Geostatistical Models for Exposure Estimation in Environmental Epidemiology

Thomas Robert Fanshawe M.Phil., M.A.

Lancaster University

Submitted for the Degree of Doctor of Philosophy

December 2009

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Geostatistical Models for Exposure Estimation in Environmental Epidemiology

Thomas Robert Fanshawe M.Phil., M.A.

Lancaster University

Submitted for the Degree of Doctor of Philosophy

December 2009

Studies investigating associations between health outcomes and exposure to environmental pollutants benefit from measures of exposure made at the individual level. In this thesis we consider geostatistical modelling strategies aimed at providing such individual-level estimates. We present three papers showing how to adapt the standard univariate stationary Gaussian geostatistical model according to the nature of the exposure under consideration. In the first paper, we show how informative spatio-temporal covariates can be used to simplify the correlation structure of the assumed Gaussian process. We apply the method to data from a historical cohort study in Newcastle-upon-Tyne, designed to investigate links between adverse birth outcomes and maternal exposure to black smoke, measured by a fixed network of monitoring stations throughout a 32-year period. In the second paper, we show how predictions in the stationary Gaussian model change when the data and prediction locations cannot be measured precisely, and are therefore subject to positional error. We demonstrate that ignoring positional error results in biased predictions with misleading prediction errors. In the third paper, we consider models for multivariate exposures, concentrating on the bivariate case. We review and compare existing modelling strategies for bivariate geostatistical data and fit a common component model to a data-set of radon measurements from a case-control study designed to investigate associations with lung cancer in Winnipeg, Canada.

Declaration

The work in this thesis is my own, and has not been submitted for the award of a higher degree elsewhere. Paper 1 has been published, and arose from a collaborative project whose research team members are listed as co-authors. Papers 2 and 3 have been submitted for publication.

T.R. Fanshawe

17th December 2009

This thesis contains three papers - one accepted for publication, one under review and one submitted - containing work carried out since October 2006. Under my supervision, which generally took the form of weekly meetings, the candidate was responsible for all literature review, theoretical work, model development, data analysis and other computational work in this thesis. He is the first author on each of the papers, and as such took the lead both in writing the first draft of each paper and in incorporating subsequent revisions.

Prof. P.J. Diggle

17th December 2009

Acknowledgements

I would like to thank my supervisor, Professor Peter Diggle, for his expertise and encouragement since I moved to Lancaster in October 2006.

Contents

List of Tables

List of Figures

List of Papers

Paper 1 : Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlton, M., Pless-Mulloli, T.

Published in *Environmetrics* (2008), **19**, 549-566.

Paper 2 : Spatial prediction in the presence of positional error.

Fanshawe, T.R., Diggle, P.J.

Under review in *Environmetrics*.

Paper 3 : Bivariate geostatistical modelling: a review and an application to spatial variation in radon concentrations.

Fanshawe, T.R., Diggle, P.J.

Submitted to *Journal of Agricultural, Biological, and Environmental Statistics*.

Chapter 1

Introduction

This thesis examines in detail some specific problems relating to estimation of exposure to environmental pollutants for use in epidemiological studies. It consists of three related papers, each of which looks at a different aspect of the exposure estimation problem and explains how standard geostatistical methods can be adapted according to the nature and context of the problem that motivates it.

1.1 Exposure Estimation for Environmental Pollutants

In order to examine associations between environmental pollutants and adverse health outcomes, an estimate of each individual's exposure to the pollutant is required. In early work in this area, ecological study designs were often used, in which average exposures and outcome rates were estimated across wide areas such as cities or counties (Pope III & Dockery (2006)).

Although some studies have noted the possibility of bias using individual exposure estimates (Tielemans *et al.* (1998)), in general designs that provide individual-level estimates have more power to detect associations with health outcomes than those that use group-level estimates. Many such studies often assume that individuals remain at a fixed location, for example their home or place of work, although more ambitious designs have also been proposed to take into account the changing locations of individuals over time. Changes in location for time-spans ranging from a single working day to a whole lifetime have been considered (Han *et al.* (2005)).

The appropriate time scale in any study will depend on both the nature of the environmental

pollutant and the health outcome under consideration. For example, fleeting exposure to certain radioactive material may be sufficient to initiate carcinogenesis (Little (2000)), while asthmatic episodes can be triggered by exposure to particulate matter for a period of minutes or hours (Pope III (2000)). Exposure to particulate matter over the course of weeks or months during pregnancy may be associated with an increased risk of low birthweight and other infant morbidities (Glinianaia *et al.* (2004)), and it may take many years of steady exposure to radon before lung cancer symptoms become visible (Whitley & Darby (1999)).

These examples motivate the work in this thesis and demonstrate the diverse nature of environmental exposure assessment. The thesis presents three self-contained papers that relate to these applications. We now provide further background information on the relevance of each of the papers to exposure estimation, before discussing their context in relation to geostatistical modelling.

Paper 1 (Fanshawe *et al.* (2008)) uses data from a historical cohort study in Newcastle-upon-Tyne over the period 1961 to 1992. The aim of the study was to investigate associations between the exposure of pregnant women to 'black smoke' (BS, also known as PM_4, particulate matter of at most 4μm diameter) and a range of adverse birth outcomes. The objective in our paper is to construct a model for a spatio-temporal BS surface using data from a network of air pollution monitors. The monitors returned weekly measures of BS at fixed locations, but provided rather sparse coverage of the study region. The resulting model can subsequently be used to predict BS levels at unmonitored spatial locations corresponding to residences of study participants. The relationship between modelled BS and study outcomes is to be reported separately from this thesis, as the main study analysis.

Paper 2 considers the problem of exposure over a much shorter time period and discusses the impact of positional error on prediction. The motivation for this work is the growing number of studies that have used Global Positioning Systems (GPS) or similar devices to monitor the position of an individual over time (Nieuwenhuijsen & Brunekeef (2008)). Such devices are typically unable to operate precisely, and are therefore subject to an error known as positional error. This error will have an impact on the estimated exposure of an individual measured at a particular location. The complexity and expense of using devices such as GPS in epidemiological work means that they are best suited for studies that investigate short-term exposures.

Paper 3 examines the scenario when more than one pollutant is of interest, or when a single pollutant is measured using more than one method. Monitoring devices routinely set up to measure variables such as air pollution quality or soil content often record several quantities concurrently. In this paper we review and discuss methods for analysing multivariate data of this type, concentrating on the bivariate case. We include an analysis from a case-control study of residential radon and lung cancer in Winnipeg, in which radon levels were measured at two distinct locations - bedroom and basement - in each household. While associations between radon and lung cancer are well-established from laboratory studies and in certain groups exposed to unusually high levels (for example, miners), there is greater contention over the magnitude of such effects in the general population (Krewski *et al.* (2006)).

These examples provide the contextual background upon which the work in this thesis is based. We now describe the basic geostatistical theory that underpins the methodology developed in the three papers.

1.2 Geostatistical Modelling

Geostatistics is a branch of the broader discipline of spatial statistics that deals with modelling data arising from a spatially continuous phenomenon measured at a finite set of locations. Other research areas in spatial statistics include spatial point patterns, which are concerned with stochastic models governing the locations of points in space as opposed to any measurement that might be made there, and lattice processes, which are concerned with data aggregated over spatial regions, often positioned on a regular grid.

Geostatistical models usually assume that there is no stochastic process governing the locations at which measurements are taken. Instead, they model the measured quantity as a function of spatial location and other covariates. In environmental epidemiology, geostatistical models are typically used to provide predictions of such quantities as pollutant levels at individual locations or across wider areas. Another goal is parameter estimation, which can provide information about the correlation structure of the modelled process, or its association with covariates.

The history of geostatistics as a discipline can be traced to the work of D.G. Krige and the problem of predicting gold concentrations in South African mines in the 1950s (Cressie (1990)). Since then, the subject has broadened substantially from its geological roots to encompass an enormous range of disciplines, including ecology, hydrology, oil production, soil and earth sciences,

climatology, agriculture, microbiology, geography, seismology and epidemiology, alongside the concomitant mathematical and statistical theory upon which it depends. We introduce this theory below.

A geostatistical data-set consists of real-valued measurements $Y_i : i = 1, ..., n$ associated with a corresponding set of spatial locations x_i in a region of interest D. A geostatistical model specifies the statistical relationship between the pairs (Y_i, x_i) and an underlying spatially continuous phenomenon $S(x) : x \in \mathbb{R}^2$. A simple yet widely-used model (Diggle & Ribeiro Jr. (2007)) is

$$Y_i = d(x_i)^T\beta + S(x_i) + Z_i : i = 1, ..., n, \tag{1.1}$$

where $d(x)$ is a set of spatially referenced explanatory variables, $S(x)$ is a zero-mean Gaussian stochastic process with variance σ^2 and correlation matrix R_ϕ, where ϕ may be a vector-valued parameter, and the Z_i are mutually independent Gaussian errors with mean zero and variance τ^2.

Let $\rho(\cdot) \equiv \rho_\phi(\cdot)$ be a function such that the matrix R_ϕ consists of entries $\rho_\phi(x_i; x_j)$. For any finite set of locations $x_1, ..., x_n$, R_ϕ must be positive definite: for any $z \in \mathbb{R}^n$, $z^T R_\phi z > 0$. If this holds, ρ is said to be a (valid) correlation function. Common simplifying assumptions are that ρ is stationary and isotropic.

ρ is stationary if it is invariant to translation: $\rho(x_1, x_2) = \rho(x_1 - x_2)$ for all x_1, x_2.

ρ is isotropic if it is invariant to rotation: $\rho(x_1, x_2) = \rho(\|x_1 - x_2\|)$ for all x_1, x_2.

Common choices are the exponential, Gaussian and Matérn correlation functions:

Exponential:

$$\rho(x) = e^{-\frac{x}{\phi}}$$

Gaussian:

$$\rho(x) = e^{-(\frac{x}{\phi})^2}$$

Matérn:

$$\rho(x) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{x}{\phi}\right)^{\kappa} K_\kappa \left(\frac{x}{\phi}\right)$$

For the Matérn function, K_κ is the modified Bessel function of order κ and ϕ is a range parameter

that indicates the rate of decay of correlation away from zero. The Matérn function has become increasingly popular in applications both because of the additional flexibility provided by the smoothness parameter κ and because it contains both the exponential and Gaussian correlation functions as special cases, for $\kappa = 0.5$ and $\kappa \to \infty$ respectively.

Parameter estimation can be achieved by maximum likelihood based on (1.1), with initial values provided by an estimate of the variogram

$$V(x_1, x_2) = \frac{1}{2}\text{Var}[S(x_1) - S(x_2)].$$

One such estimate, the empirical variogram based on data $Y_i : i = 1, ..., n$, is defined as

$$\hat{V}_{ij} = \frac{1}{2}(Y_i - Y_j)^2$$

for $i < j$. The $\hat{V}_{ij}$ are often plotted against spatial separation u after they have been 'binned' (averaging values $\hat{V}_{ij}$ for which $\|x_i - x_j\| \approx u$). The resulting plot is also useful for indicating model fit after parameter estimates have been calculated.

In the specification of (1.1), τ^2 is sometimes known as the 'nugget effect', and is equal to the sum of small-scale spatial variation and measurement error in the Y_i. Thus in (1.1) we might replace Z_i by $S^*(x_i) + Z_i^*$, where S^* represents spatial variation at distances less than $\min\|x_i - x_j\|$, $i \neq j$ and the Z_i^* are independent errors. These two quantities cannot be distinguished without replicated observations at individual locations. Schabenberger & Gotway (2005) provide a full discussion of models for disentangling these two sources of variation, but they are beyond the scope of the work in this thesis.

Much research in early classical geostatistics focused on the technique of 'kriging', whose aim is prediction of $S(x_p)$ at a new or previously-sampled location x_p. The simple kriging predictor based on data $Y_1, ..., Y_n$ is defined as $\hat{S}(x_p) = \sum_i \hat{a}_i Y_i$, where the $\hat{a}_i$ are chosen to minimise $\text{Var}[S(x_p) - \hat{S}(x_p)]$ subject to $\text{E}[S(x_p) - \hat{S}(x_p)] = 0$.

The simple kriging predictor is derived in Appendix B.1, and is equivalent to the minimum mean square error predictor under the stationary Gaussian model (1.1), as shown by Diggle & Ribeiro Jr. (2007), page 135. The prediction problem is one of the most important topics that arises from applications of geostatistical methodology, and appears in a different guise in each of

the three papers in this thesis. To address this problem, we adapt the basic model (1.1) in three distinct ways according to the aims of the three papers.

In Paper 1, we consider a spatio-temporal application in which the general formulation of the model follows the structure set out by Sahu & Mardia (2005):

$$\begin{aligned} Y(x,t) &= S(x,t) + Z_1(x,t) \\ S(x,t) &= \mu(x,t) + Z_2(x,t) \end{aligned} \tag{1.2}$$

where x indexes space and t time, μ is a mean process, and Z_1 and Z_2 are zero-mean stochastic processes. The prediction problem is then to find $\hat{S}(x_p, t_p)$ at a location-time pair (x_p, t_p).

In Paper 2, we assume model (1.1) but drop the assumption that the locations x can be measured precisely. Instead we consider x to be a realisation of a random variable X. This problem is described by Gabrosek & Cressie (2002). The choice of the joint distribution of the true location X^* and the observed location X (or the conditional distributions $[X|X^*]$ and $[X^*|X]$) affects the prediction problem. In this paper we consider both the case in which positional error affects the measurement location and the case in which positional error affects the prediction location.

In Paper 3, we generalise (1.1) to the multivariate setting, concentrating on the bivariate case. In this scenario, $S = (S_1, S_2)$ is a bivariate process with zero mean whose covariance function is defined by three components: two auto-covariance functions and a cross-covariance function. As Paper 3 shows, a major research problem in multivariate geostatistics is how to specify these three functions without violating the positive definiteness constraint. Once a model has been specified, prediction for each component, based on the data $Y_{ij} : \ i = 1, ..., n_j; \ j = 1, 2$, continues in a similar way to the univariate case.

In the simplest solution, separate univariate models could be formulated for the two components Y_1 and Y_2, but this unsatisfactory strategy would be inefficient if Y_1 and Y_2 were correlated, as is often the case for measures of environmental exposure. Scenarios in which correlated environmental exposure data might arise include the measurement of exposures both arising from a similar source (e.g. two distinct varieties of particulate matter), the measurement of a single exposure using two different instruments (e.g. a hand-held and a fixed monitoring device), or the measurement of an exposure in nominally different environments (e.g. two different rooms in a single house).

In summary, the three papers respectively investigate the implications on the basic set-up (1.1) of the existence of highly informative, spatio-temporally varying covariates d; uncertainty in the geographical locations x; and additional response variables Y.

The main part of the thesis consists of the three papers, as submitted for publication. Chapter 5 summarises the achievements of each paper, and of the thesis as a whole. It also provides extra contextual information for each paper, explains the relevance of the work with respect to the existing literature and provides suggestions for future research. Appendices A, B and C contain related background work and extensions of the technical material that appears in the three papers.

References

Cressie, N. 1990. The origins of kriging. *Mathematical Geology*, **22**, 239–252.

Diggle, P.J., & Ribeiro Jr., P.J. 2007. *Model-Based Geostatistics*. New York: Springer.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlton, M., & Pless-Mulloli, T. 2008. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics*, **19**, 549–566.

Gabrosek, J., & Cressie, N. 2002. The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, **34**, 262–285.

Glinianaia, S.V., Rankin, J., Bell, R., Pless-Mulloli, T., & Howel, D. 2004. Does particulate air pollution contribute to infant death? A systematic review. *Environmental Health Perspectives*, **112**, 1365–1371.

Han, D., Rogerson, P.A., Bonner, M.R., Nie, J., Vena, J.E., Muti, P., Trevisan, M., & Freudenheim, J.L. 2005. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. *International Journal of Health Geographics*, **4**, doi: 10.1186/1476–072X–4–9.

Krewski, D., Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Field, R.W., Klotz, J.B., Létourneau, E.G., Lynch, C.F., Lyon, J.L., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Stolwijk, J.A., Weinberg, C., & Wilcox, H.B. 2006. A combined analysis of North American case-control studies of residential radon and lung cancer. *Journal of Toxicology and Environmental Health, Part A: Current Issues*, **69**, 533–597.

Little, J.B. 2000. Radiation carcinogenesis. *Carcinogenesis*, **21**, 397–404.

Nieuwenhuijsen, M., & Brunekeef, B. 2008. *Environmental exposure assessment*. In: Environmental Epidemiology: Study methods and application. New York: Oxford University Press. Chap. 3, pages 41–71.

Pope III, C.A. 2000. Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk? *Environmental Health Perspectives*, **108**, 713–723.

Pope III, C.A., & Dockery, D.W. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association*, **56**, 709–742.

Sahu, S.K., & Mardia, K.V. 2005 (September 21-23). Recent Trends in Modeling Spatio-Temporal Data. *Pages 69–83 of: Proceedings of the special meeting on Statistics and Environment.*

Schabenberger, O., & Gotway, C.A. 2005. *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman and Hall/CRC.

Tielemans, E., Kupper, L.L., Kromhout, H., Heederik, D., & Houba, R. 1998. Individual-based and group-based occupational exposure assessment: some equations to evaluate different strategies. *The Annals of Occupational Hygiene*, **42**, 115–119.

Whitley, E., & Darby, S.C. 1999. *Quantifying the risks from residential radon*. In: Statistical aspects of health and the environment. Chichester: Wiley. Chap. 5, pages 73–89.

Chapter 2

Paper 1: Modelling Spatio-Temporal Variation in Exposure to Particulate Matter: a Two-Stage Approach

---

T.R. Fanshawe[1], P.J. Diggle[1], S. Rushton[2], R. Sanderson[2], P.W.W. Lurz[2], S.V. Glinianaia[3], M.S. Pearce[3,4], L. Parker[5], M. Charlton[6], T. Pless-Mulloli[3]

[1] Department of Medicine, Lancaster University, UK

[2] Institute for Research on Environment and Sustainability, Newcastle University, UK

[3] Institute of Health and Society, Newcastle University, UK

[4] School of Clinical Medical Sciences, Newcastle University, UK

[5] Community Health and Epidemiology/Pediatrics, Dalhousie University, Halifax, Canada

[6] National Centre for Geocomputation, National University of Ireland, Ireland

Summary

Studies investigating associations between air pollution exposure and health outcomes benefit from the estimation of exposures at the individual level, but explicit consideration of the spatio-temporal variation in exposure is relatively new in air pollution epidemiology. We address the problem of estimating spatially and temporally varying particulate matter concentrations (black smoke=BS=PM_4) using data routinely collected from 20 monitoring stations in Newcastle-upon-Tyne between 1961 and 1992. We propose a two-stage strategy for modelling BS levels. In the first stage, we use a dynamic linear model to describe the long-term trend and seasonal variation in area-wide average BS levels. In the second stage, we account for the spatio-temporal variation between monitors around the area-wide average in a linear model that incorporates a range of spatio-temporal covariates available throughout the study area, and test for evidence of residual spatio-temporal correlation. We then use the model to assign time-aggregated predictions of BS exposure, with associated prediction variances, to each singleton pregnancy that occurred in the study area during this period, guided by dates of conception and birth and mothers' residential locations. In work to be reported separately, these exposure estimates will be used to investigate relationships between maternal exposure to BS during pregnancy and a range of birth outcomes. Our analysis demonstrates how suitable covariates can be used to explain residual spatio-temporal variation in individual-level exposure, thereby reducing the need to model the residual spatio-temporal correlation explicitly.

Key words: dynamic linear model; environmental epidemiology; exposure estimation; particulate matter; spatio-temporal process

2.1 Introduction

Links between long-term or short-term exposure to particulate matter and morbidity or mortality in both children and adults are now well established (Pope III & Dockery (2006)). In particular, there is growing evidence of an association between air pollution exposures during pregnancy and adverse birth outcomes (Glinianaia *et al.* (2004b); Šrám *et al.* (2005)) or infant survival (Glinianaia *et al.* (2004a); Ritz *et al.* (2006)), especially for respiratory-related causes (Woodruff *et al.* (2006)). In order to test hypotheses relating to these associations using observational data, estimates of pollution levels to which each mother was exposed during different periods of pregnancy are needed. Some previous studies have either assumed homogeneity of exposure at any one time across large geographical areas (Woodruff *et al.* (1997); Samet *et al.* (2000)), or estimated exposure using a crude average (Bobak (2000)). Only recently has modelled city-wide variation in exposure and its impact on health outcomes been considered (Jerrett *et al.* (2005)).

In this paper, we use data from the UK Particulate Matter and Perinatal Events Research (PAMPER) study to demonstrate a method for estimating a spatio-temporal exposure surface of black smoke (BS), equivalent to PM_4, concentrations over the city of Newcastle upon Tyne for the time-period 1961-1992.

We consider data in the form of a set of time series, one for each of a number of monitoring locations within the spatial region of interest, and not necessarily providing data at a common set of times. Various approaches have been suggested in the statistical literature for analyzing environmental spatio-temporal data of this kind; for reviews, see Kyriakidis & Journal (1999) and Sahu & Mardia (2005). Key approaches to such analysis include: directly modelling the joint space-time distribution of the observations, treating time as an additional dimension (e.g. Brown *et al.* (2001)); modelling the data as a set of spatial processes correlated in time (e.g. Bogaert & Christakos (1997)); or, more commonly, as a set of time series correlated in space (e.g. Meiring *et al.* (1998)).

Most of this work has used Gaussian processes as models for the underlying spatio-temporal phenomenon, $S(x,t)$ say, with a consequent focus on the specification of valid, appropriate and computationally tractable covariance functions for $S(x,t)$ (Gneiting *et al.* (2007)). An exception is Higdon (2007), who describes a non-Gaussian kernel convolution approach. Stroud *et al.* (2001) extend state-space models of time series to the space-time domain in order to avoid making assumptions of stationarity and separability of the covariance function. In some examples,

the relatively weak dependence between observations either in space or in time has enabled the modelling process to be simplified: for example, Handcock & Wallis (1994) found a lack of temporal dependence in annual winter average temperatures in northern U.S.A. In contrast, other examples exhibit long-term temporal dependence, such as the Irish wind speed data of Haslett & Raftery (1989).

In the field of air pollution, several authors have addressed the simultaneous consideration of spatial and temporal variation of exposure. Carroll *et al.* (1997) modelled ozone exposure in Texas, U.S.A. by splitting the spatio-temporal variation into two components: a deterministic, spatially-constant component and a stationary, zero-mean Gaussian random field. Zidek *et al.* (2002) modelled the spatial covariance between residuals using a space deformation approach (Meiring *et al.* (1998)) after first fitting an AR(3) model to hourly PM_{10} levels in Vancouver, Canada (Li *et al.* (2008)). Sahu *et al.* (2006) illustrated one way in which available covariates may be used by modelling $PM_{2.5}$ monitoring data using two random spatio-temporal processes, corresponding to urban and rural areas respectively, and weighted by population density.

In this paper we demonstrate a pragmatic, two-stage modelling strategy. We first estimate the seasonally-varying temporal trend using a dynamic linear model, then account for remaining spatio-temporal variation using temporally and/or spatially varying covariates. We demonstrate that for our data, residual spatio-temporal correlation is not significant. In principle, we could include a spatio-temporally correlated residual term, at the cost of a substantial increase in computational complexity. However, in our view explicit models of spatio-temporal correlation should be used only when the possibility of obtaining an adequate explanation of spatio-temporal variation using covariate information has been exhausted. In our application, the key step was not to rely on routinely available covariate information but instead to construct a suitable surrogate using a combination of land-use information and digital images of domestic chimneys which, for the area and time-period in question, constituted a major source of BS exposure for pregnant women.

2.2 The UK PAMPER Study

The UK PAMPER study is a historical cohort study to investigate the relationship between adverse pregnancy outcomes and a range of socio-economic, meteorological and pollution-related factors. In this paper we model levels of weekly BS using data routinely, albeit spasmodically, recorded at 20 air pollution monitoring stations within the city of Newcastle-upon-Tyne (the

'study area') between October 1961 and December 1992 (the 'study period'). The data are available from the UK Air Quality Archive (http://www.airquality.co.uk/archive/data_and_statistics_home.php). Figure 2.1 shows the locations of the 20 monitoring stations within the study area, and the locations of five further monitoring stations that we will use for model validation. Pless-Mulloli *et al.* (2007) provide more details of the study's background and setting.

Figure 2.2 shows the period of time over which each monitor was in operation ('active'). Over the whole study period, the number of monitors active during any single week varied between three and ten. In our experience, the relatively sparse spatial coverage of the study area by monitors is typical, and strongly influenced our approach to the prediction problem.

Our aim is to attach to each of the 109,086 singleton births that occurred in the study area during the study period a predicted BS exposure level and associated prediction variance, both for individual weeks of the pregnancy and time-aggregated over months, trimesters and over the whole pregnancy period. Each birth is characterized by the date of birth, the estimated date of conception (for births with available gestational age) and the mother's residential location (grid reference) at which BS levels are to be estimated. In future work, we will investigate associations between this modelled exposure and a range of adverse birth outcomes, including birthweight, low birthweight, preterm birth, stillbirth, infant mortality and congenital abnormality.

2.3 Modelling the Exposure Surface

2.3.1 Exploratory Analysis

In common with other environmental applications (e.g. Brown *et al.* (2001); Zidek *et al.* (2002)), we found that a log-transformation approximately stabilises the variance of BS and gives a roughly linear time trend, i.e. city-wide average BS levels have experienced an approximately exponential decline over the study period. We therefore model the log-transformed values of BS recorded at each monitoring station.

Let Y denote log-transformed BS. Figure 2.3 shows the area-wide average, $\bar{Y}_t$ say, in each of the 1631 weeks of the study period, in each case calculated as the average of the observed log-BS levels at all monitoring stations that were active during the week in question. The scale of the overall temporal variation in $\bar{Y}_t$ is much larger than is the spatial variation between different monitors at any given time, which is typically of the order of 1 unit on the logarithmic scale, al-

though occasional recorded values fall much further than this from the corresponding city-wide average. For the subsequent modelling, we use all available data. Re-fitting the final model excluding 74 recorded values (out of 10174, i.e. around 0.7%) of log-BS more than 1.5 units away from the area-wide average has only a small impact on parameter estimates and predictions, and as we have no basis for treating these values as recording errors, we retain all of the data in the analysis presented below.

Figure 2.3 also shows that there is a strong seasonal component to average BS levels. Annual peaks and troughs occur each winter and summer respectively, albeit with some variation from year to year. This seasonal pattern is also evident from inspection of the data from individual monitors.

2.3.2 The Modelling Strategy

Our strategy is first to model the expectation of the area-wide weekly average log-transformed BS levels, $\mu_t = \mathrm{E}[\bar{Y}_t]$, ignoring any spatial variation. This results in an estimate $\hat{\mu}_t$. We then use spatio-temporally referenced covariates $\mathbf{w}$ to account for residual variation between monitors. Hence, if t denotes week and x geographical location, we model log-transformed BS, $Y_t(x)$, as

$$Y_t(x) = \hat{\mu}_t + \mathbf{w}^T\boldsymbol{\beta} + Z_t(x), \tag{2.1}$$

where $Z_t(x)$ is a residual term which may or may not exhibit temporal and/or spatial correlation, and $\hat{\mu}_t$ is treated as an offset, provided that its associated prediction variance is negligible. Note that in (2.1), time is treated as discrete, with a resolution of one week, whereas x is treated as a spatial continuum, and that $\mathbf{w}$ depends implicitly on t and x. This framework acknowledges that, although our data are confined to a discrete set of monitor locations, our aim is to predict BS at every maternal residence within the study area.

Our two-stage modelling strategy is informed by two considerations. Firstly, the exploratory analysis showed that the temporal variation in $Y_t(x)$ dominates the residual spatio-temporal variation. Secondly, and not untypically (cf de Luna & Genton (2005)), our data are temporally rich but spatially sparse. Together, these features enable relatively precise estimation of the spatially-constant component μ_t. Other authors have preferred to fit different models to the individual time series obtained from each monitor, treating periods of inactivity as missing data (Haslett & Raftery (1989); Meiring *et al.* (1998)). For our data, the extent of the incompleteness

of the time series from individual monitors, as shown in Figure 2.2, makes this a less attractive strategy. Finally, construction of the spatio-temporal part of the model is greatly helped by the availability, at both monitor and residential locations, of a set of spatio-temporal covariates that are predictive of BS levels. Hence, anticipating the results in Section 2.3.4, we do not necessarily need to build an elaborate spatio-temporal stochastic model for the residual component $Z_t(x)$.

2.3.3 Stage 1 : Modelling Area-Wide Average BS Levels

To model μ_t, we note from Figure 2.3 the approximately linear decline in log-BS levels over the study period, and the clear seasonal pattern, with higher levels occurring during the winter months. We anticipated that the seasonal pattern might be partially attributable to seasonal variation in temperature. We therefore obtained daily temperature readings from nearby weather recording stations for the whole study period, and calculated a value d_t as the average of the daily minimum temperature readings over the seven days in week t. Finally, we set $\omega = 2\pi/52$ as the frequency corresponding to an annual cycle.

A first, static regression model for the spatial average $\bar{Y}_t$ is

$$\bar{Y}_t = \alpha + \beta t + \gamma d_t + A\cos(\omega t) + B\sin(\omega t) + U_t \tag{2.2}$$

where α, β, γ, A and B are parameters and the U_t are mutually independent $\mathrm{N}(0, \sigma_U^2)$ residuals.

Figure 2.4a shows that the model (2.2) captures much of the seasonal variation in $\bar{Y}_t$; for clarity, the diagram shows only representative results from years 1984 to 1992. However, the residuals show strong evidence of short-term and long-term autocorrelation, with small peaks corresponding to one- and two-year lags indicating that a static seasonal component is inadequate (Figure 2.4b). Re-examination of Figure 2.4a suggests that the lack of fit is primarily due to year-by-year variation in the phase and amplitude of the seasonal pattern. We therefore consider instead a dynamic regression model (West & Harrison (1997)),

$$\bar{Y}_t = \alpha + \beta t + \gamma d_t + A_t\cos(\omega t) + B_t\sin(\omega t) + U_t \tag{2.3}$$

where α, β, γ and U_t are as before, but now the static parameters A and B have been replaced

by independent random walks, hence

$$\begin{aligned} A_t|A_{t-1} &\sim \mathrm{N}(A_{t-1}, \sigma_A^2) \\ B_t|B_{t-1} &\sim \mathrm{N}(B_{t-1}, \sigma_B^2) \end{aligned}$$

Given initial values A_0 and B_0, the dynamic model (2.3) can be fitted either by direct maximization of the likelihood function, or via a Kalman filter followed by Kalman smoothing using, for example, functions `kfilter` and `smoother` in the contributed R package `sspir` (`www.R-project.org`).

The estimated parameter values $\hat{\alpha}, \hat{\beta}$ and $\hat{\gamma}$ differ little between models (2.2) and (2.3), but the estimated residual variance $\hat{\sigma}_U^2$ drops from 0.14 to 0.08 and the corresponding R^2-value increases slightly, from 0.88 to 0.93. More importantly, the dynamic model provides a qualitative improvement in fit compared to the static model: the residual autocorrelation largely disappears (Figure 2.4b) as a result of the more flexible fit to the seasonal pattern (Figure 2.4a).

2.3.4 Stage 2 : Modelling Residual Spatio-Temporal Variation

If we now apply the area-wide fitted values, $\hat{\mu}_t$ say, from (2.3) to the values of log-transformed BS at individual monitors, the residuals show a strong spatial pattern, with monitors towards the south of the study area tending to have large, positive residuals. This is consistent with the fact that early in the study period this part of the study area was dominated by areas of heavy industry.

As described in Section 2.3.2, we seek to explain this effect by treating $\hat{\mu}_t$ as an offset in a linear model for monitor-specific log-transformed BS levels $Y_t(x)$ that includes spatio-temporally referenced covariates. Note that to achieve our aim of predicting BS exposure at every residential location, any covariates in the model must be available not only at monitor locations, but throughout the study area.

Covariates

To account for residual spatio-temporal variation, we constructed the following candidate covariates:

w_1 : domestic chimney count within 500 metres;

w_2 : distance to nearest industrial area;

w_3 : binary indicator of land use, either residential ($w_3 = 1$) or non-residential ($w_3 = 0$);

w_4 : binary indicator of whether the 1956 Clean Air Act (CAA) had ($w_4 = 0$) or had not ($w_4 = 1$) been implemented;

w_5 : area of industry within 500 metres.

Covariates w_1, w_2 and w_5 were derived at a time-resolution of one year from digitized annual images of the study area.

Covariate w_3 was derived as follows. For any monitoring location x within the study area and a given value of $r > 0$, we counted the number of births in each year within a radius r of location x. We then identified, by trial-and-error, a range of values of r for which the resulting count distribution was strongly bimodal, suggesting a classification of monitoring locations with high counts as residential and locations with low counts as non-residential. Using this criterion with $r = 150$ metres provided the clearest distinction between residential and non-residential locations. Moreover, by this criterion the residential status of each monitoring location did not appear to change over time. We therefore considered w_3 to be time-constant, and defined a residential area to be one for which at least 50 births occurred within a 150-metre radius throughout the study period. The majority of monitors classified in this way as non-residential were in known industrial areas, although one was in a known commercial area.

Covariate w_4 was obtained from local government records. The CAA was implemented in stages across administrative sub-areas of the city between 1959 and 1978. The assumption that implementation within a sub-area took place at a fixed date, rather than gradually over a longer period of time, is questionable. However, in the absence of more detailed information we took the pragmatic decision to define w_4 as a binary factor, changing from 1 to 0 at the nominal implementation date for the sub-area in question.

For a preliminary assessment of the importance of each candidate covariate, we compared monitor-specific average residuals and covariates as follows. For each monitor, at location x say, we defined the average residual as a time-average of $Y_t(x) - \hat{\mu}_t$ over those weeks t in which the monitor was active, and the average covariate as the corresponding time-average of the covariate at the same location. For the binary covariates, w_3 and w_4, we compared the two distributions of average residuals corresponding to $w = 0$ and to $w = 1$. For w_1, w_2 and w_5, we examined scatterplots of monitor-specific average residuals against average covariate values.

On this basis, we discarded the industry variable w_5 because it showed a relatively weak relationship with monitor-specific average residuals and a strong relationship with the other covariates. The other covariates all showed a potentially useful relationship with the monitor-specific average residuals, and were therefore retained. Figure 2.5 shows the plot for the chimney count variable, w_1. Each point represents a monitor, and is labelled according to its residential status. The plot shows a positive relationship with chimney count for monitors in residential areas, and a negative relationship in non-residential areas, suggesting a strong interaction effect between chimney count and residential status. A possible explanation for this is that within industrial areas, very few domestic chimneys would be found close to the most heavily polluting industries, whereas rather more would be found close to lighter industries. In residential areas, there is relatively little variability between levels of emission per chimney, and pollution levels therefore show a direct relationship with chimney count.

Another important interaction is between chimney count and date of implementation of the CAA. After the CAA was implemented, the emission of black smoke from any building was prohibited. Thus, as a surrogate for local levels of black smoke emission, the chimney count could be considered as being effectively zero after CAA implementation. However, as discussed below, care is needed to interpret correctly the combined effect of CAA implementation and the estimated area-wide temporal trend, $\hat{\mu}_t$.

Model Formulation

We now consider a single linear model for the data from all monitors. Taking into account the above remarks, we assume the following model:

$$Y_t(x_k) = \hat{\mu}_t + \beta_{10} w_{10} + \beta_{11} w_{11} + \beta_2 w_2 + \beta_3 w_3 + \beta_4 w_4 + Z_t(x_k), \tag{2.4}$$

where x_k is the location of monitor k, w_2, w_3 and w_4 are as defined above, and $w_{1i} = w_1 \mathrm{I}(w_3 = i)$ where $\mathrm{I}(\cdot)$ is the indicator function. We also assume that $Z_t(x_k) \sim N(0, \sigma_Z^2)$ independently for all k and t. However, to preserve the interpretation of $\hat{\mu}_t$ as the area-wide average of S_t, we also need to centre each covariate appropriately. We therefore require that, for any given t, the fitted value from the spatio-temporal model (2.4), averaged over all monitors active at t, should equal $\hat{\mu}_t$. To satisfy this condition, for each covariate w at each time t we calculate $\bar{w} = (\sum w)/m_t$, where the sum is over the m_t monitors active at time t, and subtract each value of $\bar{w}$ from the corresponding value of w before entering into equation (2.4).

Assessment of Model Fit

Including monitor-specific fixed effects would be incompatible with our goal of spatial prediction. However, as part of the assessment of the model fit, we did consider the effect of adding monitor-specific levels α_k to the right-hand side of (2.4). This resulted in only a small increase in the R^2 value, from 0.84 to 0.86, and we therefore reverted to model (2.4).

To test the assumption of independent residuals $Z_t(x_k)$, we calculate a standardized average residual for each monitor k as

$$\bar{Z}_k^* = n_k^{-0.5} \sum_t \{S_t(x_k) - \hat{Y}_t(x_k)\},$$

where n_k is the number of weeks in which monitor k was active. Under the assumed model, $\bar{Z}_k^* \sim N(0, \sigma_Z^2)$, for all k. Figure 2.6 shows the standardized residuals plotted at their corresponding monitor locations. The visual impression is of a concentration of large, negative residuals close to the southern boundary of the study area. However, visual impressions from sparse spatial data can be misleading. For a formal test, we compute for each distinct pair (i, j) of monitors $u_{ij} = \|x_i - x_j\|$ and $v_{ij} = (\bar{Z}_i^* - \bar{Z}_j^*)^2$. We then use the sample correlation between the u_{ij} and the v_{ij} as a measure of the spatial dependence and compare the observed value with that obtained after randomly re-labelling the monitoring locations. The resulting Monte Carlo test, based on 999 independent re-labellings, gives a p-value of 0.7, corresponding to no significant evidence of spatial structure. Consistent with the result of the formal test, maps of re-labelled residuals (Figure 2.7) show chance spatial concentrations of large and small residuals comparable to those seen in Figure 2.6. We conclude that the assumption of spatially independent residuals $Z_t(x_k)$ is reasonable, and that any differences between monitors are likely to reflect properties of the monitors themselves, rather than of their locations.

We also examined the temporal pattern of residuals at individual monitoring stations. Timeplots of residuals, shown in Figure 2.8, reveal clear lack of fit for some monitors over some time periods. Table 1 summarizes the fit of the model to individual monitors, including the five validation monitors located outside the study area. The R^2-values for the 20 monitors within the study area vary between 0.21 and 0.87, but the smaller values of R^2 are generally associated with monitors for which we have relatively little data.

Validation

To assess the model's external validity we used five additional monitors situated just outside the study area. The locations of these monitors are shown in Figure 2.1. Historical records were less readily available for locations outside the study area: for example, the aerial photographs needed to construct the chimney count variable were available only for the years 1966 and 1974. For this reason, we consider only data from these years in our assessment of validity. Table 1 summarizes the fit for these five monitors. The fit is rather poor for some monitors, notably Hebburn 3, and we would not recommend extrapolating the model beyond the study area. Inevitably, imposing a common model on all available monitor locations within the study area compromises the fit to any individual monitor's data, but is a necessary simplification in order to address our goal of spatio-temporal prediction at arbitrary locations. Extrapolation beyond the study area is likely to exacerbate this effect; for example, although the locations of the validation monitors are geographically close to the boundary of the study area, they differ in their historical pattern of land use.

Interpretation of the Spatio-Temporal Model Coefficients

An alternative interpretation of (2.4) is obtained by re-casting the dynamic model (2.3) to allow different area-wide average log-transformed BS levels before and after CAA implementation. We denote these by μ_{dt} ('dirty') and μ_{ct} ('clean') respectively, and let p_t be the proportion of active monitors at week t that are dirty. Then, μ_t is a weighted average of log-transformed BS levels in dirty and clean areas,

$$\mu_t = p_t\mu_{dt} + (1 - p_t)\mu_{ct}.$$

Now suppose that $\mu_{dt} = \mu_{ct} + \lambda_t$ for some function λ_t, so that

$$\begin{aligned} \mu_t &= \mu_{ct} + p_t\lambda_t \\ &= \mu_{dt} + p_t\lambda_t - \lambda_t. \end{aligned}$$

The estimated contribution to the right-hand side of (2.4) for a clean monitor is $\hat{\mu}_t - \hat{\beta}_4 p_t$, whilst the estimated contribution for a dirty monitor is $\hat{\mu}_t + \hat{\beta}_4 - \hat{\beta}_4 p_t$. These quantities estimate μ_{ct} and μ_{dt} respectively. Hence, $\hat{\beta}_4$ can be interpreted as an estimate of the difference in average log-transformed BS levels between dirty and clean areas, on the assumption that this difference is constant over time. The estimate of this difference is $\hat{\beta}_4 = 0.33$, with standard error 0.013. Figure 2.9 shows the observed average difference in log-transformed BS between dirty and clean monitors at each time, and supports the assumption that λ_t is approximately constant.

