categorical and Directional Data.* Wiley, New York.

VECCHIA, A.V. (1988). Estimation and model identification for continuous spatial processes. *J.R.Statist. Soc. B.*, **50**, 297–312.

VERLANDER, N.Q. (1990). The application of kriging to binomial data of the Southern African Bird Atlas Project. Unpublished Honours project report. Dept. Statistical Sciences, University of Cape Town.

WAHBA, G. (1990a). *Spline Models for Observational Data.* SIAM, Philadelphia.

WAHBA, G. (1990b). Letter to the editor. *American Statistician,* 44, 255–256.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, **108**, 36–57.

WARNES, J.J. and RIPLEY, B.D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74**, 640–642.

WATSON, G.S. (1971). Trend-surface analysis. *Journal of the International Association for Mathematical Geology*, **3**, 215–226.

WATSON, G.S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America, Special Paper*, **146**, 39–46.

WATSON, G.S. (1984). Smoothing and interpolation by kriging and with splines. *Math. Geol.*, **16**, 601–615.

WATSON, G.S. (1985). Interpolation and smoothing of directed and undirected line data. In *Multivariate Analysis - VI.* P.R. Krishnaiah, ed. p613–625. Elsevier Science Publishers B.V..

WHITMORE, J.S. (1968). The relationship between mean annual rainfall and locality and site factors. *S. Afr. Jour. Sci.*, **64**, 423–427.

WOLFSON, N. (1975). Topographical effects on standard normals of rainfall over Israel. *Weather*, **30**, 138–143.

WOOLHISER, D.A. (1992). Modelling daily precipitation – progress and problems. In *Statistics in the Environmental and Earth Sciences*, A.T. Walden and P. Guttorp, eds. p71–89. Edward Arnold, New York.

WOOLHISER, D.A. and PEGRAM, G.G.S. (1979). Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology*, **18**, 34–42.

YOUNG, D.S. (1987). Random vectors and spatial analysis by geostatistics for geotechnical applications. *Math. Geol.*, **19**, 467–479.

ZIMMERMAN, D.L and ZIMMERMAN, M.B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, **33**, 77–91.

ZUCCHINI, W. and ADAMSON, P.T. (1984a). The occurrence and severity of droughts in South Africa. *WRC Report* No. 91/1/84, Water Research Commission, Pretoria.

ZUCCHINI, W. and ADAMSON, P.T. (1984b). The occurrence and severity of droughts in South Africa : Appendix 6. *WRC Report* No. 91/1/84(A), Water Research Commission, Pretoria.

ZUCCHINI, W., ADAMSON, P. and McNEILL, L. (1991). A family of stochastic models for droughts. *S. Afr. J. Plant Soil*, **8**, 206–211.

ZUCCHINI, W., ADAMSON, P. and McNEILL, L. (1992). A model of southern African rainfall. *S. Afr. Jnl. Sci.*, **88**, 103–109.