UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXX

# Spatial Analysis of Geomasked and Aggregated Data

**Coordinatore del Corso:** Prof. Nicola Sartori

**Supervisore:** Prof. Giuseppe Espa

**Dottorando:** Claudio Fronterré

15/01/2018

# Abstract

New advances in global positioning systems (GPS) and geographical information systems (GIS) translated in an explosion of spatial data collection. In this thesis we tackle two well known problems of spatial data. The first one regards the quality of the geographical coordinates used as input for many spatial models. This aspect is often neglected and we provide a general framework to deal with the uncertainty present in the spatial locations. The second problem takes the name of change of support and is related to the analysis of spatial data at a scale of aggregation that is different than the one at which they are observed. Also in this case we develop a theoretical framework to figure out the inference problems inherent to this instance. For both problems the results obtained are really promising.

# Sommario

I progressi registrati negli ultimi decenni in ambito di sistemi di posizionamento globale (GPS) e sistemi informativi territoriali hanno portato ad un'esplosione del raccoglimento di dati spaziali. In questa tesi ci occupiamo di due problemi comuni a questa tipologia di dati. Il primo riguarda la qualitá delle coordinate geografiche utilizzate come input di svariati modelli spaziali. Questo aspetto é spesso trascurato e forniamo un quadro teorico generale che tenga in considerazione l'incertezza presente nelle coordinate spaziali. Il secondo problema prende il nome di *cambio di supporto* ed é legato all'analisi di dati spaziali ad un livello di aggregazione diverso da quello al quale sono stati osservati. Anche in questo caso sviluppiamo un framework teorico per risolvere i problemi di inferenza legati a questa casistica. In entrambi i casi i risultati ottenuti sono molto promettenti.

*To Sara...*

# Acknowledgements

First of all I would like to thank my supervisor, Professor Giuseppe Espa, for his full support throughout this PhD project. His patience, encouragement and advice followed me in the past three years. I also thank Prof. Peter Diggle, who supervised me during my visiting at Lancaster University, and all the members of the CHICAS group that welcomed me really warmly. I thank my colleagues from the 30th PhD cycle and, in particular, Leonardo and Umberto, with which I shared ups and down during the last three years. Finally, I thank my family for being my pillars of support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## Overview

The use of georeferenced data is nowadays pervasive in a lot of different areas: epidemiology, climate science, health studies and crime analysis to cite few. Moreover, in the last years, the quantity of available spatial data has increased considerably, and is being collected at a continuously higher level of resolution. However, an aspect often neglected is the positional accuracy of the coordinates used as input of different classes of spatial models.

Positional error can be introduced in different ways: use of imperfect measuring instruments such as GPS receivers or satellites, random displacement of the spatial locations for confidentiality reasons (geomasking) and geocoding of text addresses. Even if the process of collecting spatial data is usually not perfect, the quality of spatial coordinates is infrequently assessed. Moreover, ignoring the uncertainty present in georeferenced data can lead to flawed inferences and misleading conclusions (Jacquez, 2012). Indeed, the literature is full of studies that show how positional error can affect estimates of diseases rates (Zimmerman and Sun, 2006; Zimmerman, 2007; Goldberg and Cockburn, 2012), disease cluster statistics (Jacquez and Waller, 2000; Zimmerman *et al.*, 2010), test for space-time interaction (Malizia, 2013), exposure estimates (Zandbergen, 2007; Mazumdar *et al.*, 2008) and parameters estimates of spatial models (Gabrosek and Cressie, 2002; Arbia *et al.*, 2015). Even if the negative impact of positional error is well recognized, the current practice is to ignore the presence of positional error due to a lack of well established theories and methods to deal with it (Jacquez, 2012).

One of the goal of this research thesis is to fill this gap, providing a theoretical framework that takes into consideration the uncertainty present in the spatial locations.

Another common problem in the field of spatial statistics is what is often called COSP (change of support problem), spatial misalignment or also MAUP (modifiable areal unit problem). Spatial data are usually collected at differing scales and resolutions and many statistical issues are associated with combining such data for modelling and inference (Gotway and Young, 2002). The second goal of this work is to provide models that are able to properly combine outcome and covariates when these are misaligned, i.e. when their spatial scale is different.

# Main contributions of the thesis

### Chapter 1

1. Analyse the effects of positional error and, in particular, of geomasking on the variogram and on the linear geostatistical model.
2. Obtain equations that quantify the bias and show that geomasking is the cause of overestimation of the spatial range and underestimation of the variance of the underlying true process. Moreover it creates an artificial nugget effect.
3. Propose two types of correction. The first, following the classic geostatistical framwework, is based on a non-linear curve-fitting of the variogram. The second, is a model-based approach, and makes use of composite likelihood achieving a huge computational gain compared to the method proposed by Fanshawe and Diggle (2011).
4. Propose an approximate version of our method that allows to obtain an extra computational gain without sacrificing the efficiency of the estimators.
5. Extend our model to the case of Uniform geomasking and of heteroscedastic geomasking.
6. Application on a real dataset taken from a DHS (Demographic and Health Survey, (Burgert *et al.*, 2013)) survey conducted in Senegal in 2011.
7. Suggest some useful guidelines on the selection of displacement parameters such that both the confidentiality and the spatial structure of the data can be preserved.

### Chapter 2

1. Extend the work done in Chapter 1 to Poisson and Binomial data and show that also non Gaussian data suffer of the same problem if positional error is neglected.
2. Propose a variogram-based correction for Poisson data.
3. Propose a model based solution for Binomial data after an empirical logit transformation is applied.
4. Show the effects of geomasking on point pattern analysis and, in particular, on the detection of clusters through the Rypley's K function.
5. Suggest a possible correction in the case of Neyman-Scott processes (Neyman and Scott, 1958).

**Chapter 3**

1. Propose a geostatistical model that is able to combine outcome and covariates that are spatially misaligned.
2. Provide likelihood equations for point to area, area to area and area to point estimation.
3. Simulation study on area to point prediction, i.e. when the outcome is observed at a coarser level than the covariates that are continuously available.

# Chapter 2

# Geostatistical inference in the presence of geomasking

## 2.1 Introduction

The use of georefenced data is nowadays pervasive in a lot of different areas: epidemiology, climate science, health studies and crime analysis to cite few. Moreover, in the last years, the quantity of available spatial data has increased considerably, and is being collected at a continuously higher level of resolution. However, an aspect often neglected is the positional accuracy of the coordinates used as input of different classes of spatial models.

Positional error can be introduced in different ways. Here we identify three major sources of positional error: use of imperfect measuring instruments, geomasking and geocoding. Spatial coordinates are usually collected through the use of measuring instruments like GPS receivers or satellites. The height at which the devise is placed or other factors such as air transparency and clouding will influence the measurement process giving raise to imprecise coordinates (Devillers and Jeansoulin, 2006). Another common source of positional error is when, for confidentiality issues, the point location of the event cannot be released. In these cases a common solution is geomasking (Armstrong *et al.*, 1999), that is the random or deterministic perturbation of the observed points in a way that is not possible to go back to the original coordinates. In this case the positional error is introduced with the purpose of privacy protection. Geocoding is the process of converting text-based addresses into geographic coordinates and is very common in several disciplines. Such a process introduces positional error in the geocoded point for several reasons: incorrect street segment, incorrect offset from the street segment, incorrect placement along the street segment and positional error in

the street segment (Zandbergen, 2009). The resultant error is therefore the aggregate effect of all these factors. Several empirical studies suggest that the positional error introduced on average is neither small nor random (Dearwent *et al.*, 2001; Bonner *et al.*, 2003; Cayo and Talbot, 2003; Rushton *et al.*, 2006; Kravets and Hadden, 2007; Zinszer *et al.*, 2010).

Even if the process of collecting spatial data is usually not perfect, the quality of spatial coordinates is infrequently assessed. Moreover, ignoring the uncertainty present in georeferenced data can lead to flawed inferences and misleading conclusions (Jacquez, 2012). Indeed, the literature is full of studies that show how positional error can affect estimates of diseases rates (Zimmerman and Sun, 2006; Zimmerman, 2007; Goldberg and Cockburn, 2012), disease cluster statistics (Jacquez and Waller, 2000; Zimmerman *et al.*, 2010), test for space-time interaction (Malizia, 2013), exposure estimates (Zandbergen, 2007; Mazumdar *et al.*, 2008) and parameters estimates of spatial models (Gabrosek and Cressie, 2002; Arbia *et al.*, 2015). Even if the negative impact of positional error is well recognized, the current practice is to ignore the presence of positional error due to a lack of well established theories and methods to deal with it (Jacquez, 2012).

In a geostatistical setting, Gabrosek and Cressie (2002) examine the effect that uncertainty in the spatial lag has on the first two moments of the underlying spatial random process and show how to account for location error by adjustment of the kriging equations. They find that in presence of substantial positional error the adjusted kriging approach for location error performs better than ordinary kriging, in particular, the presence of positional error inflates both bias and mean squared prediction error of ordinary kriging. Cressie and Kornak (2003) propose new kriging equations that consider also a component of variation for the more general trend term and apply them to remote sensing data of total column ozone, where the positional error is caused by assignment of the measured value to their nearest grid-cell centers. Fanshawe and Diggle (2011) suggest a model-based solution. They obtain the likelihood function for a stationary Gaussian geostatistical model in presence of positional error and consider also the case when prediction locations contain uncertainty. Even if the approach is promising the extremely high computational burden makes it computationally infeasible. Moreover, they find that the local gradient of the surface may have a large effect on the variance of the predictive distribution and that the predictive distribution at a point is non-Gaussian, and asymmetric, in the presence of positional error, even if the underlying process is Gaussian.

In this chapter, we develop a method of inference based on the composite likelihood that overcomes the computational limits of the full likelihood method. The chapter is

structured as follows. In Section 2.2 we examine in depth the practice of geomasking and the most used geomasking methods. This will be our source of positional error even though the proposed approach can be extended to other contexts. In Section 2.3 we introduce the modeling framework. In Section 2.4 we use the variogram as a tool to asses the effects of positional error on the spatial structure of the data and on the parameters that characterize the model and the we suggest a correction based on it. Section 2.5 shows a model based solution that makes use of composite likelihood. We show also through a simulation study conducted in 2.6 that this approach is more efficient and has to be preferred to the variogram for formal parameter estimation. Section 2.7 reports the results of an analysis conducted on DHS data affected by positional error and Section 2.8 is a concluding discussion.

## 2.2 Geomasking

Geographical masking, or geomasking, was first introduced by Armstrong *et al.* (1999) as an improvement to the standard practice of aggregating health records to preserve confidentiality. Geomasking is imposed by adding stochastic or deterministic noise to the spatial coordinates. The reason is generally to protect sensible or confidential information about individuals that otherwise could be identified if the geographic information is linked with other widely available sources. In this way, geomasking allows to reduce the disclosure risk of sensible information without degrading too much the geographic properties of the data.

In this section we consider only random perturbation methods where stochastic noise, opposite to deterministic, is introduced. This choice is guided by the fact that these geomasking methods are the most used in practice. For example, the Forest Inventory Analysis Program (McRoberts *et al.*, 2005) the Living Standard Indicator Survey (Grosh *et al.*, 1996) and the Demographic and Health Surveys (Burgert *et al.*, 2013) are surveys that adopted geomasking approaches to protect respondents confidentiality so that data can still be shared publically. Even though new geomasking techniques have been proposed, such as donut geomasking (Hampton *et al.*, 2010) or gaussian bimodal displacement (Cassa *et al.*, 2006) they are not used in practice and the authors think that the extra bias introduced by these methods is not justifiable by the small reduction in risk disclosure. For a recent and complete review about geograhic masking methods we refer the reader to (Zandbergen, 2014). Figures 2.1 and 2.2 shows the effects of Gaussian and Uniform geomasking on a set of simulated points, the displacement parameters are chosen such that the expected value and variance of the positional error

models is the same.

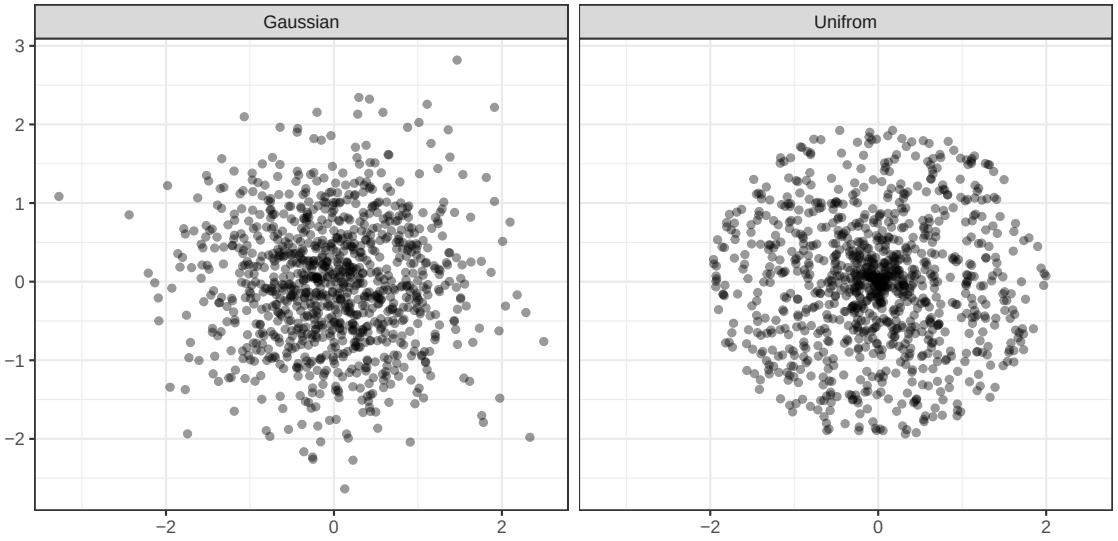


FIGURE 2.1: Repeatedely geomasking (Gaussian on the left and Uniform on the right) of one single point located at the center.

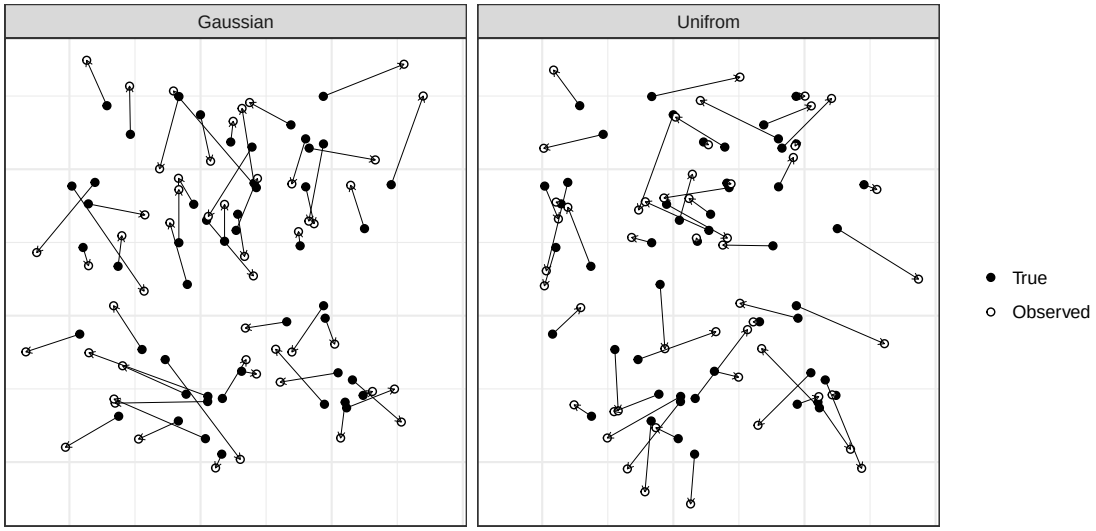


FIGURE 2.2: Gaussian and Uniform geomasking applied to a point pattern.

## 2.3 Modeling framewok

We consider a stationary Gaussian model (Diggle and Ribeiro, 2007) of the form

$$Y_i = S(x_i) + Z_i : i = 1, \ldots, n, \tag{2.1}$$

where $Y_i$ is the value observed at point $x_i \in \mathbb{R}^2$, $S(x)$ is a Gaussian process with mean 0, variance $\sigma^2$ and correlation function $\rho(u) = Corr\{S(x), S(x')\}$ where $u$ is the euclidean distance between $x$ and $x'$ and $Z_i$ are i.i.d. $N(0, \tau^2)$ independent of the spatial stochastic process $S$. We will mainly consider the class of correlation functions introduced by Matérn (1960)

$$\rho(u; \phi, \kappa) = \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1} (u/\phi)^{\kappa} K_{\kappa}(u/\phi),$$

where $\phi > 0$ is a scale parameter with the dimension of the distance, $\kappa > 0$ is a shape parameter which determines the analytic smoothness (differentiability) of the underlying process $S$ and $K_{\kappa}$ denotes the modified Bessel function of order $\kappa$. Because of their flexibility these correlation functions are widely used in practice.

## 2.4 Effects of positional error on the variogram

Due to geomasking, instead of observing the true location, say $X_i^*$, we observe a displaced location

$$X_i = X_i^* + W_i, \tag{2.2}$$

where $W_i \sim N_2(0, \delta^2 I_2)$ and $\delta^2$ is the positional error variance. Equation (2.2) represents a Gaussian geomasking. This choice allows us to obtain a nice mathematical treatment. However, we will show that the results here obtained can be generalized to other types of geomasking procedures. We want to asses the effects of positional error on the spatial structure of the observed data and, as a tool to identify it, on the variogram. The variogram is defined as

$$V_Y\left(u_{ij}^*\right) = \frac{1}{2} E\left[(Y_i - Y_j)^2\right].$$

Under stationary assumptions, $V_Y\left(u_{ij}^*\right) = \tau^2 + \sigma^2\left\{1 - \rho\left(u_{ij}^*\right)\right\}$ and summarizes the essential qualities of a geostatistical. Figure 2.3 show the shape of a standard variogram.