Direct interpretation of the other β-coefficients in (2.4) is more difficult, owing to the necessary standardization of the covariates w. Nevertheless, we note that in each case the parameter estimate has the anticipated sign (i.e. negative for $\hat{\beta}_{10}$, $\hat{\beta}_2$ and $\hat{\beta}_3$; positive for $\hat{\beta}_{11}$).

2.4 Prediction of BS Exposure at Residential Locations

Our aim is to predict BS exposure at each maternal residential location, both for individual weeks and aggregated over time within the pregnancy. Thus, to compute prediction variances we need to consider not only the prediction variance for a single week, but also the covariance between predictions made for different weeks. Each birth is associated with a single residential location, x say, so in order to estimate an individual mother's exposure we need only consider prediction at that location. To simplify notation, we therefore suppress the dependence on x and write S_t for the BS level at time t, $Y_t = \log(S_t)$, and $\mathbf{w}_t$ for the covariate vector at this location x and week t. The following discussion then holds for any location x.

Suppose that our target for prediction is the time-aggregated BS exposure over weeks $t_1, ..., t_n$. As the prediction variance of $\hat{\mu}(t)$ is small by comparison with that of Y_t (approximately 0.08 versus 0.27), we treat $\hat{\mu}_t$ as known and equal to μ_t. The predicted value of $Y_t - \mu_t$ is $\hat{Y}_t - \hat{\mu}_t$ where $\hat{Y}_t = \mathbf{w}_t^T\hat{\boldsymbol{\beta}}$, with associated prediction variance

$$\begin{aligned} V(\hat{Y}_t) = V(Y_t - \hat{Y}_t) &= V(Z_t + \mathbf{w}_t^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \\ &= V(Z_t) + V(\mathbf{w}_t^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \\ &= \sigma_Z^2 + \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_t. \end{aligned}$$

Also, for $t \neq u$,

$$\begin{aligned} \text{Cov}(\hat{Y}_t, \hat{Y}_u) &= \text{Cov}(\mathbf{w}_t^T\hat{\boldsymbol{\beta}} + \hat{Z}_t, \mathbf{w}_u^T\hat{\boldsymbol{\beta}} + \hat{Z}_u) \\ &= \text{Cov}(\mathbf{w}_t^T\hat{\boldsymbol{\beta}}, \mathbf{w}_u^T\hat{\boldsymbol{\beta}}) \\ &= \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_u. \end{aligned}$$

Under the fitted model (2.4), $\sigma_Z^2 \gg \mathbf{w}_t^T V(\hat{\boldsymbol{\beta}})\mathbf{w}_t$ in any week t (approximately 0.27 and 0.00006,

respectively), and it follows that

$$\mathrm{Var}\left(\sum_{t=t_1}^{t_n}\hat{Y}_t\right) = \sum_{t=t_1}^{t_n}\mathrm{Var}(\hat{Y}_t) + 2\sum_{t<u}\mathrm{Cov}(\hat{Y}_t,\hat{Y}_u) \approx \sum_{t=t_1}^{t_n}\mathrm{Var}(\hat{Y}_t) \approx n\sigma_Z^2.$$

We require predictions on the original scale, rather than on the log-transformed scale. At a given location, $S_t = \exp(Y_t)$ and our targets for prediction are of the form $T = n^{-1}\sum_{t=t_1}^{t_n} S_t$. Under our assumed model, each S_t follows a log-Normal distribution. Writing $\xi_t = \mathrm{E}[Y_t]$ and $\Sigma_{tu} = \mathrm{Cov}\{Y_t, Y_u\}$, it follows that

$$\mathrm{E}(S_t) = \exp(\xi_t + \Sigma_{tt}/2),$$

$$\mathrm{Var}(S_t) = \exp(2\xi_t + \Sigma_{tt})(\exp(\Sigma_{tt}) - 1),$$

and for $t \neq u$,

$$\mathrm{Cov}(S_t, S_u) = \exp(\xi_t + \xi_u + (\Sigma_{tt} + \Sigma_{uu})/2)(\exp(\Sigma_{tu}) - 1).$$

The prediction variance for the average black smoke level T, over weeks $t_1, ..., t_n$, follows as

$$\mathrm{Var}(T) = \mathrm{Var}\left(\frac{1}{n}\sum_{t=t_1}^{t_n} S_t\right) = \frac{1}{n^2}\left(\sum_{t=t_1}^{t_n}\mathrm{Var}(S_t) + 2\sum_{t<u}\mathrm{Cov}(S_t, S_u)\right) \approx \frac{1}{n^2}\sum_{t=t_1}^{t_n}\mathrm{Var}(S_t),$$

and approximate prediction intervals can be computed using a Normal approximation. For example, an approximate 95% prediction interval for T is

$$\frac{1}{n}\sum_{t=t_1}^{t_n}\exp(\hat{\xi}_t + \hat{\Sigma}_{tt}/2) \pm 1.96\sqrt{\mathrm{Var}(\hat{T})}.$$

Figure 2.10 shows a grey-scale image of predicted values on the logarithmic scale for four weeks, corresponding to summer and winter in 1969 and 1982. Non-residential locations, for which prediction is of no interest, are shown in Figure 2.10 as white areas. One feature of Figure 2.10 is the relatively low spatial variation at any one time, by comparison with the variation either between different seasons in the same year, or between different years for the same season. This pattern is consistent with our exploratory analysis of these data as reported in Section 2.3.1.

The pattern of prediction variances is qualitatively similar to that of the predictions themselves, as a consequence of the log-Normal distributional assumption for untransformed BS concentrations.

2.5 Discussion

We have demonstrated a two-stage modelling strategy for modelling spatio-temporal data using monitoring data that is temporally dense and spatially sparse, a common scenario in epidemiological studies of air pollution exposure. In the first stage, we used a dynamic model for the purely temporal trend, while in the second we used appropriately constructed covariates to take account of remaining spatio-temporal variation. Using a dynamic model in the first stage obviates the need to consider separate models for short-term and long-term correlation between observations, and in our application resulted in a materially better fit to seasonal variation in spatially averaged pollution levels than was obtainable from a static harmonic regression model.

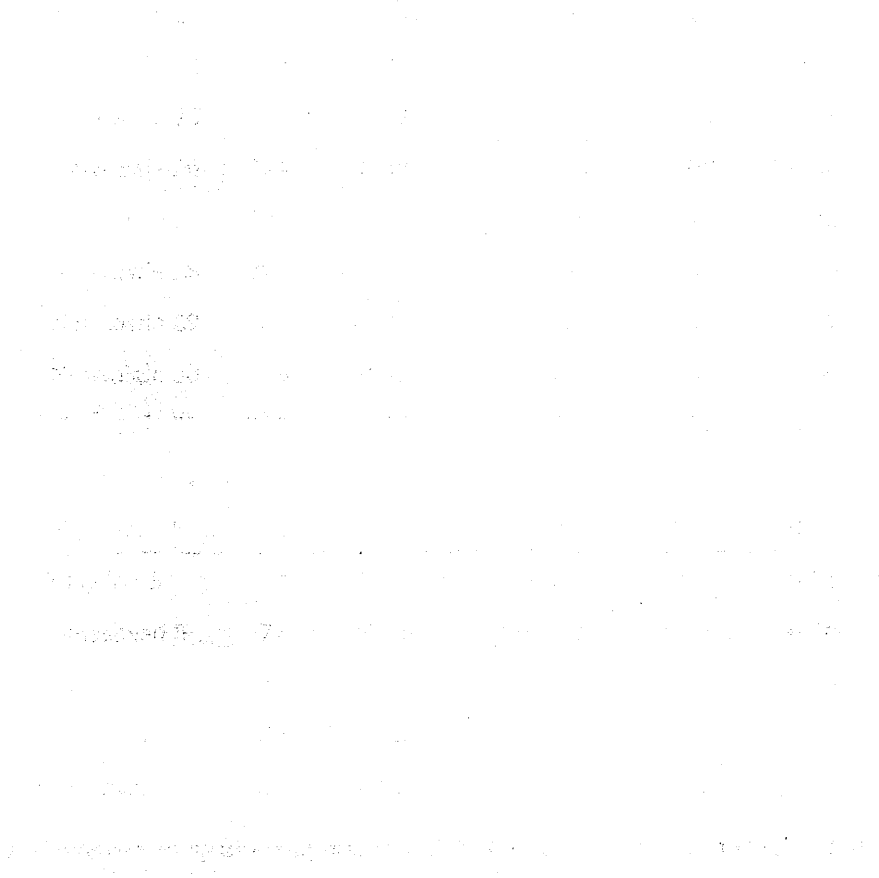
The area-wide average log-transformed BS levels given by the first-stage model are relatively precise, with prediction variance around 0.08 compared with predicted values ranging between 1.7 and 6.3. In contrast, the spatial sparsity of the data makes it important to take account of the uncertainty in the predictions at particular locations. Our exposure estimates will subsequently be used as covariates in an analysis of the relationship between exposure and adverse birth outcomes, in which context it will be necessary to check that conclusions are robust against the statistical error in the exposure estimates. We believe that these estimates, although only surrogates for the true levels of pollution to which mothers were exposed, indicate a more realistic pattern of exposure than would an assumption of homogeneity of exposures across a whole city. This seems likely to hold true both for particulate matter and for other pollutants, for which there is evidence elsewhere (Haas (1995); Meiring *et al.* (1998); Zidek *et al.* (2002)).

In our application, we have been able to model the spatio-temporal variation without the need to model spatio-temporal correlation in the residuals. This greatly eases the computational burden of computing predictions and prediction variances. In principle the methodology extends directly to models with correlated residuals, provided that we are prepared to specify a spatio-temporal covariance structure for the residual process $Z_t(x)$; see, for example, Gneiting *et al.* (2007). In problems of this kind, we would always advocate the use of relevant covariate information to explain as much as possible of the spatio-temporal variation. Nevertheless, and as the results from the validation sites indicate, extrapolation beyond the area from which the model was constructed is almost certainly unwarranted. In other settings the importance of other sources of pollution, for example traffic emissions, may require the use of different covariates. The means by which suitable covariates are identified and constructed is not necessarily straightforward and may require a degree of imagination; in our application, the construction of the chimney count

covariate and careful consideration of its interaction with both the residential/non-residential land-use classification and with the effect of the Clean Air Act were crucial to the implementation of the methodology.

Acknowledgements

The PAMPER study was funded by the Wellcome Trust, UK charity, grant no. 072465/Z/03/Z. TRF is supported by a Doctoral Training Account studentship and PJD by a Senior Fellowship from the Engineering and Physical Sciences Research Council (EPSRC).



Monitor		n	Mean residual	Standardized mean residual	Residual variance	Raw variance	R^2
1	Gosforth 1	676	0.14	3.74	0.17	0.92	0.81
2	Gosforth 2	22	0.22	1.03	0.20	0.34	0.42
3	Newburn 2	1598	-0.02	-0.78	0.20	1.58	0.87
4	Newcastle 17	1399	0.08	2.84	0.24	0.81	0.71
5	Newcastle 18	670	-0.16	-4.09	0.13	0.66	0.80
6	Newcastle 19	688	-0.09	-2.44	0.19	0.70	0.73
7	Newcastle 20	360	0.08	1.59	0.26	0.66	0.60
8	Newcastle 21	445	-0.22	-4.66	0.25	0.66	0.63
9	Newcastle 22	321	0.24	4.27	0.20	0.61	0.67
10	Newcastle 23	52	0.09	0.68	0.22	0.53	0.59
11	Newcastle 24	1064	0.20	6.38	0.26	1.61	0.84
12	Newcastle 25	339	-0.08	-1.52	0.14	0.55	0.74
13	Newcastle 26	224	-0.32	-4.86	0.09	0.44	0.79
14	Newcastle 27	1198	-0.12	-4.05	0.15	0.65	0.76
15	Newcastle 28	229	-0.23	-3.55	0.08	0.38	0.78
16	Newcastle 29	89	-0.23	-2.18	0.20	0.32	0.39
17	Newcastle 30	44	-0.06	-0.37	0.08	0.30	0.73
18	Newcastle 31	527	0.23	5.34	0.19	0.54	0.65
19	Newcastle 32	224	-0.09	-1.40	0.14	0.34	0.58
20	Newcastle 5	5	-0.31	-0.69	0.05	0.06	0.21
21	Blaydon 3	4	-1.27	-2.55	0.28	0.25	-0.11
22	Gateshead 5	77	-0.26	-2.27	0.11	0.48	0.76
23	Hebburn 3	52	-1.27	-9.17	0.27	0.33	0.18
24	Hebburn 4	52	0.62	4.51	0.26	0.43	0.38
25	Newburn 1	87	0.63	5.85	0.09	0.28	0.67

Table 2.1: Summary of spatio-temporal model fit for each of the 20 monitors used for model fitting and five monitors used for validation (see Figure 2.1). n refers to the number of weeks for which the monitor was active. $R^2 = 1 -$ (residual variance/raw variance).

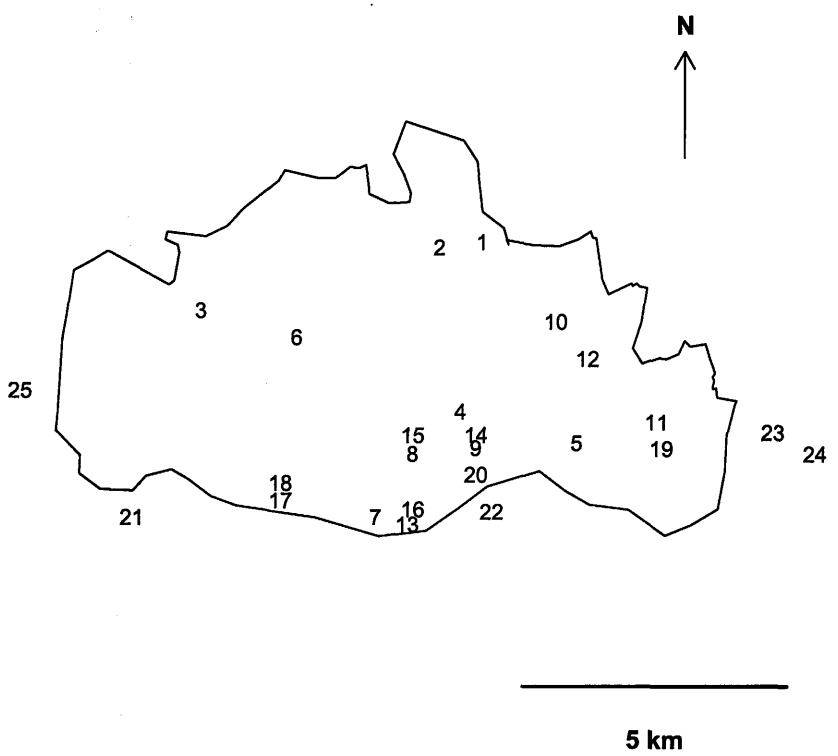


Figure 2.1: Outline of the PAMPER study area, Newcastle-upon-Tyne. Locations of black smoke monitoring stations used for modelling are numbered 1-20; those used for validation are numbered 21-25.

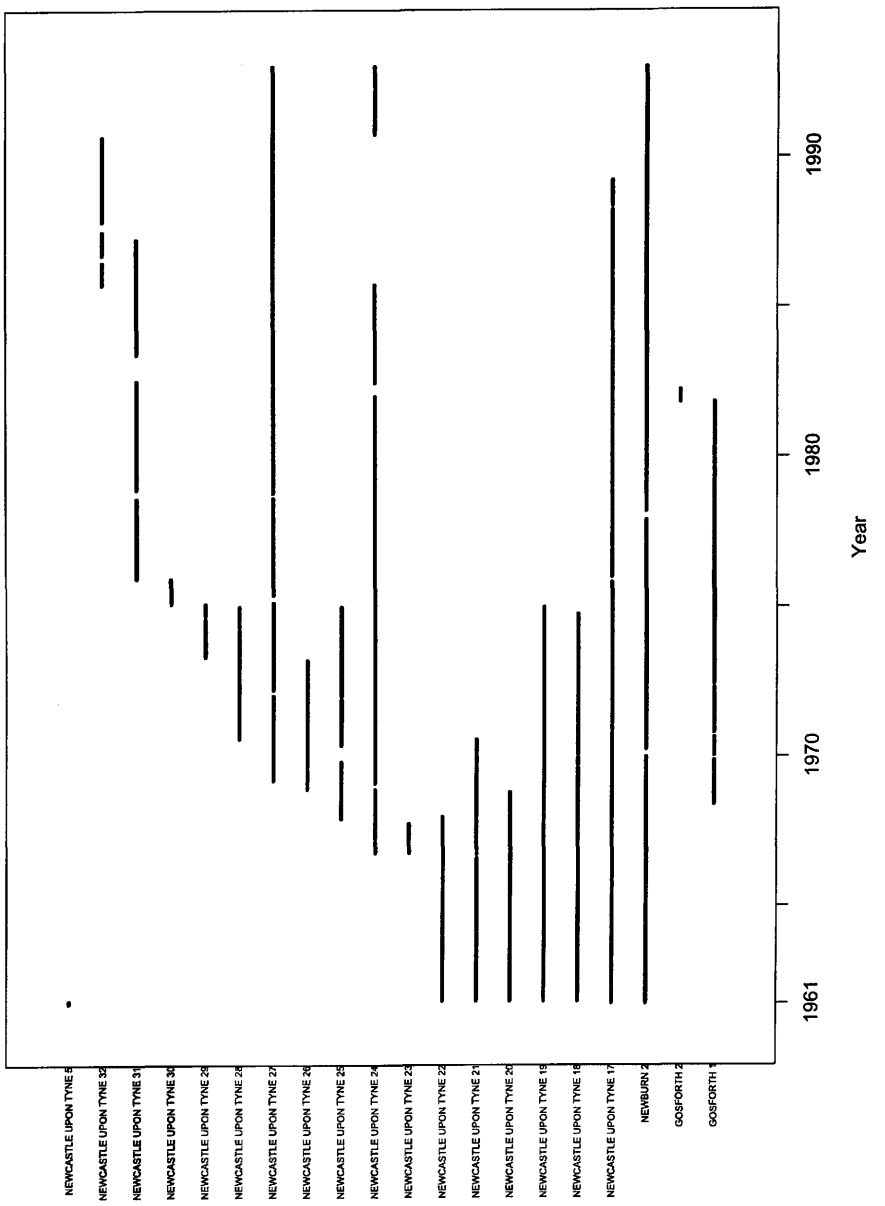


Figure 2.2: Diagram showing PAMPER monitoring station activity. Periods of activity are indicated by a black line.

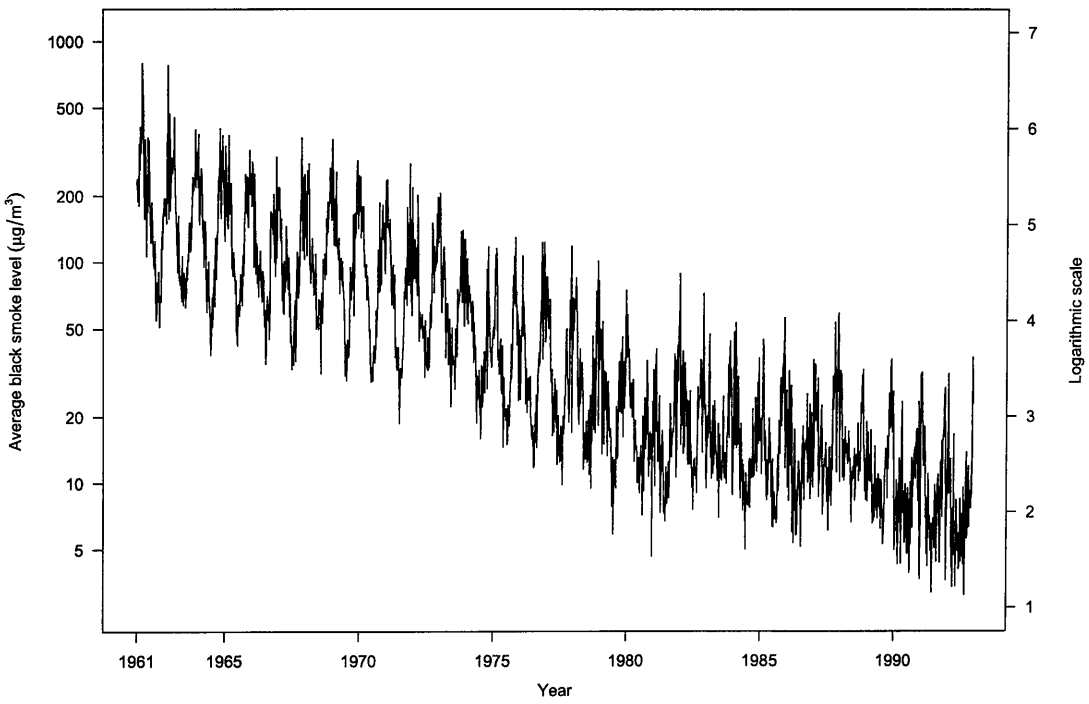


Figure 2.3: Area-wide weekly average black smoke levels, plotted as a time series. The original scale is shown on the left vertical axis and the logarithmic scale on the right vertical axis.

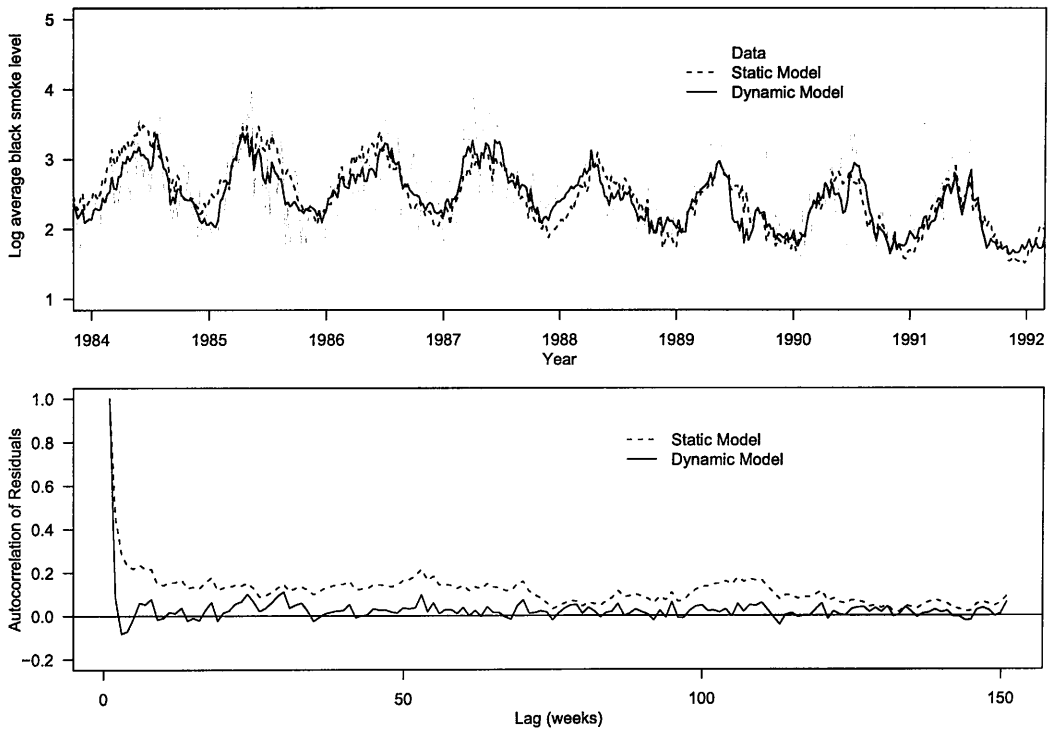


Figure 2.4: a. Fit of static (2.2) and dynamic (2.3) regression models for area-wide average black smoke levels, 1984-1992; b. Autocorrelation of residuals from static and dynamic models for area-wide average black smoke levels

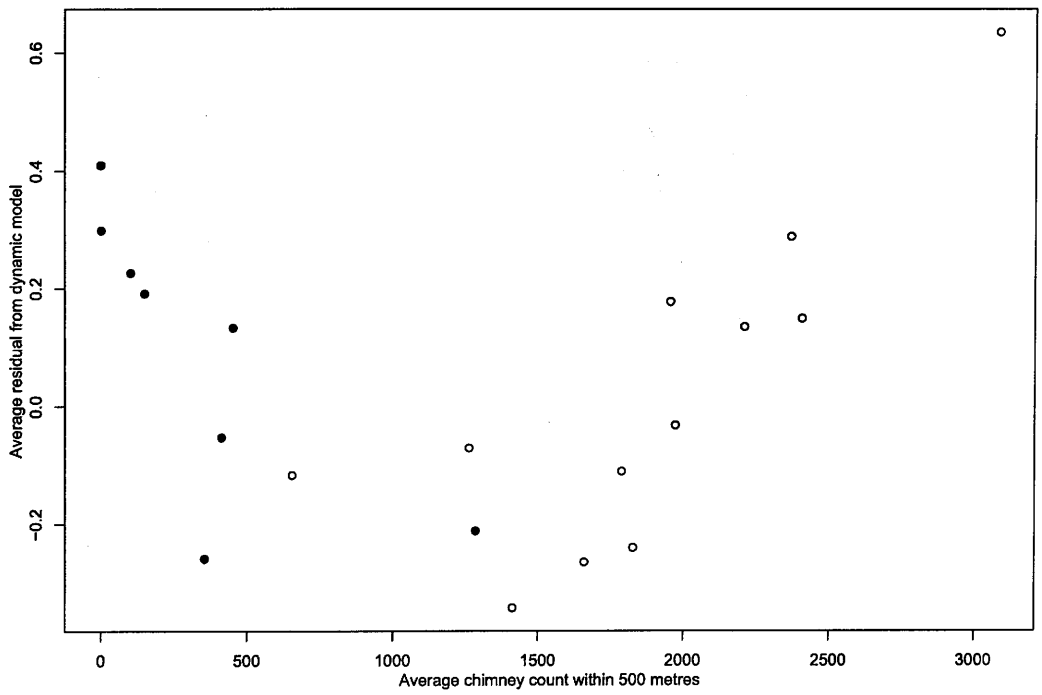


Figure 2.5: Monitor-specific average residual from dynamic model (2.3), plotted against average chimney count within 500 metres. Points are labelled according to the monitor's residential status (open circle=residential, filled circle=non-residential).

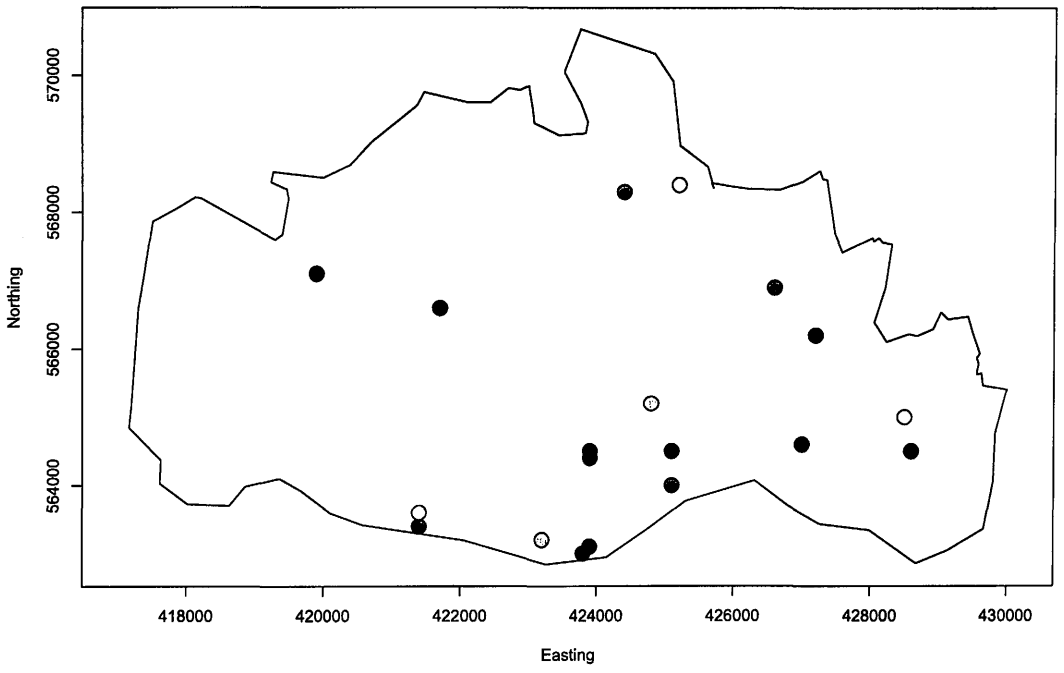


Figure 2.6: Map of standardized monitor-specific residuals from model (2.4). Darker shades indicate larger negative residuals, and lighter shades larger positive residuals

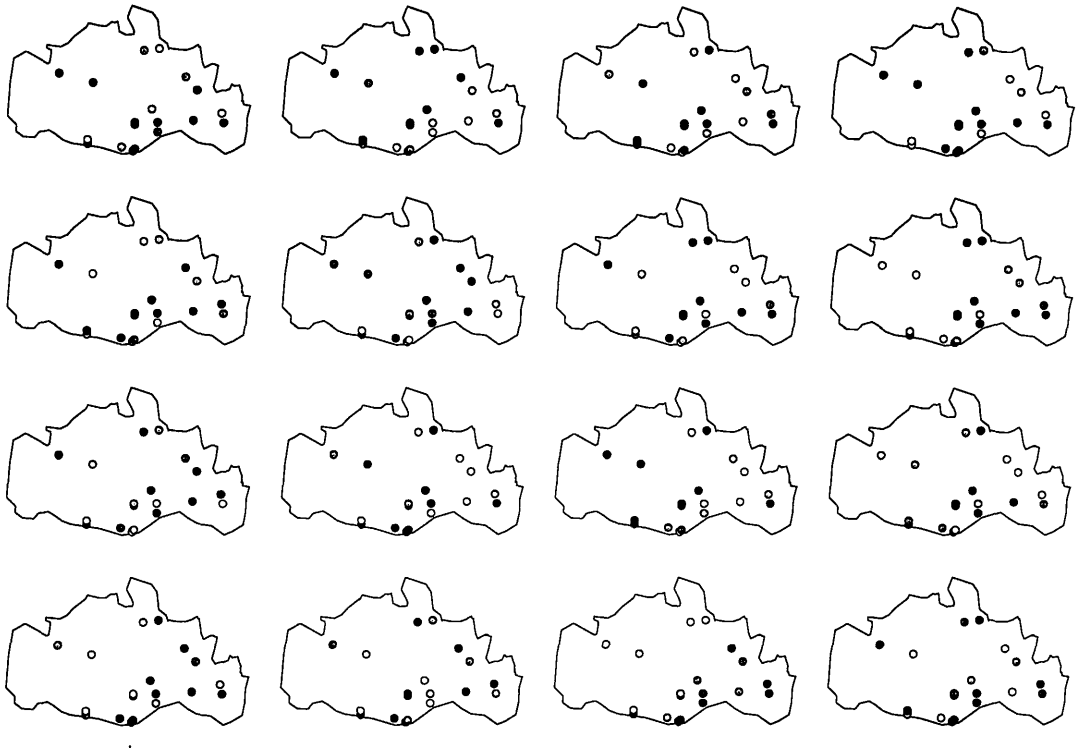


Figure 2.7: 16 replicates of a map of standardized monitor-specific residuals from model (2.4) with monitor locations randomly reassigned. Darker shades indicate larger negative residuals, and lighter shades larger positive residuals. The observed map (Figure 2.6) appears in the top left.

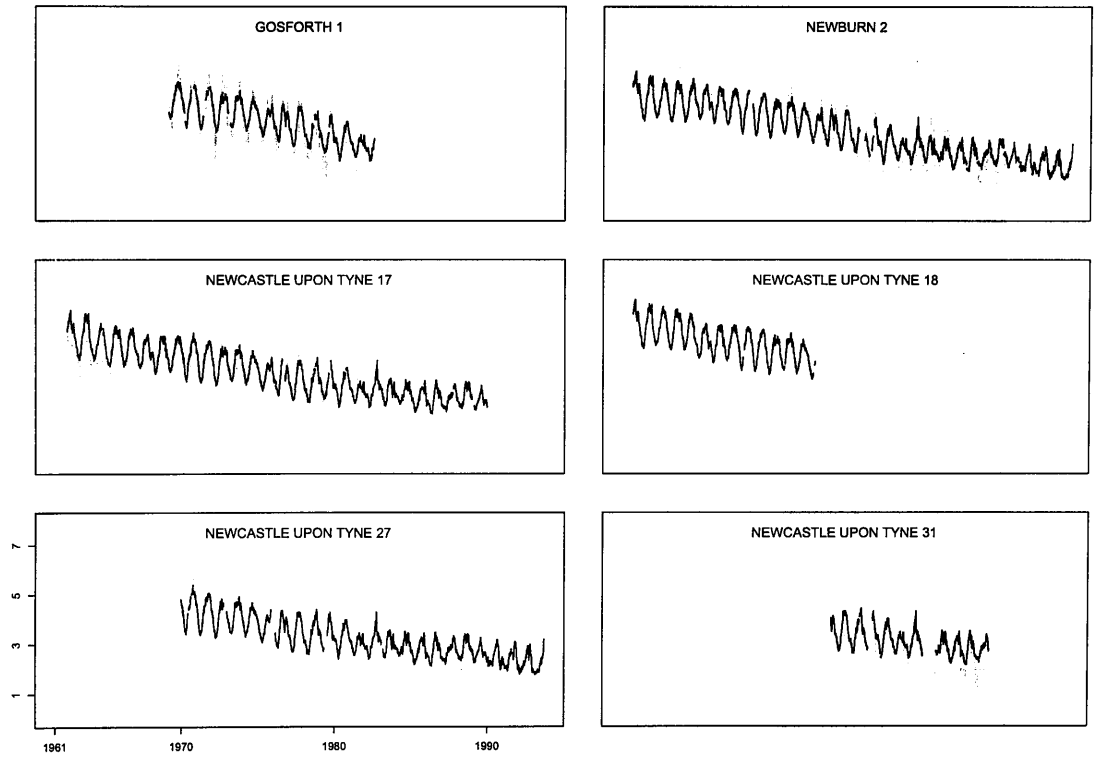


Figure 2.8: Observed (grey lines) and fitted (black lines) values from model (2.4) for six monitors used for model fitting

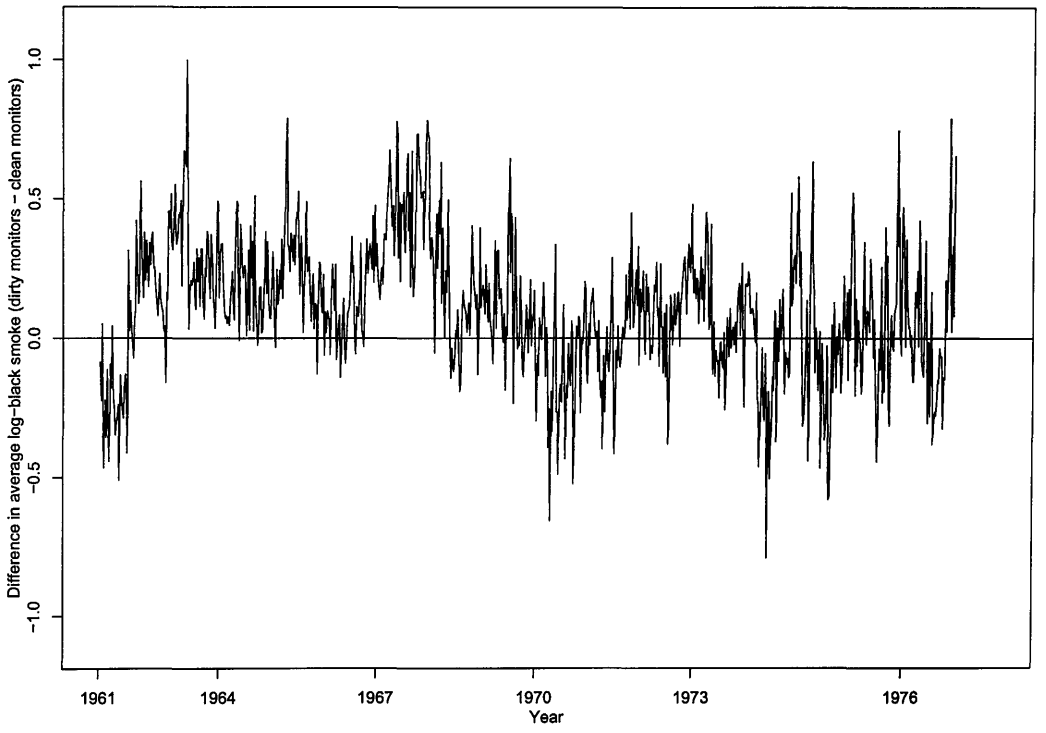


Figure 2.9: Difference between average log-black smoke levels in monitors operating in areas before ('dirty') and after ('clean') the implementation of the 1956 Clean Air Act

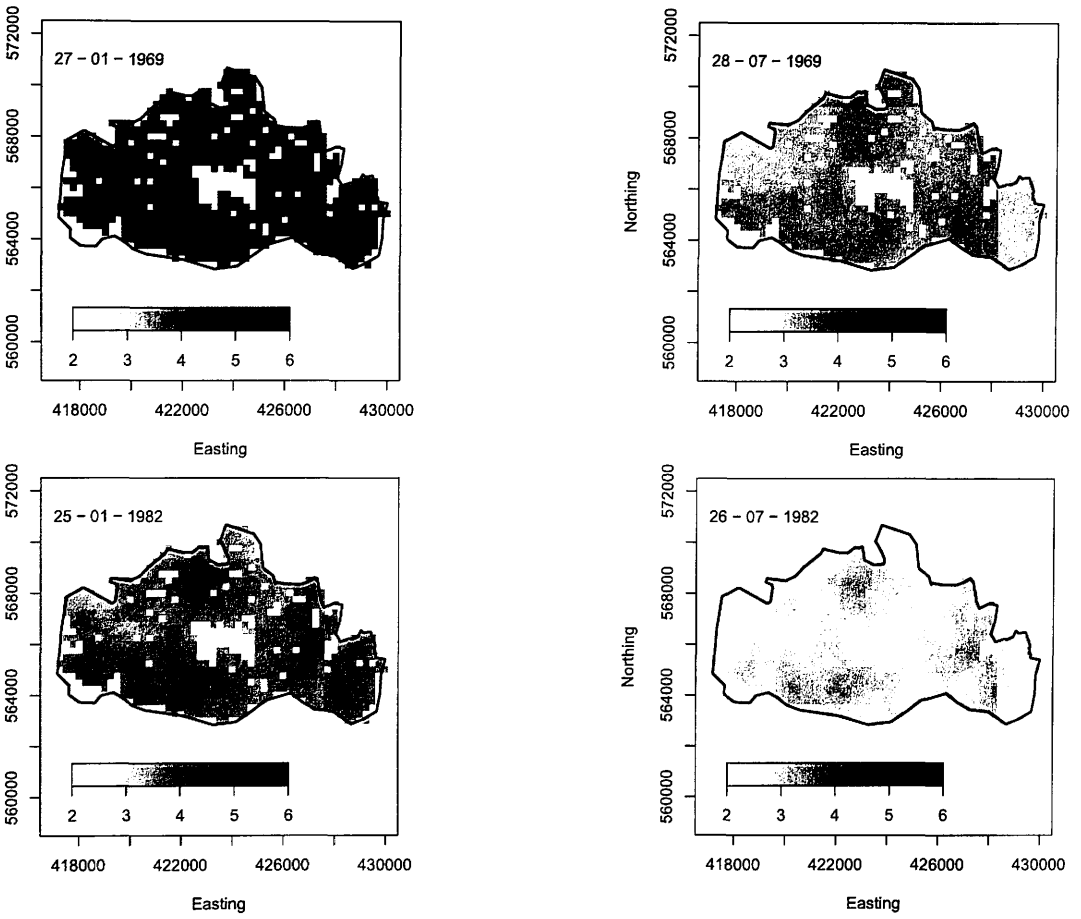


Figure 2.10: Point predictions for log-BS levels for four single weeks (dates inset) representing winter and summer, 1969 and 1982. White pixels correspond to non-residential areas, for which no prediction is made.

References

Bobak, M. 2000. Outdoor air pollution, low birth weight, and prematurity. *Environmental Health Perspectives*, **108**, 173–176.

Bogaert, P., & Christakos, G. 1997. Stochastic analysis of spatiotemporal solute content measurements using a regression model. *Stochastic Hydraulics and Hydrology*, **11**, 267–295.

Brown, P.E., Diggle, P.J., Lord, M.E., & Young, P.C. 2001. Space-time calibration of radar rainfall data. *Applied Statistics*, **50**, 221–241.

Carroll, R.J., Chen, R., George, E.I., Li, T.H., Newton, H.J., Schmiedliche, H., & Wang, N. 1997. Ozone Exposure and Population Density in Harris County, Texas. *Journal of the American Statistical Association*, **92**, 392–404.

de Luna, X., & Genton, M.G. 2005. Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica*, **15**, 547–568.

Glinianaia, S.V., Rankin, J., Bell, R., Pless-Mulloli, T., & Howel, D. 2004a. Does particulate air pollution contribute to infant death? A systematic review. *Environmental Health Perspectives*, **112**, 1365–1371.

Glinianaia, S.V., Rankin, J., Bell, R., Pless-Mulloli, T., & Howel, D. 2004b. Particulate air pollution and fetal health: a systematic review of the epidemiological evidence. *Epidemiology*, **15**, 36–45.

Gneiting, T., Genton, M.G., & Guttorp, P. 2007. *In: Statistical Methods for Spatio-Temporal Systems*. Boca Raton, Chapman and Hall/CRC. Chap. 4.

Haas, T.C. 1995. Local prediction of a spatiotemporal process with an application to wet sulfate decomposition. *Journal of the American Statistical Association*, **90**, 1189–1199.

Handcock, M.S., & Wallis, J.R. 1994. An approach to statistical spatial-temporal modelling of meteorological fields. *Journal of the American Statistical Association*, **89**, 368–378.

Haslett, J., & Raftery, A.E. 1989. Space-Time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource. *Applied Statistics*, **38**, 1–50.

Higdon, D. 2007. *In: Statistical Methods for Spatio-Temporal Systems*. Boca Raton, Chapman and Hall/CRC. Chap. 6.

Jerrett, M., Buzzelli, M., Burnett, R.T., & DeLuca, P.F. 2005. Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Social Science and Medicine*, **60**, 2845–2863.

Kyriakidis, P.C., & Journal, A.G. 1999. Geostatistical Space-Time Models: A Review. *Mathematical Geology*, **31**, 651–684.

Li, K., Le, N.D., Sun, L., & Zidek, J.V. 1999. Spatial-temporal models for ambient hourly PM_{10} in Vancouver. *Environmetrics*, **10**, 321–338.

Meiring, W., Guttorp, P., & Sampson, P.D. 1998. Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, **5**, 197–222.

Pless-Mulloli, T., Glinianaia, S.V., Rushton, S., Lurz, P.W.W., Sanderson, R., Fanshawe, T.R., Pearce, M.S., Charlton, M., Shirley, M. Rankin, J., & Diggle, P.J. 2007. Within-city space and time variation of black smoke exposure during pregnancy over 32 years. *Submitted.*

Pope III, C.A., & Dockery, D.W. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association*, **56**, 709–742.

Ritz, B., Wilhelm, M., & Zhao, Y. 2006. Air pollution and infant death in southern California, 1989-2000. *Pediatrics*, **118**, 493–502.

Sahu, S.K., & Mardia, K.V. 2005 (September 21-23). Recent Trends in Modeling Spatio-Temporal Data. *Pages 69–83 of: Proceedings of the special meeting on Statistics and Environment.*

Sahu, S.K., Gelfand, A.E., & Holland, D.M. 2006. Spatio-temporal monitoring of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 61–86.

Samet, J.M., Dominici, F., Curreiro, F.C., Coursac, I., & Zeger, S.L. 2000. Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *New England Journal of Medicine*, **343**, 1742–1749.

Stroud, J.R., Müller, P., & Sansó, B. 2001. Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society B*, **63**, 673–689.

Šrám, R.J., Binková, B., Dejmek, J., & Bobak, M. 2005. Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental Health Perspectives*, **113**, 375–382.

West, M., & Harrison, P.J. 1997. *Bayesian forecasting and dynamic models*. New York, Springer-Verlag.

Woodruff, T.J., Grillo, J., & Schoendorf, K.C. 1997. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. *Environmental Health Perspectives*, **105**, 608–612.

Woodruff, T.J., Parker, J.D., & Schoendorf, K.C. 2006. Fine particulate matter ($PM_{2.5}$) air pollution and selected causes of postneonatal infant mortality in California. *Environmental Health Perspectives*, **114**, 786–790.

Zidek, J.V., Sun, L., Le, N., & Özkaynak, H. 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics*, **13**, 595–613.

Chapter 3

Paper 2: Spatial Prediction in the Presence of Positional Error

---

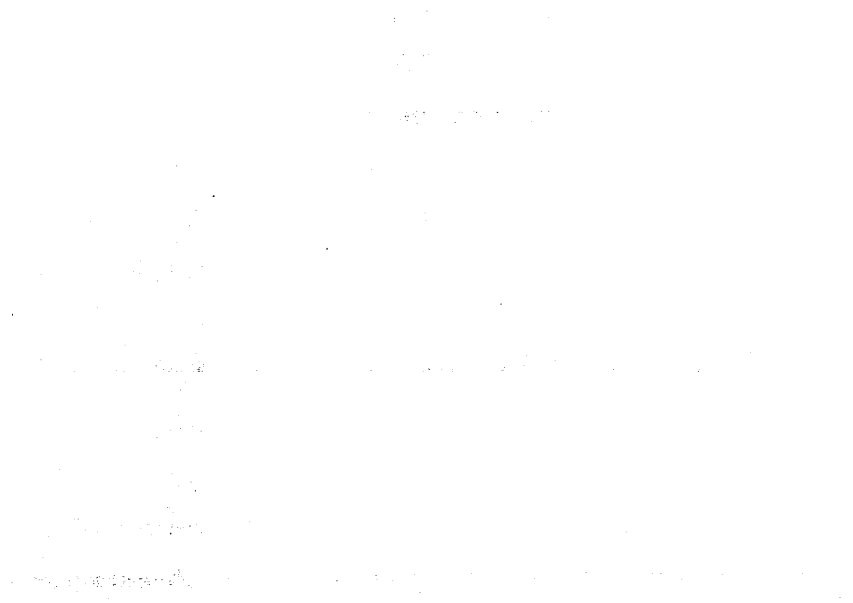
T.R. Fanshawe, P.J. Diggle

School of Health and Medicine, Lancaster University, UK

Summary

Standard analyses of spatial data assume that measurement and prediction locations are measured precisely. In this paper we consider how the problems of inference and prediction change when this assumption is relaxed and the locations are subject to positional error. We describe basic models for positional error and assess their impact on spatial prediction. Using both simulated data and lead concentration pollution data from Galicia, Spain, we show how the predictive distributions of quantities of interest change after allowing for the positional error, and describe scenarios in which positional errors may affect the qualitative conclusions of an analysis. The subject of positional error is of particular relevance when assessing the exposure of an individual to an environmental pollutant, when the position of the individual is often tracked using imperfect measurement technology.

Key words: Environmental epidemiology; Geostatistics; Measurement error; Monte Carlo inference; Location error



3.1 Introduction

In this paper, we consider the problem of spatial prediction using data that consist of real-valued measurements $Y_i : i = 1, ..., n$ associated with a corresponding set of spatial locations x_i in a region of interest D. A widely-used model specifies the statistical relationship between the pairs (Y_i, x_i) and an underlying spatially continuous phenomenon $S = \{S(x) : x \in \mathbb{R}^2\}$ as

$$Y_i = d(x_i)'\beta + S(x_i) + Z_i : i = 1, ..., n, \tag{3.1}$$

where $d(x)$ is a set of spatially referenced explanatory variables, $S(x)$ is a zero-mean Gaussian stochastic process with variance σ^2 and correlation matrix $R(\phi)$, where ϕ may be a vector-valued parameter, and the Z_i are mutually independent Gaussian errors, independent of S, with mean zero and variance τ^2. We write $\theta = (\sigma^2, \phi, \tau^2)$ for the vector of variance and covariance parameters. Model (3.1) is standard in the area of spatial statistics known as geostatistics (Chilès & Delfiner (1999); Diggle & Ribeiro Jr. (2007)).

Almost all applications require estimation of the parameter vector θ, whether the parameter estimates are of intrinsic interest themselves, or are merely a stepping-stone towards the goal of spatial prediction. In particular, the so-called nugget variance $\tau^2 = \text{Var}[Y_i|S(x_i)] : i = 1, ..., n$ can be thought of as the sum of small-scale spatial variation and the measurement error inherent in taking measurements Y_i. A more general formulation of (3.1) therefore explicitly recognises these two sources of variation, and replaces Z_i by independent processes $Z_i^{(S)}$ and $Z_i^{(M)}$, which represent small-scale spatial variation and measurement error respectively (Cressie (1991), page 112). In practice these two processes cannot be distinguished without using replicated spatial locations in the sampling design, and in the current paper we therefore adopt the more commonly-used model (3.1).

The problem of estimating the parameter τ^2 is described in most standard texts and several research papers on spatial data analysis (e.g. Cressie (1988, 1991); Jaksa *et al.* (1997); Chilès & Delfiner (1999)). In contrast, the issue of equivalent measurement error in recording the spatial locations x_i has received little attention. In practice, locations at which measurements are taken are often not recorded or stored precisely, either because of errors in recording devices or because of rounding off for computational convenience.

In this paper we consider the positional error problem in more detail, concentrating primarily on

spatial prediction but also considering how to obtain parameter estimates when positional errors are present. We consider two relevant scenarios in which positional error may occur: when the positional error affects the data locations; and when the positional error affects the prediction locations. Either or both of these scenarios may occur in applications.

Gabrosek & Cressie (2002) and Cressie & Kornak (2003) both consider the impact of positional errors in data locations on parameter inference and spatial prediction. Their approach is to define a new process consisting of pairs of measurements and locations that have been subjected to positional error, and then to derive estimates of the first two moments of this process. This approach assumes that the measurement error model has parameters that are known via *a priori* information, rather than estimated from the data. For prediction, they assume a predictor that is linear in the data values, i.e. of the form $\sum_i a_i Y_i + b$. When it is assumed that there is no positional error, the minimum mean square error predictor lies in this class of linear predictors. In the current paper we relax the assumption of linearity of the predictor and instead use a model-based approach (Diggle & Ribeiro Jr. (2007)).

Cressie & Kornak (2003) conclude that ignoring non-negligible positional error may result in predictions whose bias and mean square error tend to increase as the positional error variance increases and the range parameter ϕ decreases. They discuss an application in which locations of ozone measurements have, for convenience, been rounded off to the nearest grid coordinate. In the context of point-process data, Zimmerman (2008) considers errors derived from 'incomplete geocoding', or 'coarsening' (errors resulting from round-off to a coarse grid) using kernel smoothing methods to allow for the positional error.

Other recent papers describe applications in which the issue of positional errors arises. For example, Barber *et al.* (2006) consider the positional error in a global positioning system (G.P.S.) in a model for map calibration, using both existing maps and G.P.S. measurements to make inferences about true map locations. Related work on map calibration appears in Kiiveri (1997). Persson *et al.* (2006) consider independent positional errors in a series of G.P.S.-recorded locations that are known to form a polygon, and describe methods for estimating the true position and area of the polygon. These applications differ from the main focus of the current paper in that they aim to predict the true position *per se*, rather than the underlying spatial process S at the true position. Zandbergen & Green (2007) provide a simple comparison of four geocoding methods and a set of assumed true locations, with the aim of assessing air pollution exposure of

schoolchildren, but do not consider any specific model for the positional error.

Epidemiological studies relating air pollution exposure to adverse health outcomes have in recent years increasingly focused on providing measures of individual exposure, rather than assigning identical exposure estimates across sub-populations (Pope III & Dockery (2006); Fanshawe *et al.* (2008)). We discuss exposure estimation in this paper. Further analyses may consider associations between health outcomes and exposure estimates; these are often measured at different locations or different levels of spatial aggregation to one another, which results in a related problem known as 'spatial misalignment'. However, papers discussing this problem do not typically consider positional error in either the outcome or the exposure (Madsen *et al.* (2008); Gryparis *et al.* (2009)).

The work in the current paper is relevant to studies in which imperfect measurement devices such as G.P.S. are used to track the locations of individuals. Our results will be of particular use in studies that look at short-term pollution exposure, for example when considering the effect of positional error on spatial prediction over a trajectory, such as might be obtained by using a G.P.S. trace to monitor the movement of an individual over a short period. In this context, positional errors apply to the locations where predictions are to be made, but not necessarily the locations of data measurements, which typically occur at fixed, known locations such as air pollution monitoring sites.

The paper is structured as follows. In Section 3.2 we summarise the background to positional error models and draw parallels with measurement error models encountered in other fields. In Section 3.3 we investigate the effects of positional error on parameter estimation and prediction, discussing the effect of positional errors in both prediction and measurement locations. In Section 3.4 we illustrate our results using data from lead concentrations measured in moss samples from Galicia, Spain. Section 3.5 is a concluding discussion.

3.2 The Positional Error Model

Most analyses of spatial data assume that devices used to measure geographical location operate flawlessly. Here, we introduce a model for the positional error in the data locations x_i. Models for positional error are in many ways analogous to measurement error models used in other areas of statistics, particularly regression modelling, for which there is a wide literature (e.g. Cheng & Van Ness (1999)). Indeed, model (3.1) can be regarded as a particular type of non-linear

regression model in which the non-linear effect of the x_i on the observations Y_i acts through both the covariates d and the process S, and thus much of the general work on measurement errors in non-linear models (Carroll *et al.* (2006)) also applies in the context of spatial modelling.

We begin by considering the model specification in more detail. Positional errors can affect measurements of data locations and prediction locations. Errors in prediction locations have no impact on parameter estimation, so we will not consider them before Section 3.3.2; until then we consider the case in which positional error affects only data locations.

Let x_i^* denote the true location at which a measurement Y_i is taken, erroneously recorded at x_i because of positional error, and let X_i and X_i^* denote random variables corresponding to observed and true locations. Using the notation [ ] to mean 'distribution of', let $[Y|S, X]$ denote the conditional distribution of $Y(X^*)$, i.e. measured at location X^* but assigned to location X, given the corresponding value of the signal $S(X^*)$ and the observed, but incorrect, location X. This notation implicitly assumes that there is no stochastic dependence between the mechanisms which generate the positions X and X^* and those which govern the processes S and Y.

The model (3.1) can be written as

$$\begin{aligned}[Y, S, X, X^*] &= [Y|S, X, X^*][S, X, X^*] \\ &= [Y|S, X^*][S|X^*][X^*|X][X] \end{aligned} \tag{3.2}$$

The second line follows as $[Y|S, X, X^*] = [Y|S, X^*]$: conditional on the true location X^*, the erroneous location X carries no further information about Y. In practice we use a discrete version of S by sampling it on a finite set of locations $x_1, ..., x_n$.

In this paper we primarily consider models for $[X^*|X]$, as this conditional distribution appears naturally in (3.2). We assume that the (symmetric) distribution for positional error is bivariate Normal with uncorrelated components and variance γ^2, i.e. $X_i^*|X_i \sim \text{BVN}(X_i, \gamma^2 I_2)$, and that errors at different locations are independent. We also assume that no contextual information is available regarding the locations of either the true or observed data or prediction points, and thus the marginal distributions $[X^*]$ and $[X]$ will be identical and uniformly distributed over D. In a Bayesian setting, these marginal distributions are regarded as priors for the true and observed locations. From Bayes's rule $[X^*|X][X] = [X|X^*][X^*]$, under our assumptions it makes no difference whether we specify a positional error model in terms of $[X^*|X]$ or $[X|X^*]$. More

generally, these two conditional representations would correspond to the so-called 'Berkson' and 'classical' measurement error model classes respectively.

Using elementary properties of the Normal distribution, the assumed model for $[X_i^*|X_i]$ can also be written in the form $X_i^* = X_i + \epsilon_i$, where the ϵ_i are independent zero-mean bivariate Normal random variables with covariance matrix $\gamma^2 I_2$. The conditional distribution notation makes explicit the way in which quantities such as the likelihood depend on the assumed model for the positional error, as shown in the next section.

3.3 Inference

3.3.1 Estimation

The model specified in Section 3.2 enables the likelihood function of the standard model, $L(\theta, \beta) = [Y|\theta, \beta]$, viewed as a function of (θ, β), to be extended to allow for the additional random variable X and parameter γ. From (3.1), $[Y|S, X^*, \beta, \theta]$ is a product of $\mathrm{N}(d(X_i^*)'\beta + S(X_i^*), \tau^2)$ densities and $[S|X^*, \theta] \equiv [S(X^*)|X^*, \theta]$ is multivariate Gaussian with mean zero and covariance matrix $\sigma^2 R(X^*; \phi)$. $[X^*|X, \gamma]$ is defined by the positional error model. From (3.2), the new likelihood is $L(\theta, \beta, \gamma) = [Y, X|\theta, \beta, \gamma]$, viewed as a function of (θ, β, γ). We have

$$\begin{aligned} L(\theta, \beta, \gamma) &= \int\int [Y, S, X, X^*|\theta, \beta, \gamma] dS dX^* \\ &= \int\int [Y|S, X, X^*, \theta, \beta, \gamma][S, X, X^*|\theta, \beta, \gamma] dS dX^* \\ &= \int\int [Y|S, X^*, \theta, \beta, \gamma][S|X^*, \theta, \beta, \gamma][X^*|X, \theta, \beta, \gamma][X|\theta, \beta, \gamma] dS dX^* \\ &\propto \int\int [Y|S, X^*, \theta, \beta][S|X^*, \theta][X^*|X, \gamma] dS dX^*, \end{aligned} \tag{3.3}$$

provided $[X]$ does not depend on the parameters.

The likelihood can be evaluated by Monte Carlo integration. As the integration with respect to S can be performed exactly, (3.3) can be rewritten as $\mathrm{E}_{X^*|X}[Y|X^*, \beta, \theta]$, where the marginalised density is $[Y|X^*, \beta, \theta] \sim \mathrm{N}(d(X_i^*)'\beta, \sigma^2 R(X^*; \phi) + \tau^2)$. The likelihood can therefore be estimated by drawing n_k independent samples X_k^*, each of length n, from $[X^*|X]$, evaluating the density $f_k \equiv f(y|x_k^*)$ for each sample, and then computing $n_k^{-1} \sum_k f_k$. Computation time can be reduced by using antithetic sampling (Evans & Swartz (2000)). Note that to compute $[Y|S, X^*]$ requires the covariates $d(X_i^*)$ either to be available throughout D or on a sufficiently fine grid, or to have

been measured at the (unknown) true location x_i^*. For clarity, in the results presented below we assume a constant mean μ.

Maximisation of the likelihood can be performed using the Nelder-Mead algorithm (Nelder & Mead (1965)). Note that the positional error variance γ^2 is confounded with both the nugget variance τ^2 and the range parameter ϕ: when there is positional error, there is no way of knowing whether two observations were made at the same location, and two dissimilar observations observed at nearby locations could be explained by either a high value of γ^2, a high value of τ^2, or a low value of ϕ. In many applications, γ^2 will either be known or could be estimated by a controlled experiment, so typically the problem of interest is to estimate the other parameters assuming a fixed value of γ^2.

We now illustrate the method using simulated data. We generated a realisation of a Gaussian process on 80 randomly-chosen locations from a uniform distribution on the unit square, with mean $\mu = 0$, variance $\sigma^2 = 1$, a Matérn correlation function with scale parameter $\phi = 0.3$ and order $\kappa = 2.5$, and nugget variance $\tau^2 = 0.2$. We then subjected each data location to positional error according to the model described above firstly with $\gamma^2 = 0.03^2$, then with $\gamma^2 = 0.05^2$. Figure 3.1 shows the original and repositioned data locations.

We then found maximum likelihood estimates of the unknown parameters μ, σ^2, ϕ and τ^2, assuming known values of κ and γ^2, under the following scenarios: using the original, true locations; using the incorrect, repositioned locations, making no allowance for positional error; and using the incorrect, repositioned locations, allowing for positional error.

The resulting parameter estimates are shown in Table 3.1. Computation was slow: merely computing maximum likelihood estimates took around 72 hours on a 3GHz Intel Xeon X5450 processor core with 8Gb of RAM, making reliable estimation of standard errors impractical. However, we observe that parameter estimates obtained when allowing for positional error tend to be closer to the true values than do those obtained from the repositioned data when positional error is ignored, provided that the positional error variance is small relative to the range of the correlation. Cressie & Kornak (2003) give similar results obtained from a 'pseudolikelihood' method, and conclude that σ^2 and τ^2 are the parameters whose estimates are affected most by the presence of positional error.

3.3.2 Prediction

A typical spatial prediction problem involves making deductions about a functional $T \equiv T(S(X_p^*))$ given the data (Y_i, X_i^*), where $X_p^* = (X_{p_1}^*, X_{p_2}^*, ...)$ denotes a set of prediction locations which may or may not include the data locations X^*. A general solution to the problem is the predictive distribution of T conditional on the data, $[T|Y, X^*, X_p^*]$. For the Gaussian model, this conditional distribution can sometimes, depending on the functional form of T, be derived in closed form via $[T, Y|X^*, X_p^*]$ if there is no positional error. The appendix gives more details when $T = S(x_p^*)$ for a fixed location x_p^*.

In the most general case, the prediction locations may themselves be subject to positional error, and we assume the positional error model described in Section 3.2, i.e. $X_{p_i}^*|X_{p_i} \sim \text{BVN}(X_{p_i}, \gamma^2 I_2)$. The analogous predictive distribution is

$$\begin{aligned}[T|Y, X, X_p] &= \int\int [T|Y, X, X_p, X^*, X_p^*][X^*, X_p^*|Y, X, X_p] dX^* dX_p^* \\ &= \int\int [T|Y, X^*, X_p^*][X^*|X][X_p^*|X_p] dX^* dX_p^*. \end{aligned} \tag{3.4}$$

In general, (3.4) cannot be expressed in closed form, even for simple positional error models, and will not provide a prediction that is linear in the Y_i. It therefore differs from the predictor used by Cressie & Kornak (2003). If either the data locations X^* or the prediction locations X_p^* are known precisely, the corresponding conditional distribution term is removed from (3.4), with a reduction in the dimensionality of the integral. If all data and prediction locations are known precisely, the integral reduces to the standard predictive distribution $[T|Y, X^*, X_p^*]$.

Prediction at a point

To illustrate (3.4), we concentrate on the special case in which the data locations are assumed known precisely, and the target for prediction is the value of the process S at a single point X_p^*, i.e. $T \equiv S(X_p^*)$. This simplified scenario would arise when, for example, the prediction location is a position measured imperfectly via a G.P.S. and the data locations are fixed pollution monitors. Thus (3.4) becomes

$$[S(X_p^*)|Y, X_p] = \int [S(X_p^*)|Y, X_p^*][X_p^*|X_p] dX_p^*. \tag{3.5}$$

The integral can be evaluated approximately using Gauss-Hermite quadrature with a product

rule for two dimensions. For a given value of q, this gives an estimate of the form

$$\hat{T}_q = \sum_{i=1}^{q} \sum_{j=1}^{q} f(S(X_p^*)|Y, (X_p^*)_{i,j}) w_{i,j}, \tag{3.6}$$

where f is the density function of the given conditional distribution, $(X_p^*)_{i,j}$ is a two-dimensional vector of nodes, and $w_{i,j}$ is a weight.

The distribution of $S(X_p^*)|Y, (X_p^*)_{i,j}$ is multivariate Gaussian, albeit non-linear in $(X_p^*)_{i,j}$, as conditional on $(X_p^*)_{i,j}$ the problem reduces to the standard prediction problem. The optimal choice of nodes and weights has been studied extensively (Evans & Swartz (2000)): in general, the nodes for each dimension are the roots (ϵ_i) of the qth Hermite polynomial H_q, multiplied by γ, and for each i the accompanying weight w_i is $2^{q-1}q!\sqrt{\pi}/(q^2[H_{q-1}(\epsilon_i)]^2)$. For prediction at a single point, this quadrature rule requires q^2 function evaluations. In most practical examples, values of q around 8 or 10 give sufficient accuracy.

We illustrate the effect of positional error for simulated data in Figure 3.2. We generated a realisation of a Gaussian process on 30 randomly-chosen locations from a uniform distribution on the unit square, with mean $\mu = 0$, variance $\sigma^2 = 1$, a Matérn correlation function with scale parameter $\phi = 0.1$ and order $\kappa = 2.5$, and nugget variance $\tau^2 = 0.04$. We calculated the prediction mean and prediction variance using ordinary kriging formulae for prediction locations on a grid covering the square region, assuming zero positional error. These are shown in the two left-hand panels of Figure 3.2. Next, we used (3.6) with $q = 8$ quadrature points to calculate the prediction mean and prediction variance, considering prediction locations to be subject to independent $\text{BVN}(\mathbf{0}, \gamma^2 I_2)$ positional errors with $\gamma^2 = 0.01$, and treating all parameters as known. The results are shown in the middle two panels of Figure 3.2.

For the prediction mean (i.e. the 'best' point prediction), Figure 3.2 shows a superficially similar pattern whether or not there is positional error. However, predictions allowing for positional error tend to be more conservative near local extremes in the prediction surface. For the prediction variance, the patterns are very different. When there is no prediction error, the prediction variances depend only on the locations of the data, not on the data values, with lowest variances occurring close to observation locations. This is not necessarily the case if there is prediction error, which in this example results in a relatively high prediction variance in the region around (0.2,0.6), where there are several nearby observation locations.

In the appendix, we give a theoretical justification of the differences between the predictions before and after adjusting for positional error. Here, we summarise the main results with reference to Figure 3.2.

The naive point predictor (the prediction mean (3.8), ignoring positional error) is adjusted according to the shape of the prediction surface via second (and higher) order derivatives of the assumed correlation function (see (3.10)). As shown in Figure 3.2, the bias in this naive predictor is largest near the extremes of the prediction surface. In contrast, the prediction variance is affected by first order derivatives of the correlation function (see (3.12)), and so the largest impact of positional error is seen in regions of the surface where the gradient is steepest. For example, the prediction variance increases around (0.2,0.6) after allowing for positional error.

For similar reasons, at certain locations the prediction variance may be slightly lower in the presence of positional error than it is when the positional error is zero. In Figure 3.2 this occurs in the regions close to the pixels (0.66,0.66) and (0.94,0.7), and is caused by both a shallow gradient in S at the prediction location and a relatively large kriging variance (i.e. few nearby observations). If the measured location is at a peak in the prediction variance surface, positional error in any direction would imply that the true location would have a smaller prediction variance. Figure 3.3 shows how the prediction variance at (0.66,0.66) changes as a function of the positional error variance γ^2.

The shape of the predictive distribution is shown in Figure 3.4, for four example locations. Only if $\gamma^2 = 0$ is the predictive distribution Gaussian. If $\gamma^2 \neq 0$, the distribution is typically skewed away from the mean μ: from (3.5), it is a continuous mixture of Gaussian distributions, which in general is not Gaussian. This is most clearly seen in the figure for location (0.1,0.7), which is close to a local maximum in the surface.

For comparison, we now consider the case in which the data locations are also subject to positional error. Quadrature schemes analogous to (3.6) would require evaluation of the conditional density function f at q quadrature points in each of $2(n+n_p)$ dimensions, where n is the number of data locations and n_p is the number of prediction locations. A similar quadrature scheme that uses a product quadrature rule would require a number of function evaluations that grows exponentially with q, and is therefore prohibitive computationally.

An alternative is to approximate (3.4) using Monte Carlo integration. The right-hand side of (3.4) can be written as $\mathrm{E}_{X_p^*|X_p}[\mathrm{E}_{X^*|X}[T|Y, X^*, X_p^*]]$, and the computation can be performed in practice by repeatedly simulating values successively from the zero-mean multivariate Gaussian distributions of $X^*|X$ and $X_p^*|X_p$ and then computing the required expectation by averaging over a large number of simulations. Unlike quadrature, Monte Carlo integration allows the precision of the prediction to be calculated with little extra computational effort, via the Monte Carlo variance.

The two right-hand panels of Figure 3.2 show the prediction mean and variance calculated on a fine grid of points assuming that there is zero positional error in the prediction location, and that data locations are subject to independent bivariate Normal, mean zero positional errors with variance $\gamma^2 = 0.01$. A modest 1000 evaluations were required to ensure that the half-widths of the 95% confidence intervals for the prediction mean and variance at each prediction location were less than 0.05 (the jaggedness of the contours in the figure is a result of the approximation). Computation was much quicker than in Section 3.3.1: the computation required to produce Figure 3.2 took a total of around 30 minutes, and the quadrature calculations were near-instantaneous. The results are comparable with those in the first four panels, although prediction variances tend to be lower than when positional errors of the same variance are applied to the prediction locations.

Joint Prediction

We now return to the case in which positional error affects only the prediction locations, and consider joint prediction at two locations $X_{p_1}^*$ and $X_{p_2}^*$. The required integral is the higher-dimensional analogue of (3.5):

$$[S(X_{p_1}^*), S(X_{p_2}^*)|Y, X_{p_1}, X_{p_2}] = \int\int [S(X_{p_1}^*), S(X_{p_2}^*)|Y, X_{p_1}^*, X_{p_2}^*][X_{p_1}^*|X_{p_1}][X_{p_2}^*|X_{p_2}]dX_{p_1}^* dX_{p_2}^*. \tag{3.7}$$

This is an example of the more general integral (3.4), and can be evaluated using Monte Carlo integration.

Figure 3.5 shows the result of estimating the covariance between predicted values at the point (0.5,0.5) and all other points on the unit square for the example used in Section 3.3.2. The left-hand panel shows an image of these covariances assuming no positional error ($\gamma^2 = 0$), whereas the right-hand panel assumes $\gamma^2 = 0.01$ for both prediction locations; the panels are plotted on

a common colour scale that allows both positive and negative covariances.

We conclude that the covariance between predicted values tends to be reduced when positional error is introduced, even with a positional error variance that is small relative to the range parameter of the correlation function. Figure 3.5 illustrates this with $\gamma^2 = 0.01$ and $\phi = 0.1$, and equation (3.14) in the appendix gives a theoretical explanation.

Qualititive patterns of the two panels in Figure 3.5 are similar: the covariance between predictions at different locations depends only on the relative positions of the prediction locations and data locations, and not on the data values y (see (3.14)). This result is a consequence of the assumption that positional errors at two data locations are independent. In contrast, in the presence of positional error the prediction variance depends on the data y (see (3.12)).

For a large number of prediction locations, evaluating the high-dimensional joint distribution may be infeasible, but also unnecessary in practice. A more useful alternative may be to estimate the pairwise covariances at prediction locations using (3.7), omitting pairs of locations far apart in space relative to ϕ, for which the covariance will be negligible.

Prediction Over A Trajectory

We now consider joint prediction over a trajectory, or connected path. In the context of environmental exposure assessment, this scenario mimics the changing location of an individual moving through a pollution surface, and is a special case of the general joint prediction problem. To obviate the need to formulate an explicit model for the underlying true trajectory, we assume that it can be approximated by a sequence of connected line segments, with error-prone observations taken at the intersections between adjoining segments. As demonstrated below, this assumption allows computation of the predictive distribution over a line segment to be greatly simplified.

We assume here that data locations are known precisely. Consider a line segment connecting the true locations $x^*_{p_1}$ and $x^*_{p_2}$, subject to errors ϵ_{p_1} and ϵ_{p_2} respectively. Conditional on ϵ_{p_1} and ϵ_{p_2}, the error at any intermediate location on the line segment is known precisely, and when evaluating (3.7) it is sufficient to condition only on the errors at locations where the prediction locations are measured, and to ignore any additional form of error at intermediate locations. The integrand in (3.7) can then be evaluated in practice at locations $(1-\lambda)x_{p_1} + \lambda x_{p_2}$ using a range of values of $\lambda \in [0,1]$, not merely at the endpoints ($\lambda = 0$ and $\lambda = 1$), to improve the accuracy

of the computation. Below, we investigate the effect of changing the number of intermediate locations m on a line segment at which (3.7) is evaluated.

Consider the predictive distribution of the mean over a line segment connecting two points observed after the imposition of positional error as (0.3,0.5) and (0.7,0.5) for the previously-described example, as shown in the left-hand panel of Figure 3.6. The target for prediction is $\int_{\mathcal{L}} S(x)/|\mathcal{L}|dx$, where $\mathcal{L}$ denotes the line segment connecting true locations, with length $|\mathcal{L}|$. We treated the true values of all parameters as known and estimated the integral by Monte Carlo integration, using m equally-spaced intermediate locations along the line segment at which to evaluate the integrand in (3.7). We chose values $m = 0$ (i.e. evaluate only at the two end-points) and $m = 1, 2, 5$ and 20. 10000 simulations were sufficient to make the Monte Carlo variance of all density estimates less than 0.01. The results are shown in the right-hand panel of Figure 3.6.

The extent to which the number of intermediate points used affects the predictive distribution depends on both the length of the line segment and the parameter values. To illustrate this we carried out a simulation study in which we fixed $\gamma^2 = 0.01$ and generated, at each simulation, a realisation of a Gaussian process with zero mean and unit variance on 30 points randomly and uniformly distributed on the unit square. We assumed a Matérn correlation function with $\kappa = 2$, and varied the range parameter $\phi \in \{0.1, 0.2, 0.3, 0.5\}$. Our target for prediction was the integrated exposure over the line segment (0.3,0.5) to (0.7,0.5), with the two end-points subject to positional error. We approximated the integral using $m = 0, 1, 2, 5, 10$ and 20 intermediate locations.

The left-hand panels of Figure 3.7 show the results from 200 simulations. Results are summarised as: (i) average mean square error (M.S.E.) of the predictions relative to the 'true' integrated exposure calculated by assuming zero positional error and extending the realisation of the Gaussian process to 50 points located along the line segment (0.3,0.5) to (0.7,0.5); and (ii) the prediction variance. As anticipated, we found that predictions improved as ϕ increased, and in absolute terms the improvement in mean square error was greater for smaller values of ϕ. For all cases the improvement in mean square error and prediction variance was most marked as the first two intermediate locations were added.

We then repeated the experiment holding ϕ fixed at 0.2, and varying $\gamma \in \{0, 0.05, 0.1, 0.2\}$. Results are shown in the right-hand panels of Figure 3.7, and show that predictions gradually

worsen as γ increases. Again, the greatest gain comes from using the first two intermediate locations on the line segment.

3.4 Application

We illustrate our methods using a data-set consisting of lead pollution measurements taken from moss samples collected during July 2000 in Galicia, Spain. Fernández *et al.* (2000) and Aboal *et al.* (2006) give further details, so we provide only a brief summary here. Samples were taken on an approximately regular lattice, as shown in Figure 3.8, with measurement locations recorded using a G.P.S. and plotted on a scale of 1 unit to 100 kilometres. Lead concentrations were measured in $\mu g/g$ dry weight and are analysed here on the logarithmic scale as the log-transformed values have a distribution that is approximately Normal. Figure 3.8 also shows a histogram of the data, the smoothed sample variogram and the fitted variogram resulting from the model described below.

We fitted the Gaussian model (3.1) to the log-transformed data, with a constant mean μ and treating $S(x)$ as a stationary Gaussian process characterised by a Matérn correlation function with $\kappa = 0.5$. This value of κ resulted in better fit than other values in the discrete set $\{0.5, 1, 1.5, 2, 2.5\}$. Treating the measurement locations as fixed, we found the maximum likelihood estimates of the parameters to be $\hat{\mu} = 0.724$, $\hat{\sigma}^2 = 0.192$, $\hat{\phi} = 0.206$ and $\hat{\tau}^2 = 0$. The left-hand panels of Figure 3.9 show maps of the prediction mean and variance calculated on a grid covering the whole study region. The remaining panels of the figure show corresponding graphs when a hypothetical bivariate Normal positional error with variance γ^2 is applied independently to each of the prediction location grid cells. As anticipated from (3.10), (3.11) and (3.12), the prediction variance is much more sensitive to positional error than is the prediction mean. Appreciable differences in the prediction variance map appear when γ^2 is increased to 0.02^2, while maps for the prediction mean are superficially identical for values of γ^2 less than 0.05^2.

Figure 3.10 shows a hypothetical trajectory across the study region based on eight observed positions. The observations are assumed to be taken at regular intervals in time and subject to independent bivariate Normal positional errors with variance γ^2, and the trajectory is assumed to be piecewise linear. The right-hand panel of the figure shows the predictive distribution of average (log) moss concentration over the trajectory for different values of γ^2, computed using Monte Carlo integration with three intermediate points on each line segment. The main effect of increasing γ^2 is to increase the variance of the predictive distribution, while the mean of the

distribution remains relatively unchanged.

3.5 Conclusions

In this paper we have discussed the effects of positional error in both measurement and prediction locations in spatial problems. Little has been written on this topic in the existing literature, but it is relevant to a wide range of applications, especially in the growing number of epidemiological studies that attempt to provide measures of individual environmental exposure rather than population average exposure.

A natural question to ask is in what circumstances positional errors are likely to change the conclusions of a spatial analysis. In standard regression problems, failing to take account of measurement error in a covariate can lead to parameter estimates and predictions that are biased and have incorrect, usually over-optimistic, precision (Carroll *et al.* (2006)). The same is true of positional errors for spatial data. The extent of the problem is primarily dependent on the size of the positional error and the shape of the underlying surface. For example, pollution surfaces that change sharply (i.e. have locally large gradients), as might occur in the case of air pollution close to a point source or near a busy road, are more sensitive to the effects of positional error in the prediction location. We have shown that the local gradient of the surface may have a large effect on the variance of the predictive distribution, which would not be detected by a standard analysis in which the estimate of the nugget effect were increased. Estimating this gradient (Banerjee *et al.* (2008)) may therefore be useful in more complicated problems to determine the likely impact of positional errors. If only the prediction locations are subject to positional error then this will not affect the problem of gradient estimation.

Typically the positional error variance parameter γ^2 will either be known or could be estimated via a separate experiment, for example by taking replicate measurements at a single fixed location. Gabrosek & Cressie (2002) also make this assumption. Our methods do however differ from those provided in that paper, as we maximise the likelihood directly rather than using a pseudolikelihood approach. Our solution to the prediction problem also differs, in that we do not assume that the minimum mean square error predictor is linear in the data values; indeed, our results demonstrate that it is not. The predictive distribution at a point is non-Gaussian, and asymmetric, in the presence of positional error, even if S is a Gaussian process.

Throughout this paper, we have used a simple bivariate Normal model for positional error. This

assumption enables the required integrals with respect to the positional error distribution to be computed either by quadrature, for which computation is near-instantaneous, or by Monte Carlo integration, for which computation is slower, but not prohibitively so for data-sets of moderate size. In assuming this model we have purposely avoided the need to specify a marginal distribution for either the true location X^* or the observed location X. For the Berkson model $[X^*|X]$ this specification is not needed, while for the classical model $[X|X^*]$ a specification of this marginal distribution would be required in order to perform the integration in the computation of the likelihood.

It may be advantageous in future work to consider further positional error models. Cressie & Kornak (2003) used a uniform distribution, resulting from rounding off coordinates to the nearest grid point. We have also examined this model for our prediction task (results not shown), with very similar results to those we obtained from a bivariate Normal model with a matching positional error variance, an observation also made by Gabrosek & Cressie (2002). Thus the variance of the positional error may be of more importance than its exact distributional form. Other positional error models that warrant further investigation include those that allow correlation between measurements made sequentially, and those that take account of covariates.

Appendix

In this section we present additional results relating to Section 3.3.2 when positional error affects only the prediction locations.

Consider prediction of S at a single location x_p^* under the Gaussian model (3.1) with constant mean μ and n data points at locations $x_1, ..., x_n$. Let $r_i = \rho(||x_p - x_i||)$, where ρ is a given correlation function, R be a matrix with (i,j)th element $R_{ij} = \rho(||x_i - x_j||)$, and $Q = R^{-1}$. A standard result (e.g. Diggle & Ribeiro Jr. (2007), Chapter 6) states that the ordinary kriging prediction mean and variance, in the absence of positional error, are

$$\mu_x = \mu + r'Q(y - \mu) \equiv \mu + \sum_{i=1}^{n} a_i r_i, \tag{3.8}$$

where $a_i = \sum_{j=1}^{n} Q_{ij}(y_j - \mu)$, which does not depend on the prediction location, and

$$\sigma_x^2 = \sigma^2(1 - r'Qr) \equiv \sigma^2 \left(1 - \sum_{i,j=1}^{n} r_i Q_{ij} r_j\right). \tag{3.9}$$

The predictive distribution $[S(x_p^*)|Y] \sim \mathrm{N}(\mu_x, \sigma_x^2)$, so μ_x provides the 'best' point prediction for $S(x_p^*)$.