The observed quantities

$$v_{ij} = \frac{1}{2}(y_i - y_j)^2,$$

constitute the empirical variogram and are unbiased estimates of the corresponding variogram ordinates. From now on we will use the notation [ ] to mean "distribution of". Assuming that $U_{ij}$ and $V_{ij} = \frac{1}{2}(Y_i - Y_j)^2$ are stochastically independent given $U_{ij}^*$, the distribution of the empirical variogram ordinates conditionally on the observed distance $U_{ij}$ is

$$[V_{ij} \mid U_{ij}] = \int_0^{\infty} \left[V_{ij} \mid U_{ij}^*\right]\left[U_{ij}^* \mid U_{ij}\right] du^*, \tag{2.3}$$

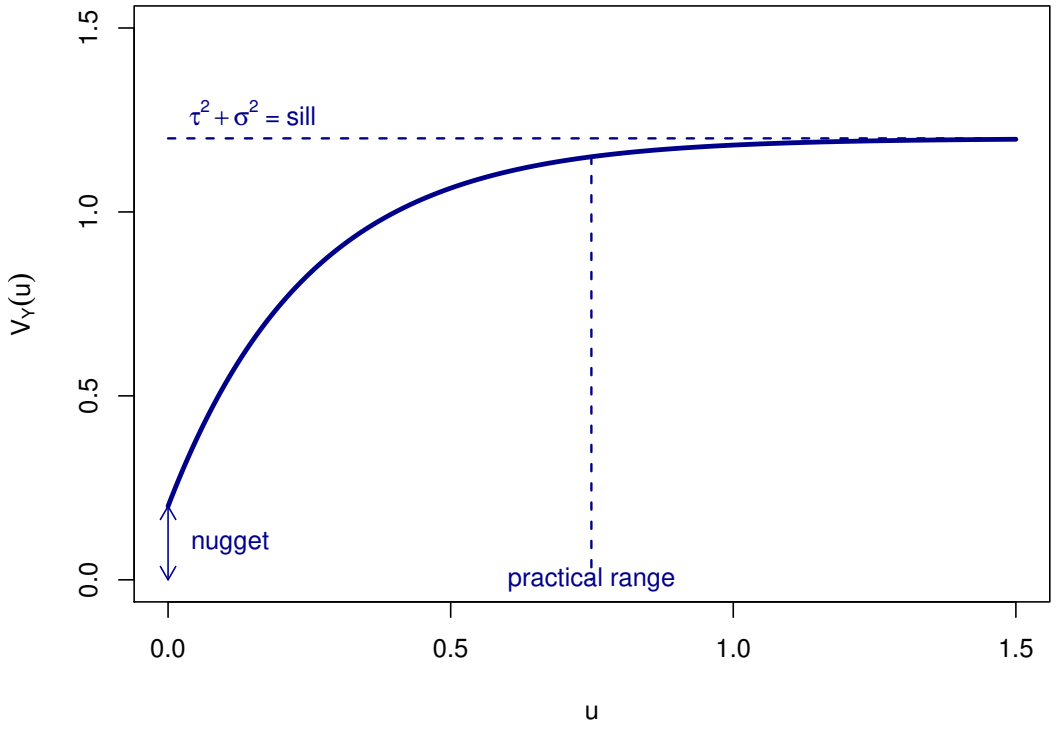


FIGURE 2.3: Standard form of the theoretical variogram. The total variance is the sum of $\sigma^2$ and $\tau^2$ and takes the name of *sill*. The *practical range* is defined as that distance $u$ such that $\rho(u) = 0.05$.

where $\left[V_{ij} \mid U_{ij}^*\right] \sim V_Y\left(u_{ij}^*\right) \chi^2_{(1)}$ and $\left[U_{ij}^* \mid U_{ij}\right] \sim Rice\left(u_{ij}, \sqrt{2}\delta\right)$. Taking the expectation on both sides of (2.3) we obtain the form of the theoretical variogram in presence of positional error

$$V_Y\left(u_{ij}\right) = \tau^2 + \sigma^2 \left\{1 - E\left[\rho\left(U_{ij}^* \mid U_{ij}\right)\right]\right\}. \tag{2.4}$$

The closed form (when exists) of $E\left[\rho\left(U_{ij}^* \mid U_{ij}\right)\right] = \int \rho\left(u_{ij}^*\right)\left[U_{ij}^* \mid U_{ij}\right] du^*$ depends on the specific correlation function used. We show that in the case of a Gaussian correlation function it exists and provides useful information on the collateral effects of positional error. It is worth noting that as $\delta \to 0$ (2.4) converges to the true variogram $V_Y\left(u_{ij}^*\right)$ and, on the other side, as $\delta \to \infty$ the points are displaced so far apart that the spatial structure is not preserved anymore and (2.4) becomes a flat line at the level of the sill $\tau^2 + \sigma^2$. As mentioned before, in the case of a Gaussian correlation function

$\rho\left(u_{ij}^{*}\right) = \exp\left\{-\left(u_{ij}^{*}/\phi\right)^{2}\right\}$ it is possible to show that

$$E\left[\rho\left(U_{ij}^{*} \mid U_{ij}\right)\right] = \frac{1}{1+(2r)^{2}} \exp\left\{-\left(\frac{u_{ij}}{\phi\sqrt{1+(2r)^{2}}}\right)^{2}\right\}, \tag{2.5}$$

where $r = \delta/\phi$. This means that the magnitude of the bias induced by geomasking depends on the ratio between the standard deviation of the positional error and the range parameter. We get another important insight looking at the behavior of (2.5) at the origin. The limiting value of (2.5) as $u_{ij} \to 0$ is $\left\{1+(2r)^{2}\right\}^{-1}$ that is smaller than 1, the value we should expect in absence of positional error. Thus, geomasking locations leads to the creation of an artificial nugget effect. More precisely, there are two forces that act in opposite directions. The first part of the equation $\{1+(2r)^{2}\}^{-1}$ that produces the artificial nugget leads also to a systematic underestimation (overestimation) of the correlation function (variogram), on the other side, $\sqrt{1+(2r)^{2}}$ increases the true value of $\phi$ leading to a systematic overestimation (underestimation) of the correlation function (variogram). The final bias is a result of the combination of these two forces. While the first effect is fixed the second one is controlled by the distance. Indeed, we can see from Figure 2.4 that while the true variogram is initially overestimated, as the distance increases this effect is soften by the other acting force leading to underestimation of the variogram for high levels of r. Likewise, this results can be generalized to any correlation function with a symmetric positional error model.

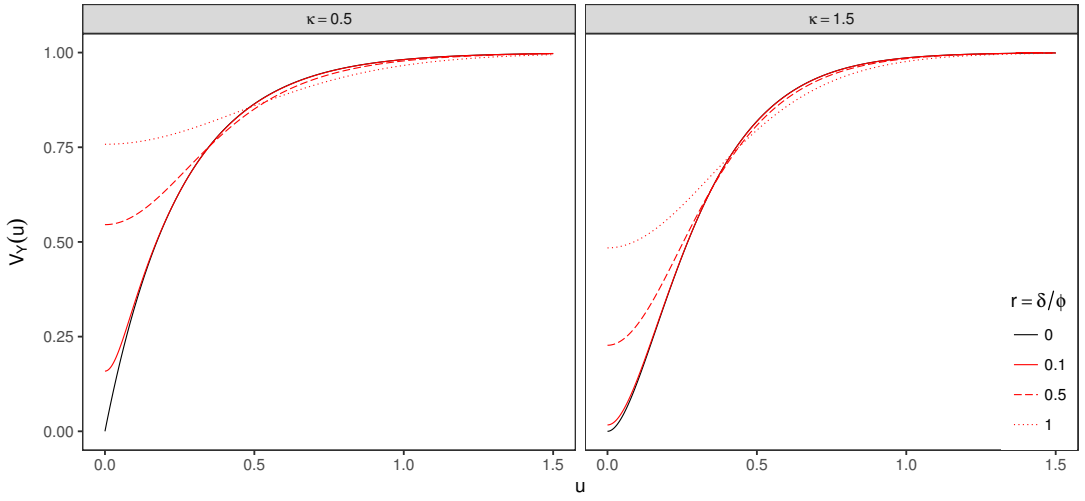


FIGURE 2.4: Departures (red lines) from the true variogram (solid black line) with $\sigma^2 = 1$ and $\tau^2 = 0$ for increasing values of $r = \delta/\phi$. Matrn correlation functions with two different shape parameters are used. For comparison purposes the scale parameter $\phi$ is chosen such that the practical range $u_0 = 0.75$, $\{u_0 : \rho(u_0) = 0.05\}$.

### 2.4.1 Variogram correction

Using the result from equation (2.4) we suggest a correction through N-weighted least squares. The vector of parameters $\theta = \{\sigma^2, \phi, \tau^2\}$ is estimated minimizing the following criterion

$$S_n(\theta) = \sum_{k=1}^{m} n_k \left\{ v_k - V_Y(u_k; \theta) \right\}^2, \tag{2.6}$$

where $v_k$ are the sample variogram ordinates, obtained averaging all $v_{ij}$ for which the corresponding $u_{ij}$ satisfies $(k-1)h < u_{ij} \leq kh$ ($h$ is the bin width), $u_k = (k - 0.5)h$ is the mid-point of the corresponding interval and $n_k$ denotes the number of empirical variogram ordinates which contributes to $v_k$. The positional error variance $\delta^2$ is assumed to be known. This is often the case with geomasking procedures. Estimating $\tau^2$ and $\delta^2$ simultaneously would not be possible because of their identifiability. Indeed, there are different combinations of $\tau^2$ and $\delta^2$ that lead to the same result. However, if we can assume that no nugget effect is present or we can estimate it from repeated measurements we are able to use this estimation procedure even with unknown $\delta^2$.

### 2.4.2 Uniform geomasking

In alternative to Gaussian geomasking, another commonly used method is uniform geomasking. Let $W = (W_1, W_2)$; we now define the positional error process as

$$\begin{cases} W_1 = R\cos\Lambda \\ W_2 = R\sin\Lambda \end{cases}, \tag{2.7}$$

where $R$ and $\Lambda$ are two independent uniform random variables in $[0, d]$, with $d$ denoting the maximum displacement distance, and $[0, 2\pi]$, respectively. However, under uniform geomasking $[U^*_{ij}|u_{ij}]$ is an intractable distribution, making computation of the likelihood function in (2.13) cumbersome.

In the application of Section 2.7, we propose to approximate $[U^*_{ij}|u_{ij}]$ under uniform geomasking with a $Rice(u_{ij}, \delta/\sqrt{6})$ since the variance for each of the components of $W$ in (2.7) is $\delta^2/6$. We can essentially well approximate a Uniform geomasking with maximum displacement distance $d$ with a Gaussian geomasking with positional error variance $\delta^2 = d^2/6$.

We illustrate the goodness of such approximation as follows. We first express $U^*_{ij}$ in terms of $R$, $\Lambda$ and $u_{ij}$ as

$$U^*_{ij} = \sqrt{u^2_{ij} + R^2 - 2u_{ij}R\sin\Lambda}. \tag{2.8}$$

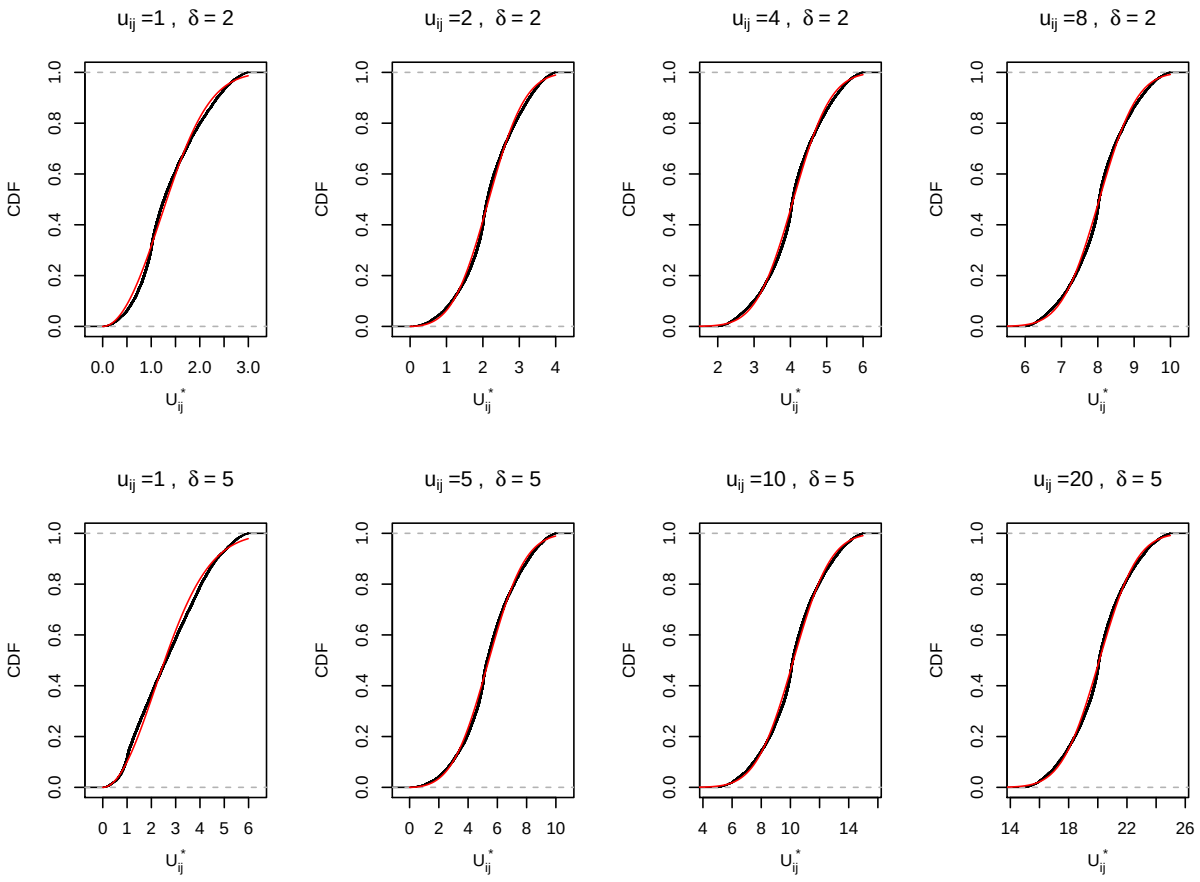


FIGURE 2.5: Each plot shows the empirical cumulative density function (CDF) based on 100,000 samples generated from $[U^*_{ij}|u_{ij}]$ under uniform geomasking (black line) and the CDF of a $Rice(u_{ij}, \delta/\sqrt{6})$ (red line). The corresponding values of $u_{ij}$ and $\delta$ are shown in the heading of each plot.

We then simulate 100,000 samples from a uniform distribution in $[0, \delta]$, setting $\delta = 2$ and $\delta = 5$ which correspond to the maximum displacement distances that were applied to the data in Section 2.7; we also simulate an equal number of samples from a uniform in $[0, 2\pi]$. For a given value of $u_{ij}$, we then compute the empirical cumulative density function (CDF) using the resulting 100,000 generated from $[U^*_{ij}|u_{ij}]$ based on (2.8).

Figure 2.5 reports the result of the simulation. The discrepancies between the empirical CDF under uniform geomasking (black line) and the CDF of a $Rice(u_{ij}, \delta/\sqrt{6})$ are small in all of the eight scenarios considered.

## 2.5 Likelihood-based inference for the linear Gaussian model

Although the results obtained through the variogram correction are promising, in general, we favour the application of principles of statistical modelling and inference to geostatistical problems. The variogram is still useful for exploratory purposes and to suggest reasonable initial values for estimation methods involving numerical optimisation. The model (2.1) can be factorized as follow

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

where, $[Y \mid S, X^*]$ is a product of $N\left(S\left(x_i^*\right), \tau^2\right)$ and $[S \mid X^*]$ is multivariate Gaussian with mean 0 and covariance matrix $\sigma^2\rho(X^*;\phi)$ and $[X_i^* \mid X_i] \sim N_2\left(X_i, \delta^2 I_2\right)$ since we are assuming Gaussian geomasking. Moreover, note that in the factorization $[Y \mid S, X, X^*] = [Y \mid S, X^*]$ we assume that $Y$ and $X$ are stochastically independent given $X^*$. This is a reasonable assumption because giving the true locations $X^*$, the observed locations do not provide further information about $Y$. The likelihood for this model accounting for positional error is $L(\theta, \delta) = [Y, X \mid \theta, \delta]$, with $\theta = (\sigma^2, \phi, \tau^2)$ the vector of parameters to be estimated, and can be written as

$$\begin{aligned}
L(\theta, \delta) &= [Y, X \mid \theta, \delta] \\
&= \int\int [Y, X, X^*, S \mid \theta, \delta]\, dS dX^* \\
&= \int\int [Y \mid X, X^*, S, \theta]\,[S, X, X^* \mid \theta, \delta]\, dS dX^* \\
&= \int\int [Y \mid X^*, S, \theta]\,[S \mid X^*, \theta]\,[X^* \mid X, \delta]\,[X]\, dS dX^* \\
&\propto \int\int [Y \mid X^*, S, \theta]\,[S \mid X^*, \theta]\,[X^* \mid X, \delta]\, dS dX^*, \qquad (2.9)
\end{aligned}$$

As the integration with respect to $S$ can be performed exactly, equation (2.9) can be rewritten as

$$E_{X^*|X,\delta}[Y \mid X^*] = \int [Y \mid X^*]\,[X^* \mid X]\, dX^*, \qquad (2.10)$$

where $[Y \mid X^*, \theta] \sim N\left(0, \sigma^2\rho\left(X^*;\phi\right) + \tau^2\right)$. Fanshawe and Diggle (2011) propose to evaluate (2.10) by Monte Carlo integration. This means that for each value of $(\theta, \delta)$, the likelihood can be therefore estimated by drawing $n_k$ independent samples $X_k^*$, each of

length $n$, from $[X^* \mid X, \delta]$, evaluating the density $f_k \equiv f\left(y \mid x_k^*, \theta\right)$ for each sample, and then computing $n_k^{-1} \sum_k f_k$. Maximization of the likelihood can then be performed using an optimization algorithm. Noting that $X^*$ appears in the variance covariance function of $f\left(y \mid x_k^*, \theta\right)$ this means that for each step of the maximization algorithm we need to do $k$ inversions of a $n \times n$ matrix. This leads to a considerable computational burden of order $O\left(kn^3\right)$, indeed merely computing maximum likelihood estimates for 80 points takes around 72 hours. As the authors highlight this also make reliable estimation of standard errors impractical.