We now investigate how these results change when x_p^* is subject to positional error under the independent Gaussian model described in Section 3.2. Our aim here is to approximate the first two moments of (3.5) using a Taylor expansion of the correlation function r. For clarity, we use the notation $\mathbf{x} = (x, y)$ for the two components of a general prediction location (removing the subscript p), with error $\epsilon = (\epsilon_x, \epsilon_y)$. The Taylor expansion for r is

$$r(\mathbf{x}+\epsilon) = r(\mathbf{x}) + \epsilon_x \frac{\partial r(\mathbf{x})}{\partial x} + \epsilon_y \frac{\partial r(\mathbf{x})}{\partial y} + \frac{1}{2}\epsilon_x^2 \frac{\partial^2 r(\mathbf{x})}{\partial x^2} + \frac{1}{2}\epsilon_y^2 \frac{\partial^2 r(\mathbf{x})}{\partial y^2} + \epsilon_x \epsilon_y \frac{\partial^2 r(\mathbf{x})}{\partial x \partial y} + \ldots$$

To simplify notation, we write a_x for $\sum_{i=1}^n a_i \, \partial r_i/\partial x$, a_{xy} for $\sum_{i=1}^n a_i \, \partial^2 r_i/\partial x \partial y$, etc, and $Q_{0,x}$ for $\sum_{i,j=1}^n r_i Q_{ij} \, \partial r_j/\partial x$, $Q_{x,xy}$ for $\sum_{i,j=1}^n \partial r_i/\partial x \, Q_{ij} \, \partial^2 r_j/\partial x \partial y$, etc, all evaluated at $\mathbf{x}$. In this notation, $\mu_{\mathbf{x}} = \mu + a_0$ and $\sigma_{\mathbf{x}}^2 = \sigma^2(1 - Q_{0,0})$.

In the derivation below, it is convenient to use the alternative form $X^* = X + \epsilon$ of the positional error model discussed in Section 3.2. The Taylor expansion corresponding to (3.8) is

$$\begin{aligned}
\mu_{\mathbf{x}+\epsilon} \;=\; & \mu_{\mathbf{x}} + \epsilon_x a_x + \epsilon_y a_y + \frac{1}{2}\epsilon_x^2 a_{xx} + \epsilon_x \epsilon_y a_{xy} + \frac{1}{2}\epsilon_y a_{yy} + \\
& \frac{1}{6}\epsilon_x^3 a_{xxx} + \frac{1}{2}\epsilon_x^2 \epsilon_y a_{xxy} + \frac{1}{2}\epsilon_x \epsilon_y^2 a_{xyy} + \frac{1}{6}\epsilon_y^3 a_{yyy} + \\
& \frac{1}{24}\epsilon_x^4 a_{xxxx} + \frac{1}{6}\epsilon_x^3 \epsilon_y a_{xxxy} + \frac{1}{4}\epsilon_x^2 \epsilon_y^2 a_{xxyy} + \frac{1}{6}\epsilon_x \epsilon_y^3 a_{xyyy} + \frac{1}{24}\epsilon_y^4 a_{yyyy} + \ldots
\end{aligned}$$

For any positive integer m, $\mathrm{E}(\epsilon^{2m}) = ((2m-1)(2m-3)\ldots 3.1)\gamma^{2m}$ and $\mathrm{E}(\epsilon^{2m-1}) = 0$, and ϵ_x and ϵ_y are assumed independent, so the mean of the predictive distribution is

$$\begin{aligned}
\mathrm{E}[S(\mathbf{x}+\epsilon)] &= \mathrm{E}_\epsilon[\mu_{\mathbf{x}+\epsilon}] \\
&= \mu_{\mathbf{x}} + \frac{1}{2}\gamma^2(a_{xx} + a_{yy}) + \frac{1}{8}\gamma^4(a_{xxxx} + 2a_{xxyy} + a_{yyyy}) + O(\gamma^6). \quad (3.10)
\end{aligned}$$

Similarly, the variance of the predictive distribution is

$$\mathrm{Var}[S(\mathbf{x}+\epsilon)] = \mathrm{E}_\epsilon[\sigma_{\mathbf{x}+\epsilon}^2] + \mathrm{Var}_\epsilon[\mu_{\mathbf{x}+\epsilon}],$$

where it can be shown that

$$\mathrm{E}_\epsilon[\sigma_{\mathbf{x}+\epsilon}^2] = \sigma_{\mathbf{x}}^2 - \sigma^2\gamma^2(Q_{x,x} + Q_{y,y} + Q_{0,xx} + Q_{0,yy}) + O(\gamma^4) \quad (3.11)$$

and

$$\mathrm{Var}_\epsilon[\mu_{\mathbf{x}-\epsilon}] = \gamma^2(a_x^2 + a_y^2) + O(\gamma^4). \tag{3.12}$$

These formulae motivate the discussion in the "prediction at a point" subsection in Section 3.3.2: the prediction mean depends on second order (and higher) derivatives of ρ, while the prediction variance depends on first order (and higher) derivatives.

We now consider the case of two prediction locations $x^*_{p_1}$ and $x^*_{p_2}$. Let r be the $(n \times 2)$ matrix with $r_{ij} = \rho(\|x_{p_j} - x_i\|)$ for $j = 1, 2$, and $c = \rho(|x_{p_1} - x_{p_2}|)$. In the case of zero positional error, standard results give

$$\mathrm{Cov}[S(x_{p_1}), S(x_{p_2})] = \sigma^2(c - (r'Qr)_{1,2}). \tag{3.13}$$

Next we present corresponding results if the prediction locations are subject to independent positional errors. Using similar notation to the above, removing the subscript p, let $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ denote the prediction locations. Write the second partial derivatives of c as $c_{x_1x_1} = \partial^2 c/\partial x_1^2$, $c_{y_1y_1} = \partial^2 c/\partial y_1^2$, etc. Also, write $Q_{x_1,0}$ for $\sum_{i,j=1}^n \partial(r_{1i})/\partial x_1 Q_{ij} r_{2j}$, Q_{0,x_2} for $\sum_{i,j=1}^n r_{1i} Q_{ij} \partial(r_{2j})/\partial x_2$, $Q_{y_1,0}$ for $\sum_{i,j=1}^n \partial(r_{1i})/\partial y_1 Q_{ij} r_{2j}$ etc. Then it can be shown that

$$\begin{aligned}\mathrm{Cov}[S(\mathbf{x}_1 + \epsilon_1), S(\mathbf{x}_2 + \epsilon_2)] &= \sigma^2\{c - Q_{0,0} + \frac{1}{2}\gamma^2(c_{x_1x_1} + c_{y_1y_1} + c_{x_2x_2} + c_{y_2y_2} - \\ &\quad Q_{x_1x_1,0} + Q_{y_1y_1,0} + Q_{0,x_2x_2} + Q_{0,y_2y_2})\} + O(\gamma^4). \end{aligned} \tag{3.14}$$

This formula motivates the discussion in the "joint prediction" subsection in Section 3.3.2: the covariance between predictions at two points tends to be reduced when positional error is present. In particular neither (3.13) nor (3.14) depends on the data y.

Acknowledgements

T.R.F. was supported by a Doctoral Training Account studentship and P.J.D. by a Senior Fellowship from the Engineering and Physical Sciences Research Council (E.P.S.R.C.). The Galicia data were made available to us by Raquel Menezes, University of Minho, Portugal. We thank Duncan Whyatt (Department of Geography, Lancaster University) and Chris Sherlock (Department of Mathematics and Statistics, Lancaster University) for helpful discussions, and an associate editor for suggestions that substantially improved the clarity of the paper.

	γ	μ	σ^2	ϕ	τ^2
True parameters	-	0	1	0.3	0.2
Scenario 1	-	0.51	0.98	0.26	0.18
Scenario 2a	0.03	0.41	0.80	0.20	0.15
Scenario 2b	0.03	0.47	0.93	0.24	0.17
Scenario 3a	0.05	0.69	1.60	0.43	0.23
Scenario 3b	0.05	0.68	1.40	0.43	0.25

Table 3.1: True parameters and parameter estimates for the data shown in Figure 3.1 under the following scenarios: 1. Using the original, true locations; 2a. Using the incorrect, repositioned locations with $\gamma^2 = 0.03^2$, making no allowance for positional error; 2b. As 2a, allowing for positional error; 3a & 3b. as 2a & 2b, with $\gamma^2 = 0.05^2$.

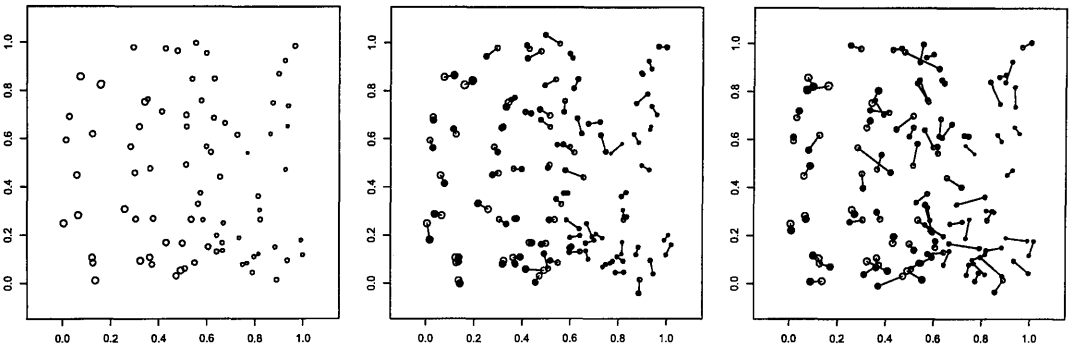


Figure 3.1: A realisation of a Gaussian process at 80 locations (left), with positional error applied to data locations (variance $\gamma^2 = 0.03^2$ (middle) and $\gamma^2 = 0.05^2$ (right)). The size of the point indicates the magnitude of the observation.

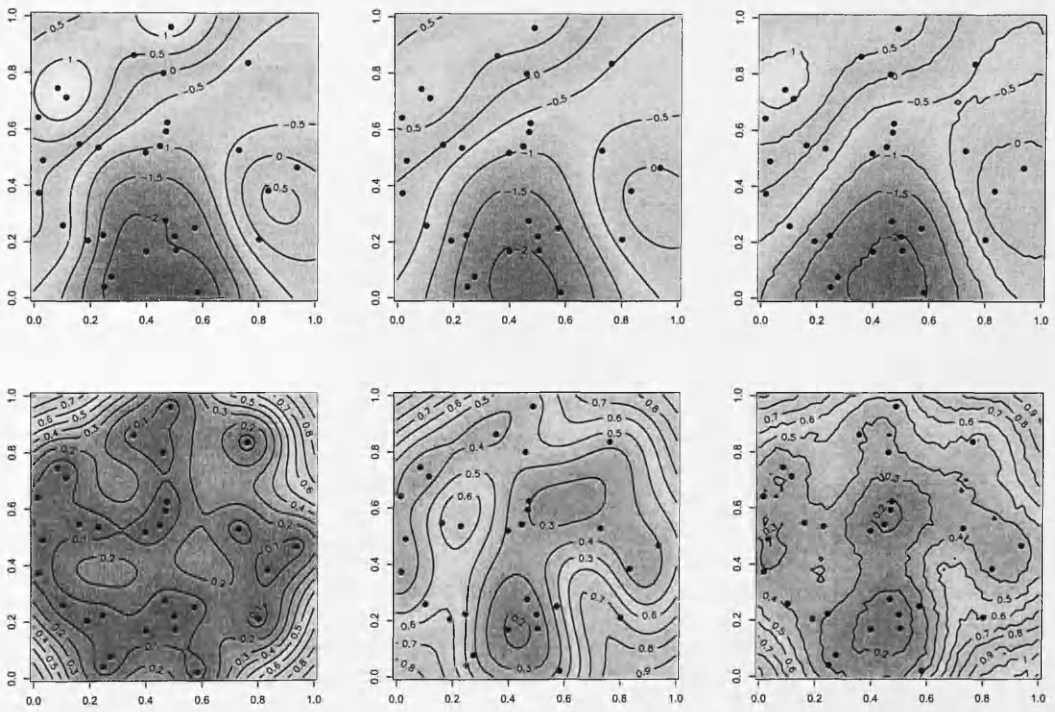


Figure 3.2: Prediction means (top row) and variances (bottom row) calculated assuming no positional error (left column), positional error in prediction locations (middle column) and positional error in data locations (right column). Data points are shown at observed locations.

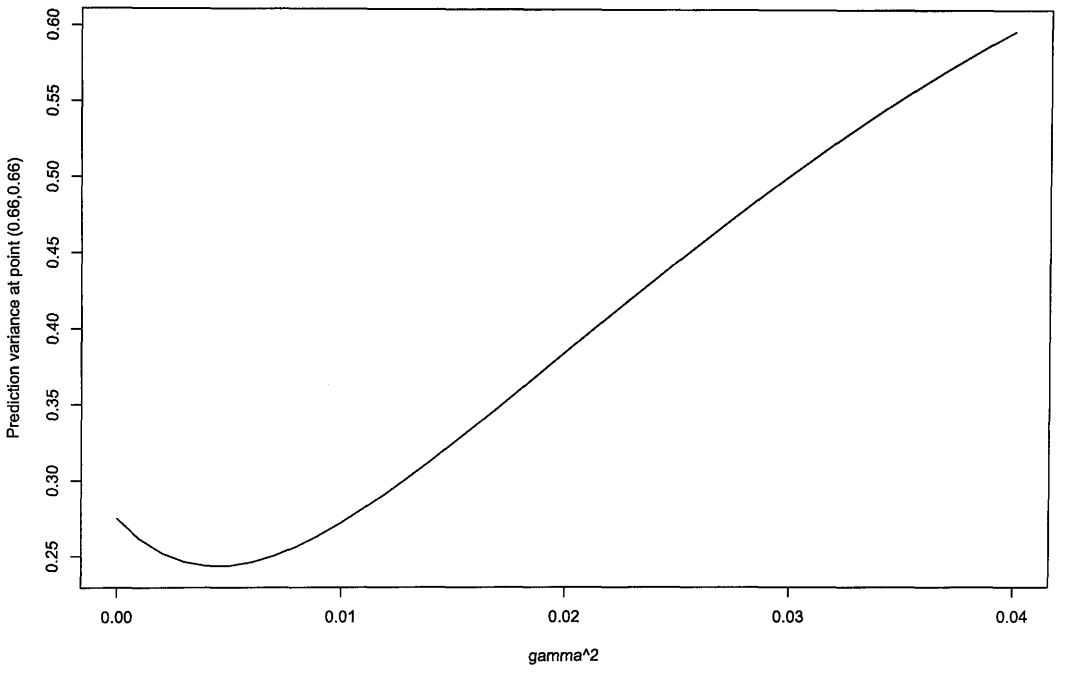


Figure 3.3: Relationship between the prediction variance at a single point and the positional error variance of the prediction location.

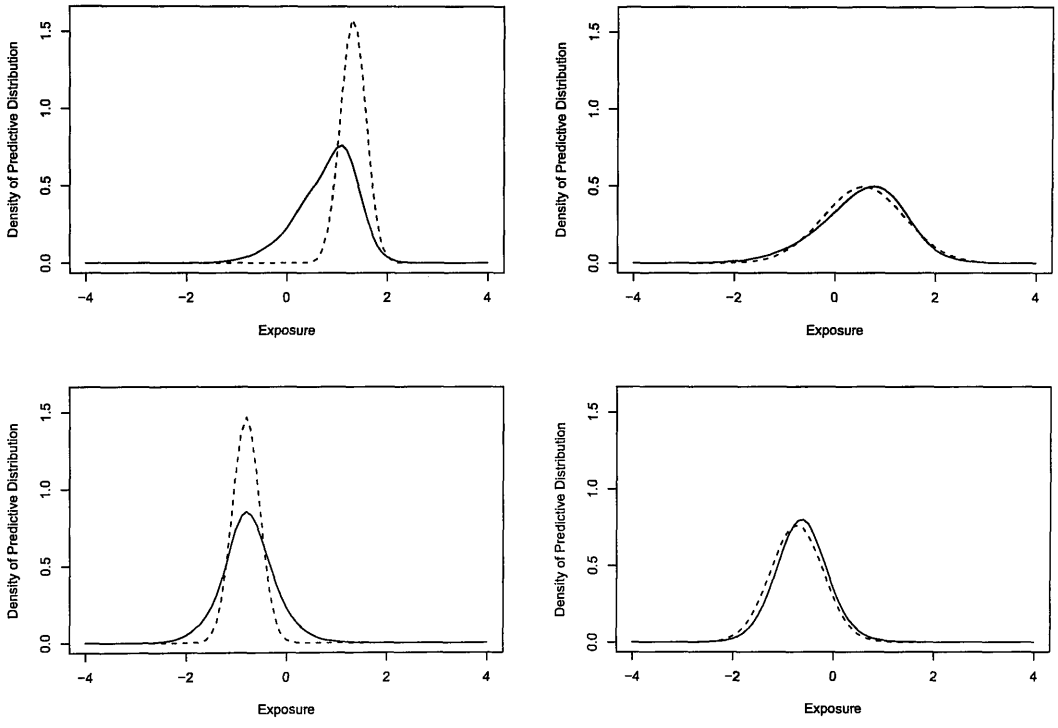


Figure 3.4: Predictive distributions at four locations assuming no positional error (dashed) and assuming positional error in the prediction location with variance $\gamma^2 = 0.01$ (solid). The four locations are (0.1,0.7) (top left), (0.1,0.95) (top right), (0.5,0.6) (bottom left) and (0.66,0.66) (bottom right).

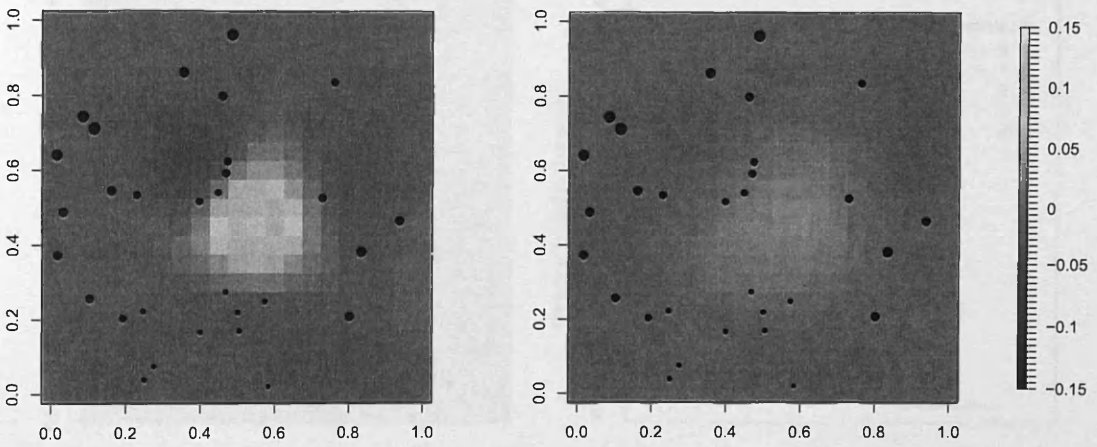


Figure 3.5: Covariance between predictions at the point (0.5,0.5) and other locations, where each prediction location is subject to no positional error (left) and positional error with variance $\gamma^2 = 0.01$ (right).

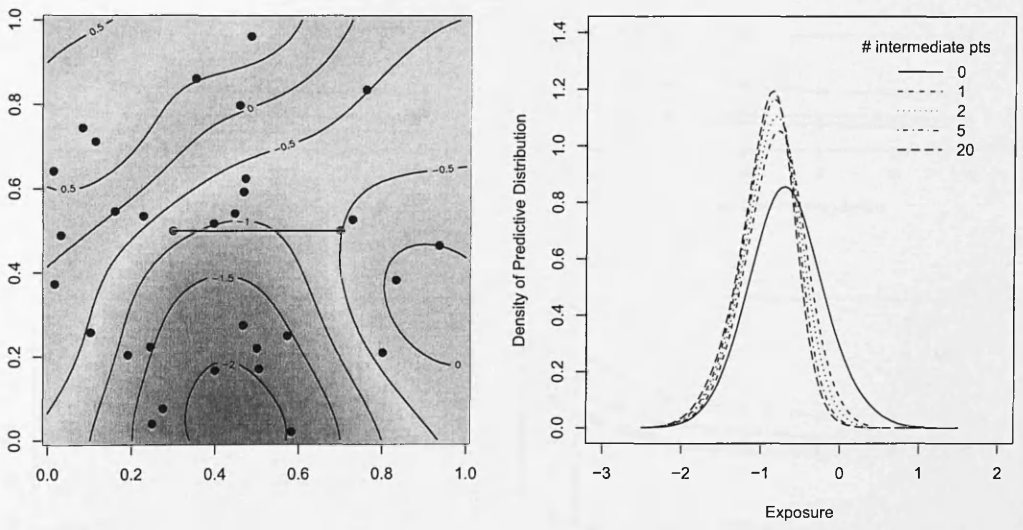


Figure 3.6: Approximation to the predictive distribution of $\int_L S(x)/|L|dx$ over the line segment L (left) as the number of intermediate locations varies (right).

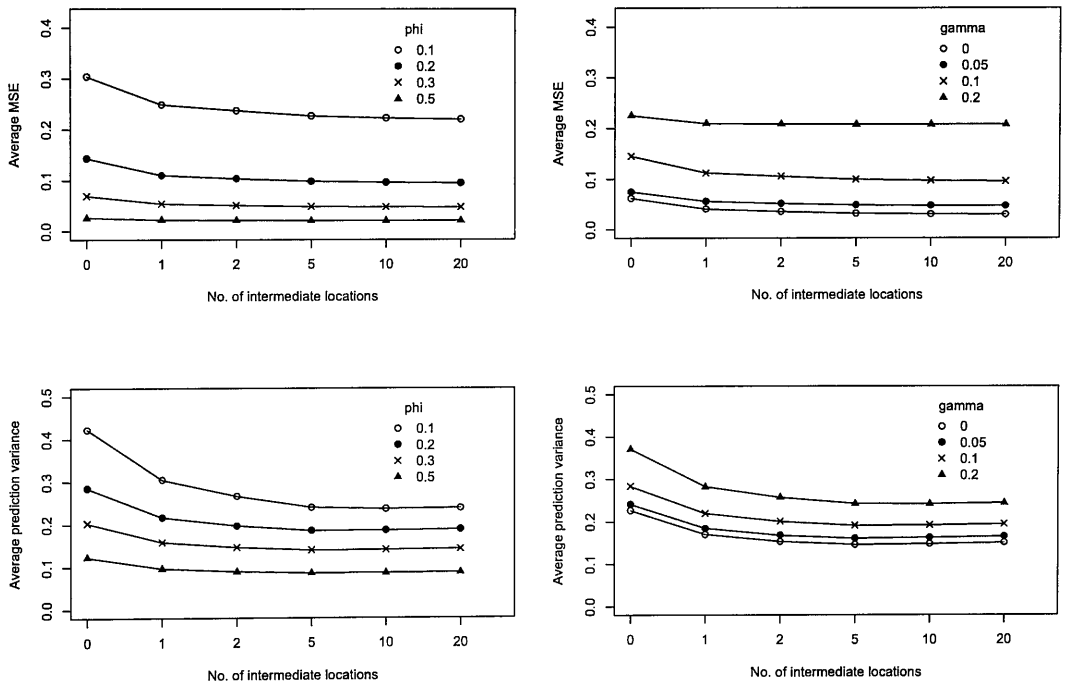


Figure 3.7: Average mean square error and prediction variance for the line integral shown in Figure 3.6 according to ϕ, γ and the number of intermediate locations used to approximate the integral.

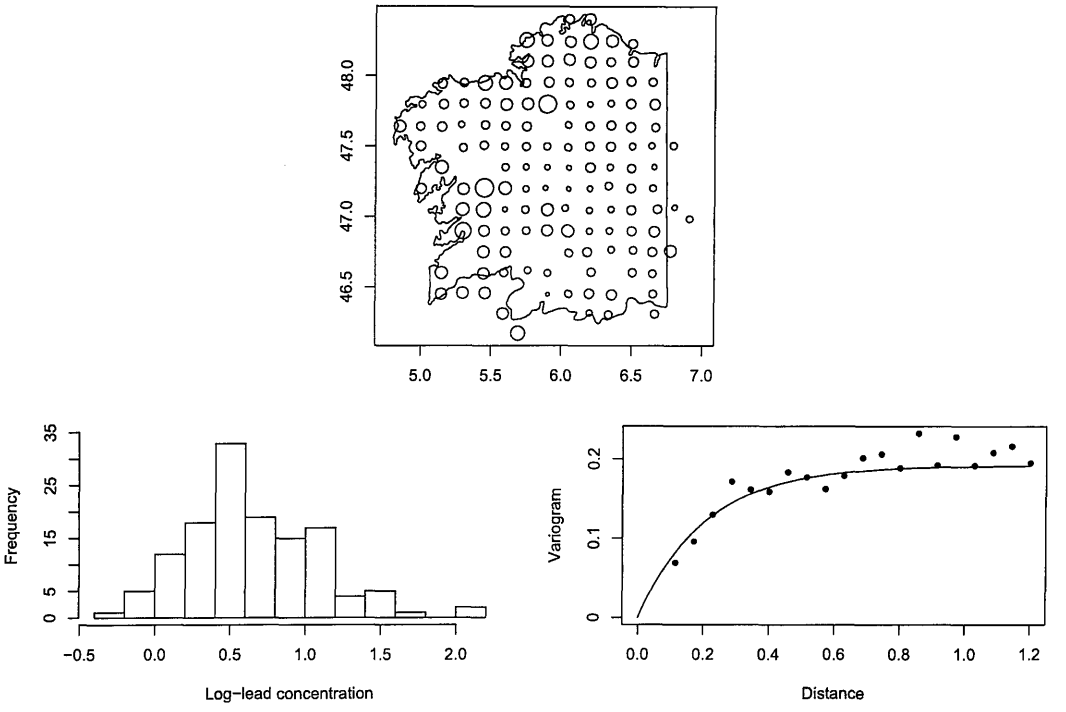


Figure 3.8: Map of lead concentrations in Galicia (top), histogram of data values (bottom left), and sample and fitted variogram (bottom right). Lead concentrations were measured in μg/g dry weight, and distances are plotted on a scale of 1 unit to 100 kilometres.

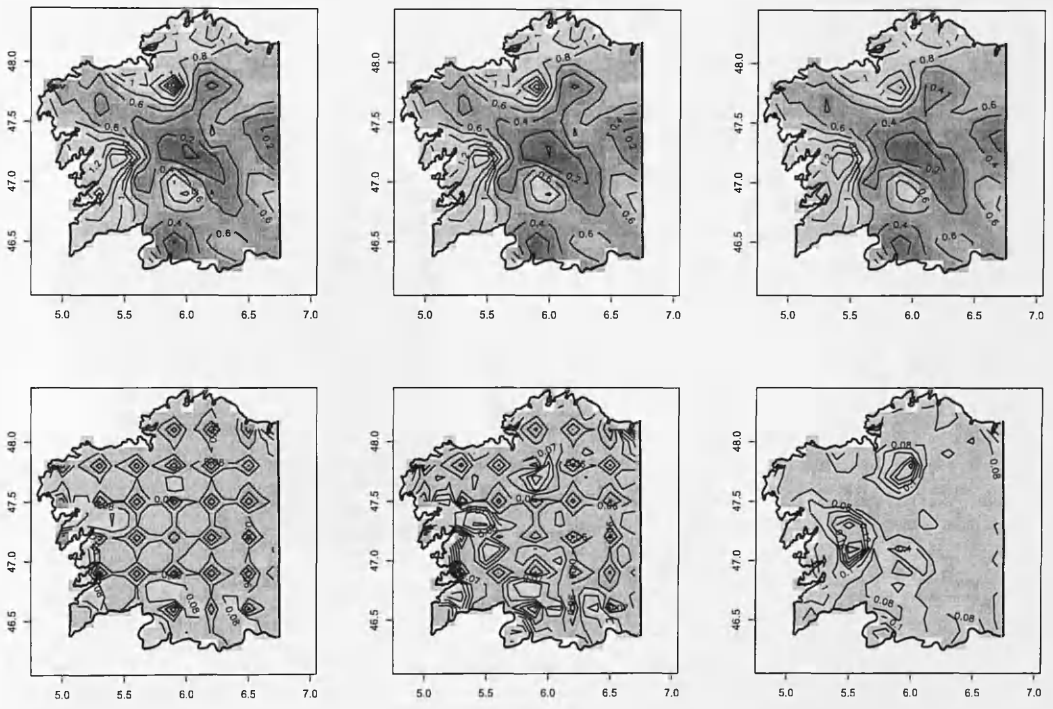


Figure 3.9: For Galicia data, prediction means (top row) and variances (bottom row) calculated assuming no positional error (left column) and positional error in prediction locations with different values of positional error variance ($\gamma^2 = 0.02^2$ (centre column); $\gamma^2 = 0.05^2$ (right column)). The three upper panels are shown on a common colour scale, as are the three lower panels.

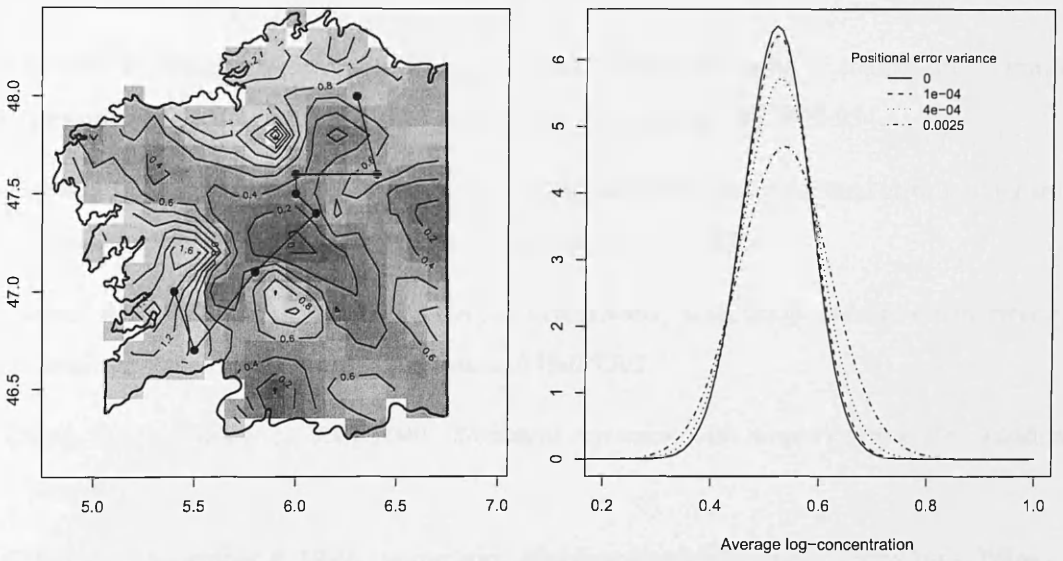


Figure 3.10: For Galicia data, example of a piecewise linear trajectory constructed using eight prediction locations; predictive distribution of mean exposure along the trajectory, assuming observations are made at equal intervals in time and assuming independent positional errors at prediction locations with different values of positional error variance γ^2.

References

Aboal, J.R., Real, C., Fernández, J.A., & Carballeira, A. 2006. Mapping the results of extensive surveys: the case of atmospheric biomonitoring and terrestrial mosses. *Science of the Total Environment*, **356**, 256–274.

Banerjee, S., Gelfand, A.E., & Sirmans, C.F. 2003. Directional rates of change under spatial process models. *Journal of the American Statistical Association*, **98**, 946–954.

Barber, J.J., Gelfand, A.E., & Silander Jr., J.A. 2006. Modelling map positional error to infer true feature location. *The Canadian Journal of Statistics*, **34**, 659–676.

Carroll, R.J., Ruppert, D., Stefanski, L.A., & Crainiceanu, C.M. 2006. *Measurement error in nonlinear models*. Boca Raton, Chapman and Hall/CRC.

Cheng, C.-L., & Van Ness, J.W. 1999. *Statistical regression with measurement error*. London, Arnold.

Chilés, J.-P., & Delfiner, P. 1999. *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. 1988. Spatial prediction and ordinary kriging. *Mathematical Geology*, **20**, 405–421.

Cressie, N. 1991. *Statistics for Spatial Data*. New York: Wiley.

Cressie, N., & Kornak, J. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, **18**, 436–456.

Diggle, P.J., & Ribeiro Jr., P.J. 2007. *Model-Based Geostatistics*. New York: Springer.

Evans, M.D., & Swartz, T. 2000. *Approximating integrals via Monte Carlo and deterministic methods*. Oxford, Oxford University Press.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlton, M., & Pless-Mulloli, T. 2008. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics*, **19**, 549–566.

Fernández, J.A., Rey, A., & Carballeira, A. 2000. An extended study of heavy metal deposition in Galicia (NW Spain) based on moss analysis. *Science of the Total Environment*, **254**, 31–44.

Gabrosek, J., & Cressie, N. 2002. The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, **34**, 262–285.

Gryparis, A., Paciorek, C.J., Zeka, A., Schwartz, J., & Coull, B.A. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, **10**, 258–274.

Jaksa, M.B., Brooker, P.I., & Kaggwa, W.S. 1997. Inaccuracies associated with estimating random measurement errors. *Journal of Geotechnical and Geoenvironmental Engineering*, **123**, 393–401.

Kiiveri, H.T. 1997. Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, **11**, 33–52.

Madsen, L., Ruppert, D., & Altman, N.S. 2008. Regression with spatially misaligned data. *Environmetrics*, **19**, 453–467.

Nelder, J.A., & Mead, R. 1965. A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Persson, A.H., Bondesson, L., & Börlin, N. 2006. Estimation of polygons and areas. *Scandinavian Journal of Statistics*, **33**, 541–559.

Pope III, C.A., & Dockery, D.W. 2006. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association*, **56**, 709–742.

Zandbergen, P.A., & Green, J.W. 2007. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives*, **115**, 1363–1370.

Zimmerman, D.L. 2008. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*, **64**, 262–270.

Chapter 4

Paper 3: Bivariate Geostatistical Modelling: A Review and an Application to Spatial Variation in Radon Concentrations

---

T.R. Fanshawe, P.J. Diggle

School of Health and Medicine, Lancaster University, UK

Summary

We present a review of multivariate geostatistical models, focusing on the bivariate case. We compare in detail three approaches, the linear model of coregionalisation, the common component model and the kernel convolution approach, and discuss similarities between them. In particular, we show how kernel convolution can be used to approximate the common component model, and demonstrate the method using a data-set of calcium and magnesium concentrations in soil samples. We then apply the common component model to a study of domestic radon concentrations in the city of Winnipeg, Canada, in which exposure was measured at two sites (bedroom and basement) in each residential location. Our analysis demonstrates that in this study the correlation between the two sites within each house dominates the short-range spatial correlation typical of the distribution of radon.

Key words: Common component model; Linear model of coregionalisation; Kernel convolution

4.1 Introduction

In this paper, and following Cressie (1993), we use the term geostatistics to mean the branch of spatial statistics that is concerned with analysing a spatially continuous phenomenon using data collected at a discrete set of spatial locations. In recent years, there have been many developments in modelling multivariate geostatistical data. Computational advances have enabled data analysts to fit models of increasing complexity, but often little attention is paid to the relative benefits and drawbacks of using each class of models. In this paper we summarise and compare the available modelling strategies, and illustrate them with examples. We focus on the bivariate case, in which the problem is to model jointly the distribution of two quantities that vary over space.

More formally, bivariate geostatistical data consist of pairs of measurements (Y_{ij}, x_{ij}), for $i = 1, ..., n_j$ and $j = 1, 2$. The locations x_{ij} are usually considered to be specified by the study design, rather than as the outcome of a spatial stochastic process. Often, the observations $Y_{ij} \equiv Y_j(x_{ij})$ can be modelled, possibly after transformation, as a realisation of a multivariate Gaussian distribution with a spatially structured covariance matrix, and we make this assumption throughout the paper.

A standard Gaussian model for univariate geostatistical data models the observations Y conditional on an underlying, unobserved, continuous spatial process S as

$$Y(x_i) = \beta^T d(x_i) + \sigma S(x_i) + Z_i : i = 1, ..., n \tag{4.1}$$

where $d(x)$ is a vector of spatially-referenced covariates, β is a vector of regression parameters, S is a Gaussian process with zero mean, unit variance and covariance function $\gamma(\cdot)$, and the Z_i are independent $\mathrm{N}(0, \tau^2)$ random variables (i.e. a white noise process with variance τ^2). Often, S is treated as stationary, meaning that $\gamma(u, v) = \mathrm{Cov}\{S(u), S(v)\}$ is a function of $\|u - v\|$, the distance between u and v.

In the bivariate extension of (4.1), $S = (S_1, S_2)$ is a bivariate process with zero mean, auto-covariance functions γ_{11} and γ_{22} and cross-covariance functions γ_{12} and γ_{21}, defined by $\gamma_{jk}(u, v) = \mathrm{Cov}(Y_j(u), Y_k(v))$. Here we make the simplifying assumption that S_1 and S_2 are stationary and also 'jointly stationary', so that $\gamma_{11}(u, v)$, $\gamma_{22}(u, v)$ and $\gamma_{12}(u, v)$ are all functions of $\|u - v\|$. This implies that $\gamma_{12} \equiv \gamma_{21}$. We model the measurement errors as a bivariate white noise process,

$(Z_1(u), Z_2(u))$, but allow its two components to be correlated, hence $\mathrm{Cov}(Z_1(u), Z_2(u))$ may be non-zero, but for any u and $v \neq u$, $\mathrm{Cov}(Z_1(u), Z_2(v)) = 0$.

Although adjustments for spatially varying covariates are important in practice, for our current purposes they are a distraction. In what follows, we therefore replace the regression model $\beta^T d(x_i)$ in (4.1) by a constant mean, and consider only bivariate models of the form

$$Y_j(x_i) = \mu_j + \sigma_j S(x_{ij}) + Z_{ij} : i = 1, ..., n_j; j = 1, 2, \tag{4.2}$$

where $Z_{ij} \equiv Z_j(x_{ij})$ denotes the measurement error term for the jth component ($j = 1, 2$) at location x_{ij}. Note that we do not require both components to be measured at the same set of locations but, as described above, at any location where both components are measured we allow the corresponding two measurement errors to be correlated. This is important in applications where the Z_{ij} are intended to capture pragmatically both pure measurement error and spatial correlation at a scale smaller than the shortest distance between any two measurememnt locations.

A key constraint on the choice of a covariance structure for the process S relates to positive definiteness: given any two sets of locations $x_{1,1}, ..., x_{1,n_1}$ and $x_{2,1}, ..., x_{2,n_2}$, the symmetric $(n_1+n_2) \times (n_1+n_2)$ block matrix with diagonal blocks given by $\gamma_{11}(x_{1,i}, x_{1,j})$, $i, j = 1, ..., n_1$ and $\gamma_{22}(x_{2,i}, x_{2,j})$, $i, j = 1, ..., n_2$, and off-diagonal block $\gamma_{12}(x_{1,i}, x_{2,j})$, $i = 1, ..., n_1$, $j = 1, ..., n_2$, must be positive definite.

The objectives of a particular analysis may include the estimation of parameters of interest, and will usually include prediction of S at new or already-sampled locations. Within the Gaussian modelling framework, both the likelihood function and the predictive distribution $[S^*|Y_1(\cdot), Y_2(\cdot)]$, where S^* denotes the values of S at the set of locations for which predictions are required, are conceptually straightforward, but may be computationally demanding.

The remainder of the paper is structured as follows. In Section 4.2 we first discuss what might constitute a desirable set of characteristics for a generally applicable class of models. We then review and compare three widely used approaches, namely the linear model of coregionalisation (LCM), the common component model (CCM) and the kernel convolution approach (KC). In Section 4.3 we consider likelihood inference, parameter estimation and prediction. In Section 4.4 we report an illustrative analysis of a data-set consisting of calcium and magnesium concen-

trations in soil samples, and an application to a study of domestic radon concentrations in the city of Winnipeg, Canada. In Section 4.5 we give a general discussion and conclusions.

In traditional geostatistics, the univariate and multivariate spatial prediction methods known as kriging and co-kriging, respectively, are presented without reference to any specific stochastic model; rather, they are justified by their delivery of best (in mean square sense) linear unbiased prediction (Chilès & Delfiner (1999)). We take the view, following Diggle & Ribeiro Jr. (2007), that linear prediction is most natural under a Gaussian model, and make this choice of model an explicit assumption at the outset. This colours our approach to inference, but does not affect our comparative discussion of different classes of model.

4.2 Review of Bivariate Models

In the traditional geostatistics literature bivariate spatial prediction is called co-kriging (Ver Hoef & Cressie (1993); Haas (1995)). In this technique, a two-dimensional point prediction $S(x) = (S_1(x), S_2(x))$ at location x is constructed as the linear combination of the data Y that minimises the mean squared prediction error. Equivalently, under our assumed stationary Gaussian model the co-kriging predictor is the mean of the Gaussian conditional distribution $[(S_1(x), S_2(x))|Y, \theta]$, calculated after 'plugging-in' estimates of the parameters θ (Diggle & Ribeiro Jr. (2007), Chapter 6). Thus, co-kriging directly relies upon specifying a suitable model for the covariance structure of S.