### 2.5.1 Composite likelihood

To overcome this problem we propose to approximate the likelihood through the use of composite likelihood. It is as method of inference that combines conditional or marginal density together to approximate the full likelihood. The resulting estimating equation obtained from the derivative of the composite log-likelihood is an unbiased estimating equation (Varin *et al.*, 2011). This approach has been applied to standard geostatistical models to make computations faster when the number of spatial locations is demanding (Vecchia 1988; Hjort *et al.* 1994; Curriero and Lele 1999; Stein *et al.* 2004; Caragea and Smith 2006, 2007; Mateu *et al.* 2007; Bevilacqua *et al.* 2012; Bevilacqua and Gaetan 2015). We refer the reader to Varin *et al.* (2011) for a thorough review on composite likelihood methods. When inference is focused on the dependence structure, we could either use composite marginal log-likelihoods based on pairwise differences or on pairwise observation

$$l_{diff}\left(\theta, y\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \log f\left(y_i - y_j; \theta\right) \tag{2.11}$$

$$l_{pair}\left(\theta, y\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \log f\left(y_i, y_j; \theta\right). \tag{2.12}$$

We have fitted the model using both (2.11) and (2.12) but since results from $l_{pair}$ are superior we will not report results obtained with $l_{diff}$. Using equation (2.12) and noting that our model depends only on the distance between pairs of observations, we

can rewrite equation (2.10) as

$$\begin{aligned} l_1(\theta,\delta) &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \log E_{U^*|U,\delta}\left[Y_i, Y_j \mid U_{ij}^*\right] \\ &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \log \int_0^{\infty} \left[Y_i, Y_j \mid U_{ij}^*\right]\left[U_{ij}^* \mid U_{ij}\right] dU_{ij}^*, \end{aligned} \tag{2.13}$$

with $\left[Y_i, Y_j \mid U_{ij}^*\right] \sim N\left(0, \sigma^2\rho\left(U_{ij}^*;\phi\right) + \tau^2 I_2\right)$, $\left[U_{ij}^* \mid U_{ij}\right] \sim Rice\left(u_{ij}, \sqrt{2}\delta\right)$ and $U_{ij}^* = \left\|X_i^* - X_j^*\right\|$. Hence, for obtaining the likelihood we need $n(n-1)/2$ computations of a one-dimensional integral with no matrix inversion involved. This results in a huge computational gain. We are able to obtain ML estimates for 80 points in only 1 minute. Moreover, as soon as $\rho\left(u_{ij}^*;\phi\right) \to 0$, the above equation will reduce to

$$l_2\left(\sigma^2,\tau^2\right) \approx \frac{1}{2\pi\left(\sigma^2+\tau^2\right)} \exp\left\{-\frac{y_i^2+y_j^2}{2\left(\sigma^2+\tau^2\right)}\right\}.$$

We can use this result to reformulate the log-likelihood in the following way

$$l(\theta,\delta) = l_1 I\left(u_{ij} \le t\right) + l_2 I\left(u_{ij} > t\right), \tag{2.14}$$

where $t$ is a threshold calculated with a numerical search algorithm such that $\rho(t;\phi) \approx 0$ and $I(\cdot)$ is an indicator function. Integral in 2.13 has shown some numerical instability. We use quasi Monte Carlo to calculate it. We avoid numerical instability and it is faster. To be more specific, we proceed as follows:

1. Decide the number of points $n$ at which we are going to evaluate the integral.
2. Generate a quasi-random low-discrepancy sequence of $n$ numbers. We choose the Halton sequence because it is suggested when the dimension of the integral is $\le 6$.
3. Convert the sequence to the actual distribution using either the quantile function of a $Rice\left(u_{ij}, \sqrt{2}\delta\right)$ or the quantile function of $N\left(x_i, \delta^2\right)$ (since we can also rewrite the integral respect to the coordinates).
4. Compute $\frac{1}{n}\sum_{i=1}^{n}\left[Y_i, Y_j \mid U_{ij}^*\right]$ with $u_{ij}^*$ the sequence obtained at step 3.

We suggest, to rewrite the one-dimensional integral as a four-dimensional integral respect to the spatial locations since evaluating four quantile functions of a Normal is faster than evaluating one quantile function of a Rice distribution.

## 2.6 Simulation study

We conduct a simulation study to quantify the effects of positional errors on parameter estimation as follows.

1. Generate $n = 1000$ locations from $[X^*]$ a homogeneous Poisson process over the square $[0, 15] \times [0, 15]$.
2. Simulate the outcome data from $Y \sim MVN\left(0, \tau^2 + \sigma^2 \left\{1 - \rho\left(u^*\right)\right\}\right)$.
3. Simulate from $[X|X^*]$ using Gaussian geomasking to obtain $X$.
4. Estimate $\theta$ to obtain $\hat{\theta}_i$ for the $i$-th simulated data-set using:
   - variogNaive, a parametric fit to the variogram that ignores positional errror using weighted least squares (WLS);
   - variogAdj, a parametric fit to the variogram that corrects for positional error using WLS;
   - geoNaive, a linear geostatistical model that ignores positional error;
   - CL, the composite likelihood method of Section 2.5.1;
   - ACL1, as CL but assuming pairs of observations $Y_i$ and $Y_j$ to be independent for values of the spatial correlation below $5 \times 10^{-2}$;
   - ACL2, as CL but assuming pairs of observations $Y_i$ and $Y_j$ to be independent for values of the spatial correlation below $5 \times 10^{-6}$;
5. Repeat from 1 to 4 for $s = 500$ times.
6. Calculate the average of the estimated parameters as

$$\frac{1}{s}\sum_{i=1}^{s}\hat{\psi}_i$$

   and the root-mean-square-error (RMSE)

$$\sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(\hat{\psi}_i - \psi\right)^2}.$$

We define the following scenarios: (a) $\sigma^2 = 1$, $\tau^2 = 0$, $\kappa = 0.5$ and $\phi = 0.25$; (b) $\sigma^2 = 1$, $\tau^2 = 0$, $\kappa = 1.5$ and $\phi = 0.16$. In both scenarios, we let $r = \delta/\phi$ vary over the set $\{0.2, 0.4, 0.6, 0.8, 1\}$. We can observe from Table 2.2 and 2.3 that if positional error is

not taken into account $\tau^2$ and $\phi$ are systematically overestimated and $\sigma^2$ is systematically underestimated. We already anticipated the first two effects from the analysis of equation (2.5). The distortion on $\sigma^2$ is a consequence of the artificial nugget effect. Indeed, the the estimated total variance $\sigma^2 + \tau^2 \approxeq 1$ is not affected by the positional error. With the corrections proposed we are able to obtain consistent estimates of $\theta$ with also a smaller RMSE compared to the naive methods. In particular, the model-based solution performs always better, in terms of efficiency, than estimation based on the sample variogram. Along with the results from the full composite likelihood (equation (2.13)) we report estimates obtained using its approximate version (equation (2.14)) and conclude that we can obtain a considerable extra computational gain without a noticeable difference in the results. However, the statical efficiency of our methods decrease with increasing $r$. In general, the drawbacks of locational uncertainty are less evident for the Matrn with$\kappa = 1.5$. This is easily explained because if the true process that has generated the data is smoother, then at a fixed distance observations will be more correlated and so less affected by a possible displacement. Table 2.4 reports results for uniform geosmaked data. The above observations still hold and our correction is suitable also for this type of geomasking.

## 2.7 Application

We analyse data on height-for-age Z-scores (HAZs) from a Demographic and Health Survey (Burgert *et al.*, 2013) conducted in Senegal in 2011. HAZs are a measure of the deviation from standard growth as defined by the WHO Growth Standards and are comparable across ages and gender. A HAZ below -2 indicates stunted growth in a child and, if close to 0, normal growth instead.

In this survey, the sampling unit are clusters of households within a predefined geographic area known as census enumeration area (EA). An EA can be a city block or apartment building in urban areas, while in rural areas this can be a village or group of villages. The estimated centre of each cluster is recorded as a latitude/longitude coordinate, obtained from a GPS receiver or derived from public online maps or gazetteers (Gething *et al.*, 2015). To preserve the confidentiality of survey respondents, uniform geomasking was applied to the cluster centres. To take into account the different population density, different values for the maximum displacement distance were applied to urban and rural locations, more specifically $\delta_{urbarn} = 2$ km and $\delta_{rural} = 5$ km.

The data consist of 384 clusters, of which 122 are urban, with 10 children per cluster on average. Our outcome of interest, $Y_i$, is the average HAZ for a cluster which we

model as

$$Y_i = \mu + S(x_i) + Z_i \tag{2.15}$$

where $Z_i \sim N(0, \tau^2/n_i)$ and $n_i$ is the number of children at $i$-th cluster. To account for positional error, we approximate uniform geomasking with its Gaussian counterpart as explained in Section 2.4.2. Moreover, we also extend our model to consider the heteroscedasticity of the geomasking applied in this case. Table 2.1 reports the results for the estimation of the model parameters from the naive geostatistical model and correction based on the composite likelihood. We were not able to obtain reliable estimates from the variogram-based correction due to the relatively high noise to signal ratio.

TABLE 2.1: Parameter estimates and corresponding 95% confidence intervals (CI) for the fitted linear geostatistical models to malnutrition data of Section 2.7. "geoNaive" is the naive approach which ignores positional error, while "CL" is the proposed approach based on the composite likelihood.

| | geoNaive | | CL | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $\mu$ | -1.303 | (-1.470, -1.137) | -1.159 | (-1.562, -0.736) |
| $\sigma^2$ | 0.117 | (0.045, 0.289) | 0.197 | (0.146, 0.257) |
| $\phi$ | 44.669 | (9.184, 80.138) | 25.860 | (17.782, 37.614) |
| $\tau^2$ | 0.536 | (0.081, 0.994) | 0.464 | (0.409, 0.521) |

Compared to our model, the naive geostatistical model estimates a bigger nugget variance, a smaller $\sigma^2$ and a bigger $\phi$. This is perfectly in line with the bias that we would expect in presence of geomasking. Moreover, the magnitude of the bias seems to be in agreement with the ratio between $\delta$ and $\phi$. We can estimate it using the average maximum displacement in our dataset $\bar{\delta} = 4.05$ and $\hat{\phi} = 25.85$ estimated from our model that leads to $\hat{r} = 0.16$.

## 2.8 Discussion

In this chapter we analysed the effects of positional error and, in particular, of geomasking on a linear geostatistical model. Using the variogram as a tool to detect the spatial structure of the data we show how this is biased when location uncertainty is present. We obtained equations that quantify the bias and found that geomasking is the cause of overestimation of the spatial range and underestimation of the variance of the underlying true process. Moreover it creates an artificial nugget effect. Two types of correction were then proposed. The first, following the classic geostatistical framwework, is based on a non-linear curve-fitting of the variogram. The second, is a model-based

approach, and makes use of composite likelihood achieving a huge computational gain compared to the method proposed by Fanshawe and Diggle (2011). We also propose an approximate version of our method that allows to obtain an extra computational gain without sacrificing the efficiency of the estimators. As expected, the likelihood-based correction performs markedly better, in terms of statistical efficiency, compared to the one based on the variogram. The corrections are suitable for different types of geomasking and consider also the case when different magnitude of displacements are applied to different categories of points (heteroscedastic geomasking).

Deciding a value for $\delta$ when applying geomasking is crucial and we have shown that the resulting bias depends both on the true scale parameter $\phi$ and the smoothness of the underlying process $\kappa$. Our suggestion for who has the role to preserve the confidentiality of spatial data is to first obtain an estimate of $\phi$ from the true data locations and apply the smallest level of $\delta$ possible providing also the resulting ratio $r$. This can be used a proxy of the level of bias that has been introduced.

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.25 | 0 | - |
| variogNaive | 0.894 (0.1115) | 0.276 (0.0365) | 0.101 (0.1011) | 0.2 |
| variogAdj | 0.950 (0.0694) | 0.263 (0.0291) | 0.014 (0.0148) | 0.2 |
| geoNaive | 0.834 (0.1663) | 0.287 (0.0377) | 0.149 (0.1494) | 0.2 |
| CL | 0.947 (0.0689) | 0.249 (0.0173) | 0.015 (0.0136) | 0.2 |
| ACL2 | 0.947 (0.0689) | 0.249 (0.0173) | 0.015 (0.0136) | 0.2 |
| ACL1 | 0.948 (0.0684) | 0.255 (0.021) | 0.015 (0.0141) | 0.2 |
| variogNaive | 0.734 (0.2656) | 0.320 (0.0703) | 0.279 (0.2794) | 0.4 |
| variogAdj | 0.947 (0.0717) | 0.263 (0.0352) | 0.002 (0.0017) | 0.4 |
| geoNaive | 0.677 (0.3229) | 0.333 (0.0833) | 0.321 (0.321) | 0.4 |
| CL | 0.948 (0.0711) | 0.248 (0.0182) | 0.002 (0.0015) | 0.4 |
| ACL2 | 0.948 (0.0711) | 0.248 (0.0183) | 0.002 (0.0015) | 0.4 |
| ACL1 | 0.949 (0.0706) | 0.253 (0.0253) | 0.002 (0.0016) | 0.4 |
| variogNaive | 0.590 (0.4098) | 0.408 (0.1585) | 0.444 (0.4443) | 0.6 |
| variogAdj | 0.945 (0.0724) | 0.274 (0.0407) | 0.007 (0.0095) | 0.6 |
| geoNaive | 0.542 (0.4575) | 0.388 (0.1384) | 0.456 (0.4565) | 0.6 |
| CL | 0.943 (0.0712) | 0.250 (0.0223) | 0.009 (0.0072) | 0.6 |
| ACL2 | 0.943 (0.0712) | 0.250 (0.0222) | 0.009 (0.0071) | 0.6 |
| ACL1 | 0.943 (0.0716) | 0.260 (0.0278) | 0.009 (0.0073) | 0.6 |
| variogNaive | 0.481 (0.5220) | 0.518 (0.2680) | 0.574 (0.574) | 0.8 |
| variogAdj | 0.933 (0.0922) | 0.287 (0.0490) | 0.030 (0.0379) | 0.8 |
| geoNaive | 0.429 (0.5706) | 0.437 (0.1867) | 0.566 (0.5664) | 0.8 |
| CL | 0.937 (0.0800) | 0.246 (0.0248) | 0.038 (0.0299) | 0.8 |
| ACL2 | 0.937 (0.0800) | 0.246 (0.0252) | 0.038 (0.0299) | 0.8 |
| ACL1 | 0.937 (0.0793) | 0.259 (0.0354) | 0.032 (0.0318) | 0.8 |
| variogNaive | 0.413 (0.6030) | 0.687 (0.4372) | 0.667 (0.667) | 1.0 |
| variogAdj | 0.934 (0.0970) | 0.302 (0.0594) | 0.023 (0.0268) | 1.0 |
| geoNaive | 0.345 (0.6548) | 0.493 (0.2426) | 0.653 (0.6528) | 1.0 |
| CL | 0.929 (0.0884) | 0.243 (0.0306) | 0.027 (0.0233) | 1.0 |
| ACL2 | 0.929 (0.0876) | 0.242 (0.0315) | 0.025 (0.0231) | 1.0 |
| ACL1 | 0.929 (0.0885) | 0.259 (0.0440) | 0.023 (0.0229) | 1.0 |

TABLE 2.2: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 0.5$.