In some applications, the specific context may suggest an equally specific model. More commonly, the model is empirical in nature and is used simply as a means to the end of spatial prediction. We consider the following properties to be desirable for a generally applicable class of such models.

Firstly, the model should have a spatially continuous interpretation, thereby allowing prediction at unsampled locations. Note, however, that once such a model has been specified, it would typically be implemented using a finely spaced grid as an approximation to the underlying spatial continuum, in which case a spatially discrete approximation as derived by Rue & Tjelmeland (2002) provides a computationally efficient means of implementation.

Secondly, the model and any associated statistical methods should allow for both common and misaligned locations at which measurements are made. Any model that is formulated as a spa-

tially continuous Gaussian process, in conjunction with likelihood-based methods for parameter estimation, automatically meets this requirement. One context in which misaligned locations would arise naturally is when scientific interest is focused on a process $S_1(\cdot)$ which is expensive to measure but is correlated with a second process $S_2(\cdot)$ which is cheap to measure. The combination of a small sample of measurements of $S_1(\cdot)$ and a large sample of measurements of $S_2(\cdot)$ might then be more cost-effective than either sampling $S_1(\cdot)$ alone, or sampling both $S_1(\cdot)$ and $S_2(\cdot)$ at a common set of locations.

Thirdly, the model should not generally depend on the labelling of the two components. An exception would be when there is a natural direction of dependence between the two components, analogous to the asymmetric formulation of a classical regression model for an explanatory variable and a response. For example, in modelling the relationship between air pollution and biodiversity, modelling biodiversity conditional on pollution would be more natural than the converse.

Fourthly, if less tangibly, the model needs to balance flexibility against parsimony. On the one hand, the choice of univariate models for the covariance structure of each component of S should not over-constrain the model for the cross-covariance between the two. But the model should also not be so flexible as to be unidentifiable, unless the scientific context justifies strong prior assumptions, tantamount to a reduction in the number of unknown parameters.

Some methods that specify a model directly as a particular multivariate Gaussian distribution fail to meet the first of the above considerations. For example, a method described by Oliver (2003) treats response vectors $Y_j = \{Y_j(x_{ij}) : j = 1, ...n_i\}$ as realisations of a multivariate Gaussian distribution by specifying

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ L_2\rho & L_2\sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}. \tag{4.3}$$

Here, the elements of Z_1 and Z_2 are independent zero-mean Normal random variables with variances τ_1^2 and τ_2^2 respectively, the elements of S_1 and S_2 are independent standard Normal random variables, and L_1 and L_2 are the generalised square roots of the covariance matrices of the multivariate Normal distributions of $Y_1 - Z_1$ and $Y_2 - Z_2$, computed using their LU factorisations, whilst the scalar ρ, in conjunction with the variance components τ_1^2 and τ_2^2, determines the cross-correlation between elements of Y_1 and Y_2. This representation cannot deal with data in which observations for the two components are not co-located, which also makes prediction

at new locations difficult. In addition, the cross-covariance is determined entirely by the two marginal covariance structures, and often does not have a standard functional form. This limits flexibility.

A conditional, or hierarchical, approach is used by some authors to represent physical processes for which an asymmetrical formulation is natural. For example, Royle & Berliner (1999) model vectors (Y_1, Y_2) according to

$$\begin{aligned} Y_1|Y_2 &\sim AY_2 + Z_1 \\ Y_2 &\sim Z_2 \end{aligned}$$

where Z_1 and Z_2 are independent multivariate Normal random variables and A is a matrix of regression coefficients parameterised by a vector of relatively low dimension. In its simplest form, the conditional part of this model is a regression of Y_1 on Y_2. In a spatially continuous setting, this idea motivates the model for top-soil geochemistry used by Calder *et al.* (2009). They represent the model for Y_1 and Y_2 and latent Gaussian processes S_1 and S_2, given parameters θ, by specifying the sequence of conditional distributions $[Y_1|S_1, \theta]$, $[Y_2|S_2, \theta]$, $[S_1|S_2, \theta]$ and $[S_2|\theta]$.

4.2.1 Linear Model of Coregionalisation

Several modelling strategies use a 'constructive' approach of forming a new covariance function as a sum of covariance functions. The most widely used example of this approach is the linear model of coregionalisation, or LCM (e.g. Goulard & Voltz (1992); D'Agostino *et al.* (1993); Goovaerts (1994); Wackernagel (1995); Chilès & Delfiner (1999); Schmidt & Gelfand (2003); Marchant & Lark (2007)). In this model a bivariate process $S^+(x) = (S_1^+(x), S_2^+(x))$ is constructed as a sum of matrix products, each of the same basic form

$$S^+(x) = \sum_{k=1}^{p} A_k S^{(k)}(x),$$

where each $S^{(k)}(x) = (S_1^{(k)}(x), S_2^{(k)}(x))$ consists of two independent Gaussian processes with zero mean and unit variance. Li *et al.* (2008) describes a sequential test for determining the optimal value of p. One appeal of the LCM is that the constructive approach guarantees validity. However, there is no guarantee that all bivariate processes can be constructed in this way (Ver Hoef & Barry (1998)).

The special case of the LCM in which $p = 1$ (the 'Single Component Model', or SCM, Mardia & Goodall (1993), Gelfand *et al.* (2003) and Schmidt & Gelfand (2003)) is specified by

$$\begin{pmatrix} S_1^+(x) \\ S_2^+(x) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} S_1(x) \\ S_2(x) \end{pmatrix}, \tag{4.4}$$

and is a special case of the more general bivariate linear process described by Jenkins & Watts (1968), p329. The choice of parameterisation may depend on the context, but for parsimony some authors (e.g. Gelfand *et al.* (2003); Schmidt & Gelfand (2003)) recommend making A lower-triangular, i.e. setting $\sigma_{12} = 0$. Calder (2008) uses this parameterisation of (4.4) to model concentrations of particulate matter ($PM_{2.5}$ and PM_{10}). The lower triangular parameterisation has a natural interpretation in this context, as particles of diameter less than 2.5μm necessarily have diameter less than 10μm, but not conversely.

4.2.2 Common Component Model

With similar notation to (4.1) and (4.2), the basic form of the common component model (Diggle & Ribeiro Jr. (2007)) is

$$Y_j(x_{ij}) = \beta^T d(x_{ij}) + \sigma_{0j} S_0(x_{ij}) + \sigma_j S_j(x_{ij}) + Z_{ij}. \tag{4.5}$$

For $j \neq j'$, let $\text{Corr}[Z_j(u), Z_{j'}(u)] = \zeta$. Marginally,

$$Y_j(x_{ij}) \sim \text{N}(\beta^T d(x_{ij}), \sigma_{0j}^2 + \sigma_j^2 + \tau_j^2).$$

We have

$$\text{Cov}[Y_j(x_{ij}), Y_{j'}(x_{ij})] = \sigma_{0j}\sigma_{0j'} + \text{I}_{\{j=j'\}}(\sigma_j^2 + \tau_j^2) + \text{I}_{\{j\neq j'\}}\zeta\tau_j\tau_{j'}$$

and, for $u > 0$,

$$\text{Cov}[Y_j(x_{ij}), Y_{j'}(x_{ij} - u))] = \sigma_{0j}\sigma_{0j'}\rho_0(u) + \text{I}_{\{j=j'\}}\sigma_j^2\rho_j(u),$$

where $I_{\{\cdot\}}$ is the indicator function. Analogously to (4.4), we can write the model for the two-dimensional process $S^+(x) = (S_1^+(x), S_2^+(x))$ as

$$\begin{pmatrix} S_1^+(x) \\ S_2^+(x) \end{pmatrix} = \begin{pmatrix} \sigma_{01} & \sigma_1 & 0 \\ \sigma_{02} & 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} S_0(x) \\ S_1(x) \\ S_2(x) \end{pmatrix} \tag{4.6}$$
$$= AS$$

One motivation for the CCM is that the 'common component' S_0 represents a shared factor that affects both Y_1 and Y_2, whereas S_1 and S_2 represent component-specific effects that are independent of each other and of S_0. Knorr-Held & Best (2001) use an analogous construction in the spatially discrete setting of disease mapping, where S_0, S_1 and S_2 represent shared and disease-specific spatially-varying latent risk-factors for two diseases. From a purely empirical standpoint, the use of three independent processes matches the need to specify three covariance functions: two auto-covariances and one cross-covariance.

The various construction-based models are compared in Figure 4.1. The single-component LCM is shown in panel (a), and the CCM in panel (b). Panel (c) shows an alternative representation of the CCM in which the shared process S_0^* represents an additional level in the hierarchy. This can be seen as equivalent to the CCM as follows.

Let $S_0^* = S_0$, $S_1^* = S_0^* + T_1$ and $S_2^* = S_0^* + T_2$, where S_0, T_1 and T_2 are independent processes. Such a representation is always possible if the S_i and S_i^* are Gaussian. In the hierarchical specification of the model in panel (c) of Figure 4.1 we would model $[S_0^*]$; $[T_1] \equiv [S_1^*|S_0^*]$ and $[T_2] \equiv [S_2^*|S_0^*]$; and finally $[(Y_1, Y_2)|S_0^*, S_1^*, S_2^*]$, which is equivalent to $[(S_1^+, S_2^+)|S_0^*, T_1, T_2]$. Thus the model in panel (c) can be rewritten as a model for $[S^+|S_0, T_1, T_2]$. Similarly, the CCM (4.5) can be written conditionally as

$$Y_j(x_{ij})|S_0(x_{ij}), S_j(x_{ij}) \sim \mathrm{N}(\beta^T d(x_{ij}) + \sigma_{0j} S_0(x_{ij}) + \sigma_j S_j(x_{ij}), \tau_j^2). \tag{4.7}$$

Further discussion of the conditional and unconditional representation is provided by Berliner (2000). The choice of representation to use may depend on the practical context of the application under consideration.

4.2.3 Kernel Convolution Approach

The kernel convolution approach is used in both spatial and spatio-temporal modelling. It leads to a reduction in computational complexity by comparison with direct specification of a bivariate Gaussian process. The approach relies on a spectral characterisation theorem for stationary Gaussian processes, which can be applied to spatial processes in much the same way as in the more familiar setting of time series (Priestley (1981)).

We begin by outlining results for the univariate case, summarising material given by Yaglom (2004) and the references therein. The key result is Bochner's Theorem, also proved by Khinchin, which states that every positive definite function is the Fourier transform of a positive finite Borel measure. Therefore, if a covariance function C is stationary, we can write

$$C(u) = \int e^{itu} F(dt) \tag{4.8}$$

for some positive definite bounded symmetric measure F. Conversely any function C which can be represented in the form (4.8) is the covariance function of a stationary Gaussian process. F is called the spectral distribution function of the underlying stationary process S. Karhunen's Theorem states that any random process S can be represented in the form

$$S(x) = \int k(x,t) F(dt), \tag{4.9}$$

where k is a deterministic, square-integrable kernel function and F is an orthogonal random measure. We assume that F is a Gaussian measure with mean zero and independent increments such that $\mathrm{Var}(F(dt)) = v(dt)$, say. If $v(dt) = dt$, it can be shown that the resulting process S, defined by (4.9), is Gaussian with mean zero and covariance function

$$C(u) = \int k(t) k(t-u) dt. \tag{4.10}$$

Higdon (2002) summarises the method for determining the kernel from a given covariance function: first determine the spectrum of C as the Fourier transform $\mathcal{F}(C)$; then, by the convolution theorem, take the inverse Fourier transform of the square root of $\mathcal{F}(C)$ to find a kernel corresponding to the covariance function C. The choice of kernel is not unique, as either the positive or the negative square root could be taken, but there is at most one positive definite kernel that corresponds to a given covariance function. However, as discussed by Xia & Gelfand (2005), it may not be possible to find such a kernel in closed form. Kern (2000) gives further details

and derives the result that the Gaussian covariance function corresponds to a Gaussian kernel. Similarly, the Matérn covariance function gives rise to a Matérn kernel, provided the smoothness parameter κ is sufficiently large (Xia & Gelfand (2005)).

Adaptations for bivariate spatial data are considered by Barry & Ver Hoef (1996), Ver Hoef & Barry (1998) and Ver Hoef *et al.* (2004). Define S_0, S_1 and S_2 as independent white noise processes, with

$$T_j(x_{ij}) = \rho_j S_0(x_{ij} - \Delta_j) + \sqrt{1-\rho_j^2} S_j(x_{ij}) \tag{4.11}$$

for $j = 1, 2$. Here, the parameters ρ_j and Δ_j represent the strength and shift in the spatial cross-correlation between the processes T_1 and T_2.

The process S^+ is constructed as

$$S_j^+(x_{ij}) = \int k_j(t - x_{ij}) T_j(t) dt, \tag{4.12}$$

where the k_j are kernel functions, and the observation process is

$$Y_j(x_{ij}) = S_j^+(x_{ij}) + Z_{ij}.$$

The analogous results to (4.10) are

$$\begin{aligned} C_{jj}(u) &= \int k_j(t) k_j(t-u) dt \quad (j = 1, 2) \\ C_{12}(u) &= \rho_1 \rho_2 \int k_1(t) k_2(t - u + \Delta_2 - \Delta_1) dt \end{aligned}$$

and

$$\mathrm{Corr}(S_1^+(x), S_2^+(x)) = \rho_1 \rho_2.$$

The use of the parameters ρ_j in (4.11) resembles the formulation of Oliver (2003) in (4.3), where ρ_j is interpreted as the correlation between the processes S_1^+ and S_2^+ at a common location. However, Ver Hoef *et al.* (2004) note that the parameters ρ_j and Δ_j in such models as (4.11) may not be identifiable. If the auto-covariance and cross-covariance functions are all Gaussian, so are the corresponding kernel functions (Boyle & Frean (2005)).

This type of construction permits a hierarchy of models of increasing complexity and generality,

analogous to the SCM and CCM, to be drawn up. Hence, for $j = 1, 2$

$$S_j^+(x_{ij}) = \int k_j(t - x_{ij}) S_j(t) dt \tag{4.13}$$

$$S_j^+(x_{ij}) = \int k_0(t - x_{ij}) S_0(t) dt + \int k_j(t - x_{ij}) S_j(t) dt \tag{4.14}$$

$$S_j^+(x_{ij}) = \int k_{0j}(t - x_{ij}) S_0(t) dt + \int k_j(t - x_{ij}) S_j(t) dt \tag{4.15}$$

In (4.13) S_1 and S_2 are correlated white noise processes (the correlation being represented by a single parameter), whilst in (4.14) and (4.15) S_0, S_1 and S_2 are independent white noise processes.

To implement the kernel convolution method in practice, the integrals are approximated as sums over a finite set of m fixed locations, called knots. For example,

$$\begin{aligned} S(x) &= \int k(x,t) W(dt) \\ &\approx \sum_{j=1}^{m} k(x, t_j) W(t_j), \end{aligned} \tag{4.16}$$

where W is a white noise process. Such an approximation raises the question of the choice of m and the positions of the knots, t_j. We give further details on implementation in Section 4.3.3.

4.2.4 Other Approaches

In a slightly different modelling strategy, Majumdar & Gelfand (2007) create cross-covariances for a bivariate process by convolving the covariance functions of independent univariate processes, hence

$$\mathrm{Cov}(S_j^+(x_{ij}), S_{j'}^+(x_{ij} - u)) = \sigma_j \sigma_{j'} \int C_j(u - t) C_{j'}(t) dt,$$

where $\sigma_j C_j(\cdot)$ is the covariance function of the process S_j^+. This yields a valid covariance function for the multivariate process S^+, and the integral can be evaluated using a Monte Carlo approximation. In further work, Majumdar *et al.* (2007) review methods for constructing multivariate non-stationary processes using the kernel convolution approach.

Several authors have tackled the problem of non-stationarity of S using process convolution methods. Higdon (1998) and Higdon (2002) develop such a method for spatio-temporal data by allowing the parameters of a separable kernel k to vary over space and/or time. A simpler example for a purely spatial process is provided by Lee *et al.* (2005), who allow the convolved

process W to be more general than white noise. They treat W as an "intrinsically stationary" process, such as a random walk or Markov random field on a discrete set of locations, while fixing k as a Gaussian kernel. The covariance of S is thus modified via the dependence structure of W according to spatial location.

In a series of papers, Fuentes (2001, 2002a, 2002b) uses the representation

$$S(x) = \int k(x,t)W_{\theta(t)}(t)dt,$$

where $W_{\theta(t)}$ is a stationary process with spatially-varying parameter $\theta(t)$. The $W_{\theta(t)}$ can therefore themselves be represented in the form of a kernel convolution, in this case a convolution of white noise processes.

Banerjee *et al.* (2008) consider a "predictive process model", in which the process S at a location x is approximated by the simple kriging predictor based on a realisation of S over a discrete set of knots. Y is expressed as a linear combination of the (unobserved) values of S at the knot locations, which reduces the dimensionality of the model.

Bárdossy (2006) and Bárdossy & Li (2008) introduce an entirely different method based on copulas. Copulas are parametrically-specified joint distributions generated from given marginals of the bivariate components, and have univariate $\mathrm{U}[0,1]$ marginal distributions (Frees & Valdez (1998)). Any m-variate cumulative distribution function F with margins $F_1, ...F_m$ can be expressed in the parametric form

$$F(x_1, ..., x_m) = C(F_1(x_1), ..., F_m(x_m)),$$

where C is a copula. A variety of commonly-used functional forms for C is available. This approach is used when the margins of the distribution have a known form but the joint distribution is unknown. However, it treats the data as a realisation of a multivariate distribution, rather than a multivariate process, and therefore lacks a spatially continuous interpretation.

4.3 Inference

4.3.1 Exploratory Methods

Most analyses begin with simple data exploration to suggest appropriate models and initial parameter estimates. For univariate geostatistical data, a commonly-used exploratory device for model (4.1) is the variogram,

$$V(x_1, x_2) = \frac{1}{2}\text{Var}(S(x_1) - S(x_2)).$$

In the bivariate setting, there are two different definitions of the cross-variogram for a stationary process $Y(x) = (Y_1(x), Y_2(x))$. The first is

$$V_{12}^*(u) = \frac{1}{2}\text{Cov}(Y_1(x) - Y_1(x-u), Y_2(x) - Y_2(x-u)).$$

The second is the "variance-based cross-variogram"

$$V_{12}(u) = \frac{1}{2}\text{Var}(Y_1(x) - Y_2(x-u)),$$

in which Y_1 and Y_2 are standardised so as to make $V_{12}(u)$ well-defined when the two variables are measured in different units (Cressie & Wikle (1998)). Method of moments estimators for these quantities are, respectively,

$$\hat{V}_{12}^*(u) = \frac{1}{2N(u)}\sum_i \{Y_1(x_i) - Y_1(x_i - u)\}\{Y_2(x_i) - Y_2(x_i - u)\}$$

and

$$\hat{V}_{12}(u) = \frac{1}{2N(u)}\sum_i \{(Y_1(x_i) - \bar{Y}_1) - (Y_2(x_i - u) - \bar{Y}_2)\}^2,$$

where $\bar{Y}_i$ is the sample mean of the Y_i, and the sums are taken over the $N(u)$ data pairs separated by a spatial distance of u. Robust estimators of the cross-variograms are provided by Lark (2002, 2003).

When the data-locations are irreglarly distributed, the empirical estimates are typically averaged ('binned') over discrete ranges of u before being plotted against u. These non-parametric estimates can then be compared to the corresponding theoretical quantities for a candidate model,

$$V_{12}^*(u) = \sigma_1\sigma_2\{\rho_{12}(0) - (\rho_{12}(u) - \rho_{12}(-u))/2\}$$

and

$$V_{12}(u) = (\sigma_1^2 + \sigma_2^2)/2 - \sigma_1\sigma_2\rho_{12}(u),$$

where $\sigma_i^2 = \mathrm{Var}(Y_i)$ and $\rho_{12}(u) = \mathrm{Corr}(Y_1(x), Y_2(x-u))$. An example is shown in Section 4.4.1.

A simple additional plot for co-located bivariate data is a scatter-plot of the data from the two components, which provides a check on whether $\rho_{12}(0)$ depends on Y_1 and Y_2.

4.3.2 Likelihood-Based Inference and Parameter Estimation

In principle, likelihood-based inference based on the LCM and CCM is simple, as the likelihood is simply the density function of a high-dimensional Gaussian distribution. To demonstrate parameter estimation for the CCM (4.6), we generated bivariate data from this model for 50 randomly-chosen locations on the unit square, using parameters $\mu_1 = \mu_2 = 0$, $\tau_1^2 = \tau_2^2 = 0$, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1$ and assigning to each of the S_i a covariance function of Matérn form with $\phi = 0.2$ and $\kappa = 0.5$. For each of 100 realisations of the model, we treated τ_1^2, τ_2^2 and κ as known and estimated the remaining parameters by maximum likelihood. The results are shown in Figure 4.2. Zhang (2004) provides similar results for the univariate model.

The models described in Section 4.2 may require constraints on their parameters to ensure identifiability. For example, the LCM specified in (4.4) is not identifiable if the covariance structures of S_1 and S_2 are identical, as it has one too many variance parameters. Context often indicates a reasonable parameterisation. For example, the CCM (4.6) may be most useful when modelling quantities that are physically related to each other and which are measured on the same scale, possibly after transformation. In such cases it may be appropriate to assume that $\sigma_{01}^2 = \sigma_{02}^2$ and/or that $\tau_1^2 = \tau_2^2$.

4.3.3 Implementation of the Kernel Convolution Approach

As noted above, when analysing large data-sets the kernel convolution construction (4.16) yields computational savings compared to direct specifications such as the LCM or CCM, which require inversion of a covariance matrix V of dimension $n_1 + n_2$ by $n_1 + n_2$, irrespective of the number of locations at which predictions are to be made. Fitting can usually be performed using standard software routines for Gaussian random effects models, treating the W terms in (4.16) as independent and identically distributed and the $k(\cdot, \cdot)$ terms as fixed multipliers.

Implementation requires choices to be made for the form of the kernel and the number m and locations of the knots. These choices will depend on the nature of the data being analysed, but some general points can be made.

Firstly, it is common practice to use a standard functional form of kernel, such as the Gaussian (Higdon (1998)) or uniform ("small rectangle moving average", Ver Hoef & Barry (1998)), that is chosen for computational convenience, rather than to match the empirical covariance structure of the data. See Calder (2007) and Gelfand *et al.* (2003).

Secondly, in general it is not possible to determine the analytic form of the kernel that corresponds to a given covariance function. Instead, the analyst has two choices: either to specify the kernel functions directly, or to find by numerical experimentation a set of kernels whose convolutions (4.10) approximate the required forms for the auto-covariance and cross-covariance functions. For example, the Matérn family of kernel functions provides a degree of flexibility that is sufficient for many applications; we give an example in Section 4.4.1.

Thirdly, the most commonly-used layout for knot locations is a rectangular or hexagonal grid, in which case a stationary process is obtained in the limit as the grid-spacing shrinks to zero. Other designs use an increased density of knots in certain regions to capture non-stationarity (Nychka & Saltzman (1998)), or in regions with a high density of data-points (Stroud *et al.* (2001)).

Finally, a formal measure of the effect of the number and locations of the knots on the performance of the approximation can be obtained, based on the Kullbeck-Leibler divergence (Xia & Gelfand (2005)). In practice, little improvement is gained by locating the knots on a grid whose spacing is less than the standard deviation of the kernel (Calder & Cressie (2007)). The convolution model with knots on a regular lattice can be used as an approximation to the limiting form as the lattice spacing approaches zero. In this case, the number of knots can be chosen pragmatically by decreasing the lattice spacing until the fitted covariance stucture and associated predictions stabilise (Rodrigues & Diggle (In press)).

4.4 Applications

4.4.1 Calcium/Magnesium Soil Data

To compare the modelling strategies described in Section 4.2, we use a dataset of 178 pairs of soil chemistry readings. These data are available in the contributed R (R Development Core Team (2008)) package geoR (Ribeiro Jr & Diggle (2001)), where further information about the source of the data is available. We use calcium and magnesium measurements in the 0-20cm soil layer. The data are shown in the right panel of Figure 4.5, in which one unit represents one kilometre. Empirical variograms and cross-variograms are shown in Figure 4.3.

We fit the CCM and SCM to the data, assuming Matérn correlation functions with known $\kappa = 0.5$. This gave the largest likelihood amongst values of $\kappa \in \{0.5, 1.5, 2.5\}$, although point predictions using different values of κ were almost identical. We fit the CCM both with $\tau_1^2 = \tau_2^2$ (Model 1) and with $\tau_1^2 \neq \tau_2^2$ (Model 2), although there was no significant improvement in likelihood using the latter (-425.3 versus -425.0). The fitted variograms for Model 1, shown in Figure 4.3, suggest reasonable fit. Parameter estimates for these two models, the SCM (4.4) with $\sigma_{12} = 0$ (Model 3), and two univariate models (Model 4) are shown in Table 4.1.

In the CCM, the correlation of $\mathrm{Corr}(Y_1(x), Y_2(x))$ at a single location x is

$$\frac{\sigma_{01}\sigma_{02}}{\sqrt{(\sigma_{01}^2 + \sigma_1^2 + \tau_1^2)(\sigma_{02}^2 + \sigma_2^2 + \tau_2^2)}},$$

which is estimated for Model 1 as 0.39, close to the sample (co-located) correlation between the calcium and magnesium components, 0.33.

Figure 4.5 shows predictions of $S(x)$ and the prediction variances, calculated by plugging in the parameter estimates from Model 1. Predictions from the SCM are superficially identical to these. Figure 4.5 also shows the predictions resulting from an analysis using kernel convolution. For this, we first used the parameter estimates from the CCM with $\kappa = 2.5$ to determine kernel functions that, after convolution, gave a good approximation to the fitted auto- and cross-covariance functions. The best fit was obtained with a Gaussian kernel for k_0 and Matérn kernels for k_1 and k_2, and the approximation was found by minimising

$$\int \{(C_{11}(x) - \tilde{k}_{11}(x))^2 + (C_{22}(x) - \tilde{k}_{22}(x))^2 + (C_{12}(x) - \tilde{k}_{12}(x))^2\} dx,$$

where $\tilde{k}_{ij}(x)$ denotes the convolution between k_i and k_j, and the $C_{ij}(x)$ denote auto- and cross-covariance functions. Figure 4.4 shows the approximation of the two auto-covariance functions and one cross-covariance function. As might be expected, the point predictions from the kernel convolution model (middle panel of Figure 4.5) strongly resemble those from the CCM, although there are some minor differences for the magnesium component.

4.4.2 Winnipeg Radon Data

These data consist of radon measurements taken as part of a case-control study in Winnipeg, Canada to investigate epidemiological associations with lung cancer (Létourneau *et al.* (1992)). Radon dosimeters were placed in bedrooms and basements of current and former residences of study participants. As radon concentrations are seasonal (Whitley & Darby (1999)) and highly variable over short time intervals (Brabec & Jílek (2009)), measurements were taken over a whole year to produce a total exposure measurement in Bq/m^3 at each site. Here we use data from 1901 dwellings for which a bedroom measurement was recorded; for 1622 of these, a basement measurement was also available. Reasons for missing data include equipment failure, refusal of house owners to allow installation of dosimeters and the absence of a basement from the property.

The original study used exposure for each individual aggregated across their lifetime residential locations, and found no evidence of an association with lung cancer (Létourneau *et al.* (1994)). For example, the odds of lung cancer in cases relative to controls was estimated as 0.97 (95% confidence interval 0.81 to 1.15), where the units are per 3750 Bq/m^3-years for radon measured in the bedroom. Nevertheless, there is now strong epidemiological evidence from other cohort and case-control studies to support the link between prolonged radon exposure and lung cancer incidence (Krewski *et al.* (2006)). Also, as radon gas tends to become trapped within buildings, in many areas the home is a substantially greater source of exposure than the outdoors (Steck *et al.* (1999)).

The original analysis of Létourneau *et al.* (1994) assigned an observed bedroom and basement exposure reading to each household without examining the spatial variation in exposure. Here we investigate this spatial variation, treating bedroom and basement measurements as the two components in a bivariate common component model and using data from residences of participants in the control group. Data locations are shown in the right panel of Figure 4.6.

We analyse the radon data on the logarithmic scale, for consistency with the multiplicative model used by Brabec & Jílek (2009). Preliminary analysis suggested short-range spatial autocorrelation for measurements made in the bedroom (Y_1) and in the basement (Y_2), with a large nugget effect. However, bedroom and basement measurements made within the same dwelling were highly correlated ($r = 0.80$), and at least as high in the basement as the bedroom for 1469 of the 1622 dwellings (91%) for which data were available at both sites.

Other studies have demonstrated that house-specific factors such as detachment, double glazing, floor type and date of construction are associated with radon build-up (Hunter *et al.* (2009)). For our study, the only available covariate was altitude, which did not vary substantially over the study region and showed no association with radon.

As discussed by Whitley & Darby (1999), radon typically accumulates in the basements of houses before dispersing to upstairs rooms and the outside atmosphere. This suggests spatially-varying terms of the form $\exp\{\sigma_0 S_0(\cdot)\}$ (on the original scale) for the basement component and $\exp\{\sigma_0 S_0(\cdot)\}\exp\{\sigma_1 S_1(\cdot)\}$ for the bedroom component, where the S_1 term is interpreted as the proportion of basement radon that is detected in the bedroom at a given location. We incorporate house-specific effects via a correlated nugget effect (correlation parameter ζ), and fit the common component model

$$\log\{Y_j(x_{ij})\} = \beta_j + \sigma_0 S_0(x_{ij}) + \mathrm{I}_{\{j=1\}}\sigma_1 S_1(x_{ij}) + Z_{ij}, \tag{4.17}$$

where $S_0(\cdot)$ and $S_1(\cdot)$ are standardised Gaussian processes with exponential correlation functions, $\rho_0(u) = \exp(-|u|/\phi_0)$ and $\rho_1(u) = \exp(-|u|/\phi_1)$. We also specify that, for any location at which both measurements are made, (Z_{i1}, Z_{i2}) has a zero mean bivariate Gaussian distribution with covariance.

$$\begin{pmatrix} \tau^2 & \zeta\tau^2 \\ \zeta\tau^2 & \tau^2 \end{pmatrix}.$$

The maximum likelihood estimates for the Winnipeg data are $\hat{\beta}_1 = 4.65$, $\hat{\beta}_2 = 5.05$, $\hat{\sigma}_0^2 = 0.15$, $\hat{\sigma}_1^2 = 0.015$, $\hat{\phi}_0 = 0.009$, $\hat{\phi}_1 = 0.014$, $\hat{\tau}^2 = 0.47$ and $\hat{\zeta} = 0.56$. The fitted correlation between bedroom and basement measurements made in the same house is

$$\frac{(\hat{\sigma}_0^2 + \hat{\zeta}\hat{\tau}^2)}{\sqrt{(\hat{\sigma}_0^2 + \hat{\sigma}_1^2 + \hat{\tau}^2)(\hat{\sigma}_0^2 + \hat{\tau}^2)}} = 0.66.$$

The variances corresponding to the two S-processes are small compared to τ^2, and the spatial correlation decays rapidly; for example, at a distance of 1km ($u \approx 0.014°$), $\hat{\rho}_0(u) = 0.21$ and $\hat{\rho}_1(u) = 0.37$. Figure 4.6 shows the predicted bedroom radon surface across the Winnipeg region; units of distance are degrees latitude/longitude ($0.1° \approx 7.1$km at this latitude). There is evidence that radon levels are higher in some areas of the city than others, but the variation in Y_1 between individual houses dominates the spatial variation.

To test the value of the bivariate model for prediction, we deleted 50 Y_2 observations from the data-set, refit the model and made predictions of the deleted Y_2 values. Figure 4.7 shows a comparison of predicted and observed values. There is good agreement between the two, as a result of the extra information provided by the observed Y_1 at the prediction locations, albeit with some shrinkage towards the overall mean. We also fitted a univariate model to the basement data alone. Figure 4.7 shows that predictions based on this model compare much less favourably to the observed values. This is essentially because, in the absence of useful explanatory variables the rapidly decaying fitted spatial correlation structure forces the predictions at unmeasured locations to revert rapidly to the fitted mean as the distance to the nearest measured location increases.

4.5 Discussion

We have presented a comparison of the many models available for analysing bivariate geostatistical data. In particular, we have discussed in detail the properties of the common component model and a corresponding model in the frequency domain, represented as a kernel convolution.

The advantage of the kernel convolution approach is its computational tractability, but at the cost that the form of the kernel k is often chosen arbitrarily. We suggest that a suitable kernel can usually be found by assuming a Matérn functional form, estimating parameters by maximum likelihood and checking that the resulting covariance structure gives a good fit to the empirical covariance structure of the data. Alternatively, a Bayesian approach could be used, as discussed in detail by Higdon (1998) and Calder (2008).

For many bivariate geostatistical models, an appropriate choice of parameterisation is often neither obvious nor unique, and identifiability can cause difficulties in obtaining meaningful parameter estimates. This has been especially noted for the linear model of coregionalisation with several components (Zhang (2007)), but also applies to simpler related models. For example, a

simple reformulation of (4.17) is

$$\begin{aligned} Y_2(x_{i2}) &= \beta_2 + \sigma_0 S_0(x_{i2}) + Z_{i2}, \\ Y_1(x_{i1}) &= (\beta_1 - \beta_2) + Y_2(x_{i1}) + \sigma_1 S_1(x_{i1}) + Z'_{i1} \end{aligned}$$

where now Z'_{i1} is a random variable with $\mathrm{Var}(Z'_{i1}) \neq \mathrm{Var}(Z_{i2})$. More equivalent formulations can often be found for multivariate models of higher dimension.

The choice of model for a particular application is dependent on context. In the radon example, our choice of an asymmetric model was suggested by the typical pattern of radon flow in buildings (Whitley & Darby (1999)). Our results are broadly consistent with other models for the distribution of indoor radon: there is short-range spatial correlation, and a strong house-specific effect that is not easily distinguished from measurement error without replicated observations. Although the original lung cancer analysis of the radon data did not account for the spatial distribution of radon (Létourneau *et al.* (1994)), the relatively uniform nature of radon exposure in Winnipeg suggests that a change in the qualitative conclusions of that study would be unlikely.

Finally, the good predictive performance of our radon model by comparison with the predictions obtained from a univariate model illustrates the benefit of fitting a bivariate model when the data are incomplete. The extent of the improvement is not well documented for geostatistical models, but it will depend on the strength of the correlation between Y_1 and Y_2 and the degree of incompleteness in the data. We intend to expore these issues in future research.

Acknowledgements

TRF was supported by a Doctoral Training Account studentship and PJD by a Senior Fellowship from the Engineering and Physical Sciences Research Council (EPSRC). We thank Patrick Brown for helpful discussions and Jan Zielinski for provision of the radon data files.

Parameter	Model 1	Model 2	Model 3	Model 4
μ_1	50.0	50.5	50.1	50.1
μ_2	25.1	25.1	25.1	25.1
σ_{01}^2	32.3	31.2	143	-
σ_{02}^2	32.3	31.2	7.53	-
σ_1^2	110	101	-	135
σ_2^2	2.94	4.53	27.9	35.2
ϕ_0	0.13	0.13	0.14	-
ϕ_1	0.14	0.19	-	0.16
ϕ_2	0.13	0.12	0.13	0.13
τ_1^2	8.93	19.6	8.81	16.8
τ_2^2	8.93	8.26	8.81	8.30

Table 4.1: Comparison of parameter estimates from four models fit to the soil data. Model 1: CCM with $\tau_1^2 = \tau_2^2$; Model 2: CCM with $\tau_1^2 \neq \tau_2^2$; Model 3: SCM with $\tau_1^2 = \tau_2^2$; Model 4: two independent univariate models, one for each component.

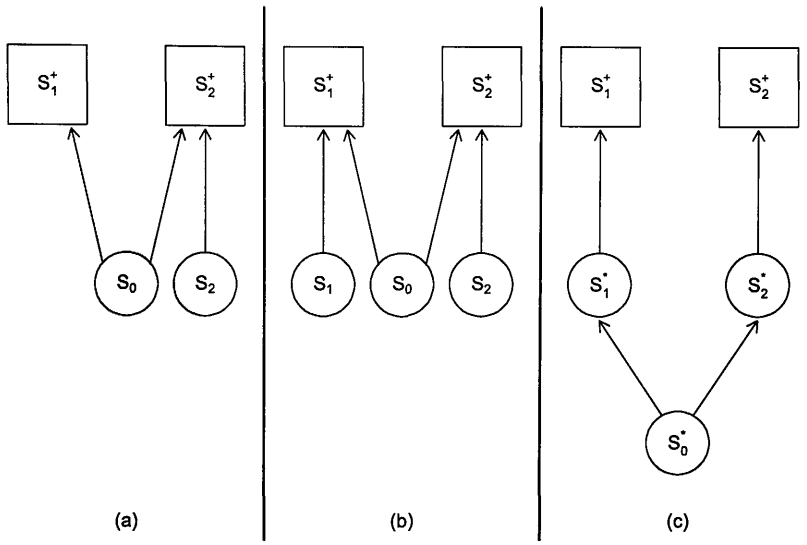


Figure 4.1: Graphical representation of (a) the single component LCM; (b) the CCM; (c) an alternative representation of the CCM.

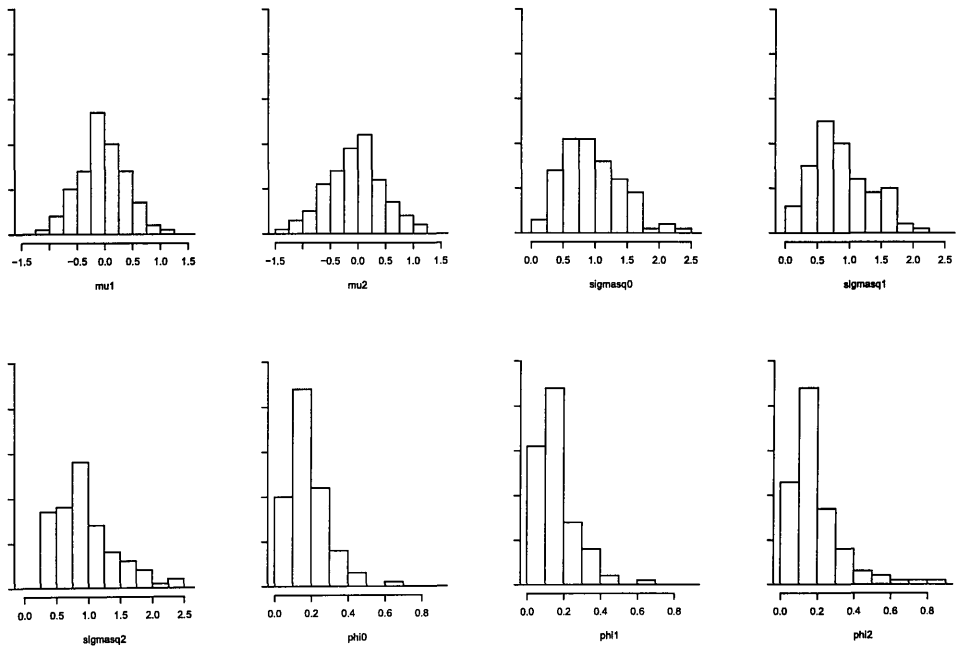


Figure 4.2: Parameter estimates for simulated data from the CCM. Correct parameter values are $\mu_1 = \mu_2 = 0$, $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1$, $\phi_0 = \phi_1 = \phi_2 = 0.2$. See Section 4.3.2 for further details. N.B. This figure is Supplementary Material.

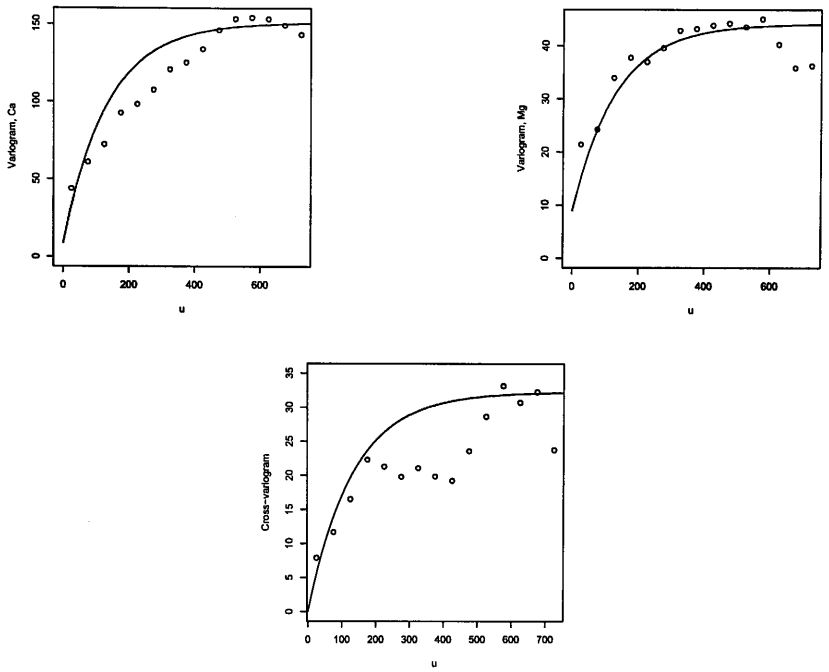


Figure 4.3: Empirical variograms for calcium (top left) and magnesium (top right) data, and the empirical cross-variogram $V_{12}^*(u)$. Curves show the fitted variograms from the CCM, Model 1 in Table 4.1. N.B. This figure is Supplementary Material.