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.16 | 0 | - |
| variogNaive | 0.951 (0.0949) | 0.169 (0.0233) | 0.051 (0.0737) | 0.2 |
| variogAdj | 0.967 (0.0856) | 0.167 (0.0219) | 0.034 (0.0593) | 0.2 |
| geoNaive | 0.947 (0.0786) | 0.168 (0.0123) | 0.048 (0.0511) | 0.2 |
| CL | 0.962 (0.0840) | 0.160 (0.0107) | 0.035 (0.0587) | 0.2 |
| ACL2 | 0.962 (0.0841) | 0.160 (0.0107) | 0.035 (0.0588) | 0.2 |
| ACL1 | 0.962 (0.0851) | 0.162 (0.0124) | 0.035 (0.0600) | 0.2 |
| variogNaive | 0.876 (0.1542) | 0.178 (0.0300) | 0.124 (0.1412) | 0.4 |
| variogAdj | 0.950 (0.1017) | 0.169 (0.0241) | 0.049 (0.0795) | 0.4 |
| geoNaive | 0.853 (0.1603) | 0.180 (0.0241) | 0.141 (0.1448) | 0.4 |
| CL | 0.948 (0.0987) | 0.161 (0.0115) | 0.049 (0.0783) | 0.4 |
| ACL2 | 0.948 (0.0987) | 0.161 (0.0115) | 0.049 (0.0783) | 0.4 |
| ACL1 | 0.947 (0.1010) | 0.162 (0.0139) | 0.050 (0.0813) | 0.4 |
| variogNaive | 0.788 (0.2299) | 0.191 (0.0438) | 0.217 (0.2281) | 0.6 |
| variogAdj | 0.952 (0.1049) | 0.169 (0.0296) | 0.050 (0.0919) | 0.6 |
| geoNaive | 0.757 (0.2522) | 0.195 (0.0391) | 0.240 (0.2437) | 0.6 |
| CL | 0.949 (0.1044) | 0.160 (0.0129) | 0.049 (0.0865) | 0.6 |
| ACL2 | 0.949 (0.1049) | 0.160 (0.0129) | 0.049 (0.0871) | 0.6 |
| ACL1 | 0.948 (0.1074) | 0.162 (0.0191) | 0.050 (0.0904) | 0.6 |
| variogNaive | 0.688 (0.3238) | 0.211 (0.0612) | 0.325 (0.3325) | 0.8 |
| variogAdj | 0.941 (0.1185) | 0.173 (0.0322) | 0.066 (0.1139) | 0.8 |
| geoNaive | 0.655 (0.3519) | 0.214 (0.0576) | 0.345 (0.3486) | 0.8 |
| CL | 0.937 (0.1195) | 0.161 (0.0140) | 0.063 (0.1064) | 0.8 |
| ACL2 | 0.937 (0.1195) | 0.161 (0.0140) | 0.063 (0.1065) | 0.8 |
| ACL1 | 0.935 (0.1249) | 0.164 (0.0203) | 0.065 (0.1122) | 0.8 |
| variogNaive | 0.600 (0.4082) | 0.239 (0.0902) | 0.420 (0.4256) | 1.0 |
| variogAdj | 0.924 (0.1465) | 0.182 (0.0422) | 0.088 (0.1454) | 1.0 |
| geoNaive | 0.567 (0.4377) | 0.234 (0.0777) | 0.433 (0.4355) | 1.0 |
| CL | 0.918 (0.1441) | 0.161 (0.0208) | 0.083 (0.1341) | 1.0 |
| ACL2 | 0.917 (0.1449) | 0.161 (0.0208) | 0.083 (0.1349) | 1.0 |
| ACL1 | 0.914 (0.1524) | 0.166 (0.0217) | 0.087 (0.1427) | 1.0 |

TABLE 2.3: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 1.5$.

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.16 | 0 | - |
| variogNaive | 0.948 (0.0964) | 0.166 (0.0222) | 0.047 (0.0698) | 0.2 |
| variogAdj | 0.964 (0.0864) | 0.165 (0.0210) | 0.031 (0.0553) | 0.2 |
| geoNaive | 0.944 (0.0824) | 0.166 (0.0123) | 0.047 (0.0505) | 0.2 |
| CL | 0.961 (0.0830) | 0.160 (0.0113) | 0.032 (0.0551) | 0.2 |
| ACL2 | 0.961 (0.0831) | 0.160 (0.0113) | 0.032 (0.0552) | 0.2 |
| ACL1 | 0.960 (0.0852) | 0.161 (0.0133) | 0.033 (0.0575) | 0.2 |
| variogNaive | 0.896 (0.1372) | 0.176 (0.0272) | 0.113 (0.1320) | 0.4 |
| variogAdj | 0.966 (0.0928) | 0.168 (0.0214) | 0.042 (0.0735) | 0.4 |
| geoNaive | 0.861 (0.1527) | 0.180 (0.0238) | 0.140 (0.1442) | 0.4 |
| CL | 0.961 (0.0900) | 0.161 (0.0114) | 0.043 (0.0721) | 0.4 |
| ACL2 | 0.961 (0.0900) | 0.161 (0.0114) | 0.043 (0.0720) | 0.4 |
| ACL1 | 0.960 (0.0923) | 0.163 (0.0141) | 0.044 (0.0750) | 0.4 |
| variogNaive | 0.790 (0.2280) | 0.190 (0.0401) | 0.220 (0.2306) | 0.6 |
| variogAdj | 0.957 (0.1046) | 0.168 (0.0252) | 0.050 (0.0900) | 0.6 |
| geoNaive | 0.755 (0.2540) | 0.195 (0.0387) | 0.246 (0.2488) | 0.6 |
| CL | 0.953 (0.1034) | 0.159 (0.0145) | 0.050 (0.0873) | 0.6 |
| ACL2 | 0.953 (0.1035) | 0.159 (0.0145) | 0.050 (0.0873) | 0.6 |
| ACL1 | 0.952 (0.1054) | 0.161 (0.0181) | 0.051 (0.0899) | 0.6 |
| variogNaive | 0.679 (0.3323) | 0.213 (0.0631) | 0.331 (0.3385) | 0.8 |
| variogAdj | 0.931 (0.1304) | 0.175 (0.0331) | 0.074 (0.1222) | 0.8 |
| geoNaive | 0.649 (0.3578) | 0.215 (0.0583) | 0.349 (0.3519) | 0.8 |
| CL | 0.928 (0.1286) | 0.161 (0.0155) | 0.070 (0.1136) | 0.8 |
| ACL2 | 0.928 (0.1288) | 0.161 (0.0155) | 0.071 (0.1138) | 0.8 |
| ACL1 | 0.926 (0.1323) | 0.164 (0.0195) | 0.072 (0.1178) | 0.8 |
| variogNaive | 0.588 (0.4198) | 0.241 (0.0935) | 0.429 (0.4346) | 1.0 |
| variogAdj | 0.911 (0.1595) | 0.182 (0.0469) | 0.098 (0.1599) | 1.0 |
| geoNaive | 0.557 (0.4478) | 0.235 (0.0795) | 0.440 (0.4433) | 1.0 |
| CL | 0.906 (0.1585) | 0.158 (0.0234) | 0.091 (0.1452) | 1.0 |
| ACL2 | 0.907 (0.1585) | 0.158 (0.0234) | 0.091 (0.1453) | 1.0 |
| ACL1 | 0.904 (0.1627) | 0.162 (0.0291) | 0.093 (0.1501) | 1.0 |

TABLE 2.4: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 1.5$. Locations are displaced using Uniform geomasking.

# Chapter 3

# Effects of positional errors on spatial GLM and point-pattern analysis

## 3.1 Introduction

If we consider the taxonomy of spatial processes we can separate discrete spatial variation from continuous spatial variation. This primary distinction is between a phenomenon that is defined on a finite (or countably infinite) set of locations, and one that is defined on a continuous spatial region, $A \in \mathbb{R}^2$. If we consider the first category (discrete spatial variation) it is obvious that positional error is not a problem since we work with areal data instead of point data. Within the second category, continuous spatial variation, we can further distinguish real-valued processes, $\{S(x) \in \mathbb{R}^2\}$, from point processes whose realizations are a countable sets of points, $\mathcal{X} = \{x_i \in \mathbb{R}^2 : i = 1, 2, \ldots\}$. The secondary distinction between spatially continuous real-valued processes and point processes is made because the tools needed to analyse data from the two types of process turn out to be somewhat different. In the previous chapter we have studied the effect of positional errors, and in particular of geomasking, on Gaussian data generated from a real-valued spatial process. In this chapter we further extend our analysis to two other frameworks: non-Gaussian data whose underlying spatial variation comes from a real-valued spatial process and point processes.

If in the case of Gaussian data the solution, reported in the literature, for the presence of positional errors are very limited, when we talk about non-Gaussian there are no solution at all. Hence, the corrections provided here are of great help. Instead, in the case of point pattern analysis a first attempt to explore the effects of positional errors has been made by Arbia *et al.* (2017). Starting from a homogeneous point process they show that patterns of clustering or inhibition may be observed not as genuine phenomena but

only as the effect of data imperfections.

## 3.2 Generalized linear geostatisical models

Data whose stochastic variation is known to be non-Gaussian are very frequent in a lot of contexts. In particular, they are standard output in epidemiological and health studies, where they usually arise as disease counts or as prevalence data (Woodward, 2013). Observations of this type can be treated as either spatially indexed Poisson or binomial counts conditional on an unobserved spatially varying intensity (or relative risk surface). Diggle *et al.* (1998) extended the framework of generalized linear models, as introduced by Nelder and Wedderburn (1972) for independently replicated data, to geostatistical data to deal with non-Gaussian distributional assumptions. The class of models they introduced is based on two general assumptions:

1. The spatially varying outcome is linked by a one-to-one function to a Gaussian random field with certain parametric mean and covariance functions.

2. For any set of locations the observations of the response variable at these locations are conditionally independent given the values of the Gaussian random field at these locations.

Let $\{S(x) : x \in \mathbb{R}^2\}$ be the Gaussian random field that is functionally related to the spatially varying attribute of interest, and $S = (S(x_1), \ldots, S(x_n))^T$. Each observed value $Y_i$ is then stochastically related to the attribute of interest at $x_i$. The general model can be then hierarchically specified as follows:

$$\begin{aligned} Y_i \mid S(x_i) &\sim p(\cdot \mid \mu_i), \ \ i = 1, \ldots, n \\ \mu_i &= m_i g^{-1}(S(x_i)) \\ S &\sim N_n\left(D\beta, \Sigma + \tau^2 I_n\right) \end{aligned} \tag{3.1}$$

where:

- $\{Y_i : i = 1, \ldots, n\}$ are conditionally independent given $S$, and have marginal probability density function $p(\cdot \mid \mu_i)$.

- $\mu_i = E[Y_i \mid S(x_i)]$, $g(\cdot)$ is a known one-to-one link function and $m_i$ is an offset.

- $D = (1, d_1, \ldots, d_p)$ is a known $n \times (p+1)$ design matrix, with 1 the $n \times 1$ vector of ones and $d_j = (d_j(x_1), \ldots, d_j(x_n))^T$, where $d_j(x_i)$ is the value of

the $j$-th spatial varying covariate measuread at the $i$-th sampling location, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$are unkown regression parameteres.

- $\Sigma = (\sigma_{ij})$ is a $n \times n$ positive definite variance-covariance matrix with $\sigma_{ij} = \sigma^2 \rho(u_{ij})$ if $i \neq j$ and $\sigma_{ij} = \sigma^2 + \tau^2$ otherwise, where $\sigma^2 > 0$ is the unknown constant variance of the Gaussian random field, $\tau^2$ is the so called nugget effect, $\rho(u_{ij})$ is a parametric isotropic correlation function and $u_{ij} = \|x_i - x_j\|$ is the Euclidean distance.

Note that we have specified the nugget inside the variance covariance function of the spatial random effects $S$. This is equivalent to add some non-structured normal random effects $Z$ with mean zero and variance $\tau^2$ to the inverse of the link function $g^{-1}(S(x_i) + Z_i)$. If we use the idendity link function $g(\mu) = \mu$ and set $m_i = 1$ then we obtain the standard linear geostatistical model used in Chapter 2.

Although the above general framework can be used to model different types of non-Gaussian spatial data, we will concentrate on spatial count data and, in particular, on Poisson and binomial counts as already anticipated. We will consider the Poisson-lognormal and the bionomial-logitnormal spatial models. Their hierarchical specificiation is

$$\begin{aligned} Y_i \mid S(x_i) &\sim \text{Poisson}(\mu_i), \quad i = 1, \ldots, n \\ \mu_i &= m_i \exp(S(x_i)), \end{aligned} \tag{3.2}$$

for the Poisson model and

$$\begin{aligned} Y_i \mid S(x_i) &\sim \text{Binomial}(m_i, \mu_i/m_i), \quad i = 1, \ldots, n \\ \mu_i &= m_i \frac{\exp(S(x_i))}{1 + \exp(S(x_i))}. \end{aligned} \tag{3.3}$$

### 3.2.1 Variogram

If assumpations 1 and 2 stated in Section 3.2 hold, using the same definition of variogram given in Section 2.4 we can obtain the theoretical form of the variogram for the generalized geostatistial models

$$\begin{aligned} V_Y\left(u_{ij}^*\right) &= \frac{1}{2} var\{Y_i\} + \frac{1}{2} var\{Y_j\} - cov\{Y(x_i), Y(x_j)\} \\ &= \frac{1}{2} E_S[var_Y\{Y_i \mid S(x_i)\}] + \frac{1}{2} E_s[var_Y\{Y_j \mid S(x_j)\}] + \frac{1}{2} var_S\{E_Y[Y_i \mid S(x_i)]\} \\ &\quad + \frac{1}{2} var_S\{E_Y[Y_j \mid S(x_j)]\} - cov_S\{E_Y[Y_i \mid S(x_i)], E_Y[Y_j \mid S(x_j)]\}, \end{aligned} \tag{3.4}$$

using the fact that $E_S\left[cov_Y\left\{Y_i, Y_j \mid S\right\}\right] = 0$, since the observations are indipendent conditionally on $S$. Writing $\mu_i = E_Y\left[Y_i \mid S\left(x_i\right)\right]$ and $v_i = var_Y\left\{Y_i \mid S\left(x_i\right)\right\}$, equation (3.4) simplifies to

$$\begin{aligned} V_Y\left(u_{ij}^*\right) &= \frac{1}{2}\left[E_S\left\{v_i + v_j\right\} + var_S\left\{\mu_i\right\} + var_S\left\{\mu_j\right\}\right] - cov_S\left\{\mu_i, \mu_j\right\}. \\ &= \frac{1}{2}\left[2E_S\left\{v_i\right\} + var_S\left\{\mu_i - \mu_j\right\}\right] \\ &= \frac{1}{2}\left[2E_S\left\{v_i\right\} + E_S\left\{\left(\mu_i - \mu_j\right)^2\right\}\right]. \end{aligned} \tag{3.5}$$

The first term of equation (3.5) is a constant, which we can write as $2\bar{\tau}^2$ to emphasise that it is the average of the conditional variance over the distribution of $S$. Indeed, $\bar{\tau}^2$ can be interpreted as the analogous to the nugget varince in the stastionary Gaussian model studied in Chapter 2. Then, if we assume that the Gaussian random field is stationary with constant mean $\alpha$ we can write $\mu_i = g^{-1}\left(\alpha + S\left(x_i\right)\right)$ and using a first-order Taulor series approximation $g^{-1}\left(\alpha + S\right) \approx g\left(\alpha\right) + Sg^{-1\prime}\left(\alpha\right)$ we obtain an helpful equation of the variogram

$$V_y\left(u_{ij}^*\right) \approx g^{-1\prime}\left(\alpha\right)^2 V_S\left(u_{ij}^*\right) + \bar{\tau}^2.$$

Therefore we can conclude that the variogram on the non-Gaussian observations is approximately proportinal to the variogram of the Gaussian process $S$ plus an intercept which represents an average nugget effect induced by the variance of the error distribution of the model. It is obvious that also the theoretical variogram for a generalised linear geostatistical model will be biased in presence of positional error since it is a by product of the variogram of the Gaussian process that we have already shown to be biased in the previous Chapter. We expect to observe the same effects on the vector of parameters $\theta = \left(\sigma^2, \phi, \tau^2\right)^T$ and this will be confirmed through simulation in Section 3.2.4.

### 3.2.2 Correction for Poisson data

It is possible to obtain a closed form equation for (3.5) when we have Gaussian or Poisson distributed data. We introduce here the variogram for the Poisson case since for the Gaussian case it has been already analysed in Section 2.4. If we set

$\mu_i = \exp\{\alpha + S(x_i) + Z_i\}$, where $Z_i$ are *i.i.d.* $N(0, \tau^2)$ then (3.5) becomes

$$V_Y\left(u_{ij}^*\right) = \exp\left(\alpha + \frac{\sigma^2+\tau^2}{2}\right) + \exp\left(2\alpha + \sigma^2 + \tau^2\right)\left[\exp\left(\sigma^2+\tau^2\right) - \exp\left\{\sigma^2\rho\left(u_{ij}^*\right)\right\}\right]. \tag{3.6}$$

Using the same arguments of Section 2.4 we can obtain the theoretical variogram in presence of positional error for Poisson data

$$V_Y(u_{ij}) = \exp\left(\alpha + \frac{\sigma^2+\tau^2}{2}\right) + \exp\left(2\alpha + \sigma^2 + \tau^2\right)\left[\exp\left(\sigma^2+\tau^2\right) - \exp\left\{\sigma^2 E\left[\rho\left(U_{ij}^* \mid U_{ij}\right)\right]\right\}\right], \tag{3.7}$$

wehre $E\left[\rho\left(U_{ij}^* \mid U_{ij}\right)\right] = \int \rho\left(u_{ij}^*\right)\left[U_{ij}^* \mid U_{ij}\right] du^*$ and $\left[U_{ij}^* \mid U_{ij}\right] \sim Rice\left(u_{ij}, \sqrt{2}\delta\right)$. We can then estimate the vector of parameters $\theta$ using this result with the same procedure of Section 2.4.1, N-weighted least squares

$$S_n(\theta) = \sum_{k=1}^{m} n_k \left\{v_k - V_Y(u_k;\theta)\right\}^2, \tag{3.8}$$

all the quantities inside equation (3.8) are the defined as in Chapter 2 apart from $V_Y(\cdot)$ that is replaced with equation (3.7).