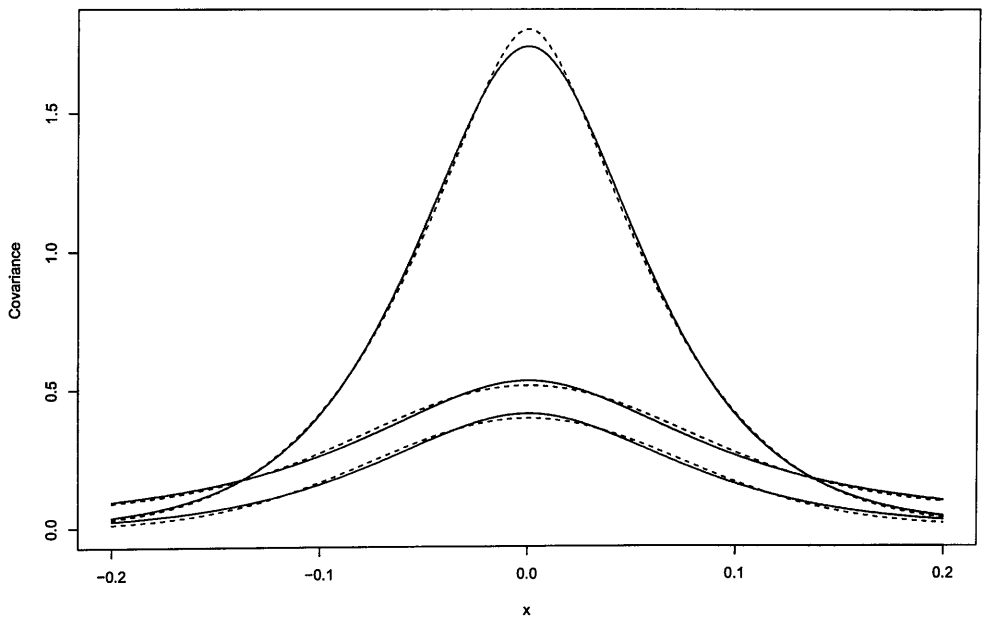


Figure 4.4: For the soil data analysis, comparison of the covariance functions (solid lines) $C_{11}(x)$ (top), $C_{22}(x)$ (middle) and $C_{12}(x)$ (bottom) with convolved kernel function approximations (dotted lines).

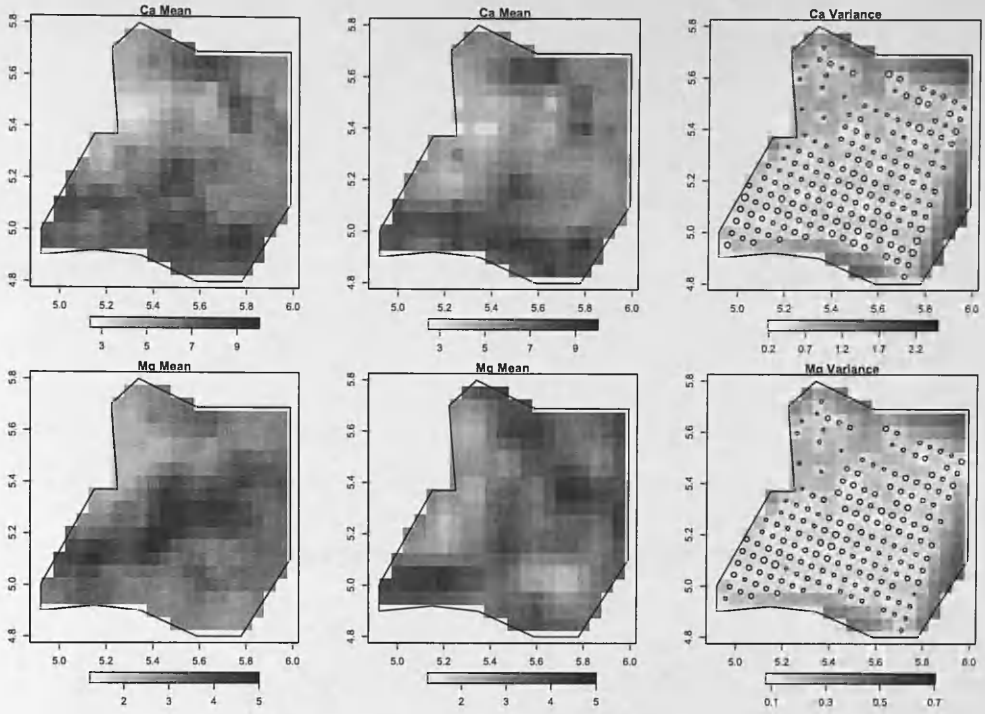


Figure 4.5: Point predictions from the CCM (left) and kernel convolution method (centre), and prediction variances from the CCM (right). The top row corresponds to the first component (calcium) and the bottom row to the second (magnesium).

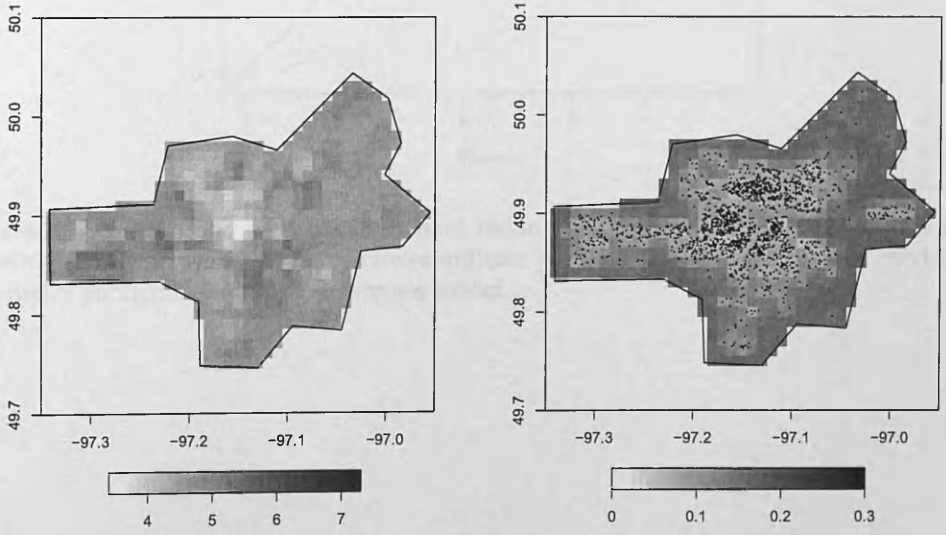


Figure 4.6: Bedroom radon point predictions (left) and prediction variances (right) from the CCM (4.17). Points indicate data locations. Distances are shown in degrees latitude/longitude.

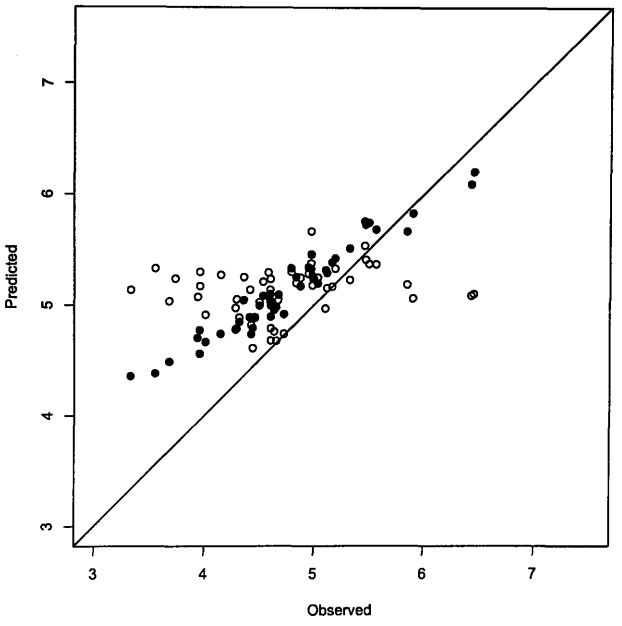


Figure 4.7: Observed and predicted basement radon measurements at 50 locations for which bedroom data were available. Filled circles indicate predictions from the bivariate model, and open circles predictions from the univariate model.

References

Banerjee, S., Gelfand, A.E., Finley, A.O., & Sang, H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society B*, **70**, 825–848.

Bárdossy, A. 2006. Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, **42**. W11416, doi:10.1029/2005WR004754.

Bárdossy, A., & Li, J. 2008. Geostatistical interpolation using copulas. *Water Resources Research*, **44**. W07412, doi:10.1029/2007WR006115.

Barry, R.P., & Ver Hoef, J.M. 1996. Blackbox kriging: Spatial prediction without specifying variogram models. *Journal of Agricultural, Biological and Environmental Statistics*, **1**, 297–322.

Berliner, L.M. 2000. Hierarchical Bayesian modeling in the environmental sciences. *Allgemeines Statistische Archiv*, **84**, 141–153.

Boyle, P., & Frean, M. 2005. *Multiple output Gaussian process regression.* Tech. rept. School of Mathematical and Computing Sciences, Victoria University of Wellington.

Brabec, M., & Jílek, K. 2009. Dynamical model for indoor radon concentration monitoring. *Environmetrics*, **20**, 718–729.

Calder, C.A. 2007. Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. *Environmental and Ecological Statistics*, **14**, 229–247.

Calder, C.A. 2008. A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics*, **19**, 39–48.

Calder, C.A., & Cressie, N. 2007 (August 22-29). Some topics in convolution-based spatial modeling. *In: Proceedings of the 56th Session of the International Statistics Institute.*

Calder, C.A., Craigmile, P.F., & Zhang, J. 2009. Regional spatial modelling of topsoil geochemistry. *Biometrics*, **65**, 206–215.

Chilès, J.-P., & Delfiner, P. 1999. *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. 1993. *Statistics for Spatial Data*. New York: Wiley.

Cressie, N., & Wikle, C.K. 1998. The variance-based cross-variogram: you can add apples and oranges. *Mathematical Geology*, **30**, 789–799.

D'Agostino, V., Greene, E.A., Passarella, G., & Vurro, M. 1993. Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization. *Environmental Geology*, **36**, 285–295.

Diggle, P.J., & Ribeiro Jr., P.J. 2007. *Model-Based Geostatistics*. New York: Springer-Verlag.

Frees, E.W., & Valdez, E.A. 1998. Understanding relationships using copulas. *North American Actuarial Journal*, **2**, 1–25.

Fuentes, M. 2001. A high frequency kriging approach for nonstationary environmental processes. *Environmetrics*, **12**, 469–483.

Fuentes, M. 2002a. Interpolation of nonstationary air pollution processes: a spatial spectral approach. *Statistical Modelling*, **2**, 281–298.

Fuentes, M. 2002b. Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Gelfand, A.E., Schmidt, A.M., & Sirmans, C.F. 2003. *Multivariate spatial process models: conditional and unconditional Bayesian approaches using coregionalization*. Tech. rept. Institute of Statistics and Decision Sciences, Duke University.

Goovaerts, P. 1994. On a controversial method for modeling a coregionalization. *Mathematical Geology*, **26**, 197–204.

Goulard, M., & Voltz, M. 1992. Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, **24**, 269–286.

Haas, T.C. 1996. Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. *Environmetrics*, **7**, 145–165.

Higdon, D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5**, 173–190.

Higdon, D. 2002. *Space and space-time modeling using process convolutions*. In: Quantitative Methods for Current Environmental Issues. New York: Springer-Verlag. Pages 37–56.

Hunter, N., Muirhead, C.R., Miles, J.C.H., & Appleton, J.D. 2009. Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. *Radiation Protection Dosimetry*, **136**, 17–22.

Jenkins, G.M., & Watts, D.G. 1968. *Spectral Analysis and its Applications*. San Francisco: Holden-Day.

Kern, J. 2000. *Bayesian process-convolution approaches to specifying spatial dependence structure*. Ph.D. thesis, Duke University.

Knorr-Held, L., & Best, N.G. 2001. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society A*, **164**, 73–85.

Krewski, D., Lubin, J.H., Zielinski, J.M., Alavanja, M., Catalan, V.S., Field, R.W., Klotz, J.B., Létourneau, E.G., Lynch, C.F., Lyon, J.L., Sandler, D.P., Schoenberg, J.B., Steck, D.J., Stolwijk, J.A., Weinberg, C., & Wilcox, H.B. 2006. A combined analysis of North American case-control studies of residential radon and lung cancer. *Journal of Toxicology and Environmental Health, Part A: Current Issues*, **69**, 533–597.

Lark, R.M. 2002. Robust estimation of the pseudo cross-variogram for cokriging soil properties. *European Journal of Soil Science*, **53**, 253–270.

Lark, R.M. 2003. Two robust estimators of the cross-variogram for multivariate geostatistical analysis of soil properties. *European Journal of Soil Science*, **54**, 187–201.

Lee, H.K.H., Higdon, D.M., Calder, C.A., & Holloman, C.H. 2005. Efficient models for correlated data via convolutions of intrinsic processes. *Statistical Modelling*, **5**, 53–74.

Létourneau, E.G., Zielinski, J.M., Krewski, D., & McGregor, R.G. 1992. Levels of radon gas in Winnipeg homes. *Radiation Protection Dosimetry*, **45**, 531–534.

Létourneau, E.G., Krewski, D., Choi, N.W., Goddard, M.J., McGregor, R.G., Zielinski, J.M., & Du, J. 1994. Case-control study of residential radon and lung cancer in Winnipeg, Manitoba, Canada. *American Journal of Epidemiology*, **140**, 310–322.

Li, B., Genton, M.G., & Sherman, M. 2008. Testing the covariance structure of multivariate random fields. *Biometrika*, **95**, 813–829.

Majumdar, A., & Gelfand, A.E. 2007. Multivariate spatial modeling for geostatistical data using convolved covariance functions. *Mathematical Geology*, **39**, 225–245.

Majumdar, A., Paul, D., & Bautista, D. 2007. *A generalized convolution model for multivariate nonstationary spatial processes*. Tech. rept. Department of Mathematics and Statistics, Arizona State University.

Marchant, B.P., & Lark, R.M. 2007. Estimation of linear models of coregionalization by residual maximum likelihood. *European Journal of Soil Science*, **58**, 1506–1513.

Mardia, K.V., & Goodall, C.R. 1993. *Spatial-temporal analysis of multivariate environmental monitoring data*. In: Multivariate Environmental Statistics. Amsterdam: North-Holland. Pages 347–386.

Nychka, D., & Saltzman, N. 1998. *Design of air-quality monitoring networks*. In: Case studies in Environmental Statistics. New York: Springer-Verlag. Chap. 4, pages 51–76.

Oliver, D.S. 2003. Gaussian cosimulation: modelling of the cross-covariance. *Mathematical Geology*, **35**, 681–698.

Priestley, M.B. 1981. *Spectral Analysis and Time Series*. New York: Academic Press.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ribeiro Jr, P.J., & Diggle, P.J. 2001. geoR: a package for geostatistical analysis. *R-NEWS*, **1**(2), 14–18. ISSN 1609-3631.

Rodrigues, A., & Diggle, P. In press. A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics*.

Royle, J.A., & Berliner, L.M. 1999. A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 29–56.

Rue, H., & Tjelmeland, H. 2002. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, **29**, 31–49.

Schmidt, A.M., & Gelfand, A.M. 2003. A Bayesian coregionalization approach for multivariate pollution data. *Journal of Geophysical Research - Atmospheres*, **108 (D24)**, 8783.

Steck, D.J., Field, R.W., & Lynch, C.F. 1999. Exposure to atmospheric radon. *Environmental Health Perspectives*, **107**, 123–127.

Stroud, J.R., Müller, P., & Sansó, B. 2001. Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society B*, **63**, 673–689.

Ver Hoef, J.M., & Barry, R.P. 1998. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275–294.

Ver Hoef, J.M., & Cressie, N. 1993. Multivariable spatial prediction. *Mathematical Geology*, **25**, 219–240.

Ver Hoef, J.M., Cressie, N., & Barry, R.P. 2004. Flexible spatial models for kriging and cokriging using moving averages and the Fast Fourier Transform (FFT). *Journal of Computational and Graphical Statistics*, **13**, 265–282.

Wackernagel, H. 1995. *Multivariate Geostatistics*. Berlin: Springer-Verlag.

Whitley, E., & Darby, S.C. 1999. *Quantifying the risks from residential radon*. In: Statistical aspects of health and the environment. Chichester: Wiley. Chap. 5, pages 73–89.

Xia, G., & Gelfand, A.E. 2005. *Stationary process approximation for the analysis of large spatial datasets*. Tech. rept. Institute of Statistics and Decision Sciences, Duke University.

Yaglom, A.M. 2004. *An Introduction to the Theory of Stationary Random Functions*. Mineola, New York: Dover.

Zhang, H. 2004. Inconsistent estimation and asymptopically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

Zhang, H. 2007. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, **18**, 125–139.

Chapter 5

Conclusions

In this section we summarise the findings of each of the three papers separately, discuss common themes between them, and suggest possible topics for future work. As each paper contains its own set of conclusions and discussion, we focus here on the broader context in which the work is based.

5.1 Paper 1

The main aim of this paper was to develop a model for weekly levels of black smoke (BS) across the city of Newcastle-upon-Tyne over a 31-year period, based on data from a relatively sparse spatial network of air pollution monitoring stations. Our modelling strategy arose from the context, availability and nature of both the monitoring data and a series of temporally- and spatially-varying covariates that we derived from other sources.

Our final modelling strategy allowed us to take advantage of the approximately constant temporal decline in log-BS levels over the study period. We did this by first modelling the overall temporal trend μ_t, ignoring spatial variability, and then modifying this to take account of spatial characteristics using covariates $\mathbf{w}$. Thus our model for log-BS $Y_t(x)$ at time t and location x was

$$Y_t(x) = \hat{\mu}_t + \mathbf{w}^T\boldsymbol{\beta} + Z_t(x), \tag{5.1}$$

where $Z_t(x)$ is a residual term. The nature of the covariates was such that, to a large extent, they accounted for the residual spatio-temporal correlation from the model for μ_t, and thus we

were able to make the simplifying assumption that the $Z_t(x)$ are independent $\mathrm{N}(0, \sigma^2)$ random variables. Other papers to have treated the residual term in this way, without the need for further modelling, include Handcock & Wallis (1994), for winter temperature data, Holland *et al.* (1999), for sulphur dioxide data, and Smith *et al.* (2003), for weekly $\mathrm{PM}_{2.5}$ data.

As a major objective of the work was to predict BS levels at unsampled locations and times, it was important that candidate covariates be available at both monitoring and prediction locations. Otherwise it would be necessary to construct a spatio-temporal model for the covariates, a task that, while not infeasible (e.g. Zhu *et al.* (2003)), is little different from the problem of modelling $Y_t(x)$ itself (Zidek *et al.* (2002)).

It is interesting to consider which modelling approach may have been appropriate if the correlation between BS and the covariates had been weaker. As in the general model formulation (1.2), it would then have been necessary to consider an extra spatio-temporal process $S_t(x)$ in (5.1).

A simple form for such an $S_t(x)$ would assume separability of the space and time components (e.g. Haas (1995); Glasbey *et al.* (2001)). This assumption would allow the process to be written as $S_t(x) = S_{1,t}S_2(x)$, where S_1 and S_2 are processes to be specified. Separability guarantees positive-definiteness of the resulting spatio-temporal covariance function if the covariance functions of S_1 and S_2 are themselves positive definite (Gneiting *et al.* (2007)). In our application, we "de-trended" the data from each station by subtracting the dominant average temporal trend; separability does not appear to be an unreasonable assumption for the resulting residual process.

Our choice to use a two-stage modelling approach was driven by the nature of the decline in particulate matter in the study region between the 1960s and 1990s. During this time, implementation of the Clean Air Act and a rapid downturn in levels of industrial activity in Newcastle caused weekly levels of PM_{10} to drop from over $500\mu\mathrm{g/m}^3$ to under $10\mu\mathrm{g/m}^3$. This suggested that a model for the temporal change alone would be a good starting point for a full spatio-temporal analysis, and the fit was improved by including a covariate relating to weekly temperature. The temperature covariate was also suggested by the context: people consume more polluting fuels in unseasonably cold weather. Similar methodological approaches, known variously as "de-trending" or "pre-whitening", have been used successfully by several other authors, including Zidek *et al.* (2002) and Meiring *et al.* (1998).

The final model allowed predictions to be made for each spatial location, for each week during the study period. The predictions are used as surrogate measures of exposure to particulate matter for pregnant women throughout the course of pregnancy, and are being used in ongoing analyses of the PAMPER study. This is an unmatched historical cohort study, so a plausible model is the logistic regression

$$\text{logit}(p_i) = \alpha + \beta f(Y_{\mathbf{t}_i}(x_i)) + \gamma \mathbf{u}_i + \epsilon_i, \tag{5.2}$$

where p_i is the probability of a 'positive' outcome (e.g. low birthweight) for birth i, $\text{logit}(p) = \log(p/(1-p))$, $f(\cdot)$ is a function of maternal exposure, $Y_{\mathbf{t}_i}(x_i)$ is a vector of predicted BS values at location x_i and times $t_{i_1}, ..., t_{i_n}$, the weeks in the pregnancy period of i, $\mathbf{u}_i$ denotes additional risk factors (e.g. age, socio-economic status and smoking status) and ϵ_i is an error term. Wakefield & Shaddick (2006) discuss inference for similar models that link environmental exposures and area-aggregated health outcomes.

In practice, predictions $\hat{Y}_{\mathbf{t}_i}(x)$ would be used in (5.2) in place of $Y_{\mathbf{t}_i}(x_i)$. As Paper 1 shows, the $\hat{Y}_{\mathbf{t}_i}(x)$ are subject to considerable prediction error, which may affect inference in the cohort study. A simple sensitivity analysis might involve simulating from the (Gaussian) predictive distribution of $\hat{Y}_{\mathbf{t}_i}(x)$ and refitting model (5.2); a more principled approach would treat the prediction error as a form of Berkson measurement error (Gryparis *et al.* (2009)). This might employ one of the many solutions for dealing with measurement error in non-linear models, such as regression calibration (Carroll *et al.* (2006)). We discuss measurement error in a different context in Paper 2.

Determining the functional form of f is another challenge. In Paper 1 we calculated predicted exposure weekly and time-aggregated over trimesters and the whole pregnancy period, but other authors have suggested that different functions of maternal exposure may be responsible for an increased risk of adverse birth outcomes (Šrám *et al.* (2005)). Such functions might include the maximum exposure during the whole pregnancy or a particular trimester. There is evidence that, for developmental defects, both the timing and the magnitude of exposure are significant factors (Axelrod *et al.* (2001)). This argument suggests that a function such as $\int w(\mathbf{t}_i) Y_{\mathbf{t}_i}(x_i) d\mathbf{t}_i$, where $w(\mathbf{t}_i)$ is a weighting function, may be an appropriate choice for f.

A more general issue arising in epidemiological studies of environmental exposures is whether a predictor such as $\hat{Y}_t(x)$ truly represent the exposure to which an individual is subjected. There are at least two substantial objections to the use of such predictors: firstly, that individuals typi-

cally do not remain in a fixed location for a prolonged time period; and secondly, that even at the purported fixed location, the individual's uptake of the pollutant would not equal the ambient level of the pollutant in the atmosphere.

Such jeremiads are neither new, unfounded, nor easily resolved (Cox (2000)). In the context of historical studies such as PAMPER, little retrospective information is available to inform a more accurate measure of personal exposure. For such studies, predictions such as $\hat{Y}_t(x)$ should therefore be interpreted as surrogate measures of exposure, and associations with outcome status interpreted in relative, not absolute, terms. The numerical value of such an exposure for an individual is less important than the relative values for different individuals.

Collecting more accurate personalised measures of exposure is infeasible for historical studies and studies for which the relevant exposure accumulates over a long time period, but can be attempted in prospective designs. To calculate a 'gold standard' measure of exposure, it is necessary to track the locations of individuals over time, and this motivates the work in Paper 2.

5.2 Paper 2

In this paper we relaxed an assumption made implicitly in virtually all geostatistical analyses, that the data locations and prediction locations can be measured precisely. Very little research had previously been conducted around this issue, and the papers by Gabrosek & Cressie (2002) and Cressie & Kornak (2003) were the only ones to assess its impact on geostatistical modelling.

Our solution to the problem was to formulate two models: a geostatistical model, typically of the form (3.1), assuming all locations are known; and a positional error model for $[X^*|X]$ or $[X|X^*]$ that specifies the conditional relationship between the true and observed locations, X^* and X respectively. We showed that the likelihood can then be constructed in the form of an integral with respect to X^* (3.3), from which likelihood-based inference can proceed.

Our results show that point predictions are biased, and prediction variances incorrect, if there is positional error that is not taken into account in the analysis. This is not a startling conclusion, as similar results are seen in the broader class of non-linear regression models (Carroll *et al.* (2006)). More surprising is the finding that in certain circumstances the variance of the prediction can *decrease* if positional error is present.

These results, described and elaborated upon in Appendix B, suggest that if sizeable positional error exists, a principled analysis should take its effects into account. Nevertheless, there is a substantial computational cost in doing such an analysis. Monte Carlo routines to evaluate (3.3) for likelihood maximisation require a large number of iterations to discern a likelihood surface that is often flat even for standard geostatistical models (Zhang (2004)).

Thus while we have demonstrated the method for data-sets of moderate size, extra work is required to enable these ideas to bear fruit in larger studies. Possible alternative techniques include Markov chain Monte Carlo methods, although it is far from clear how to implement an appropriate algorithm for the required integral (e.g. Example 3 of Evans & Swartz (1995)), and the "pseudolikelihood" approach of Cressie & Kornak (2003). The latter method assumes local normality of the likelihood surface, but there is no guarantee that this would hold even approximately for standard geostatistical models (Warnes & Ripley (1987)).

Throughout Paper 2 we used simple independent Gaussian positional errors, with a Berkson error structure, in a model for $X^*|X$. Extensions of this work could incorporate alternative positional error models. For example, in an application that tracks the position of an individual over a trajectory, it is plausible that errors $(\epsilon_1, ..., \epsilon_n)$ in a time series of consecutive measurements might be auto-correlated ($\mathrm{Corr}(\epsilon_i, \epsilon_{i+1}) = \alpha$ for $i = 1, ..., n-1$ and $\mathrm{Corr}(\epsilon_i, \epsilon_j) = 0$ for all other i, j). Alternatively, one might represent the correlation as a function of the time between measurements. While adding extra complexity, such models easily fit into the framework described in Paper 2, at least in principle, although determining a suitable positional error model for a given application constitutes a separate research topic in its own right.

The impact of positional error on inference is greatest when the positional error variance γ^2 is large and the underlying surface S is changing rapidly. In the setting of exposure estimation, this work is therefore likely to be of most importance for short-term exposures, when replicate positional measurements cannot be made, and for exposure surfaces that are discontinuous or have short-range spatial correlation. An example is the pollution surface around a point source such as an incinerator, or a line source such as a road. Traffic studies in particular demonstrate a sharp drop in exhaust fume diffusion beyond a certain distance from the roadside (Zhu *et al.* (2002)).

In many scenarios, data locations, such as monitoring stations in air pollution studies, are effec-

tively fixed. In studies that record the changing positions of individuals over time, there is scope for considerable error in the recorded locations. This was the original motivation for our work. Allied to methodological differences, this provides a contrast to the earlier work by Gabrosek & Cressie (2002) in which error in the prediction location was not considered.

As a postscript, we note that under very precise circumstances (Appendix B.5), the problems of positional error in data and prediction locations can be regarded as theoretically equivalent. In practice, though, they are rather different, and arise from distinct applications.

5.3 Paper 3

In this paper we considered the effect of multiple outcome variables Y_j on the basic geostatistical model (4.1), concentrating on the bivariate case. We illustrated three key approaches, the linear model of coregionalisation (LCM), the common component model (CCM) and the kernel convolution method, using data-sets relating to soil chemistry and radon exposure. We also demonstrated how kernel convolution can be used to approximate the CCM, as a computationally cheap alternative for large data-sets. This is explained further in Appendix C.1.

One of the main new findings of this paper is that a suitable parameterisation of kernels of Matérn form can often be found to approximate a given set of auto- and cross-covariance functions via kernel convolution. This is consistent with the general theory that such kernels exist, but that provides no means by which they might be found (Xia & Gelfand (2005)).

A second conclusion is that the common component-type models are well-suited to modelling stationary Gaussian processes. Three key considerations suggest that this is a fruitful class of models.

Firstly, they have the desirable property that the spatially continuous phenomenon $S = (S_1, S_2)$ characterised by the covariance functions of the components of the CCM is indeed a valid bivariate process, in the sense that it has a positive definite covariance function. This is described in Section 4.2. All members of the 'constructive' class of models, which also contains the LCM, share this property.

Secondly, and importantly, the common component model is sufficiently rich to allow the specification of two auto-covariance functions and a cross-covariance function for the two components.

This is a key consideration as in general there appears to be no compelling reason to assume that the functional forms of any two of these are the same. With this increased flexibility, however, it may become difficult to estimate parameters precisely, and careful consideration of the parameterisation of such models is required in order to maintain parameter interpretation. The identifiability problem may not greatly affect prediction, as we found for the soil data example in Paper 3; this may explain why the linear model of coregionalisation has historically been the model of choice in multivariate geostatistics, for which the primary goal is often point prediction.

Thirdly, the common component model is easily adapted to allow functional constraints between the two components of S, and includes as a special case the single component model (SCM). It is therefore well-suited for modelling environmental exposures, for which the physical nature of the exposure often dictates model choices. Examples include the radon analysis of Paper 3 and the analysis of particulate matter concentrations of Calder (2008).

In the radon analysis, we eventually settled on a form of the SCM, in which a single process S_0 was common to both components (bedroom and basement exposure) and an additive second, independent process S_1 affected only the bedroom component. This was motivated by the physical process of radon generation, in which the gas originates in the ground and enters a house primarily via its basement.

We included an extra parameter ζ to model the correlation between the nugget effects of the two components at a common location. This was necessary in order to allow for the large within-house correlation of radon measurements, and allows vastly improved prediction at locations at which Y_2 was unobserved.

Our estimate of ζ is consistent with the correlation found in other studies of radon distribution. For example, Zhu *et al.* (1998) report a correlation of 0.68 between log-radon levels in the cellar and the first floor in a study in Belgium. The high estimate of ζ provides evidence that a large proportion of this nugget effect is not error related to the measurement of radon *per se*. It seems likely to represent either extremely short-range correlation in Winnipeg's radon surface or, more probably, house-specific factors (Hunter *et al.* (2009)).

In our analysis we did not allow for such factors directly, as they were not available for our application. The one available covariate was altitude. Theoretically, radon concentrations are

inversely related to height above sea level because of the diffusion pattern of the gas from the Earth's crust and as it is denser than air (Wilkening (1990)). Barros-Dios *et al.* (2007) demonstrate this in an observational study. However, altitude varies rather little over the inner Winnipeg area and showed no appreciable association with radon in this study (Appendix C.3). Any appropriate covariates would be included as fixed effects in the radon model, which may help to simplify the structure of the S- and Z-processes. Paper 1 (Fanshawe *et al.* (2008)) shows the value of including such covariates in models for environmental exposures.

It is interesting to consider extensions of the suggested models to allow for non-Gaussian processes, non-stationary processes, and multivariate processes of dimension greater than two. Here we briefly discuss each topic, and suggest possibilities for future research.

Non-Gaussian Processes

One avenue for further work in modelling bivariate non-Gaussian processes follows from the discussion of generalised linear geostatistical models in papers such as Diggle *et al.* (1998), Section 3. In a bivariate extension, one might consider the process $S = (S_1, S_2)$, itself modelled using one of the methods described in Paper 3.

A possible model is to treat Y_{ij} $(i = 1, ..., n_j,\ j = 1, 2)$ as conditionally independent, given S, with some distribution $f(y_j; M_{ij})$. Here, M_{ij} is the conditional expectation $\mathrm{E}[Y_{ij}|S(x_{ij})]$, and $h_j(M_{ij}) = d(x_{ij})^T\beta + S_j(x_{ij})$, where the h_j are link functions to be specified. An initial analysis for data arising from similar distributions might assume $h_1 = h_2$.

Non-Stationary Processes

Section 4.2.4 contains a discussion of existing methods for modelling non-stationary geostatistical data. Two strategies appear to have been used most in applications. Both extend the kernel convolution method described above by generalising the form of the convolved process W.

The first is an adaptation of the kernel convolution approach, in which W is not merely white noise, but has its own correlation structure (Lee *et al.* (2005)). In general, this produces a non-stationary process even if W is itself a stationary process. The second allows W to itself be represented in the form of a kernel convolution, usually a convolution of an additional white noise process (Fuentes (2002)).

A further idea, which has received relatively little attention in the multivariate geostatistics literature, is that of spatial deformation, introduced by Sampson & Guttorp (1992). The method relies on assuming a stationary model for data located in a spatial domain that has undergone a transformation relative to the original coordinate system. So if x and y are spatial locations, but the covariance function $\gamma(x, y)$ is not stationary, one seeks a function f such that a new covariance function $\gamma'(f(x), f(y))$, in the transformed coordinate system, is stationary. Sampson & Guttorp (1992) choose such a function f using a combination of multidimensional scaling and fitting a thin-plate spline.

This idea has been developed for both spatial and spatio-temporal modelling (Dryden *et al.* (2005)). For bivariate geostatistical data at locations $x_{1,\cdot}$ and $x_{2,\cdot}$, it might be adapted further by finding three functions f_{11}, f_{22} and f_{12} such that the two auto-covariance functions and the cross-covariance function, i.e. the functions

$$\gamma_{11}(f_{11}(x_{1,i}), f_{11}(x_{1,j}))$$
$$\gamma_{22}(f_{22}(x_{2,i}), f_{22}(x_{2,j}))$$
$$\gamma_{12}(f_{12}(x_{1,i}), f_{12}(x_{1,j}))$$

are all stationary for any i, j. This suggestion clearly needs further work to find suitable f-functions and to ensure that a valid bivariate process emerges.

Multivariate Processes

In principle, an analysis for multivariate processes of dimension $d > 2$ might proceed along similar lines to a bivariate analysis. Issues such as practical identifiability, however, are likely to become more pronounced in the multivariate case: in the LCM, for example, even using a lower-triangular restriction on the multiplying matrices as in (4.4) leaves $pd(d+1)/2$ variance parameters. Considerable care is needed to obtain a realistic yet parsimonious model, especially for large d.

The natural extension of the CCM requires $d+1$ processes $S_0, S_1, ..., S_d$ to model a multivariate

process of dimension d. The analogue of (4.6) in the multivariate setting is

$$\begin{pmatrix} S_1^+(x) \\ S_2^+(x) \\ \vdots \\ S_d^+(x) \end{pmatrix} = \begin{pmatrix} \sigma_{01} & \sigma_1 & 0 & \dots & 0 \\ \sigma_{02} & 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{0d} & 0 & 0 & \dots & \sigma_d \end{pmatrix} \begin{pmatrix} S_0(x) \\ S_1(x) \\ S_2(x) \\ \vdots \\ S_d(x) \end{pmatrix}.$$

This has fewer variance parameters than the LCM, but further simplification may be useful. For example, physical reasons might suggest including shared processes that influence only subsets of $\{S_1^+, ..., S_d^+\}$.

For the kernel convolution method, we have described in Paper 3 a procedure for determining approximate forms of kernels of the Matérn family. These kernels are selected so that, after convolution, they approximate a pre-specified set of auto- and cross-covariance functions. This method also holds for $d > 2$, and in general $d+1$ kernels would be required. However, optimising over an increasingly large set of kernel parameters may become burdensome for large d, and this may force the analyst to specify some of the kernels *a priori*.

One final research topic in this area relates to the usefulness of multivariate, as opposed to univariate, modelling in the geostatistical setting. Our work, other anecdotal evidence (e.g. Webster & Oliver (2000), Chapter 9), and analogies with the corresponding problem in multivariate linear modelling all suggest that the most benefit is likely to accrue when Y_1 and Y_2 are highly correlated but one component is sparsely sampled, but this is not widely discussed in the geostatistics literature. How to document this observation with numerical evidence would be another worthwhile future research question.

References

Axelrod, D., Lee Davis, D., Hajek, R.A., & Jones, L.A. 2001. It's time to rethink dose: the case for combining cancer and birth and developmental defects. *Environmental Health Perspectives*, **109**, A246–A249.

Barros-Dios, J.M., Ruano-Ravina, A., Gastelu-Iturri, J., & Figueiras, A. 2007. Factors underlying residential radon concentrations: results from Galicia, Spain. *Environmental Research*, **103**, 185–190.

Calder, C.A. 2008. A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics*, **19**, 39–48.

Carroll, R.J., Ruppert, D., Stefanski, L.A., & Crainiceanu, C.M. 2006. *Measurement error in nonlinear models*. Boca Raton, Chapman and Hall/CRC.

Cox, L.H. 2000. Statistical issues in the study of air pollution involving airborne particulate matter. *Environmetrics*, **11**, 611–626.

Cressie, N., & Kornak, J. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, **18**, 436–456.

Diggle, P.J., Tawn, J.A., & Moyeed, R.A. 1998. Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.

Dryden, I.L., Markus, L., Taylor, C.C., & Kovacs, J. 2005. Non-stationary spatiotemporal analysis of karst water levels. *Applied Statistics*, **54**, 673–690.

Evans, M., & Swartz, T. 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **10**, 254–272.

Fanshawe, T.R., Diggle, P.J., Rushton, S., Sanderson, R., Lurz, P.W.W., Glinianaia, S.V., Pearce, M.S., Parker, L., Charlton, M., & Pless-Mulloli, T. 2008. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics*, **19**, 549–566.

Fuentes, M. 2002. Spectral methods for nonstationary spatial processes. *Biometrika*, **89**, 197–210.

Gabrosek, J., & Cressie, N. 2002. The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, **34**, 262–285.

Glasbey, C.A., Graham, R., & Hunter, A.G.M. 2001. Spatio-temporal variability of solar energy across a region: a statistical modelling approach. *Solar Energy*, **70**, 373–381.

Gneiting, T., Genton, M.G., & Guttorp, P. 2007. *Geostatistical space-time models, stationarity, separability, and full symmetry*. In: Statistical methods for spatio-temporal systems. Chapman and Hall/CRC. Chap. 4.

Gryparis, A., Paciorek, C.J., Zeka, A., Schwartz, J., & Coull, B.A. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, **10**, 258–274.

Haas, T.C. 1995. Local prediction of a spatiotemporal process with an application to wet sulfate decomposition. *Journal of the American Statistical Association*, **90**, 1189–1199.

Handcock, M.S., & Wallis, J.R. 1994. An approach to statistical spatial-temporal modelling of meteorological fields. *Journal of the American Statistical Association*, **89**, 368–378.

Holland, D.M., De Oliveira, V., Cox, L.H., & Smith, R.L. 1999. Estimation of regional trends in sulfur dioxide over the eastern United States. *Applied Statistics*, **48**, 345–362.

Hunter, N., Muirhead, C.R., Miles, J.C.H., & Appleton, J.D. 2009. Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. *Radiation Protection Dosimetry*, **136**, 17–22.

Lee, H.K.H., Higdon, D.M., Calder, C.A., & Holloman, C.H. 2005. Efficient models for correlated data via convolutions of intrinsic processes. *Statistical Modelling*, **5**, 53–74.

Meiring, W., Guttorp, P., & Sampson, P.D. 1998. Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, **5**, 197–222.

Sampson, P.D., & Guttorp, P. 1992. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108–119.

Smith, R.L., Kolenikov, S., & Cox, L.H. 2003. Spatio-temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research - Atmospheres*, **108**, D24 9004.

Šrám, R.J., Binková, B., Dejmek, J., & Bobak, M. 2005. Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental Health Perspectives*, **113**, 375–382.

Wakefield, J., & Shaddick, G. 2006. Health-exposure modeling and the ecological fallacy. *Biostatistics*, **7**, 438–455.

Warnes, J.J., & Ripley, B.D. 1987. Problems with likelihood estimation of covarariance functions of spatial Gaussian processes. *Biometrika*, **74**, 640–642.

Webster, R., & Oliver, M.A. 2000. *Geostatistics for environmental scientists*. New York, Wiley.