### 3.2.3 Correction for Binomial data

Unfortunately, if the observed data have a Binomial distribution it is not possible to obtain a closed form of the variogram. In this case we suggest to use a trans-Gaussian approximation of the model. Before the introduction of generalised linear model for spatial data, a common technique to deal with non-Gaussian data was trans-Gaussian kriging (Cressie, 1993, pages 137-138). It consists of applying a marginal non-linear function $g(\cdot)$ to the data such that the resulting transformation $g(Y_i)$ is approximately Gaussian and standard Gaussian methods can so be used. With Binomial data a suitable function $g(\cdot)$ is the empirical logit. Hence, we propose to apply the variogram or the composite likelihood correction introduced in Chapter 2 to the empirical logit transformation of the data,

$$\tilde{Y}_i = \log\left(\frac{Y_i + 1/2}{m_i - Y_i + 1/2}\right)$$

where we assume that $\tilde{Y}_i \mid S(x_i) \sim N\left(d(x_i)^T\beta + S(x_i), \tau^2\right)$ with $S(x)$ having the same properties as previously defined. Caution should be exercised when applying this

transformation to Binomial data. Useful guidelines are that the goodness of the Gaussian approximation deteriorates both as the binomial denominators decrease and/or the overall prevalence of the outcome of interest becomes very small or very large, that is, approaches either zero or one (Stanton and Diggle, 2013).

### 3.2.4 Simulation study

Similar to what we have done in Section 2.6, also here we conduct a simulation study to asses the effects of geomasking on parameters estimation of spatial GLM. We will proceed as follows:

1. Generate $n = 1000$ locations from $[X^*]$ a homogeneous Poisson process over the square $[0, 15] \times [0, 15]$.
2. Simulate the outcome data from models (3.2) and (3.3).
3. Generate the observed locations $X$ from $[X|X^*]$ using Gaussian geomasking.
4. Estimate $\theta$ to obtain $\hat{\theta}_i$ for the $i$-th simulated data-set using:
   - variogNaive, a parametric fit to the variogram that ignores positional errror using weighted least squares (WLS);
   - variogAdj, a parametric fit to the variogram that corrects for positional error using WLS;

   for both the Poisson model and the Binomial model and
   - geoNaive, a linear geostatistical model that ignores positional error;
   - CL, the composite likelihood method of Section 2.5.1;
   - ACL1, as CL but assuming pairs of observations $Y_i$ and $Y_j$ to be independent for values of the spatial correlation below $5 \times 10^{-2}$;
   - ACL2, as CL but assuming pairs of observations $Y_i$ and $Y_j$ to be independent for values of the spatial correlation below $5 \times 10^{-6}$;

   only for the Binomial model.
5. Repeat from 1 to 4 for $s = 500$ times.
6. Calculate the average of the estimated parameters as

$$\frac{1}{s}\sum_{i=1}^{s}\hat{\psi}_i$$

and the root-mean-square-error (RMSE)

$$\sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(\hat{\psi}_i - \psi\right)^2}.$$

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.25 | 0 | - |
| Naive Variogram | 0.961 (0.0546) | 0.266 (0.0338) | 0.031 (0.0314) | 0.1 |
| Adjusted Variogram | 0.978 (0.0492) | 0.263 (0.0304) | 0.004 (0.0044) | 0.1 |
| Naive Variogram | 0.902 (0.1075) | 0.283 (0.0491) | 0.101 (0.1014) | 0.2 |
| Adjusted Variogram | 0.971 (0.0614) | 0.267 (0.0364) | 0.007 (0.0068) | 0.2 |
| Naive Variogram | 0.802 (0.1981) | 0.318 (0.0721) | 0.198 (0.1978) | 0.3 |
| Adjusted Variogram | 0.939 (0.0763) | 0.272 (0.0330) | 0.053 (0.0533) | 0.3 |
| Naive Variogram | 0.731 (0.2688) | 0.348 (0.0984) | 0.276 (0.2757) | 0.4 |
| Adjusted Variogram | 0.957 (0.0712) | 0.277 (0.0480) | 0.016 (0.0164) | 0.4 |
| Naive Variogram | 0.656 (0.3443) | 0.383 (0.1335) | 0.344 (0.3440) | 0.5 |
| Adjusted Variogram | 0.952 (0.0739) | 0.281 (0.0427) | 0.027 (0.0271) | 0.5 |
| Naive Variogram | 0.589 (0.4108) | 0.424 (0.1739) | 0.425 (0.4253) | 0.6 |
| Adjusted Variogram | 0.938 (0.0818) | 0.286 (0.0476) | 0.024 (0.0244) | 0.6 |
| Naive Variogram | 0.532 (0.4680) | 0.458 (0.2084) | 0.463 (0.4631) | 0.7 |
| Adjusted Variogram | 0.935 (0.0867) | 0.287 (0.0497) | 0.037 (0.0366) | 0.7 |
| Naive Variogram | 0.461 (0.5392) | 0.554 (0.3045) | 0.526 (0.5260) | 0.8 |
| Adjusted Variogram | 0.920 (0.0865) | 0.297 (0.0566) | 0.057 (0.0572) | 0.8 |
| Naive Variogram | 0.425 (0.5749) | 0.597 (0.3468) | 0.564 (0.5643) | 0.9 |
| Adjusted Variogram | 0.866 (0.1337) | 0.289 (0.0409) | 0.108 (0.1076) | 0.9 |
| Naive Variogram | 0.389 (0.6110) | 0.681 (0.4310) | 0.612 (0.6115) | 1.0 |
| Adjusted Variogram | 0.859 (0.1409) | 0.304 (0.0562) | 0.137 (0.1367) | 1.0 |

TABLE 3.1: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (Naive Variogram) and the correction proposed (Adjusted Variogram) for increasing levels of $r$. The true correlation function is Matrn with $\kappa = 0.5$. Data were generated from model (3.2).

We define the following scenarios: (a) $\sigma^2 = 1$, $\tau^2 = 0$, $\kappa = 0.5$ and $\phi = 0.25$; (b) $\sigma^2 = 1$, $\tau^2 = 0$, $\kappa = 1.5$ and $\phi = 0.16$; (c) $\sigma^2 = 1$, $\tau^2 = 0$, $\kappa = 2.5$ and $\phi = 0.13$. In all scenarios, we let $r = \delta/\phi$ vary over the set $\{0.2, 0.4, 0.6, 0.8, 1\}$. The value for $\phi$ has been chosen such that the practical range is approximately 0.74 for the three

scenarios. In this way the results are comparable. Output from simulation is reported in Tables 3.1, 3.2 and 3.3 for the Poisson model and in Tables 3.6, 3.7 and 3.8 for the Binomial model. As anticipated in Section 3.2.1, also for non-Gaussian data ignoring the uncertainty hidden in the spatial location due to geomasking leads to biased estimates: inflation of the nugget $\tau^2$ and the scale parameter $\phi$ and underestimation of the spatial variance $\sigma^2$. We can appreciate how the proposed corrections help us to correct the bias. However, the variogram based correction for Poisson data, even if preferable to the naive approach, is not able to provide consistent estimates when the positional error is too high.

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.16 | 0 | - |
| Naive Variogram | 0.959 (0.0656) | 0.161 (0.0148) | 0.011 (0.0114) | 0.1 |
| Adjusted Variogram | 0.968 (0.0628) | 0.161 (0.0147) | 0.003 (0.0028) | 0.1 |
| Naive Variogram | 0.945 (0.0625) | 0.166 (0.0151) | 0.037 (0.0366) | 0.2 |
| Adjusted Variogram | 0.975 (0.0590) | 0.163 (0.0142) | 0.001 (0.0010) | 0.2 |
| Naive Variogram | 0.909 (0.0943) | 0.172 (0.0160) | 0.073 (0.0729) | 0.3 |
| Adjusted Variogram | 0.966 (0.0603) | 0.164 (0.0138) | 0.011 (0.0112) | 0.3 |
| Naive Variogram | 0.872 (0.1282) | 0.180 (0.0210) | 0.120 (0.1196) | 0.4 |
| Adjusted Variogram | 0.957 (0.0638) | 0.171 (0.0194) | 0.009 (0.0093) | 0.4 |
| Naive Variogram | 0.827 (0.1730) | 0.187 (0.0272) | 0.158 (0.1578) | 0.5 |
| Adjusted Variogram | 0.956 (0.0654) | 0.168 (0.0127) | 0.000 (0.0003) | 0.5 |
| Naive Variogram | 0.779 (0.2212) | 0.197 (0.0367) | 0.203 (0.2033) | 0.6 |
| Adjusted Variogram | 0.958 (0.0618) | 0.170 (0.0148) | 0.000 (0.0000) | 0.6 |
| Naive Variogram | 0.744 (0.2562) | 0.211 (0.0511) | 0.254 (0.2542) | 0.7 |
| Adjusted Variogram | 0.956 (0.0650) | 0.174 (0.0208) | 0.002 (0.0016) | 0.7 |
| Naive Variogram | 0.694 (0.3058) | 0.219 (0.0591) | 0.295 (0.2953) | 0.8 |
| Adjusted Variogram | 0.948 (0.0670) | 0.177 (0.0181) | 0.002 (0.0017) | 0.8 |
| Naive Variogram | 0.650 (0.3499) | 0.228 (0.0685) | 0.345 (0.3453) | 0.9 |
| Adjusted Variogram | 0.955 (0.0650) | 0.175 (0.0177) | 0.006 (0.0056) | 0.9 |
| Naive Variogram | 0.620 (0.3803) | 0.241 (0.0807) | 0.380 (0.3801) | 1.0 |
| Adjusted Variogram | 0.968 (0.0682) | 0.177 (0.0196) | 0.005 (0.0045) | 1.0 |

TABLE 3.2: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (Naive Variogram) and the correction proposed (Adjusted Variogram) for increasing levels of $r$. The true correlation function is Matrn with $\kappa = 1.5$. Data were generated from model (3.2).

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.13 | 0 | - |
| Naive Variogram | 0.959 (0.0593) | 0.133 (0.0093) | 0.009 (0.0088) | 0.1 |
| Adjusted Variogram | 0.961 (0.0595) | 0.133 (0.0090) | 0.005 (0.0056) | 0.1 |
| Naive Variogram | 0.954 (0.0617) | 0.132 (0.0079) | 0.020 (0.0200) | 0.2 |
| Adjusted Variogram | 0.963 (0.0581) | 0.131 (0.0075) | 0.001 (0.0007) | 0.2 |
| Naive Variogram | 0.935 (0.0715) | 0.134 (0.0092) | 0.039 (0.0391) | 0.3 |
| Adjusted Variogram | 0.966 (0.0570) | 0.132 (0.0083) | 0.000 (0.0003) | 0.3 |
| Naive Variogram | 0.913 (0.0970) | 0.138 (0.0110) | 0.066 (0.0664) | 0.4 |
| Adjusted Variogram | 0.963 (0.0582) | 0.135 (0.0083) | 0.000 (0.0003) | 0.4 |
| Naive Variogram | 0.878 (0.1241) | 0.142 (0.0135) | 0.099 (0.0994) | 0.5 |
| Adjusted Variogram | 0.956 (0.0568) | 0.136 (0.0093) | 0.001 (0.0008) | 0.5 |
| Naive Variogram | 0.850 (0.1538) | 0.145 (0.0166) | 0.135 (0.1348) | 0.6 |
| Adjusted Variogram | 0.950 (0.0646) | 0.137 (0.0097) | 0.000 (0.0000) | 0.6 |
| Naive Variogram | 0.813 (0.1871) | 0.150 (0.0206) | 0.170 (0.1705) | 0.7 |
| Adjusted Variogram | 0.955 (0.0590) | 0.136 (0.0096) | 0.000 (0.0000) | 0.7 |
| Naive Variogram | 0.772 (0.2280) | 0.157 (0.0270) | 0.205 (0.2046) | 0.8 |
| Adjusted Variogram | 0.953 (0.0564) | 0.137 (0.0092) | 0.001 (0.0009) | 0.8 |
| Naive Variogram | 0.742 (0.2579) | 0.160 (0.0303) | 0.239 (0.2394) | 0.9 |
| Adjusted Variogram | 0.951 (0.0585) | 0.137 (0.0099) | 0.000 (0.0000) | 0.9 |
| Naive Variogram | 0.695 (0.3051) | 0.171 (0.0414) | 0.285 (0.2850) | 1.0 |
| Adjusted Variogram | 0.937 (0.0776) | 0.141 (0.0136) | 0.004 (0.0035) | 1.0 |

TABLE 3.3: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (Naive Variogram) and the correction proposed (Adjusted Variogram) for increasing levels of $r$. The true correlation function is Matrn with $\kappa = 2.5$. Data were generated from model (3.2).

### 3.2.5 Application

We illustrate further our methods using real data that consist of *Loa loa* (eyeworm) prevalence from a series of surveys undertaken in 197 villages in Cameroon and southern Nigeria. *Loa loa* is a filarial disease that is of interest to the African Programme for Onchocerciasis (APOC, see WHO (2013)), because individuals with high filarial loadings of these parasites can experience serious adverse reactions to the onchocerciasis prophylactic, ivermectin. As a result, APOC has declared a policy objective of identifying areas of high *Loa loa* prevalence within the 19 countries taking part in APOC.Specifically,

APOC policy states that in areas where *Loa loa* prevalence exceeds 20%, precautionary measures should be put in place before mass prophylactic treatment with ivermectin. This data-set has been already extensively analysed by Diggle *et al.* (2007).

For each of the $n$ villages we have the longitude and latitude $x_i^*$ of the $i$th study village, the number of individuals $m_i$ tested for the *Loa loa* infection (median 132, range 24 to 432) and the number of blood samples $Y_i$ that tested positive for the *Loa loa* parasite. Observed village-level prevalence (proportion of positive samples) ranges from 0 to 0.53, with median 0.12. The distance between villages in the study region ranges from 0.01 km to 1500 km, with a median distance of 895 km. Figure 3.1 shows the locations of sampled villages with the observed prevalence.

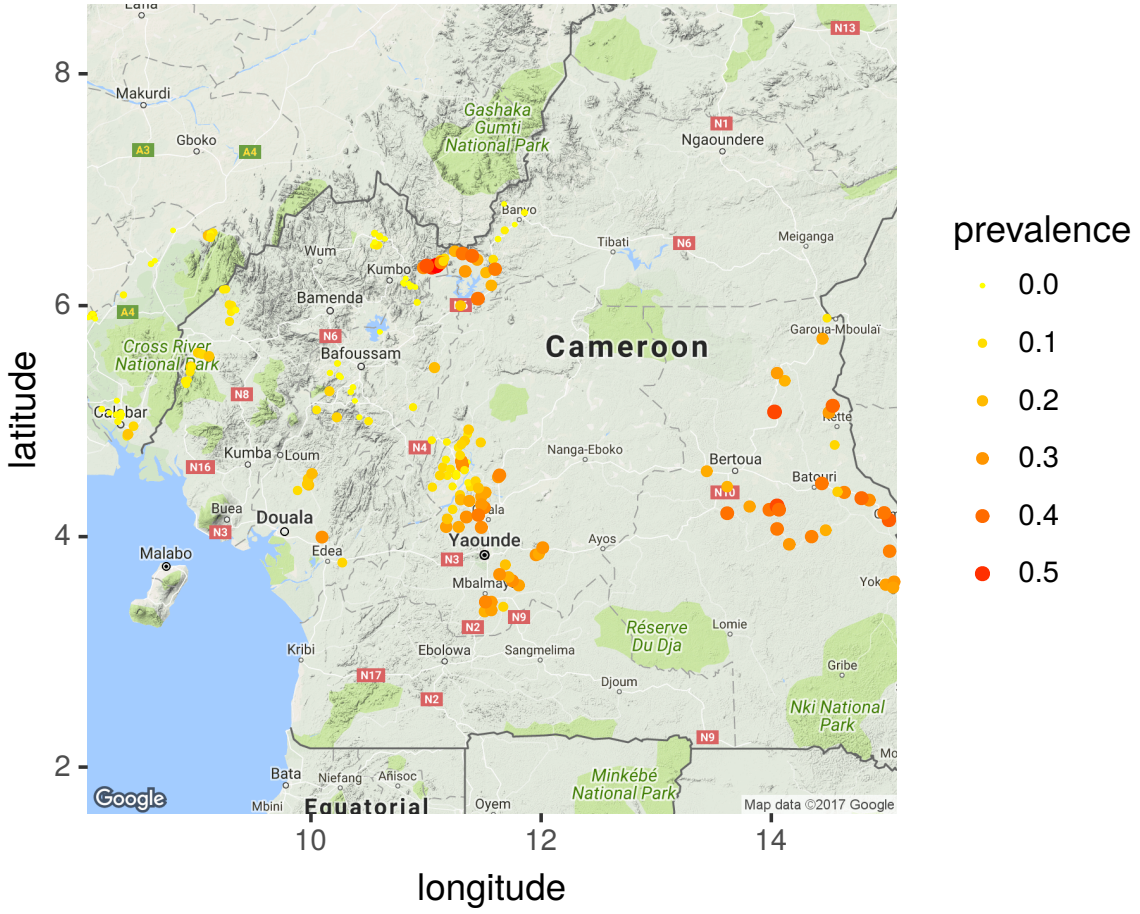


FIGURE 3.1: Sampling locations for the *Loa loa* data. Size and colour of the points indicates the level of prevalence observed.

We fitted the Gaussian model (2.1) to the logit transformed data, assuming a constant mean $\mu$ and treating $S(x)$ as a stationary Gaussian process characterised by a Matérn correlation function with $\kappa = 0.5$. This value has been chosen from a discrete set of candidate values, which we compared by evaluating the profile likelihood for $\kappa$ based

on the empirical logit transformation of the observed prevalence as reported in Figure 3.2. Since the maximum likelihood estimate is very close to 1/2, we then fix the shape parameter $\kappa$ at this value for the subsequent analysis.