Wilkening, M. 1990. *Radon in the environment*. Amsterdam, Elsevier.

Xia, G., & Gelfand, A.E. 2005. *Stationary process approximation for the analysis of large spatial datasets*. Tech. rept. Institute of Statistics and Decision Sciences, Duke University.

Zhang, H. 2004. Inconsistent estimation and asymptopically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

Zhu, H.-C., Charlet, J.M, & Tondeur, F. 1998. Geological controls to the indoor radon distribution in southern Belgium. *The Science of the Total Environment*, **220**, 195–214.

Zhu, L., Carlin, B.P., & Gelfand, A.E. 2003. Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, **14**, 537–557.

Zhu, Y., Hinds, W.C., Kim, S., & Sioutas, C. 2002. Concentration and size distribution of ultrafine particles near a major highway. *Journal of the Air & Waste Management Association*, **52**, 1032–1042.

Zidek, J.V., Sun, L., Le, N., & Özkaynak, H. 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics*, **13**, 595–613.

Appendix A

Appendices for Paper 1

A.1 Fitting the Dynamic Model

In this section we provide further details on the model-fitting procedure for the dynamic model described in Section 2.3.3 of Paper 1, elaborating on the statement "the dynamic model can be fitted either by direct maximisation of the likelihood function, or via a Kalman filter followed by Kalman smoothing".

We concentrate primarily on the technique of direct maximisation of the likelihood function, as this method is not easily found in the literature in the context of dynamic model-fitting. The model (2.3) under consideration is

$$Y_t = \alpha + \beta t + \gamma d_t + A_t \cos(\omega t) + B_t \sin(\omega t) + U_t, \tag{A.1}$$

where

$$\begin{aligned} A_t | A_{t-1} &\sim \mathrm{N}(A_{t-1}, \sigma_A^2) \\ B_t | B_{t-1} &\sim \mathrm{N}(B_{t-1}, \sigma_B^2) \\ U_t &\sim \text{IID } \mathrm{N}(0, \sigma_U^2) \end{aligned} \tag{A.2}$$

Equivalently, we can write

$$\begin{aligned} A_t &= A_{t-1} + \eta_t \\ B_t &= B_{t-1} + \zeta_t, \end{aligned}$$

where $\eta_t \sim$ IID $\mathrm{N}(0, \sigma_A^2)$ and $\zeta_t \sim$ IID $\mathrm{N}(0, \sigma_B^2)$. For convenience, we use here the notation Y_t for the value at time t of the log-transformed black smoke levels, averaged over all active monitors at each instant in time (Section 2.3.3 uses $\bar{Y}_t$). Let Y denote the time series $(Y_1, ..., Y_n)$, and A and B the corresponding time series for the two dynamic coefficients. As each Y_t in (A.1) is composed of a sum of independent multivariate Normal random variables, $Y \sim \mathrm{MVN}(\mu, \Sigma)$ for some mean vector μ and covariance matrix Σ.

Given starting values $A_0 = a_0$ and $B_0 = b_0$ of the series A and B, for each $t > 0$, $\mathrm{E}(A_t) = a_0 \cos(\omega t)$ and $\mathrm{E}(B_t) = b_0 \sin(\omega t)$, so $\mathrm{E}(Y_t) = \alpha + \beta t + \gamma d_t + a_0 \cos(\omega t) + b_0 \sin(\omega t)$. Then

$$\begin{aligned} A_t \cos(\omega t) &= (A_{t-1} + \eta_t) \cos(\omega t) \\ &= (A_{t-2} + \eta_t + \eta_{t-1}) \cos(\omega t) \\ &\vdots \\ &= (a_0 + \sum_{i=1}^{t} \eta_i) \cos(\omega t) \end{aligned}$$

and for each i, $\eta_i \cos(\omega t) \sim \mathrm{N}(0, \sigma_A^2 \cos^2(\omega t))$ independently, so

$$A_t \cos(\omega t) \sim \mathrm{N}(a_0 \cos(\omega t), t\sigma_A^2 \cos^2(\omega t)).$$

Similarly, $B_t \sin(\omega t) \sim \mathrm{N}(b_0 \sin(\omega t), t\sigma_B^2 \sin^2(\omega t))$, and as A_t, B_t and U_t are, by assumption, mutually independent for each t,

$$Y_t \sim \mathrm{N}(\alpha + \beta t + \gamma d_t + a_0 \cos(\omega t) + b_0 \sin(\omega t), t(\sigma_A^2 \cos^2(\omega t) + \sigma_B^2 \sin^2(\omega t)) + \sigma_U^2). \qquad \text{(A.3)}$$

Also, for any s and t,

$$\begin{aligned} \mathrm{Cov}(Y_s, Y_t) &= \mathrm{Cov}(A_s \cos(\omega s) + B_s \sin(\omega s), A_t \cos(\omega t) + B_t \sin(\omega t)) \\ &= \cos(\omega s) \cos(\omega t) \mathrm{Cov}(A_s, A_t) + \sin(\omega s) \sin(\omega t) \mathrm{Cov}(B_s, B_t) \\ &= \cos(\omega s) \cos(\omega t) \mathrm{Cov}\left(\sum_{i=1}^{s} \eta_i, \sum_{i=1}^{t} \eta_i\right) + \sin(\omega s) \sin(\omega t) \mathrm{Cov}\left(\sum_{i=1}^{s} \zeta_i, \sum_{i=1}^{t} \zeta_i\right) \\ &= \min(s, t)[\sigma_A^2 \cos(\omega s) \cos(\omega t) + \sigma_B^2 \sin(\omega s) \sin(\omega t)] \qquad \text{(A.4)} \end{aligned}$$

Thus, for any t, μ_t and $\Sigma_{t,t}$ are as in (A.3), and for $s \neq t$, $\Sigma_{s,t}$ is given by (A.4). Maximum likelihood estimates $\hat{\mu}$ and $\hat{\Sigma}$ are then found by minimizing, with respect to μ and Σ,

$$-2l(\mu, \Sigma) = \text{const.} + \log(|\Sigma|) + (Y - \mu)^T \Sigma^{-1} (Y - \mu)$$

In order to apply the second stage of the modelling procedure, described in Section 2.3.4, we need to calculate predicted values of Y obtained from the dynamic model. These can be obtained by substituting the maximum likelihood estimates of the parameters into (A.1):

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta} t + \hat{\gamma} d_t + \hat{A}_t \cos(\omega t) + \hat{B}_t \sin(\omega t) \tag{A.5}$$

To obtain values of $\hat{A}_t$ and $\hat{B}_t$ to use in (A.5), we use

$$\begin{bmatrix} A \\ B \\ Y \end{bmatrix} \sim \text{MVN} \left\{ \begin{bmatrix} a_0 \cos(\omega \mathbf{t}) \\ b_0 \sin(\omega \mathbf{t}) \\ \mu_{\mathbf{t}} \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & 0 & \Sigma_{AY} \\ 0 & \Sigma_{BB} & \Sigma_{BY} \\ \Sigma_{AY}^T & \Sigma_{BY}^T & \Sigma \end{bmatrix} \right\}, \tag{A.6}$$

where $\mu_{\mathbf{t}} = \alpha \mathbf{1} + \beta \mathbf{t} + \gamma d_{\mathbf{t}} + a_0 \cos(\omega \mathbf{t}) + b_0 \sin(\omega \mathbf{t})$, $\mathbf{1}$ is a vector of ones of length n, $\mathbf{t} = (t_1, ..., t_n)'$ and the covariance matrix is composed of blocks, each of dimension $n \times n$, defined by

$$\begin{aligned} (\Sigma_{AA})_{s,t} &= \text{Cov}(A_s, A_t) = \min(s,t)\sigma_A^2 \\ (\Sigma_{BB})_{s,t} &= \text{Cov}(B_s, B_t) = \min(s,t)\sigma_B^2 \\ (\Sigma_{AY})_{s,t} &= \text{Cov}(A_s, Y_t) = \min(s,t)\sigma_A^2 \cos(\omega t) \\ (\Sigma_{BY})_{s,t} &= \text{Cov}(B_s, Y_t) = \min(s,t)\sigma_B^2 \cos(\omega t). \end{aligned}$$

Estimates of matrices Σ_{AA}, Σ_{BB}, Σ_{AY}, Σ_{BY} and Σ can be computed by direct substitution of the maximum likelihood estimates $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$ and $\hat{\sigma}_U^2$.

Standard properties of the multivariate Normal distribution enable estimates of A and B to be calculated:

$$\begin{pmatrix} \hat{A} \\ \hat{B} \end{pmatrix} = \begin{pmatrix} a_0 \cos(\omega \mathbf{t}) \\ b_0 \sin(\omega \mathbf{t}) \end{pmatrix} + \begin{pmatrix} \hat{\Sigma}_{AY} \\ \hat{\Sigma}_{BY} \end{pmatrix} \hat{\Sigma}^{-1} (Y - \hat{\mu}_{\mathbf{t}})$$

Likelihood maximisation in R (R Development Core Team (2005)) was performed using the general-purpose optimisation function `optim`, using as starting values a_0 and b_0 the maximum likelihood estimates from the static model (in any case, predicted values of A and B were found to be robust to the choice of starting value owing to the length of the time series in this application). Convergence of maximum likelihood estimates took approximately four hours of computation time, the most time-consuming step being the inversion of the $n \times n$ matrix Σ.

Computation time can be reduced substantially by instead using the Kalman filter, as described by West & Harrison (1997), followed by Kalman smoothing. This quicker process can be implemented in R using the functions `kfilter` and `smoother` in the `sspir` package (Dethlefsen & Lundbye-Christensen (2006)). The Kalman filter is a widely-used method for estimating the parameters of state space models, and in view of the large literature on the subject and its application to time series models, we present only a brief summary here.

With notation similar to that used in the tutorial paper of Meinhold & Singpurwalla (1983), we recast model (A.1-A.3) in the general matrix form

$$\begin{aligned} Y_t &= F_t\theta_t + \eta_t \\ \theta_t &= G_t\theta_{t-1} + \zeta_t \end{aligned}$$

where for our application F is the design matrix whose ith row consists of the values of the covariates at time i, G is the identity matrix, the η_t are independent univariate $\mathrm{N}(0, \sigma_V^2)$ random variables and the ζ_t are independent $\mathrm{MVN}(0, \Psi_t)$, where the first three diagonal elements of Ψ_t are constrained to be zero for each t . These elements correspond to the three fixed effects in (A.1).

Kalman filtering is a recursive procedure that relies on repeated evaluation of a version of Bayes's Theorem,

$$[\theta_t|Y_1, ..., Y_t] \propto [Y_t|\theta_t, Y_1, ..., Y_{t-1}][\theta_t|Y_1, ..., Y_{t-1}],$$

to update the estimated mean and covariance matrix of the parameter vector θ_t for increasing values of t. At each stage the required conditional distributions are multivariate Normal, a consequence of the Gaussian assumption in (A.3), which is the key observation in reducing the complexity of the computation. Meinhold & Singpurwalla (1983) derive the exact form of these conditional distributions.

The improved fit of the dynamic model relative to the static model in which $A_t = A$, $B_t = B$ for all t is shown in Figure 2.4, in which both long-term and short-term autocorrelation are seen to be largely eliminated.

A.2 Additional Covariate Information

In Section 2.3.4, we described five covariates considered for modelling the spatio-temporal component of variation in black smoke levels at monitor locations. We add here several supplementary figures relating to covariate selection that were not included in Paper 1 because of lack of space.

Figure A.1, similar to Figure 2.5, justifies the decision to use covariate w_2 (distance to industry) in the model, and to exclude covariate w_5 (area of industry within a 500-metre radius). There is an inverse relationship between the average residual from the dynamic model and distance to industry, but no clear relationship with area of industry.

Figure A.2 refers to covariate w_3, and indicates which monitors were assigned residential status. It shows the expected pattern: most non-residential areas of Newcastle-upon-Tyne lie in the centre of the city and along the River Tyne at the southern edge of the study region.

Figure A.3 refers to covariate w_4, and indicates the years in which the Clean Air Act was implemented across sub-areas of the city. Areas with missing information for this covariate (shown as white in the figure) were assigned a date of 1978, by which time the whole of the city was officially 'smoke-free'.

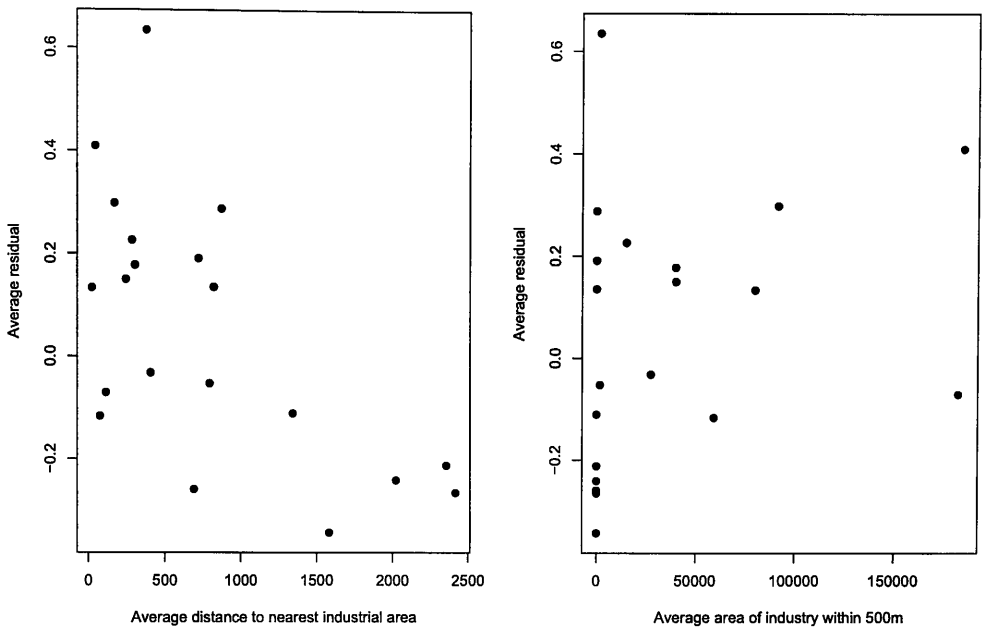


Figure A.1: Monitor-specific average residual from the dynamic model, plotted against (a) distance from monitor to nearest industrial area; and (b) area within 500m of monitor used in industry.

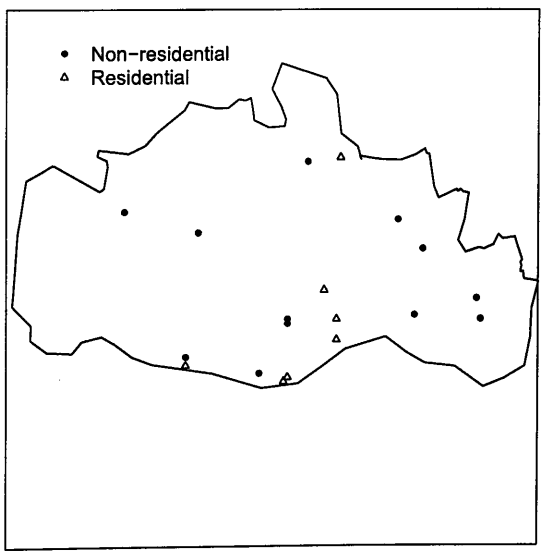


Figure A.2: Map of study region showing residential status of the twenty air pollution monitors.

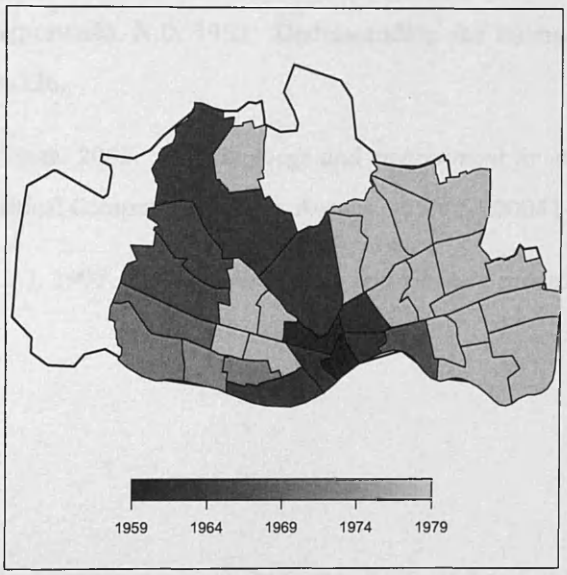


Figure A.3: Map of study region showing date of implementation of the Clean Air Act across sub-areas of the city. Areas shown in white were developed after the nominal end of Clean Air Act implementation.

References

Dethlefsen, C., & Lundbye-Christensen, S. 2006. Formulating state space models in R with focus on longitudinal regression models. *Journal of Statistical Software*, **16**, 1–15.

Meinhold, R.J., & Singpurwalla, N.D. 1983. Understanding the Kalman Filter. *The American Statistician*, **37**, 123–126.

R Development Core Team. 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

West, M., & Harrison, P.J. 1997. *Bayesian forecasting and dynamic models*. New York, Springer-Verlag.

Appendix B

Appendices for Paper 2

B.1 Standard Results in Geostatistical Prediction

In this section we derive some standard results used in geostatistical prediction problems, used throughout Paper 2.

B.1.1 The Distribution of the Minimum Mean Square Error Predictor

Consider the geostatistical model (3.1) with constant mean μ:

$$Y_i = \mu + S(x_i) + Z_i \ : \ i = 1, ..., n. \tag{B.1}$$

Here, $S(x)$ is a zero-mean Gaussian process with variance σ^2 and correlation function ρ and the Z_i are independent $\mathrm{N}(0, \tau^2)$ errors. Let the target for prediction be $T = (S(x_1^*), S(x_2^*))'$. Here we derive the 'best' point predictor for T, i.e. $\mathrm{E}[T|Y]$, and its variance $\mathrm{Var}[T|Y]$.

T follows a bivariate Normal distribution with mean vector $\mu \mathbf{1}_2$ and covariance matrix

$$\sigma^2 V^* = \begin{bmatrix} \sigma^2 & \sigma^2 \rho(\|x_1^* - x_2^*\|) \\ \sigma^2 \rho(\|x_1^* - x_2^*\|) & \sigma^2 \end{bmatrix}.$$

Thus (T, Y) has a multivariate Normal distribution with mean $\mu \mathbf{1}_{n+2}$ and covariance matrix

$$\begin{bmatrix} \sigma^2 V^* & \sigma^2 r' \\ \sigma^2 r & \sigma^2 V \end{bmatrix},$$

where r is an $n \times 2$ matrix with $r_{ij} = \rho(\|x_j^* - x_i\|)$ for $j = 1, 2$ and $i = 1, ..., n$, and $V_{ij} = \rho(\|x_i - x_j\|)$ for $i, j = 1, ..., n$.

Standard properties of the multivariate Normal distribution show that $T|Y$ has a multivariate Normal distribution with

$$\mathrm{E}[T|Y] = \hat{T} = \mu \mathbf{1}_2 + r' V^{-1}(Y - \mu \mathbf{1}_n) \tag{B.2}$$

and

$$\mathrm{Var}[T|Y] = \sigma^2(V^* - r' V^{-1} r), \tag{B.3}$$

whose off-diagonal component thus provides the covariance between the predictions at the two locations, used in the subsection on joint prediction in Section 3.3.2. These results easily generalise to the case in which there are more than two prediction locations.

B.1.2 The Generalised Least Squares Estimator of the Mean

In 'simple kriging', μ is replaced by its method-of-moments estimator $n^{-1} \sum Y_i$ in (B.2). In 'ordinary kriging', the Generalised Least Squares (GLS) estimator $\hat{\mu}$ is used. This estimator is of the form $b'Y$, where b is a vector chosen to minimise $\mathrm{Var}(\hat{T} - T)$ under the constraint $\mathrm{E}(\hat{T} - T) = 0$. Here we derive the form of this estimator.

Consider the special case in which prediction is required at a single location, in which case $V^* = 1$ in (B.3). Let $a' = r' V^{-1}$ and $s = a' \mathbf{1}_n = r' V^{-1} \mathbf{1}$, where $\mathbf{1}$ is a vector of ones of length n. The unbiasedness constraint implies that

$$\begin{aligned} 0 &= \mathrm{E}(b'Y + a'Y - (a'\mathbf{1})b'Y) - \mu \\ &= \mu(b'\mathbf{1} + a'\mathbf{1} - (a'\mathbf{1})(b'\mathbf{1}) - 1), \end{aligned}$$

and so we must have $\sum b_i = 1$, i.e. $\hat{\mu}$ is a weighted mean of the data values.

We have $\mathrm{Var}(T) = \sigma^2$,

$$\begin{aligned} \mathrm{Var}(\hat{T}) &= \mathrm{Var}((a + b - sb)'Y) \\ &= \sigma^2 (a + b - sb)' V (a + b - sb) \end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(T,\hat{T}) &= \mathrm{Cov}(T,(a+b-sb)'Y)\\ &= \sigma^2(a+b-sb)'r,\end{aligned}$$

so

$$\mathrm{Var}(\hat{T}-T) = \sigma^2(1-2(a+(1-s)b)'r+(a+(1-s)b)'V(a+(1-s)b)).$$

Let $\tilde{V} = \mathrm{Var}(\hat{T}-T) - \lambda(b-1)$, where λ is a Lagrange multiplier. Then, as $Va = r$,

$$\begin{aligned}\frac{d\tilde{V}}{db} &= 2\sigma^2(1-s)(-r+V(a+(1-s)b)) - \lambda 1\\ &= 2\sigma^2(1-s)^2 b - \lambda 1.\end{aligned}$$

This is zero when

$$b = \frac{\lambda}{2\sigma^2(1-s)^2}V^{-1}1,$$

which corresponds to a minimum of $\tilde{V}$. To ensure $1'b = 1$, we need

$$\lambda = \frac{2\sigma^2(1-s)^2}{1'V^{-1}1},$$

from which

$$b = \frac{V^{-1}1}{1'V^{-1}1}.$$

Thus the GLS estimator is

$$\hat{\mu} = \frac{1'V^{-1}Y}{1'V^{-1}1}.$$

B.1.3 Mean and Variance of the Generalised Least Squares Estimator

The estimate $\hat{T} = \hat{\mu} + r'V^{-1}(Y-\hat{\mu}1) = b'Y + a'Y - (a'1)b'Y$, where $\hat{\mu}$ is the GLS estimator of the mean, and a and b are as previously defined, is by construction unbiased ($\mathrm{E}(\hat{T}-T) = 0$). To

find its variance, let $\alpha = 1'V^{-1}1$. Then $\mathrm{Var}(T) = \sigma^2$,

$$\begin{aligned}
\mathrm{Var}(\hat{T}) &= \mathrm{Var}(\alpha^{-1}1'V^{-1}Y + r'V^{-1}Y - \alpha^{-1}(r'V^{-1}1)(1'V^{-1}Y)) \\
&= (\alpha^{-1}1'V^{-1} + r'V^{-1} - \alpha^{-1}(r'V^{-1}1)1'V^{-1})\sigma^2 V(\alpha^{-1}1'V^{-1} + \\
&\quad r'V^{-1} - \alpha^{-1}(r'V^{-1}1)1'V^{-1})' \\
&= \sigma^2\alpha^{-2}(\alpha + \alpha 1'V^{-1}r - \alpha 1'V^{-1}r + \alpha 1'V^{-1}r + \alpha^2 r'V^{-1}r - \\
&\quad \alpha(1V^{-1}r)^2 - \alpha 1'V^{-1}r - \alpha(1'V^{-1}r)^2 + \alpha(1'V^{-1}r)^2) \\
&= \sigma^2\alpha^{-1}(1 - (1'V^{-1}r)^2 + \alpha r'V^{-1}r)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(T, \hat{T}) &= \mathrm{Cov}(T, (\alpha^{-1}1'V^{-1} + r'V^{-1} - \alpha^{-1}r'V^{-1}1(1'V^{-1}))Y) \\
&= \sigma^2\alpha^{-1}(1V^{-1}r + \alpha r'V^{-1}r - (1V^{-1}r)^2),
\end{aligned}$$

so

$$\begin{aligned}
\mathrm{Var}(\hat{T} - T) &= \mathrm{Var}(T) + \mathrm{Var}(\hat{T}) - 2\mathrm{Cov}(T, \hat{T}) \\
&= \sigma^2(1 - r'V^{-1}r) + \sigma^2\alpha^{-1}(1 - (1'V^{-1}r))^2,
\end{aligned}$$

the first term of this expression being equal to 'simple kriging' variance (B.3).

For prediction at two locations x_1^* and x_2^*, partition the matrix r^* into its two columns r_1 and r_2. Then similar steps to the above show that

$$\begin{aligned}
\mathrm{Cov}(\hat{T}_1, \hat{T}_2) &= \mathrm{Cov}(\hat{\mu} + (r_1)'V^{-1}(Y - \hat{\mu}1), \hat{\mu} + (r_2)'V^{-1}(Y - \hat{\mu}1)) \\
&= \sigma^2\alpha^{-1}(1 + \alpha(r_1'V^{-1}r_2) - (1'V^{-1}r_1)(1'V^{-1}r_2))
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(\hat{T}_1 - T_1, \hat{T}_2 - T_2) &= \mathrm{Cov}(\hat{T}_1, \hat{T}_2) - \mathrm{Cov}(\hat{T}_1, T_2) - \mathrm{Cov}(\hat{T}_2, T_1) + \mathrm{Cov}(T_1, T_2) \\
&= \sigma^2\alpha^{-1}(1 - (1'V^{-1}r_1)(1'V^{-1}r_2) + \alpha(r_1'V^{-1}r_2)) - \\
&\quad \sigma^2\alpha^{-1}(1'V^{-1}r_2 + \alpha r_1'V^{-1}r_2 - (1'V^{-1}r_1)(1'V^{-1}r_2)) - \\
&\quad \sigma^2\alpha^{-1}(1'V^{-1}r_1 + \alpha r_1'V^{-1}r_2 - (1'V^{-1}r_1)(1'V^{-1}r_2)) + \sigma^2 r_{12} \\
&= \sigma^2(r_{12} - r_1'V^{-1}r_2) + \sigma^2\alpha^{-1}(1 - 1'V^{-1}r_1)(1 - 1'V^{-1}r_2),
\end{aligned}$$

where $r_{12} = \rho(\|x_1^* - x_2^*\|)$. The first term is equal to the covariance of the 'simple kriging' predictor, the off-diagonal element of (B.3), while the second is of similar form to the corresponding term in the expression for $\mathrm{Var}(\hat{T} - T)$ for prediction at a single location.

To summarise, in matrix notation we have

$$\mathrm{Var}(\hat{T} - T) = \sigma^2(V^* - (r^*)'V^{-1}r^*) + \sigma^2\alpha^{-1}(1 - \mathbf{1}'V^{-1}r^*)'(1 - \mathbf{1}'V^{-1}r^*).$$

B.2 First and Second Moments of the Predictive Distribution in the Presence of Positional Error

In the appendix in Paper 2, we presented several results to illustrate how the mean and variance of the predictive distribution change when there is positional error in the prediction location (but not in the data locations). In this appendix we give further similar results and their derivations, covering both the scenario of a one-dimensional and two-dimensional domain. In Appendix B.2.1 we assume that the underlying exposure surface S is known precisely. Although this scenario may be unrealistic in practice, results obtained under this assumption resemble and motivate similar results given in the appendix in Paper 2 and Appendix B.2.2, in which a correlation function ρ is assumed and a geostatistical model is fitted to the data.

We assume that all necessary derivatives exist, and that all error distributions are bivariate Normal with zero mean and covariance matrix $\gamma^2 I_2$. In two dimensions, we write the two components of the error distribution ϵ as (ϵ_x, ϵ_y). For the error distributions $\epsilon_1 = (\epsilon_{x_1}, \epsilon_{y_1})$ and $\epsilon_2 = (\epsilon_{x_2}, \epsilon_{y_2})$ we will consider separately the cases when ϵ_1 and ϵ_2 are uncorrelated and when $\mathrm{Cov}(\epsilon_{x_1}, \epsilon_{x_2}) = \mathrm{Cov}(\epsilon_{y_1}, \epsilon_{y_2}) = \alpha\gamma^2$ (i.e. $\mathrm{Corr}(\epsilon_{x_1}, \epsilon_{x_2}) = \mathrm{Corr}(\epsilon_{y_1}, \epsilon_{y_2}) = \alpha$). The derivations below use the following basic properties.

For a univariate $\mathrm{N}(0, \gamma^2)$ distribution ϵ, for any positive integers n and m:

$$\mathrm{E}(\epsilon^{2n}) = ((2n-1)(2n-3)...3.1)\gamma^{2n}$$

$$\mathrm{E}(\epsilon^{2n-1}) = 0$$

$$\mathrm{Cov}(\epsilon^m, \epsilon^n) = 0 \text{ if } m + n \text{ is odd}$$

For two univariate $\mathrm{N}(0, \gamma^2)$ distributions ϵ_1 and ϵ_2 with covariance $\alpha\gamma^2$, for any positive integers

n and m:

$$\mathrm{E}[\epsilon_1^m \epsilon_2^n] = 0 \text{ if } m+n \text{ is odd}$$

$$\mathrm{E}[\epsilon_1 \epsilon_2] = \alpha\gamma^2$$

$$\mathrm{E}[\epsilon_1^3 \epsilon_2] = 3\alpha\gamma^4, \text{ so } \mathrm{Cov}[\epsilon_1^3, \epsilon_2] = 3\alpha\gamma^4$$

$$\mathrm{E}[\epsilon_1^2 \epsilon_2^2] = \gamma^4 + 2\alpha^2\gamma^4, \text{ so } \mathrm{Cov}[\epsilon_1^2, \epsilon_2^2] = 2\alpha^2\gamma^4$$

Where necessary to avoid ambiguity, in two dimensions we will write a single prediction location as $\mathbf{x} = (x, y)$ and a pair of prediction locations as $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$. Vector notation for ϵ is suppressed, but implied by the context. Subscripts on S, a, c and Q (defined later) denote partial derivatives, as in the appendix in Paper 2.

B.2.1 Exposure Surface Known

One dimension - Prediction at a single location

Using the Taylor series expansion

$$S(x+\epsilon) = S + \epsilon S_x + \frac{1}{2}\epsilon^2 S_{xx} + \ldots$$

we have

$$\mathrm{E}[S(x+\epsilon)] = S + \frac{1}{2}\gamma^2 S_{xx} + \frac{1}{8}\gamma^4 S_{xxxx} + O(\gamma^6)$$

and

$$\begin{aligned} \mathrm{Var}[S(x+\epsilon)] &= \mathrm{Var}(\epsilon) S_x^2 + \mathrm{Var}(\epsilon^2)\frac{1}{4}S_{xx}^2 + \mathrm{Cov}(\epsilon^3, \epsilon)\frac{2}{6} S_x S_{xxx} + \ldots \\ &= \gamma^2 S_x^2 + \gamma^4(\frac{1}{2}S_{xx}^2 + S_x S_{xxx}) + O(\gamma^6) \end{aligned}$$

The bias in the point prediction using the naive predictor $S(x)$ is given, to the lowest-order approximation, by $\frac{1}{2}\gamma^2 S_{xx}$, which depends on the magnitude of the error variance and the second derivative of the surface. In particular, at local maxima ($S_{xx} < 0$) the predictor $S(x)$ will tend to give an overestimate of the true exposure, and at local minima ($S_{xx} > 0$) it will tend to give an underestimate. The prediction variance will always be underestimated if the positional error is ignored, as in this case the true prediction variance would be zero if S were known precisely. The largest difference in prediction variances will occur at points where the gradient of the surface, S_x, is largest in magnitude. Appendix B.3 contains an illustrative example based on similar results in Appendix B.2.2.

One dimension - Joint prediction at two locations

If errors ϵ_1, ϵ_2 are independent then the predictions are also independent, so the problem reduces to the single-location case described above. If the errors are dependent, with the correlation structure described above, then

$$\begin{aligned}\mathrm{Cov}[S(x_1+\epsilon_1), S(x_2+\epsilon_2)] &= \mathrm{Cov}(\epsilon_1,\epsilon_2)S_{x_1}S_{x_2} + \mathrm{Cov}(\epsilon_1,\epsilon_2^3)\frac{1}{6}S_{x_1}S_{x_2x_2x_2} + \\ &\quad \mathrm{Cov}(\epsilon_1^2,\epsilon_2^2)\frac{1}{4}S_{x_1x_1}S_{x_2x_2} + \mathrm{Cov}(\epsilon_1^3,\epsilon_2)\frac{1}{6}S_{x_1x_1x_1}S_{x_2} + ... \\ &= \alpha\{\gamma^2 S_{x_1}S_{x_2} + \frac{1}{2}\gamma^4(S_{x_1x_1x_1}S_{x_2} + S_{x_1}S_{x_2x_2x_2} + \\ &\quad \alpha S_{x_1x_1}S_{x_2x_2}) + O(\gamma^6)\}\end{aligned}$$

Two dimensions - Prediction at a single location

We use the two-dimensional Taylor series expansion

$$S(\mathbf{x}+\epsilon) = S + \epsilon_x S_x + \epsilon_y S_y + \frac{1}{2}\epsilon_x^2 S_{xx} + \frac{1}{2}\epsilon_y^2 S_{yy} + \epsilon_x\epsilon_y S_{xy} + ...$$

with general term $\frac{1}{m!n!}\epsilon_x^m \epsilon_y^n S_{x^{(m)}y^{(n)}}$.

Then

$$\mathrm{E}[S(\mathbf{x}+\epsilon)] = S + \frac{1}{2}\gamma^2(S_{xx}+S_{yy}) + \frac{1}{8}\gamma^4(S_{xxxx} + 2S_{xxyy} + S_{yyyy}) + O(\gamma^6)$$

and

$$\begin{aligned}\mathrm{Var}[S(\mathbf{x}+\epsilon)] &= \gamma^2(S_x^2+S_y^2) + \gamma^4(\frac{1}{2}S_{xx}^2 + S_{xy}^2 + \frac{1}{2}S_{yy}^2 + S_xS_{xxx} + \\ &\quad S_xS_{xyy} + S_yS_{xxy} + S_yS_{yyy}) + O(\gamma^6).\end{aligned}$$

Two dimensions - Joint prediction at two locations

As in the one-dimensional case, if errors at two prediction locations are independent then the correlation between the predictions is zero. If errors are not independent, under the assumed correlation structure, $\mathrm{Cov}[S(\mathbf{x}_1+\epsilon_1), S(\mathbf{x}_2+\epsilon_2)]$ is obtained simply by adding together the terms similar to the one-dimensional case corresponding to the x- and y-directions:

$$\begin{aligned}\mathrm{Cov}[S(\mathbf{x}_1+\epsilon_1), S(\mathbf{x}_2+\epsilon_2)] &= \alpha\{\gamma^2(S_{x_1}S_{x_2} + S_{y_1} + S_{y_2}) + \frac{1}{2}\gamma^4(S_{x_1x_1x_1}S_{x_2} + S_{x_1}S_{x_2x_2x_2} + \\ &\quad S_{y_1y_1y_1}S_{y_2} + S_{y_1}S_{y_2y_2y_2} + \\ &\quad \alpha(S_{x_1x_1}S_{x_2x_2} + S_{y_1y_1}S_{y_2y_2})) + O(\gamma^6)\}\end{aligned}$$

B.2.2 Exposure Surface Unknown

In this section we give corresponding results when the surface S is not known everywhere, expanding on the results given in the appendix in Paper 2, and using the same notation. We assume the stationary Gaussian model (B.1), and let x_p be a prediction location (for convenience, the subscript is dropped in much of the succeeding derivation), $r_i = \rho(\|x_p - x_i\|)$, R be a matrix with (i,j)th element $R_{ij} = \rho(|x_i - x_j\|)$, and $Q = R^{-1}$.

From (3.8) and (3.9), in the absence of positional error, $S(x)$ has distribution $\mathrm{N}(\mu_x, \sigma_x^2)$, where

$$\mu_x = \mu + r'Q(y - \mu) \equiv \mu + \sum_{i=1}^{n} a_i r_i, \tag{B.4}$$

and

$$\sigma_x^2 = \sigma^2(1 - r'Qr) \equiv \sigma^2 \left(1 - \sum_{i,j=1}^{n} r_i(x) Q_{ij} r_j(x)\right). \tag{B.5}$$

One dimension - Prediction at a single location

Conditional on ϵ, $S(x+\epsilon) \sim \mathrm{N}(\mu_{x+\epsilon}, \sigma_{x+\epsilon}^2)$, where

$$\mu_{x+\epsilon} = \mu_x + \epsilon a_x + \frac{1}{2}\epsilon^2 a_{xx} + \ldots$$

and

$$\sigma_{x+\epsilon}^2 = \sigma_x^2 - \sigma^2(2\epsilon Q_{0,x} + \epsilon^2 Q_{x,x} + \epsilon^2 Q_{0,xx} + \frac{1}{3}\epsilon^3 Q_{0,xxx} + \epsilon^3 Q_{x,xx} + \ldots)$$

using the symmetry of Q ($Q_{i,j} = Q_{j,i}$). Thus

$$\mathrm{E}[S(x+\epsilon)] = \mathrm{E}_\epsilon[\mu_{x+\epsilon}] = \mu_x + \frac{1}{2}\gamma^2 a_{xx} + \frac{1}{8}\gamma^4 a_{xxxx} + O(\gamma^6)$$

and

$$\mathrm{Var}[S(x+\epsilon)] = \mathrm{E}_\epsilon[\sigma_{x+\epsilon}^2] + \mathrm{Var}_\epsilon[\mu_{x+\epsilon}],$$

where

$$\mathrm{E}_\epsilon[\sigma_{x+\epsilon}^2] = \sigma_x^2 - \sigma^2(\gamma^2(Q_{x,x} + Q_{0,xx}) + \gamma^4(\frac{1}{4}Q_{0,xxxx} + Q_{x,xxx} + \frac{3}{4}Q_{xx,xx}) + O(\gamma^6))$$

and

$$\mathrm{Var}_\epsilon[\mu_{x+\epsilon}] = \gamma^2 a_x^2 + \gamma^4(\frac{1}{2}a_{xx}^2 + a_x a_{xxx}) + O(\gamma^6). \tag{B.6}$$

These formulae resemble those in Appendix B.2.1; when the surface S is known everywhere the term $\mathrm{E}_\epsilon[\sigma^2_{x+\epsilon}]$ is zero.

Two dimensions - Prediction at a single location

The analogous results in two dimensions, also derived in the appendix in Paper 2, but reproduced here for completeness, are

$$\mathrm{E}[S(\mathbf{x}+\epsilon)] = \mu_x + \frac{1}{2}\gamma^2(a_{xx}+a_{yy}) + \frac{1}{8}\gamma^4(a_{xxxx}+2a_{xxyy}+a_{yyyy}) + O(\gamma^6)$$

and

$$\mathrm{Var}[S(\mathbf{x}+\epsilon)] = \mathrm{E}_\epsilon[\sigma^2_{x+\epsilon}] + \mathrm{Var}_\epsilon[\mu_{x+\epsilon}],$$

where

$$\begin{aligned}\mathrm{E}_\epsilon[\sigma^2_{x+\epsilon}] &= \sigma^2_x - \sigma^2(\gamma^2(Q_{x,x}+Q_{y,y}+Q_{0,xx}+Q_{0,yy}) + \\ &\quad \gamma^4(\frac{1}{4}Q_{0,xxxx}+\frac{1}{4}Q_{0,yyyy}+Q_{x,xxx}+Q_{y,yyy}+\frac{3}{4}Q_{xx,xx}+ \\ &\quad \frac{3}{4}Q_{yy,yy}+\frac{1}{2}Q_{xx,yy}+Q_{xy,xy}+Q_{x,xyy}+Q_{y,xxy})+O(\gamma^6))\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_\epsilon[\mu_{x+\epsilon}] &= \gamma^2(a_x^2+a_y^2)+\gamma^4(\frac{1}{2}a_{xx}^2+a_{xy}^2+\frac{1}{2}a_{yy}^2+a_xa_{xxx}+ \\ &\quad a_xa_{xyy}+a_ya_{xxy}+a_ya_{yyy})+O(\gamma^6).\end{aligned}$$

One dimension - Joint prediction with independent errors in locations

Consider two prediction locations x_1 and x_2, so r is now an $n \times 2$ matrix. For each location x_k, $\mathrm{E}[S(x_k+\epsilon)]$ is unaffected by the addition of further prediction locations: the kth element of μ_x is $\mu+\sum_{i=1}^n a_i r_{ik}$, which uses no columns of r except the kth. For this reason, $\mu_{x+\epsilon}$ uses no columns of r except the kth. Thus, for each k, the results already given for a single prediction location can be used to approximate $\mathrm{E}[S(x_k+\epsilon)]$.