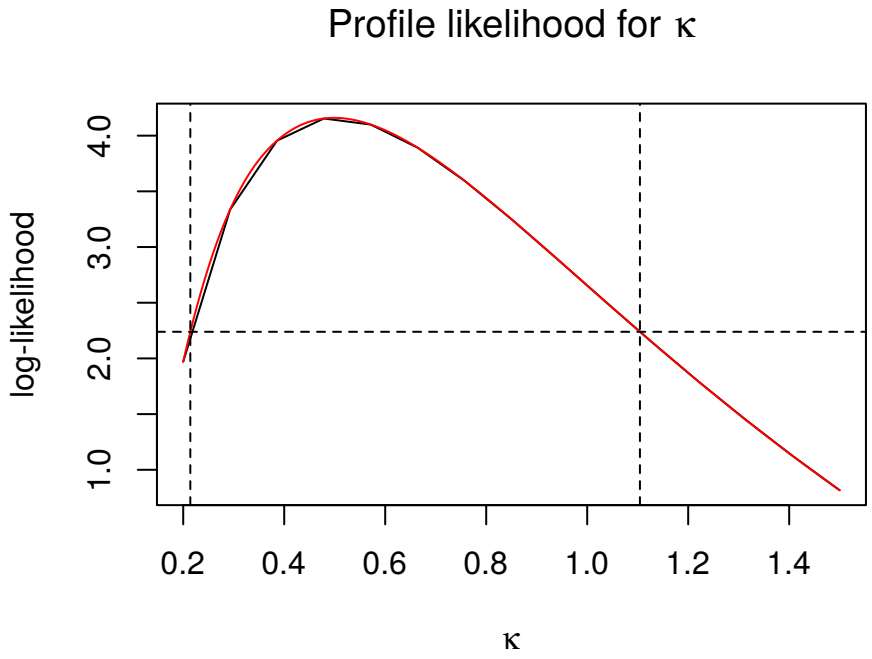


FIGURE 3.2: Profile likelihood for the shape parameter $\kappa$ of the Matérn covariance function. The profile likelihood (black solid line) is interpolated by a spline (red solid line), which is then used to obtain a confidence interval of coverage 95% (vertical dashed lines).

| **Scenario** | $\mu$ | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|---|
| **1** | -2.299 | 2.451 | 0.844 | 0.369 | - |
| **2a** | -2.345 | 2.158 | 12.848 | 1.659 | 0.5 |
| **2b** | -2.214 | 2.548 | 0.697 | 0.463 | 0.5 |

TABLE 3.4: Parameter estimates for the *Loa loa* data-set shown in Figure 3.1 under the following scenarios: (1) Using the original, true locations; (2a) Using the incorrect, geomasked locations with $\delta = 0.422$, making no allowance for positional error; (2b) As 2a, but correcting for positional error.

Treating the measurement locations as fixed, we found the maximum likelihood estimates of the parameters to be $\mu = -2.299$, $\sigma^2 = 2.451$, $\phi = 0.844$ and $\tau^2 = 0.369$. We then impose a Gaussian geomasking on the observed locations using a positional error standard deviation $\delta = 0.422$ such that $r \approxeq 0.5$. Using these new set of coordinates $x_i$ we refit the previous model and calculate the MLE from the composite likelihood

correction introduced in Section 2.5.1. Results are reported in Table 3.4. As we can see, ignoring positional error it's not a wise decision since it leads to biased estimates. With our correction applied to the empirical logit transformation of the observed prevalence we are able to recover the true parameters.

## 3.3 Point Pattern Analysis

A spatial point pattern, following the definition provided by Diggle (2013), is a countable set of locations $x_i$ irregularly distributed in a region, lets say $A$, that arise as the realizations of some stochastic mechanism. The region $A$ can be defined in $\mathbb{R}^d$ with $d \geq 1$ but we will consider only planar regions, hence $d = 2$. Indeed, this is usually the standard framework for the majority of real applications. The goal of point pattern analysis is to understand the spatial distribution of a certain variable and o try to individuate phenomena like clustering or repulsion. Usually, the strategy adopted is to compare the observed point pattern with a benchmark that is the homogeneous Poisson process. This type of process is characterised by the two following conditions

1. The numer of events in a study region $A$ with area $|A|$ follow a Poisson distribution with mean $\lambda |A|$.
2. Given $n$ events $x_i$ in a region $A$, the points $x_i$ are independent random samples from a Uniform distribution over $A$.

The constant $\lambda$ is the intensity or average number of points for unit area. The first condition implies that the intensity of the events does not vary spatially. The second condition guaranties that the $n$ events are independent and don't interact in any possible way. A pattern with these characteristics is called a CSR (complete spatial random) pattern. The hypothesis of complete spatial randomness is often unrealistic in practical applications but it is used as the null hypothesis to individuate statistical significant deviations from it. There are two wide classes of point processes that constitute violations to conditions 1 and 2 (and so deviations from the CSR hypothesis):

- cluster processes,
- inhibitory or regular processes.

In the following Section we will focus on cluster processes since this will be the object of study for our work. Figure 3.3 shows the realisations of an homogeneous Poisson process (left), a cluster Poisson process (centre) and a inhibitory Poisson process (right).

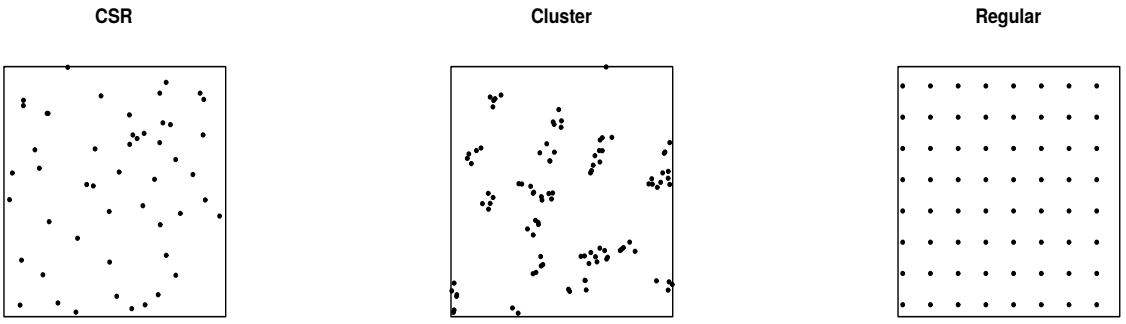


FIGURE 3.3: The three main typologies of spatial point patterns.

### 3.3.1 Poisson cluster processes

The violation of one or both the hypotheses stated in Section 3.3 can lead to the formation of a clustered pattern. If the first condition does not hold we could assist to clustering of events due to an apparent diffusion (or contagion). We use the term apparent because in this case the observed clusters are the results of the presence of spatial heterogeneity that makes the intensity of the process non constant. Instead, if te second condition is violated then we can observe a real diffusion phenomenon because it is the presence of an event in a specific part of the region that attracts other events (the condition on independence does not hold any more). In the first case the underlying process is called a inhomogeneous Poisson process, in the second case we deal with Poisson cluster processes.

If we define with $N(A) = n$ the random variable that generates the $n$ events in a finite planar region $A$, a inhomogeneous Poisson process is then defined through the two following properties:

1. $N(A) \sim Poisson\left(\int_A \lambda(x)\, dx\right)$.
2. The $n$ events in $A$ constitute an indipendent random sample from the distribution on $A$ having pdf proportional to $\lambda(x)$.

Poisson cluster processes were introduced by Neyman and Scott (1958) and incorporate an explicit form of spatial clustering. They are generated through the following three steps:

1. Parent events form a Poisson process with intensity $\xi$.
2. Each parent produces a random number $T$ of i.i.d. offspring, realized for each parent according to a probability distribution $p_t : t = 0, 1, \ldots, n$.

3. The positions of the offspring $x_t$ are i.i.d. realisations from a bivariate pdf $h(\cdot)$ (usually a normal or uniform distribution).

Poisson cluster processes as defined here are stationary, with intensity $\lambda = \xi\mu$ where $\mu = E[T]$. If the offspring of each parent point are uniformly distributed in a disc of radius $R$ centred around the parent we then have what is called a Matèrn cluster process (Matern, 1986). The spatial scale of the clusters is controlled by the radius $R$. Instead, in Thomas cluster process (Thomas, 1949), the probability density of offspring locations $h(\cdot)$ is an isotropic Gaussian density. Effectively, each offspring is randomly displaced from its parent, with the displacement vectors having an isotropic Gaussian distribution $N(0, \sigma^2 I)$ with standard deviation $\sigma$ along each coordinate axis. The spatial scale of the clusters is controlled by $\sigma$. This type of cluster process is extensively used in ecological and environmental studies to test for the presence of clustering. The way in which offspring are generated resemble the geomasking process. Indeed, we will exploit this fact to suggest a possible correction.

### 3.3.2 First and second moment properties

A spatial point process is mainly characterised by the first and second moment properties. Before introducing them is useful to specify when a spatial point process is stationary and isotropic:

- The process is stationary if, for every number $k$ and every region $A_i$ $(i = 1, \ldots, k)$, the joint distribution of $N(A_i), \ldots, N(A_k)$ is invariant under translation. This means that all the property of the process won't change after a translation of the plane.

- The process is isotropic if, for every number $k$ and every region $A_i$ $(i = 1, \ldots, k)$, the joint distribution of $N(A_i), \ldots, N(A_k)$ is invariant under rotation. This means that all the property of the process won't change after a rotation of the plane of an arbitrary angle $\theta$.

This characteristics will be reflected on the first and second moment properties as we will see soon. First moment properties describe how the expected value of the process varies spatially, instead, second moment properties describe the covariance (and correlation) between events of the process over the region $A$.

First order properties are described in terms of intensity function, $\lambda(x)$, of the process, as an indicator of the mean number of events per unit area. The intensity function

is defined through the following limit

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\},$$

where $dx$ is an infinitesimal region that contain the point $x$, $|dx|$ is its area and $N(dx)$ is the number of points that lies in the region $dx$. Hence, $\lambda(x)\, dx$ is the probability that an event is located in an infinitesimal region with area $|dx|$ and with centre the point $x$.

Second order properties, or spatial dependence, of a spatial point process summarize the relation between the number of events observed in couples of subregions. The second order intensity function is defined through the following limit

$$\lambda_2(x, y) = \lim_{|dx|,|dy| \to 0} \left\{ \frac{E[N(dx)\, N(dy)]}{|dx|\, |dy|} \right\},$$

where $x$ and $y$ denote the coordinates of two dinstinct points and $\lambda_2(x, y)\, dxdy$ is the probability that two points lie in two infinitesimal regions centered at $x$ and $y$ and with area equal to $|dx|$ and $|dy|$, respectively. Note that if $N(dx)$ and $N(dy)$ are uncorrelated $\lambda_2(x, y) = \lambda(x)\, \lambda(y)$. Another useful quantity is the covariance density of the process

$$\gamma(x, y) = \lambda_2(x, y) - \lambda(x)\, \lambda(y), \tag{3.9}$$

and, if we divide (3.9) by $\lambda(x)\, \lambda(y)$ and sum by 1 we obtain what is called the pair correlation function $g(x, y) = \lambda_2(x, y) / \lambda(x)\, \lambda(y)$. If we assume that the point process is stationary and isotropic, then it follows that

1. $\lambda(x) = \lambda = E[N(A)] / |A|$, the intensity does not vary spatially and is constant over the region $A$.

2. $\lambda_2(x, y) = \lambda_2(u) / \lambda^2$, with $u = \|x - y\|$ the Euclidean distance between $x$ and $y$. This indicates that the second order intensity function depends only on the distance between the two points and not on their locations in absolute terms.

While the first order intensity function is easily interpretable, we cannot say the same about the second order intensity. Indeed, for a stationary and isotropic spatial point process, the second order properties are described through a more easily interpretable function: the Ripley's K function (Ripley, 1976, 1977).

### 3.3.3 Ripley's K function

The $K$ function is a summary of the pairwise distances in the point pattern dataset, normalised to enable us to compare different datasets. It is defined as

$$K(u) = 2\pi\lambda^{-2} \int_0^u t\lambda_2(t)\, dt, \tag{3.10}$$

in particular, the quantity $\lambda K(u)$ is the expected number of further events within distance $u$ of an arbitrary event. This result gives the $K$ function a tangible interpretation as a scaled expectation. For a stationary Poisson process, is possible to show that

$$K(u) = \pi u^2.$$

This is a very useful result since we now have a value that can be used as a benchmark to validate the hypothesis of CSR and the deviations from it. Figure 3.4 shows the $K$ function for the three main types of point patterns. Positive deviations from the

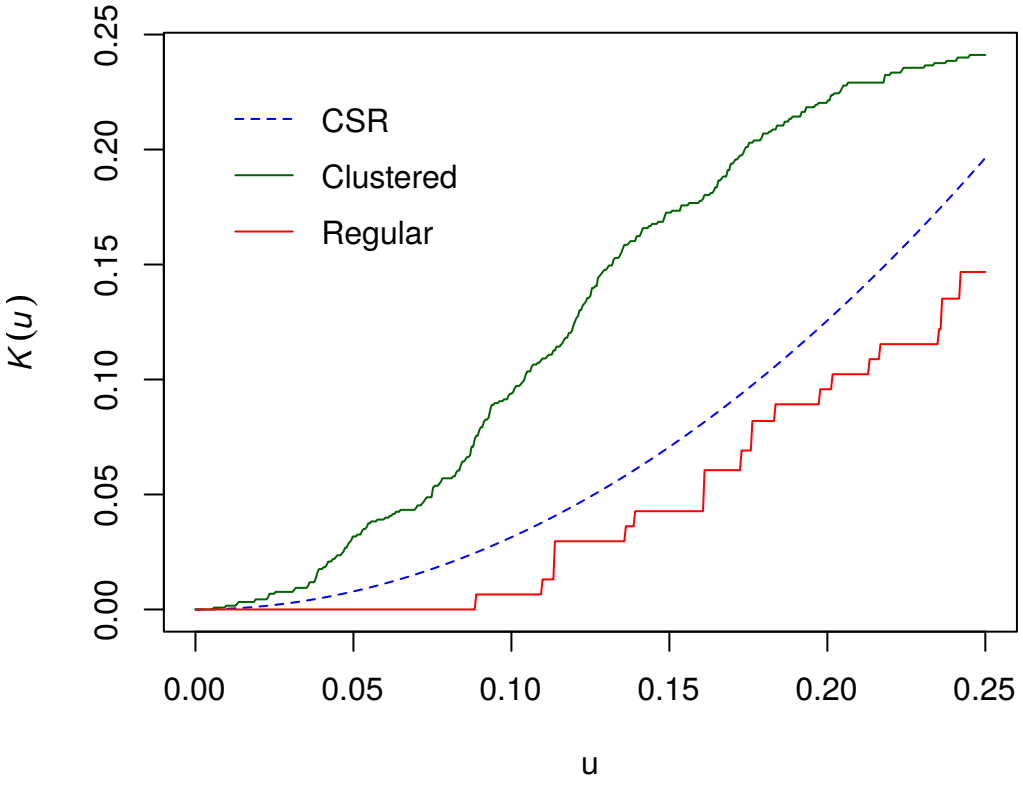


FIGURE 3.4: Empirical $K$ functions for the three patterns in Figure 3.3. Green line: clustered pattern. Blue line: independent pattern. Red line: regular pattern.

benchmark value $\pi u^2$ are evidence of some clustering going on; vice versa, negative deviations from $\pi u^2$ are signal of the presence of inhibitory patterns.

For a general Poisson cluster process, if the random number of offspring $T$ follows a Poisson distribution, the resulting $K$ function is

$$K(u) = \pi u^2 + \frac{H_2(u)}{\xi},$$

where $H_2(u)$ is the cumulative distribution function of the vector difference between the positions of the offspring from the same parent. For a Thomas process $H_2(u)$ is available in closed form and the $K$ function becomes

$$K_\psi(u) = \pi u^2 + \frac{1}{\xi}\left\{1 - \exp\left(-\frac{u^2}{4\sigma^2}\right)\right\}, \tag{3.11}$$

where $\psi = (\xi, \sigma)$. These results suggest a useful way of identifying whether a Poisson cluster process might be a reasonable model for an observed pattern, and if so a mean of obtaining preliminary parameter estimates.

### 3.3.4 Effects of positional error

In this section we consider what happens to a point pattern generated by a Poisson cluster process when the locations $x_i$ are affected by positional error. In particular, we will consider a practical case that is when random displacement (geomasking) is applied to the true original coordinates $x_i^*$ (see Section 2.2). Our guess is that as the magnitude of the displacement increases (parameter $\delta$ and $R$ in equations (2.2) and (2.7)) the spatial structure of the clusters will be destroyed and the observed point pattern will converge to a CSR pattern as if it was generated by a homogeneous Poisson process.

Following Diggle (1993), if the locations are displaced according to some symmetric positional error function $f(\cdot)$, then the resulting $K$ function is

$$K(u) = \pi u^2 + 2\pi\lambda^{-2}\int_0^\infty tP(u,t)\,\gamma(t)\,dt, \tag{3.12}$$

where $P(u,t) = \int_{\|x\|\le u} f(x-z)\,dz$ is the probability that the displacement induced by $f(\cdot)$ will move an event originally at the point $z$ to a point somewhere in the disc $\|x\| \le u$. We now compare equation (3.12) with $K^*(u) = \pi u^2 + 2\pi\lambda^{-2}\int_0^\infty t\gamma(t)\,dt$, that is the $K$ function of the true process. In (3.12) the integrand is attenuated by the function $P(u,t)$. If the perturbation distribution degenerates to a zero perturbation with probability 1, i.e. the positional error standard deviation $\delta = 0$, then $P(u,t) = 1$

if $t \leq u$ and $P(u,t) = 0$ otherwise, and $K(u) = K^*(u)$. Instead, if the perturbation distribution is highly dispersed, i.e. high values of $\delta$, then for any fixed $u$, $P(u,t) \approx 0$ for all $u$, and $K(u) \approx \pi u^2$. In intermediate cases, typically for any value of $u$, we have that $0 < P(u,t) < 1$ and $P(u,t)$ is monotone decreasing in $t$. Since typically $\gamma(t)$ is also positive and monotone decreasing in $t$, the effect of positional error is to reduce the value of the covariance integral so that,

$$\pi u^2 < K(u) < K^*(u).$$

The practical implication of this result is that second-moment analyses of randomly perturbed data are likely to be conservative, in the sense that they are likely to underestimate the true extent of spatial heterogeneity or clustering.

At the moment of writing a correction has not been implemented yet. However, we suggest the following solution. We can consider the observed point pattern after geomasking as a Poisson cluster process where the parents are the offspring of the true point pattern and each parent as one and one only child (offspring) generated applying the geomasking procedure. If we consider a Thomas process and Gaussian geomasking, it should then be possible to obtain a closed form for the $K$ function in presence of positional error. Model fitting and hypothesis testing is then straightforward. I

### 3.3.5 Simulation study

Here we provide further evidence of the drawbacks of ignoring the presence of positional error in point pattern analysis. We proceed as follows:

- Generate a clustered point pattern on the unit square as a realisation of a Thomas process with $\xi = 10$, $\mu = 4$ and $\sigma = 0.05$.
- for $i : 1, \ldots, 1000$:
  - introduce the positional error using Gaussian geomasking;
  - obtain an estimate of $K_i(u)$
- calculate the empirical Bias, the RMSE and the Type II error where the null hypothesis is the one of CSR.

Results are reported in Figure 3.5 and in Table 3.5. Also a small displacement of the true locations is enough to move the empirical $K$ function to the region of acceptance of the null hypothesis. This is even more clear if we look at the Type II error rate in Table 3.5. It monotonically increase with the positional error standard deviation $\delta$, this means

that we wrongly don't reject the null hypothesis of CSR. This is a great limitation in practice because makes the individuation of clusters with gemoasked a difficult task. Moreover, if the displacement is relatively high there is the additional risk to arrive to the opposite conclusion, that is to infer a spurious regular pattern.

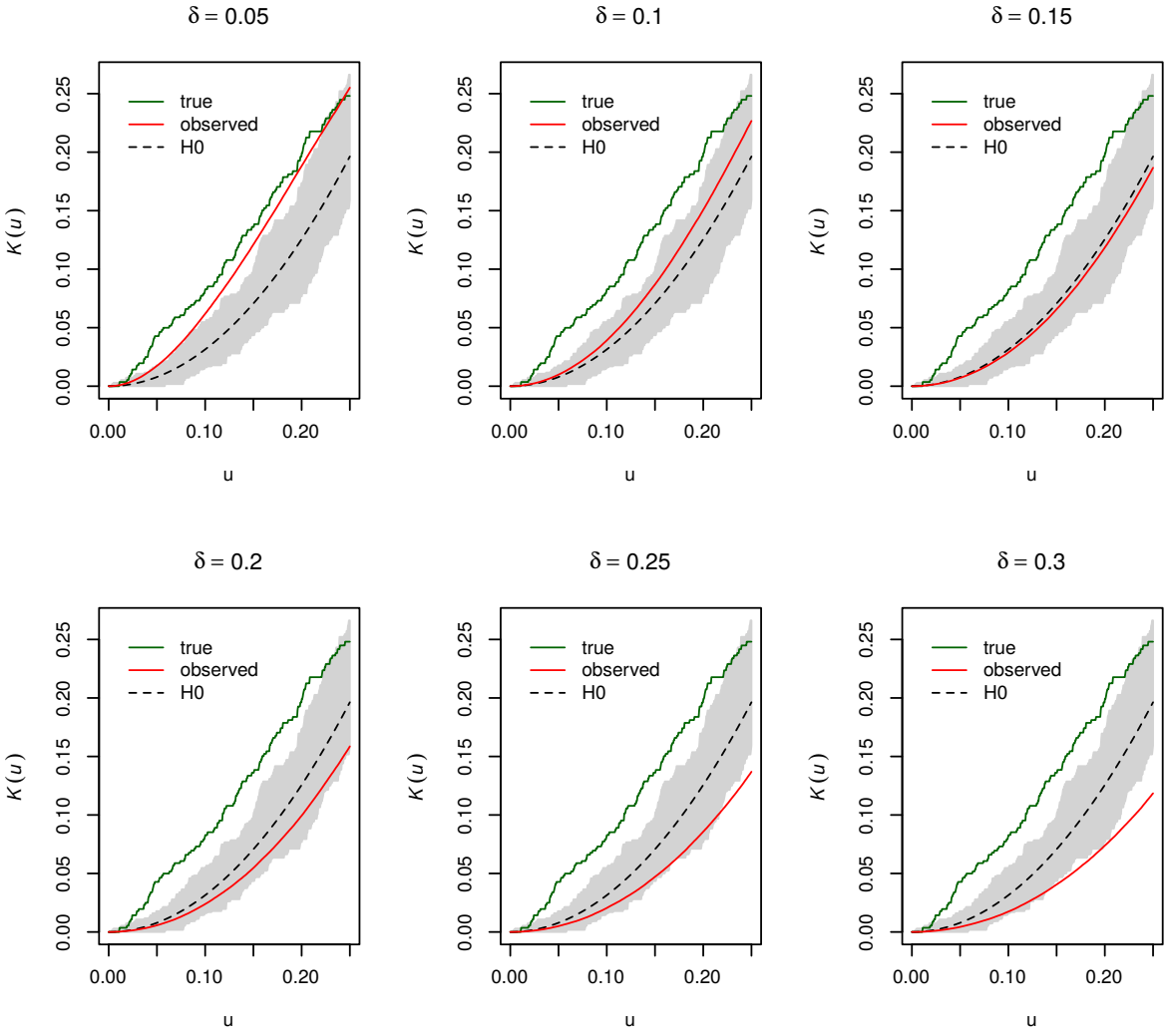


FIGURE 3.5: Empirical $K$ function for the true point pattern (green line). $K$ function calculated as the average estimate at each distance $u$ from Monte Carlo simulations (red lines). $K$ function for the null hypothesis (dashed black line) and confidence bands in grey.

## 3.4 Conclusions

In this chapter we extended the work done in Chapter 2 in two different directions. Since our study was constrained to Gaussian data we relaxed this assumption and provided corrections also for Poisson and binomial data. In the case of Poisson data

| $\delta$ | $\delta/\max(u)$ | Bias | RMSE | Type II error rate |
|---|---|---|---|---|
| 0.05 | 3.5% | -0.0136 | 0.0158 | 0.092 |
| 0.10 | 7.1% | -0.0369 | 0.0398 | 0.637 |
| 0.15 | 10.6% | -0.0542 | 0.0597 | 0.888 |
| 0.20 | 14.1% | -0.0640 | 0.0718 | 0.899 |
| 0.25 | 17.7% | -0.0713 | 0.0809 | 0.896 |
| 0.30 | 21.2% | -0.0774 | 0.0888 | 0.852 |

TABLE 3.5: Empirical bias, RMSE of the $K$-function estimator and type II error rate for the CSR test under locational errors generated by random geomasking.

we are able to obtain a closed form expression of the variogram and so we applied the variogram-based correction. Instead, if the observations are binomial distributed a closed form of the variogram does not exist. Hence, our suggestion was to apply an empirical logit transformation to the observed data that will now be approximately Gaussian and both the variogram-based and the model-based solutions can be applied to the transformed data. We then showed with a simulation study that either for Poisson data or binomial data our corrections lead to consistent estimation of the model parameters.

We then moved from real-valued continuous processes (geostatistical data) to point processes. In particular, we showed the effects of geomasking on a clustered point pattern. We found that, as the positional error variance increases, the clusters are obfuscated and their spatial structure is destroyed. Using the standard summary statistics for these type of processes, such as the Ripley's $K$ function, this leads to misleading inferences, i.e. the null hypothesis of complete spatial randomness won't be (wrongly) rejected, inflating the Type II error.

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.25 | 0 | - |
| variogNaive | 0.940 (0.1019) | 0.273 (0.0373) | 0.142 (0.1417) | 0.2 |
| variogAdj | 1.003 (0.0876) | 0.260 (0.0325) | 0.039 (0.0394) | 0.2 |
| geoNaive | 0.851 (0.1488) | 0.285 (0.0346) | 0.192 (0.1917) | 0.2 |
| CL | 1.056 (0.0577) | 0.238 (0.0181) | 0.002 (0.0019) | 0.2 |
| ACL2 | 1.056 (0.0577) | 0.238 (0.0181) | 0.002 (0.0019) | 0.2 |
| ACL1 | 1.056 (0.0583) | 0.245 (0.0179) | 0.002 (0.0017) | 0.2 |
| variogNaive | 0.744 (0.2564) | 0.324 (0.0739) | 0.338 (0.3379) | 0.4 |
| variogAdj | 0.990 (0.0887) | 0.263 (0.0375) | 0.058 (0.0609) | 0.4 |
| geoNaive | 0.684 (0.3157) | 0.341 (0.0912) | 0.377 (0.3772) | 0.4 |
| CL | 1.051 (0.0529) | 0.240 (0.0189) | 0.002 (0.0019) | 0.4 |
| ACL2 | 1.051 (0.0529) | 0.241 (0.0189) | 0.002 (0.0019) | 0.4 |
| ACL1 | 1.051 (0.0530) | 0.246 (0.0158) | 0.002 (0.0017) | 0.4 |
| variogNaive | 0.605 (0.3983) | 0.427 (0.1773) | 0.507 (0.5069) | 0.6 |
| variogAdj | 1.003 (0.1025) | 0.271 (0.0422) | 0.073 (0.0730) | 0.6 |
| geoNaive | 0.534 (0.4663) | 0.381 (0.1312) | 0.509 (0.5092) | 0.6 |
| CL | 1.054 (0.0564) | 0.240 (0.0206) | 0.002 (0.0020) | 0.6 |
| ACL2 | 1.054 (0.0564) | 0.240 (0.0206) | 0.002 (0.0020) | 0.6 |
| ACL1 | 1.056 (0.0570) | 0.244 (0.0188) | 0.002 (0.0017) | 0.6 |
| variogNaive | 0.502 (0.5022) | 0.501 (0.2513) | 0.621 (0.6206) | 0.8 |
| variogAdj | 1.003 (0.0838) | 0.271 (0.0476) | 0.003 (0.0032) | 0.8 |
| geoNaive | 0.436 (0.5641) | 0.434 (0.1836) | 0.624 (0.6235) | 0.8 |
| CL | 1.049 (0.0532) | 0.241 (0.0179) | 0.002 (0.0018) | 0.8 |
| ACL2 | 1.049 (0.0532) | 0.241 (0.0179) | 0.002 (0.0019) | 0.8 |
| ACL1 | 1.049 (0.0531) | 0.255 (0.0216) | 0.002 (0.0017) | 0.8 |
| variogNaive | 0.442 (0.6250) | 0.655 (0.4049) | 0.722 (0.7221) | 1.0 |
| variogAdj | 1.008 (0.1038) | 0.264 (0.0608) | 0.062 (0.0632) | 1.0 |
| geoNaive | 0.345 (0.6553) | 0.477 (0.2269) | 0.702 (0.7019) | 1.0 |
| CL | 1.057 (0.0588) | 0.239 (0.0229) | 0.002 (0.0019) | 1.0 |
| ACL2 | 1.057 (0.0588) | 0.239 (0.0229) | 0.002 (0.0019) | 1.0 |
| ACL1 | 1.057 (0.0586) | 0.252 (0.0269) | 0.002 (0.0018) | 1.0 |

TABLE 3.6: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 0.5$. Data were generated from model (3.3).

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.16 | 0 | - |
| variogNaive | 0.965 (0.1009) | 0.166 (0.0229) | 0.090 (0.1170) | 0.2 |
| variogAdj | 0.987 (0.0932) | 0.163 (0.0219) | 0.068 (0.0986) | 0.2 |
| geoNaive | 0.952 (0.0801) | 0.166 (0.0146) | 0.098 (0.1029) | 0.2 |
| CL | 1.047 (0.0773) | 0.130 (0.0638) | 0.008 (0.0171) | 0.2 |
| ACL2 | 1.051 (0.0837) | 0.151 (0.0158) | 0.004 (0.0155) | 0.2 |
| ACL1 | 1.050 (0.0808) | 0.152 (0.0151) | 0.005 (0.0101) | 0.2 |
| variogNaive | 0.890 (0.1461) | 0.175 (0.0303) | 0.165 (0.1845) | 0.4 |
| variogAdj | 0.975 (0.1014) | 0.165 (0.0258) | 0.079 (0.1159) | 0.4 |
| geoNaive | 0.861 (0.1539) | 0.179 (0.0238) | 0.187 (0.1913) | 0.4 |
| CL | 1.045 (0.0763) | 0.132 (0.0614) | 0.008 (0.0160) | 0.4 |
| ACL2 | 1.050 (0.0837) | 0.152 (0.0159) | 0.003 (0.0142) | 0.4 |
| ACL1 | 1.050 (0.0823) | 0.154 (0.0145) | 0.004 (0.0093) | 0.4 |
| variogNaive | 0.796 (0.2235) | 0.186 (0.0370) | 0.260 (0.2696) | 0.6 |
| variogAdj | 0.989 (0.1095) | 0.162 (0.0251) | 0.064 (0.1080) | 0.6 |
| geoNaive | 0.765 (0.2464) | 0.194 (0.0377) | 0.290 (0.2936) | 0.6 |
| CL | 1.050 (0.0788) | 0.129 (0.0674) | 0.006 (0.0130) | 0.6 |
| ACL2 | 1.055 (0.0849) | 0.153 (0.0148) | 0.002 (0.0020) | 0.6 |
| ACL1 | 1.053 (0.0832) | 0.156 (0.0141) | 0.003 (0.0091) | 0.6 |
| variogNaive | 0.683 (0.3328) | 0.208 (0.0617) | 0.375 (0.3837) | 0.8 |
| variogAdj | 0.955 (0.1384) | 0.169 (0.0366) | 0.097 (0.1549) | 0.8 |
| geoNaive | 0.655 (0.3518) | 0.213 (0.0565) | 0.394 (0.3977) | 0.8 |
| CL | 1.046 (0.0770) | 0.135 (0.0597) | 0.006 (0.0146) | 0.8 |
| ACL2 | 1.049 (0.0807) | 0.154 (0.0162) | 0.003 (0.0083) | 0.8 |
| ACL1 | 1.050 (0.0801) | 0.156 (0.0158) | 0.002 (0.0029) | 0.8 |
| variogNaive | 0.596 (0.4126) | 0.233 (0.0859) | 0.467 (0.4731) | 1.0 |
| variogAdj | 0.939 (0.1496) | 0.175 (0.0409) | 0.116 (0.1770) | 1.0 |
| geoNaive | 0.574 (0.4311) | 0.231 (0.0767) | 0.476 (0.4788) | 1.0 |
| CL | 1.047 (0.0795) | 0.137 (0.0597) | 0.006 (0.0136) | 1.0 |
| ACL2 | 1.051 (0.0827) | 0.155 (0.0168) | 0.002 (0.0028) | 1.0 |
| ACL1 | 1.050 (0.0823) | 0.159 (0.0170) | 0.003 (0.0049) | 1.0 |

TABLE 3.7: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 1.5$. Data were generated from model (3.3).

| **Method** | $\sigma^2$ | $\phi$ | $\tau^2$ | $r$ |
|---|---|---|---|---|
| **True Parameters** | 1 | 0.13 | 0 | - |
| variogNaive | 0.991 (0.0840) | 0.130 (0.0144) | 0.065 (0.0810) | 0.2 |
| variogAdj | 1.005 (0.0843) | 0.129 (0.0142) | 0.050 (0.0682) | 0.2 |
| geoNaive | 0.982 (0.0650) | 0.132 (0.0079) | 0.074 (0.0757) | 0.2 |
| CL | 1.013 (0.0698) | 0.128 (0.0106) | 0.042 (0.0484) | 0.2 |
| ACL2 | 1.013 (0.0695) | 0.128 (0.0106) | 0.043 (0.0487) | 0.2 |
| ACL1 | 1.042 (0.0789) | 0.121 (0.0125) | 0.014 (0.0228) | 0.2 |
| variogNaive | 0.943 (0.1027) | 0.134 (0.0155) | 0.111 (0.1221) | 0.4 |
| variogAdj | 1.001 (0.0849) | 0.129 (0.0144) | 0.053 (0.0727) | 0.4 |
| geoNaive | 0.922 (0.1018) | 0.139 (0.0124) | 0.132 (0.1338) | 0.4 |
| CL | 1.023 (0.0735) | 0.126 (0.0116) | 0.029 (0.0422) | 0.4 |
| ACL2 | 1.023 (0.0732) | 0.126 (0.0116) | 0.030 (0.0423) | 0.4 |
| ACL1 | 1.049 (0.0844) | 0.123 (0.0109) | 0.004 (0.0087) | 0.4 |
| variogNaive | 0.868 (0.1600) | 0.143 (0.0208) | 0.190 (0.1991) | 0.6 |
| variogAdj | 0.991 (0.0959) | 0.132 (0.0162) | 0.065 (0.0936) | 0.6 |
| geoNaive | 0.841 (0.1748) | 0.147 (0.0206) | 0.213 (0.2165) | 0.6 |
| CL | 1.015 (0.0806) | 0.128 (0.0121) | 0.040 (0.0591) | 0.6 |
| ACL2 | 1.016 (0.0803) | 0.127 (0.0121) | 0.039 (0.0573) | 0.6 |
| ACL1 | 1.049 (0.0857) | 0.124 (0.0109) | 0.006 (0.0130) | 0.6 |
| variogNaive | 0.788 (0.2306) | 0.150 (0.0276) | 0.271 (0.2773) | 0.8 |
| variogAdj | 0.990 (0.1064) | 0.132 (0.0179) | 0.066 (0.1030) | 0.8 |
| geoNaive | 0.769 (0.2411) | 0.154 (0.0272) | 0.284 (0.2868) | 0.8 |
| CL | 1.017 (0.0798) | 0.127 (0.0132) | 0.036 (0.0603) | 0.8 |
| ACL2 | 1.017 (0.0797) | 0.127 (0.0131) | 0.036 (0.0603) | 0.8 |
| ACL1 | 1.050 (0.0868) | 0.124 (0.0117) | 0.003 (0.0064) | 0.8 |
| variogNaive | 0.699 (0.3138) | 0.162 (0.0386) | 0.360 (0.3670) | 1.0 |
| variogAdj | 0.970 (0.1295) | 0.134 (0.0203) | 0.085 (0.1349) | 1.0 |
| geoNaive | 0.681 (0.3257) | 0.166 (0.0383) | 0.374 (0.3769) | 1.0 |
| CL | 1.022 (0.0885) | 0.127 (0.0128) | 0.031 (0.0560) | 1.0 |
| ACL2 | 1.021 (0.0899) | 0.127 (0.0130) | 0.033 (0.0588) | 1.0 |
| ACL1 | 1.051 (0.0844) | 0.125 (0.0110) | 0.003 (0.0066) | 1.0 |

TABLE 3.8: Average of Monte Carlo simulations and RMSE in parentheses for the naive methods (variogNaive and geoNaive) and their respective corrections (variogAdj and CL) for increasing levels of $r$. ACL2 and ACL1 reports results from (2.14) where $t$ has been chosen such that $\rho(t;\phi) = 5 \times 10^{-}6$ and $\rho(t;\phi) = 5 \times 10^{-}2$ respectively. The true correlation function is Matrn with $\kappa = 0.5$. Data were generated from model (3.3).