Similarly, for each k, $\mathrm{Var}[S(x_k+\epsilon)]$ is unaffected by the addition of further prediction locations: the (k,k)th element of σ^2_x is $\sigma^2(1-\sum_{i,j=1}^n r_{ik}Q_{ij}r_{jk})$, which uses no columns of r except the kth. Therefore it is only the covariance between the two predictions that needs further consideration.

From (B.3), the off-diagonal element of $\mathrm{Var}[T|Y]$ is $\sigma^2(\rho(\|x_1-x_2\|)-\sum_{i,j=1}^n r_{1i}Q_{ij}r_{j2})$, which

we write as $\sigma^2(c - Q_{0,0})$, where $c = \rho(\|x_1 - x_2\|)$.

The Taylor series expansion of the element corresponding to $\sum_{i,j=1}^{n} r_{i1} Q_{ij} r_{j2}$ after addition of error yields

$$Q_{0,0} + \epsilon_1 Q_{x_1,0} + \epsilon_2 Q_{0,x_2} + \epsilon_1\epsilon_2 Q_{x_1,x_2} + \tfrac{1}{2}\epsilon_1^2 Q_{x_1x_1,0} + \tfrac{1}{2}\epsilon_2^2 Q_{0,x_2x_2} +$$
$$\tfrac{1}{2}\epsilon_1^2\epsilon_2 Q_{x_1x_1,x_2} + \tfrac{1}{2}\epsilon_1\epsilon_2^2 Q_{x_1,x_2x_2} + \tfrac{1}{6}\epsilon_1^3 Q_{x_1x_1x_1,0} + \tfrac{1}{6}\epsilon_2^3 Q_{0,x_2x_2x_2} + \tfrac{1}{4}\epsilon_1^2\epsilon_2^2 Q_{x_1x_1,x_2x_2} + \ldots$$

Using the general result that, for any random variables X, Y and Z,

$$\mathrm{Cov}[X, Y] = \mathrm{Cov}[\mathrm{E}(X|Z), \mathrm{E}(Y|Z)] + \mathrm{E}[\mathrm{Cov}(X, Y|Z)],$$

we have

$$\mathrm{Cov}[S(x_1 + \epsilon_1), S(x_2 + \epsilon_2)] = \mathrm{Cov}[\mu_{x_1+\epsilon_1}, \mu_{x_2+\epsilon_2}] + \mathrm{E}[(\sigma_{x+\epsilon})_{1,2}], \tag{B.7}$$

where $(\sigma_{x+\epsilon})_{1,2}$ denotes the (1,2)th element of the matrix $\sigma_{x+\epsilon}$.

For the reasons given above, $\mathrm{Cov}[\mu_{x_1+\epsilon_1}, \mu_{x_2+\epsilon_2}] = 0$. An approximation for $\mathrm{Cov}[S(x_1+\epsilon_1), S(x_2+\epsilon_2)]$ is therefore given by the following approximation of $\mathrm{E}[(\sigma_{x+\epsilon})_{1,2}]$:

$$\begin{aligned}\mathrm{Cov}[S(x_1 + \epsilon_1), S(x_2 + \epsilon_2)] \;=\; & \sigma^2\{c + \frac{1}{2}\gamma^2(c_{x_1x_1} + c_{x_2x_2}) + \\ & \frac{1}{8}\gamma^4(c_{x_1x_1x_1x_1} + 2c_{x_1x_1x_2x_2} + c_{x_2x_2x_2x_2}) - \\ & (Q_{0,0} + \frac{1}{2}\gamma^2(Q_{x_1x_1,0} + Q_{0,x_2x_2}) + \\ & \frac{1}{8}\gamma^4(Q_{x_1x_1x_1x_1,0} + 2Q_{x_1x_1,x_2x_2} + Q_{0,x_2x_2x_2x_2})) + O(\gamma^6)\}\end{aligned}$$

Like the variance of the simple kriging predictor (B.3) in classical geostatistics, this expression depends on the data locations, but not the data. However, from (B.6), $\mathrm{Var}[S(x + \epsilon)]$ is affected by the data via its dependence on a and its derivatives.

One dimension - Joint prediction with dependent errors in locations

Again using (B.7), we have

$$\begin{aligned}\mathrm{Cov}[\mu_{x_1+\epsilon_1}, \mu_{x_2+\epsilon_2}] \;=\; & \alpha\{\gamma^2 a_{x_1} a_{x_2} + \frac{1}{2}\gamma^4(a_{x_1} a_{x_2x_2x_2} + \\ & a_{x_1x_1x_1} a_{x_2} + \alpha a_{x_1x_1} a_{x_2x_2}) + O(\gamma^6)\}\end{aligned} \tag{B.8}$$

and

$$\begin{aligned}\mathrm{E}[c(x_1+\epsilon_1, x_2+\epsilon_2)] &= c+\frac{1}{2}\gamma^2(c_{x_1x_1}+2\alpha c_{x_1x_2}+c_{x_2x_2})+ \\ &\quad \frac{1}{8}\gamma^4(c_{x_1x_1x_1x_1}+c_{x_2x_2x_2x_2}+2(1+2\alpha^2)c_{x_1x_1x_2x_2}+ \\ &\quad 4\alpha(c_{x_1x_1x_1x_2}+c_{x_1x_2x_2x_2}))+O(\gamma^6). \qquad \text{(B.9)} \\ &\quad x_1x_1x_1x_2 + c_{x_1x_2x_2x_2}\end{aligned}$$

The expectation of the element corresponding to $Q_{0,0}$ after addition of error is

$$\begin{aligned}&Q_{0,0}+\tfrac{1}{2}\gamma^2(Q_{x_1x_1,0}+Q_{0,x_2x_2}+2\alpha Q_{x_1,x_2})+\tfrac{1}{8}\gamma^4(Q_{x_1x_1x_1x_1,0}+ \\ &Q_{0,x_2x_2x_2x_2}+2(1+2\alpha^2)Q_{x_1x_1,x_2x_2}+4\alpha(Q_{x_1x_1x_1,x_2}+Q_{x_1,x_2x_2x_2}))+O(\gamma^6). \qquad \text{(B.10)}\end{aligned}$$

Then $\mathrm{Cov}[S(x_1+\epsilon_1), S(x_2+\epsilon_2)]$ is given by ((B.8)+σ^2((B.9)-(B.10))).

Two dimensions - Joint prediction with independent errors in locations

As for the one-dimensional case, the expressions for the mean and variance are the same as for prediction at a single point.

Again using (B.7), with $\mathrm{Cov}[\mu_{x_1+\epsilon_1}, \mu_{x_2+\epsilon_2}] = 0$ if ϵ_1 and ϵ_2 are independent, we have

$$\begin{aligned}\mathrm{Cov}[S(\mathbf{x}_1+\epsilon_1), S(\mathbf{x}_2+\epsilon_2)] &= \sigma^2\{c+\frac{1}{2}\gamma^2(c_{x_1x_1}+c_{y_1y_1}+c_{x_2x_2}+c_{y_2y_2})+ \\ &\quad \frac{1}{8}\gamma^4(c_{x_1x_1x_1x_1}+c_{y_1y_1y_1y_1}+c_{x_2x_2x_2x_2}+c_{y_2y_2y_2y_2}+ \\ &\quad 2(c_{x_1x_1y_1y_1}+c_{x_1x_1x_2x_2}+c_{x_1x_1y_2y_2}+ \\ &\quad c_{y_1y_1x_2x_2}+c_{y_1y_1y_2y_2}+c_{x_2x_2y_2y_2}))- \\ &\quad (Q_{0,0}+\frac{1}{2}\gamma^2(Q_{x_1x_1,0}+Q_{y_1y_1,0}+Q_{0,x_2x_2}+Q_{0,y_2y_2})+ \\ &\quad \frac{1}{8}\gamma^4(Q_{x_1x_1x_1x_1,0}+Q_{y_1y_1y_1y_1,0}+Q_{0,x_2x_2x_2x_2}+Q_{0,y_2y_2y_2y_2}+ \\ &\quad 2(Q_{x_1x_1,y_1y_1}+Q_{x_1x_1,x_2x_2}+Q_{x_1x_1,y_2y_2}+ \\ &\quad Q_{y_1y_1,x_2x_2}+Q_{y_1y_1,y_2y_2}+Q_{x_2x_2,y_2y_2})))+O(\gamma^6)\}\end{aligned}$$

Two dimensions - Joint prediction with dependent errors in locations

Similar reasoning to the one-dimensional case yields the convoluted expressions

$$\begin{aligned}\mathrm{Cov}[\mu_{x_1+\epsilon_1}, \mu_{x_2+\epsilon_2}] &= \alpha\{\gamma^2(a_{x_1a_{x_2}} + a_{y_1a_{y_2}}) + \\ &\frac{1}{2}\gamma^4(a_{x_1x_1x_1a_{x_2}} + a_{x_1x_1y_1a_{y_2}} + a_{x_1y_1y_1a_{x_2}} + a_{x_1a_{x_2x_2x_2}} + \\ &a_{x_1a_{x_2y_2y_2}} + a_{y_1y_1y_1a_{y_2}} + a_{y_1a_{x_2x_2y_2}} + a_{y_1a_{y_2y_2y_2}} + \\ &\alpha(a_{x_1x_1a_{x_2x_2}} + a_{y_1y_1a_{y_2y_2}} + 2a_{x_1y_1a_{x_2y_2}})) + O(\gamma^6)\} \qquad \text{(B.11)}\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}[c(\mathbf{x}_1+\epsilon_1, \mathbf{x}_2+\epsilon_2)] &= p + \frac{1}{2}\gamma^2(c_{x_1x_1} + c_{y_1y_1} + c_{x_2x_2} + c_{y_2y_2} + 2\alpha(c_{x_1x_2} + c_{y_1y_2})) + \\ &\frac{1}{8}\gamma^4(c_{x_1x_1x_1x_1} + c_{y_1y_1y_1y_1} + c_{x_2x_2x_2x_2} + c_{y_2y_2y_2y_2} + \\ &4\alpha(c_{x_1x_1x_1x_2} + c_{x_1x_1y_1y_2} + c_{x_1y_1y_1x_2} + c_{x_1x_2x_2x_2} + \\ &c_{x_1x_2y_2y_2} + c_{y_1y_1y_1y_2} + c_{y_1y_2y_2y_2} + c_{y_1x_2x_2y_2}) + \\ &2(1+2\alpha^2)(c_{x_1x_1x_2x_2} + c_{y_1y_1y_2y_2}) + \\ &2(c_{x_1x_1y_1y_1} + c_{x_1x_1y_2y_2} + c_{y_1y_1x_2x_2} + c_{x_2x_2y_2y_2}) + O(\gamma^6), \qquad \text{(B.12)}\end{aligned}$$

and the expectation of the element corresponding to $Q_{0,0}$ after the addition of error is

$$\begin{aligned}&Q_{0,0} + \tfrac{1}{2}\gamma^2(Q_{x_1x_1,0} + Q_{y_1y_1,0} + Q_{0,x_2x_2} + Q_{0,y_2y_2} + 2\alpha(Q_{x_1,x_2} + Q_{y_1,y_2})) + \\ &\tfrac{1}{8}\gamma^4(Q_{x_1x_1x_1x_1,0} + Q_{y_1y_1y_1y_1,0} + Q_{0,x_2x_2x_2x_2} + Q_{0,y_2y_2y_2y_2} + \\ &4\alpha(Q_{x_1x_1x_1,x_2} + Q_{x_1x_1y_1,y_2} + Q_{x_1y_1y_1,x_2} + Q_{x_1,x_2x_2x_2} + \\ &+Q_{x_1,x_2y_2y_2} + Q_{y_1y_1y_1,y_2} + Q_{y_1,y_2y_2y_2} + Q_{y_1,x_2x_2y_2}) + \\ &2(1+2\alpha^2)(Q_{x_1x_1,x_2x_2} + Q_{y_1y_1,y_2y_2}) + \\ &2(Q_{x_1x_1,y_1y_1} + Q_{x_1x_1,y_2y_2} + Q_{y_1y_1,x_2x_2} + Q_{x_2x_2,y_2y_2})) + O(\gamma^6). \qquad \text{(B.13)}\end{aligned}$$

$\mathrm{Cov}[S(\mathbf{x}_1+\epsilon_1), S(\mathbf{x}_2+\epsilon_2)]$ is given by ((B.11)+σ^2((B.12)-(B.13))).

B.3 Illustration of Approximations of Moments of the Predictive Distribution

In this section we illustrate the results of Appendix B.2.2 using a simple example in one dimension, showing how individual terms resulting from the Taylor series expansion each contribute

to the approximation of the mean and variance of the predictive distribution.

We generated a realisation of a Gaussian process in one dimension on $\{0, 1, ..., 8\}$ with parameters $\mu = 0$, $\sigma^2 = 1$ and nugget effect $\tau^2 = 0$, and used a Gaussian correlation function with range parameter $\phi = 1$. This is shown in the top panel of Figure B.1. Assuming that the true parameter values were known, we then estimated the prediction mean and variance at a series of prediction locations in $[0, 8]$, with prediction locations assumed to be subject to independent $\mathrm{N}(0, \gamma^2)$ positional errors.

In the lower panel of Figure B.1 we compare the prediction mean for $\gamma^2 = 0$ and $\gamma^2 = 0.25$. It shows the difference between the prediction means for the two scenarios, and also demonstrates how the approximation for $\gamma^2 = 0.25$ improves as extra terms are added, up to terms of order γ^4.

Figure B.2 is similar, but uses $\gamma^2 = 0.04$ for greater clarity. The prediction variance is clearly much more sensitive to changes in γ^2 than the prediction mean. This is to be expected from (3.10) and (3.12): the leading term of $\mathrm{E}[S(\mathbf{x} + \epsilon)]$ contains only second- and higher-order derivatives of a.

The lower panel of Figure B.2 also highlights a result stated in Section 3.3.2, that in some circumstances the prediction variance may be smaller when $\gamma^2 > 0$ than when $\gamma^2 = 0$, which occurs when the curve drops below the zero line. For the one-dimensional example, this tends to occur midway between observation locations, when the prediction variance in the absence of positional error is largest. The periodic shape of the prediction variance in Figure B.2 when $\gamma^2 = 0$ is a consequence of the standard result in geostatistics that the prediction variance depends on the data locations, but not on the observations themselves. When $\gamma^2 > 0$, the prediction variance does depend on the observations, as (B.6) shows.

Figure B.3 shows the leading terms in the Taylor approximation of the prediction variance when $\gamma^2 = 0.04$ (see the figure legend for details). Curve a shows the prediction variance when $\gamma^2 = 0$; the other curves contribute to the prediction variance when $\gamma^2 > 0$. In particular the effect of the term $\gamma^2 a_x^2$ in the approximation, strongly related to the gradient of the underlying surface, is clearly seen (curve d). Curve e is a balancing higher order term that also depends on the gradient.

B.4 Higher Moments of the Predictive Distribution in the Presence of Positional Error

Here we demonstrate, in one dimension, that the skewness of the predictive distribution in the presence of positional error is $O(\gamma^4)$. This result provides a justification to concentrate on the first two moments of the predictive distribution, which are $O(\gamma^2)$, if there is positional error.

Using the results $\mathrm{E}(\epsilon^2) = \gamma^2$, $\mathrm{E}(\epsilon^4) = 3\gamma^4$ and $\mathrm{E}(\epsilon^6) = 15\gamma^6$, we have

$$\begin{aligned}
\mathrm{E}[\{S(x+\epsilon) - \mathrm{E}(S(x+\epsilon))\}^3] &= \mathrm{E}[(\epsilon S_x + \frac{1}{2}S_{xx}(\epsilon^2 - \gamma^2) + \\
&\quad \frac{1}{6}S_{xxx}\epsilon^3 + \frac{1}{24}S_{xxxx}(\epsilon^4 - 3\gamma^4) + ...)^3] \\
&= \frac{3}{2}\epsilon^2 S_x^2 S_{xx}(\epsilon^2 - \gamma^2) + ... \\
&= 3S_x^2 S_{xx}\gamma^4 + O(\gamma^6).
\end{aligned}$$

The skewness of a random variable X is defined as

$$\mathrm{Skew}[X] = \frac{\mathrm{E}[(X - \mathrm{E}[X])^3]}{(\mathrm{Var}[X])^{3/2}},$$

so

$$\mathrm{Skew}[S(x+\epsilon)] = \frac{3S_x^2 S_{xx}\gamma^4 + ...}{(S^2 + \gamma^2(S_x^2 + SS_{xx} + \gamma^4(\frac{3}{4}S_{xx}^2 + S_x S_{xxx} + ...)^{3/2}}. \tag{B.14}$$

Using the result

$$ax^4(b + cx^2)^{-3/2} = \frac{ax^4}{b^{3/2}} + O(x^6),$$

provided $|cx^2/b| < 1$, (B.14) is approximately equal to

$$\frac{3S_x^2 S_{xx}\gamma^4}{S^3} + O(\gamma^6),$$

provided $|(S_x^2 + SS_{xxx})\gamma^2/S^2| < 1$.

B.5 The Relationship Between Error in Prediction Location and Error in Data Location

As a postscript to the work in Paper 2, we note the following relationship between the scenarios in which positional error affects only the prediction location and that in which it affects only the

data locations.

Proposition: Consider the stationary model (3.1) with constant mean. Given known d-dimensional data locations $X_1, ..., X_n$, if the single location X_p is subject to a Berkson positional error ϵ with symmetric distribution F, predictions of $T = S(X_p^*)$ based on the following are identical:

$$\int [T|Y, X^*, X_p^*][X_p^*|X_p]dX_p^* \tag{B.15}$$

$$\int [T|Y, X^*, X_p^*][X^*|X]dX^* \tag{B.16}$$

where in (B.15) X^* is fixed and in (B.16) X_p^* is fixed. In (B.16), the integral has dimension d and $X^*|X \sim F(X, \epsilon I_n)$.

In other words, predictions are the same as if an *identical* (and not just identically-distributed) error term had been applied to each data location.

Proof: From Section B.1.1, the predictive distribution $[T|Y]$ depends on the X_i and X_p only through $\|X_p^* - X_j\|$ and $\|X_i - X_j\|$, for $i, j = 1, ..., n$. Using the representation $X_p^* = X_p + \epsilon$, we have $\|X_p^* - X_j\| = \|X_p - (X_j - \epsilon)\|$ for each j. If the distribution of ϵ is symmetric, then $-\epsilon$ has the same distribution F. This corresponds to the identical error ϵ being added to each data location. As $\|(X_i - \epsilon) - (X_j - \epsilon)\| = \|X_i - X_j\|$ for all i, j, (B.15) and (B.16) are identical.

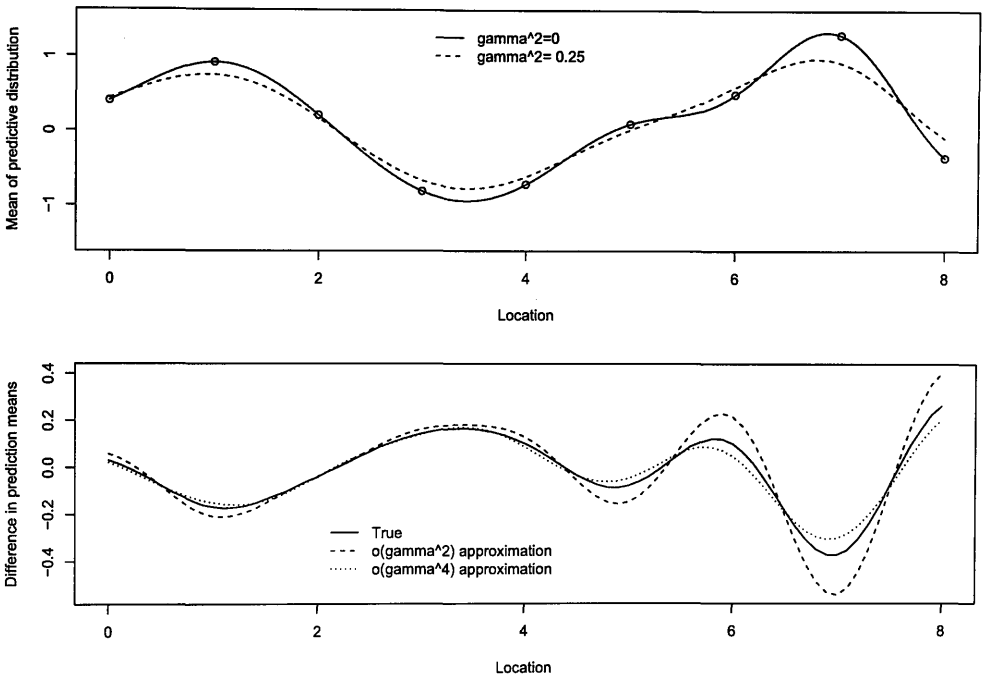


Figure B.1: (Top) Realisation of a one-dimensional Gaussian Process on $\{0, 1, ..., 8\}$ with the mean of the predictive distribution plotted assuming a positional error variance of $\gamma^2 = 0$ and $\gamma^2 = 0.25$ in the prediction location. (Bottom) Difference between the means of the predictive distributions ($\gamma^2 = 0.25$ minus $\gamma^2 = 0$), with $O(\gamma^2)$ and $O(\gamma^4)$ approximations derived from Taylor series expansions (see Appendix B.2.2).

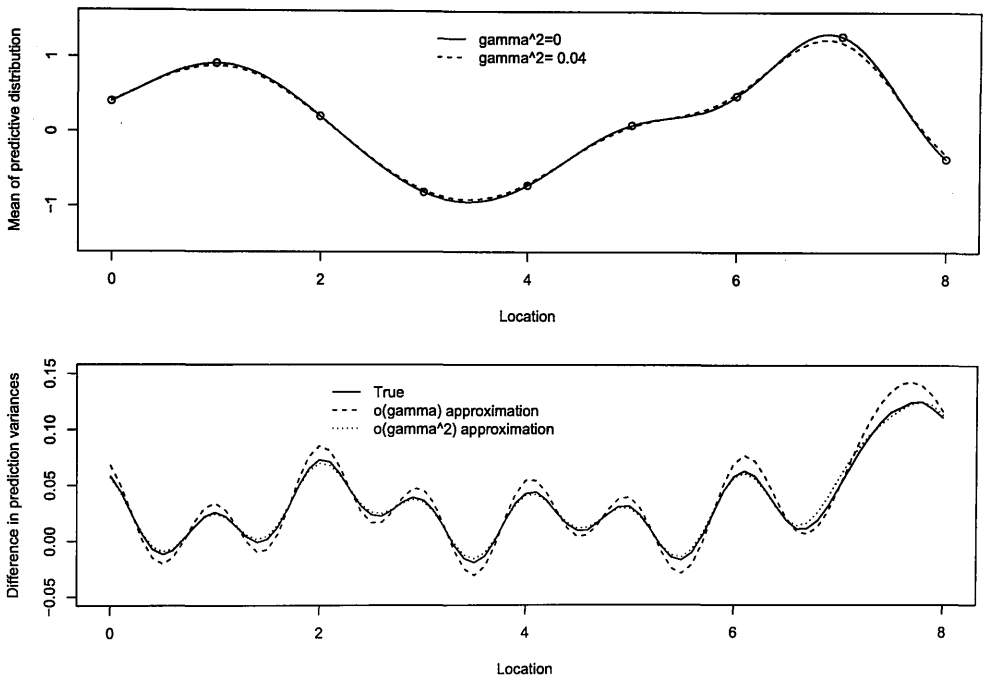


Figure B.2: (Top) Realisation of a one-dimensional Gaussian Process on $\{0, 1, ..., 8\}$ with the mean of the predictive distribution plotted assuming a positional error variance of $\gamma^2 = 0$ and $\gamma^2 = 0.04$ in the prediction location. (Bottom) Difference between the variances of the predictive distributions ($\gamma^2 = 0.04$ minus $\gamma^2 = 0$), with $O(\gamma)$ and $O(\gamma^2)$ approximations derived from Taylor series expansions (see Appendix B.2.2).

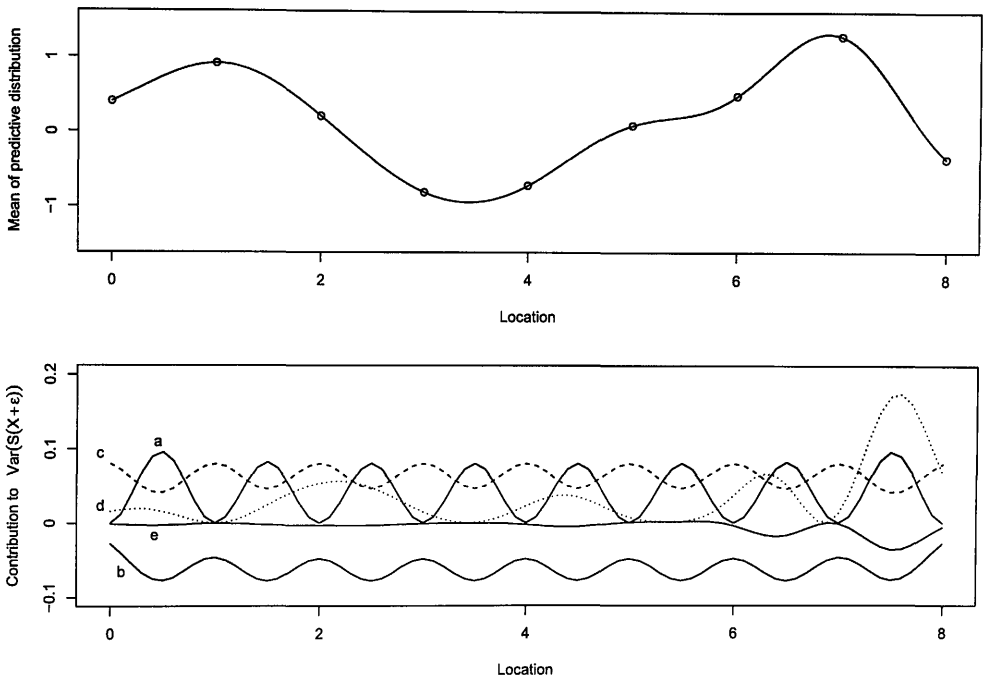


Figure B.3: (Top) Realisation of a one-dimensional Gaussian Process on $\{0, ..., 8\}$ with the mean of the predictive distribution plotted assuming a positional error variance of $\gamma^2 = 0$ in the prediction location. (Bottom) Contribution to the prediction variance of various components, assuming a positional error variance of $\gamma^2 = 0.04$: curve a: σ_x^2; b: $-\sigma^2\gamma^2 Q_{x,x}$; c: $-\sigma^2\gamma^2 Q_{0,xx}$; d: $\gamma^2 a_x^2$; e: $\gamma^4 a_x a_{xxx}$ (see Appendix B.2.2 for explanations of notation and formulae).

Appendix C

Appendices for Paper 3

C.1 Approximating the Kernels

In Section 4.3.3, we described a method for finding kernels which, after convolution, correspond to a specified set of auto- and cross-covariance functions. In the soil data application, we used three kernels, k_0 of Gaussian form, and k_1 and k_2 of Matérn functional form, and maximised over the eight parameters σ_i^2, ϕ_i $(i = 0, 1, 2)$ and κ_i $(i = 1, 2)$. Figure 4.4 shows the good fit of the approximation.

Gaussian kernels have been routinely used in many applications, including Higdon (1998). For comparison, Figure C.1 presents similar results if the functional form of two or more of the kernels is Gaussian. It shows the specified two auto-covariance functions and one cross-covariance function for the soil data (each of Matérn form), and the estimates of the covariance functions based on kernel convolution. In the left-hand panel, each of the k_i was specified as Gaussian, while in the right-hand panel, k_0 was specified as Matérn and k_1 and k_2 as Gaussian.

The left panel shows that the approximation is poor if each of the kernels is Gaussian. In this case, the approximations of $C_{22}(x)$ and $C_{12}(x)$ coincide. The right panel shows that the approximation is much improved if k_0 is of Matérn form, and is nearly as good as that shown in Figure 4.4. In general, the approximation will be at least as good when Matérn kernels are used rather than Gaussian ones, as the latter are a special case of the former.

C.2 Non-Positive Definite Kernels

In the kernel convolution representation (4.9), the kernel that corresponds to a given covariance function is not uniquely defined: if $k(x)$ is a kernel used in (4.9), then $-k(x)$ gives rise to a process with the same covariance function. More generally, so does the discontinuous kernel $\tilde{k}(x)$, where

$$\tilde{k}(x) = k(x)(-1)^{\mathrm{I}(\|x\|>\alpha)},$$

as $\{k(x)\}^2 = \{\tilde{k}(x)\}^2$. Here, $\mathrm{I}(\cdot)$ is the indicator function, and α is a fixed threshold, to be chosen.

To illustrate the usefulness of kernels of the form $\tilde{k}(x)$, we repeated the one-dimensional example described by Higdon (2002), Section 4.1.

We generated 30 data points at equally-spaced locations $x_i \in [1, 10]$ from the model

$$y(x_i) = \sin(2\pi x_i/10) + 0.2\cos(2\pi x_i/2.5) + \epsilon_i,$$

where the ϵ_i are independent $\mathrm{N}(0, 0.01)$ random variables. Figure C.2 shows the raw data and the predictions resulting from an analysis by kernel convolution. In this analysis, we used 7 independent Gaussian kernels with standard deviation 2, centred on 7 equally-spaced locations in $[-1, 12]$.

We then repeated the analysis using independent and identically-distributed kernels of the form $\tilde{k}(x)$, for various values of α. Figure C.3 shows the results. The discontinuous nature of $\tilde{k}(x)$ causes ugly discontinuities in the resulting predictions; this is resolved only when α is set smaller than $\min\|x_i - x_j\|$, $i \neq j$. In this extreme case, equivalent to using $-k(x)$, predictions are identical to those obtained using $k(x)$.

This simple example demonstrates that, while the kernel convolution method theoretically admits non-positive definite kernels, only positive definite kernels produce plausible predictions in practice. Thus when approximating the covariance functions in Paper 3, we considered only positive definite kernels.

C.3 Altitude

In Section 4.4.2 and Chapter 5.3, we discussed the relationship between altitude and radon exposure. Here we provide further details. Figure C.4 is an altitude map of the Winnipeg region. Residential locations in the radon study are shown as points. Clearly visible are the Red River (north-south), Assiniboine River (east-west) and Red River Floodway (straight lines to the east of the city). There is little variation in altitude in the city of Winnipeg compared to neighbouring regions.

Figure C.5 is a plot of radon measurement against altitude for the 1622 homes that had complete radon data. The figures show the weak relationship between altitude and radon in this study. The sample correlations are $r = -0.07$ and $r = -0.10$ for bedroom and basement radon respectively. On the basis of these plots, we decided not to use the altitude covariate in the analysis in Paper 3.

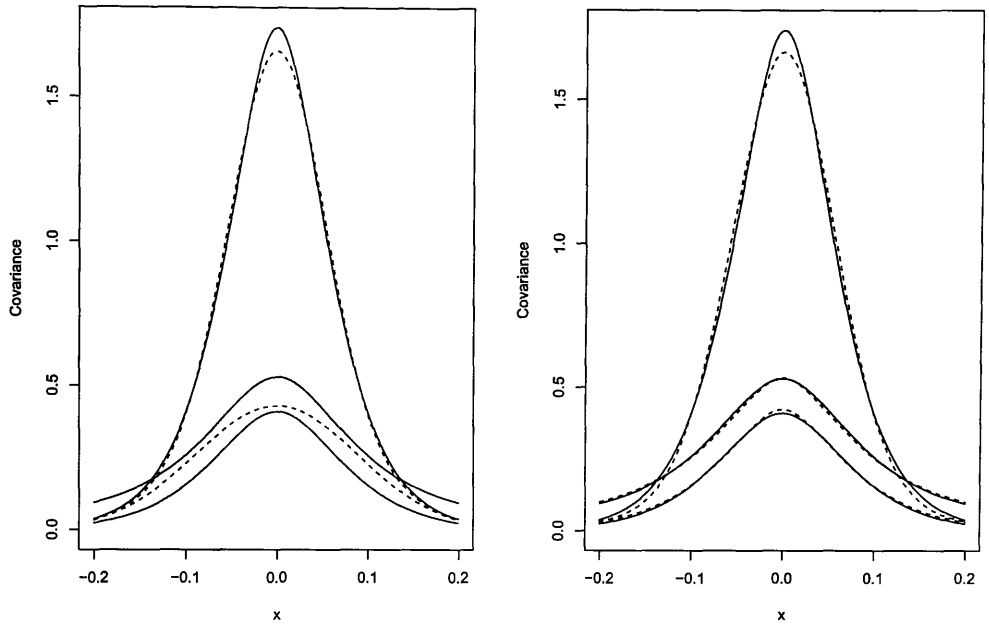


Figure C.1: For the soil data analysis, comparison of the covariance functions (solid lines) $C_{11}(x)$ (top), $C_{22}(x)$ (middle) and $C_{12}(x)$ (bottom) with convolved kernel function approximations (dotted lines). Left panel: k_0, k_1 and k_2 all Gaussian kernels. Right panel: k_0 Matérn kernel, k_1 and k_2 Gaussian kernels.

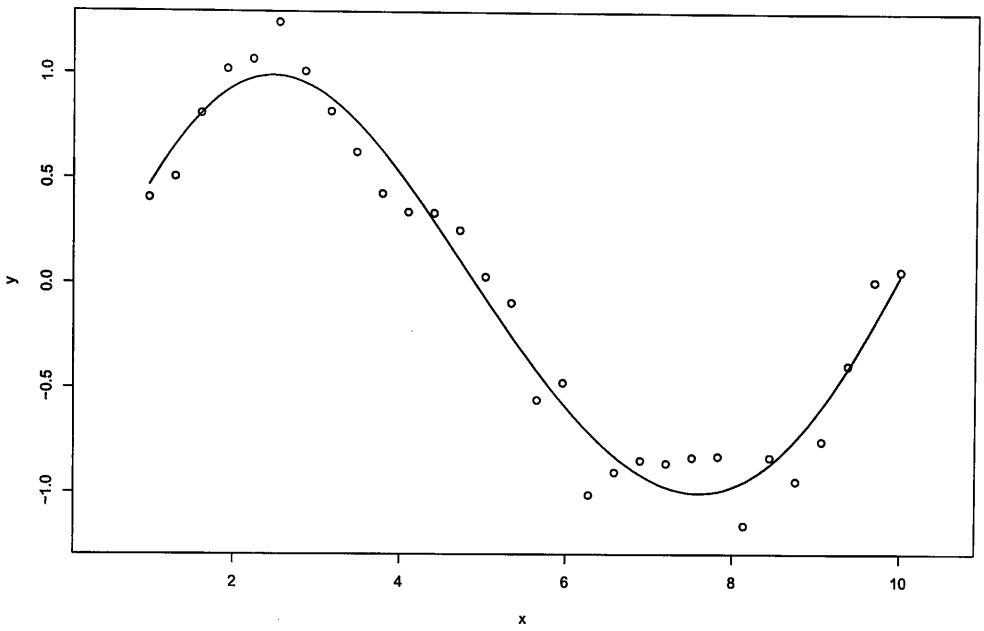


Figure C.2: Realisation of a one-dimensional Gaussian process using the model described by Higdon (2002). The solid curve indicates predictions obtained from the kernel convolution method with 7 independent Gaussian kernels.

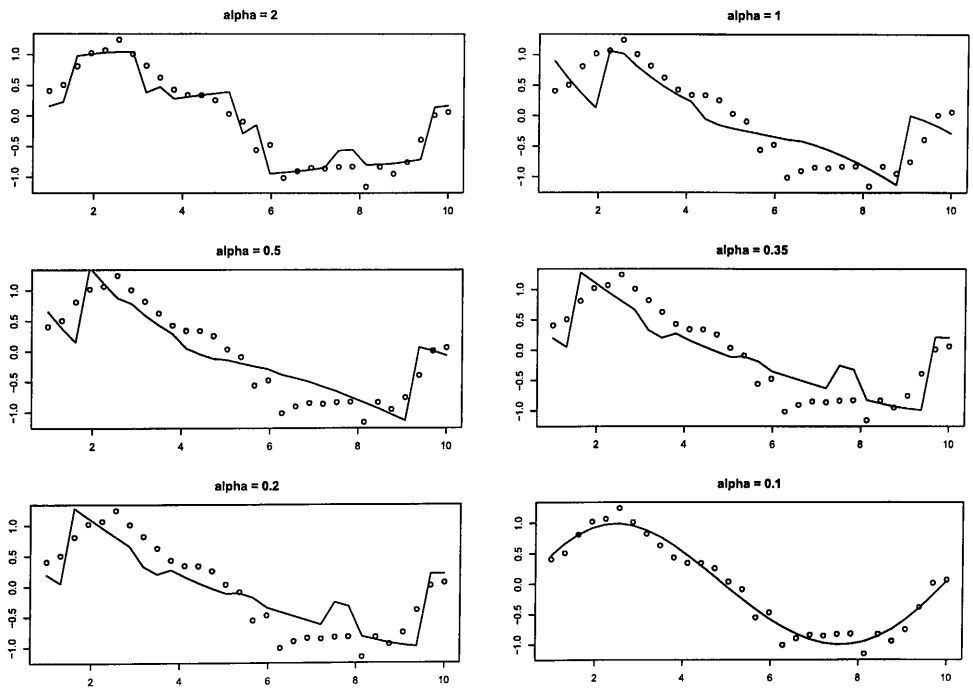


Figure C.3: Predictions resulting from use of a non-positive definite kernel $\tilde{k}(x)$, with varying values of the threshold parameter α.

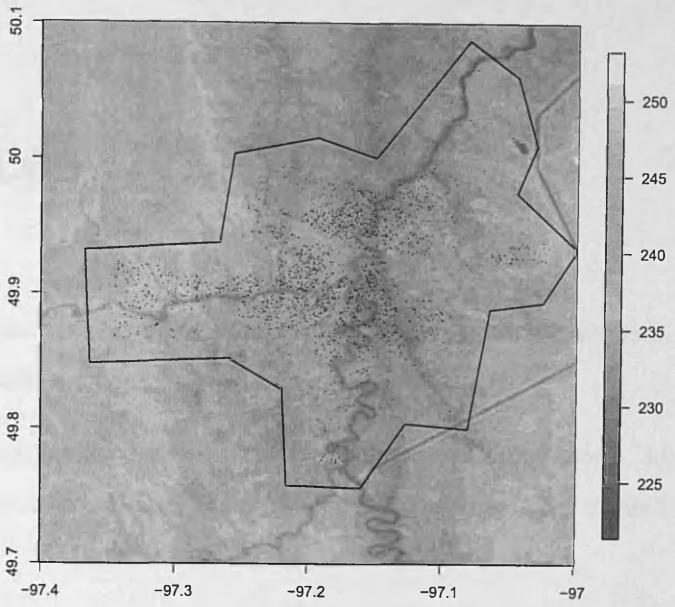


Figure C.4: Altitude map for the Winnipeg region. Distances are in kilometres, and altitude in metres.

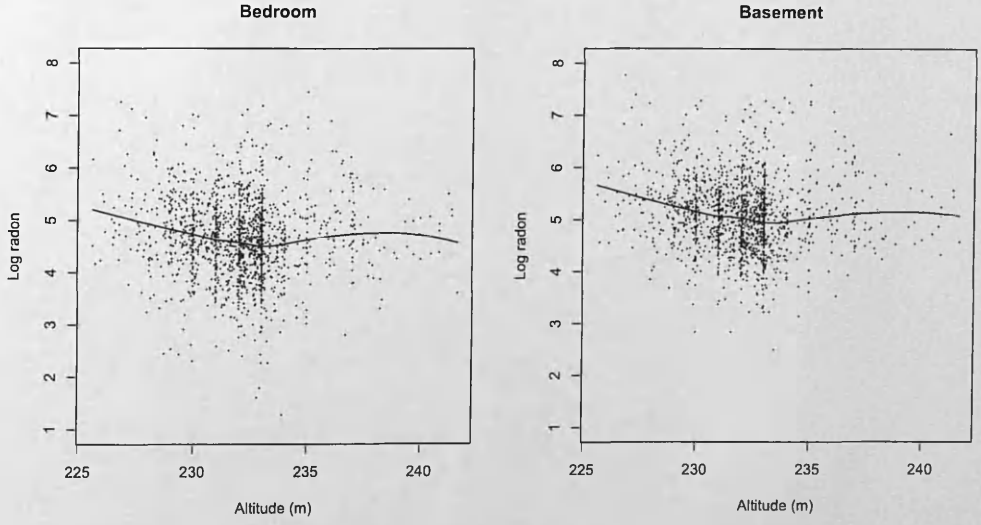


Figure C.5: Plot of bedroom and basement radon against altitude, with loess curves.

References

Higdon, D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5**, 173–190.

Higdon, D. 2002. *Space and space-time modeling using process convolutions*. In: Quantitative Methods for Current Environmental Issues. New York: Springer-Verlag. Pages 37–56.