# Chapter 4

# Geostatistics for aggregated data

## 4.1 Introduction

The analysis of spatial data collected at different spatial scales is a challenging task in the field of spatial statistics. Nowadays it is often the case that different spatial data layers are collected at different scales. For example, we may have one layer at point level, another at the regional level or vice versa. These types of spatial data are often called *misaligned* and the inference problem related with them takes the name of *change of support problem* The change of support problem (COSP) is concerned with inference about the values of a variable at points or blocks different from those at which it has been observed. Gotway and Young (2002) provides a really nice review about the problem. Table 4.1, modified from Gotway and Young (2002) reports the most common examples of COSPs.

| **We observe or analyse** | **But the nature of the process is** | **Examples** |
|---|---|---|
| Point | Point | Point kriging or model based geostatistcs |
| Area | Point | Ecological inference; quadrant counts |
| Point | Line | Contouring |
| Point | Area | Use of areal centroids; spatial smootghin; |
| block kriging | | |
| Area | Area | The modifiable areal unit problem (MAUP); |
| | | areal interpolation; incompatible/misaligned zones |
| Point | Surface | Trend surface analysis; environmental monitoring; |
| | | exposure assessment |
| Area | Surface | Remote sensing; multiresolution images; image analysis |

Table 4.1: Examples of COSPs

Changing the support implies that a new random variable is created whose distribution may be developed from the original one but, in any event, has different statistical and spatial properties.A naive approach when the observed value is aggregated over a certain region $A$ is to attach it to the centroid of the region $x_a$ and then fit a standard geostatistical model to $Y(x_a)$. This approach uses a single centroid value to represent the outcome level in the entire region, and fails to properly capture variability and spatial association. In the following sections we introduce a model based treatment to this problem that allows to obtain better predictions.

## 4.2 Methodological framework

In these section we provide a detailed theoretical framework for the case in which we observe the output at the areal level but the nature of the process is continuous and we aim to obtain point predictions. A as real example we might have a very low-resolution global climate model for weather prediction, and seek to predict more locally (i.e., at higher resolution).

Let $Y(x_i)$ for $i = 1, \ldots, n$ denotes the spatial process of some continuous measurement. We recall the stationary Gaussian model used in the previous Chapters

$$Y(x) = \mu(x; \beta) + S(x) + Z \tag{4.1}$$

where $\mu(x; \beta)$ is the mean function including some covariates, $S(x)$ is a Gaussian process with zero mean and variance covariance matrix $\sigma^2 P(x; \phi) = \sigma^2 \rho(x - x'; \phi)$ and $Z$ is a multivariate normal, independent from $S(x)$, with zero mean and variance covariance matrix $\tau^2 I$. Here $\phi$ denotes the vector of parameters that define the correlation function $\rho(\cdot)$. We can summarize model (4.1) saying that

$$Y(x) \sim N\left(\mu(x; \beta), \sigma^2 P(x; \phi) + \tau^2 I\right) \tag{4.2}$$

Instead of observing data at point locations, we observe block data and we assume they arise as block averages. That is, for a block $A_m \subset D$ for $m = 1, \ldots, M$ where $D \subset \mathbb{R}^2$ is the observed region

$$Y(A_m) = |A_m|^{-1} \int_{A_m} Y(x) dx. \tag{4.3}$$

The above integral is an average of random variables, hence, a random or stochastic integral. Thus, the assumption of an underlying spatial process is only appropriate for block data that can be sensibly viewed as an averaging over point data; examples of

this would include rainfall, pollutant level, temperature, and elevation. It would be inappropriate for, say, population, since there is no population at a particular point. Our goal is to make inference on the true process at the finest spatial resolution using block averages data and covariates collected at point level (blocks to points prediction). This translates to find the distribution of $Y(x) \mid Y(A)$. Using (4.3) we can state that

$$Y(A) \sim N\left(\mu_A(\beta), \sigma^2 P_A(\phi) + \tau^2 I_A\right) \tag{4.4}$$

with $\mu_A(\beta) = |A|^{-1} \int_A \mu(x;\beta)\, dx$ and $P_A(\phi) = |A|^{-2} \int_A \int_A \rho(x - x';\phi)\, dx dx'$. Since $Y(x)$ and $Y(A)$ are jointly normal we obtain a $n+M$ - dimensional multivariate normal

$$\begin{pmatrix} Y(x) \\ Y(A) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(x;\beta) \\ \mu_A(\beta) \end{pmatrix}, \begin{pmatrix} \sigma^2 P(x;\phi) + \tau^2 I & P_{x,A}(\phi) \\ P_{x,A}^T(\phi) & \sigma^2 P_A(\phi) + \tau^2 I_A \end{pmatrix}\right) \tag{4.5}$$

where

$$\begin{aligned} \left(\mu_A(\beta)\right)_m &= E\left[Y(A_m)\right] = |A_m|^{-1} \int_{A_m} \mu(x;\beta)\, dx, \\ \left(P_A(\theta)\right)_{mm'} &= |A_m|^{-1} |A_{m'}|^{-1} \int_{A_m} \int_{A_{m'}} \rho(x - x';\phi)\, dx dx', \\ \left(P_{x,A}(\theta)\right)_{im} &= |A_m|^{-1} \int_{A_m} \rho(x_i - x';\phi)\, dx'. \end{aligned}$$

Noting that the above equations are nothing but an expectation with respect to a uniform distribution, we can use MC integration to estimate them. For each block $A_m$ we can draw a set of locations $x_{m,l}$ for $l = 1, \ldots, L_m$, distributed independently and uniformly over $A_m$. Hence we can replace the preceding formulas with

$$\begin{aligned} \left(\hat{\mu}_A(\beta)\right)_m &= L_m^{-1} \sum_l \mu(x_{m,l};\beta), \\ \left(\hat{P}_A(\phi)\right)_{mm'} &= L_m^{-1} L_{m'}^{-1} \sum_l \sum_{l'} \rho(x_{ml} - x_{m'l'};\phi), \\ \left(\hat{P}_{x,A}(\phi)\right)_{im} &= L_m^{-1} \sum_l \rho(x_i - x_{ml};\phi). \end{aligned}$$

From (4.5) we can obtain the distribution of $Y(x) \mid Y(A)$ that is $N\left(\mu_{x|A}, \Sigma_{x|A}\right)$ where

$$\begin{aligned} \mu_{x|A} &= \mu(x;\beta) + P_{x,A}(\phi)\left(\sigma^2 P_A(\phi) + \tau^2 I_A\right)^{-1}\left(Y(A) - \mu_A(\beta)\right) \\ \Sigma_{x|A} &= \sigma^2 P(x;\phi) - P_{x,A}(\phi)\left(\sigma^2 P_A(\phi) + \tau^2 I_A\right)^{-1} P_{x,A}^T(\phi) \end{aligned}$$

As far as we are able to fit the model (4.3) and then sample from $[Y(x) \mid Y(A)]$ we should end up with consistent point level predictions.

What we have shown here is for the case of area or block to point prediction. Since everything is based on conditional expectation of multivariate Normal random variable, it is straightforward to extend the theory to the point to area, and area to area case.

### 4.2.1 Inference

We first specify the mean function $\mu(x;\beta) = D\beta$ with $D$ a $n \times p$ matrix with geo-referenced covariates as entries $D_{i,j} = d_j(x_i)$ and $\beta$ a $p \times 1$ vector of parameters. Since we consider the covariates as a deterministic component, the mean vector of $Y_A$ can be redefined as $\mu_A(\beta) = \bar{D}\beta$, where $\bar{D}$ is a $M \times p$ matrix whose entries are $\bar{D}_{mj} = \frac{1}{\#x_i \in A_m} \sum_{x_i \in A_m} d_j(x_i)$.

Before proceeding to calculate the MLE estimates, it is convenient to express the nugget variance parameter, $\tau^2$ in relative terms $r = \tau^2/\sigma^2$, thus we have

$$Cov(Y_A) = \sigma^2 \left( \hat{P}_A(\phi) + rI_A \right)$$

where $P(\phi)$ is the correlation matrix. The log-likelihood for model (4.3) is

$$l(\beta,\theta) = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\left( \left| \sigma^2 \left( \hat{P}_A(\phi) + rI_A \right) \right| \right) - \frac{1}{2\sigma^2} \left( Y_A - \bar{D}\beta \right)^T \left( \hat{P}_A(\phi) + rI_A \right)^{-1} \left( Y_A - \bar{D}\beta \right).$$

Maximum likelihood estimates for $\beta$ and $\sigma^2$ are

$$\hat{\beta} = \left( \bar{D}^T \left( \hat{P}_A(\phi) + rI_A \right)^{-1} \bar{D} \right)^{-1} \bar{D}^T \left( \hat{P}_A(\phi) + rI_A \right)^{-1} Y_A$$

$$\hat{\sigma}^2 = \frac{\left( Y_A - \bar{D}\beta \right)^T \left( \hat{P}_A(\phi) + rI_A \right)^{-1} \left( Y_A - \bar{D}\beta \right)}{M}.$$

Substituting $\hat{\beta}$ and $\hat{\sigma}^2$ into the log-likelihood , we get the profile likelihood

$$l_p(r,\phi) = -\frac{1}{2}\left\{ \log\left( \left| \hat{\sigma}^2 \left( \hat{P}_A(\phi) + rI_A \right) \right| \right) + M\left( \log\left( 2\pi\hat{\sigma}^2 \right) + 1 \right) \right\}.$$

An optimization algorithm can then be used to estimate the noise $r$ and the scale parameter $\phi$.

## 4.3 Simulation study

We use a simulation study to asses the goodness of the predictions obtained from our model compared to the naive approach. We proceed as follows:

- for $i : 1, \ldots, 1000$:
  - generate $n = 100$ points from model (4.1) over the unit square;
  - aggregate the observed values in a number $M$ of blocks using the empirical version of (4.3);
  - calculate the parameters using the proposed method and the naive geostatistical approach that consider each averaged value over region $A$ as the true value located at the centroid of $A$;
  - Calculate the mean squared prediction error using 4-fold cross validation. We randomly sample the 75% of the data points and use it as the training set for parameter estimation and then we validate the accuracy of prediction on the remaining 24% of points left out.

We define the following scenarios: (a) $\sigma^2 = 1$, $\tau^2 = 0.2$, $\kappa = 0.5$ and $\phi = 0.1$; (b) $\sigma^2 = 1$, $\tau^2 = 0.2$, $\kappa = 1.5$ and $\phi = 0.7$. In both scenarios, we let the number of blocks $M$ vary over the set $\{10, 30, 50\}$. Results are reported in Table 4.2. Our method outperforms the naive approach since we obtain a smaller MSPE in all the considered scenarios.

| **Method** | $\boldsymbol{\phi = 0.1}$ | | | $\boldsymbol{\phi = 0.7}$ | | |
|---|---|---|---|---|---|---|
| | $\boldsymbol{M = 50}$ | $\boldsymbol{M = 30}$ | $\boldsymbol{M = 10}$ | $\boldsymbol{M = 50}$ | $\boldsymbol{M = 30}$ | $\boldsymbol{M = 10}$ |
| Naive | 0.667 | 0.935 | 1.691 | 0.444 | 0.832 | 1.174 |
| Adjusted | 0.221 | 0.654 | 1.218 | 0.104 | 0.591 | 0.983 |

TABLE 4.2: MSPE for the Naive and the Adjusted model calculated from 1000 Monte Carlo simulations.

## 4.4 Conclusions

This chapter takes into consideration the change of support problem for spatial data. We propose a geostatistical model that is able to produce consistent block to point prediction. The approach introduced could also be used as a first step when dealing with spatially misaligned data, i.e. to bring all the spatial layers to the same level of resolution and then fit a regression model. We first provided likelihood equations for block to point prediction i.e. when the outcome is observed at a coarser level than the

covariates that are continuously available. We compared standard spatial models with our method and found that we can obtain predictions with a smaller mean squared prediction error.

# Appendix

## A.1 Mathematical Proofs

In this section we provide mathematical proofs needed to understand some of the equation reported in the thesis.

### A.1.1 Proof for the variogram in presence of positional error

The distribution of $V_{ij} \mid U_{ij}$ reported in (2.3) is obtained through the following calculations

$$\begin{aligned}
[V_{ij} \mid U_{ij}] &= \frac{\int \left[V_{ij}, U_{ij}^*, U_{ij}\right] dU_{ij}^*}{[U_{ij}]} \\
&= \frac{\int \left[V_{ij} \mid U_{ij}^*, U_{ij}\right] \left[U_{ij}^*, U_{ij}\right] dU_{ij}^*}{[U_{ij}]} \\
&= \frac{\int \left[V_{ij} \mid U_{ij}^*\right] \left[U_{ij}^* \mid U_{ij}\right] [U_{ij}] \, dU_{ij}^*}{[U_{ij}]} \\
&= \int \left[V_{ij} \mid U_{ij}^*\right] \left[U_{ij}^* \mid U_{ij}\right] dU_{ij}^*,
\end{aligned}$$

Now we need to obtain $\left[V_{ij} \mid U_{ij}^*\right]$ and $\left[U_{ij}^* \mid U_{ij}\right]$. Let's start from the distribution of $V_{ij} = (Y_i - Y_j)^2 \mid U_{ij}^*$. Since $S\left(x_i^*\right)$ and $Z_i$ are independent by assumption (see the general model reported in (2.1)), it follows that $Y_i \sim N(0, \sigma^2 + \tau^2)$. To obtain the distribution of $Y_i - Y_j$ we need first to calculate the covariance between these two variables

$$\begin{aligned}
Cov\left(Y_i, Y_j\right) &= Cov\left(S\left\{x_i^*\right\} + Z_i, S\left\{x_j^*\right\} + Z_j\right) \\
&= Cov\left(S\left\{x_i^*\right\}, S\left\{x_j^*\right\}\right) \\
&= \sigma^2 \rho\left(u_{ij}^*\right).
\end{aligned}$$

Hence, $Y_i - Y_j \sim N\left(0, 2\left[\tau^2 + \sigma^2\left\{1 - \rho\left(u_{ij}^*\right)\right\}\right]\right)$. Let's define $\alpha^2 = 2\left[\tau^2 + \sigma^2\left\{1 - \rho\left(u_{ij}^*\right)\right\}\right]$, it follows that $(Y_i - Y_j)^2/\alpha^2 \sim \chi^2_{(1)}$. Thus, we can conclude that $V_{ij} \mid U_{ij}^* \sim \alpha^2\chi^2_{(1)}$. Before we turn to the calculations of $\left[U_{ij}^* \mid U_{ij}\right]$, we need to introduce the *Rice* distribution. A random variable $U$ follows a $Rice(\nu, \sigma)$ if its density function is

$$f\left(u; \nu, \sigma\right) = \frac{u}{\sigma^2}\exp\left(-\frac{u^2 + \nu^2}{2\sigma^2}\right) I_0\left(\frac{u\nu}{\sigma^2}\right),$$

with $I_k\left(\cdot\right)$ is the modified Bessel function of the first kind with order $k$.

The mean of $U$ is

$$E[U] = \sigma\sqrt{\frac{\pi}{2}}L(\nu^2/2\sigma^2)$$

where

$$L(x) = e^{x/2}\left[(1 - x)I_0(x/2) - xI_1(x/2)\right];$$

the variance is

$$\mathrm{Var}[U] = 2\sigma^2 + \nu^2 - \frac{\pi\sigma^2}{2}L^2(-\nu^2/2\sigma^2).$$

To calculate $\left[U_{ij}^* \mid U_{ij}\right]$ we start from the distribution of $X_i^* \mid X_i = x_i$. From (2.2) it's easy to deduce that it is a bivariate normal with mean the observed point $x_i$ and covariance matrix $\delta^2 I_2$,

$$\begin{pmatrix} X_{i1}^* \mid X_{i1} \\ X_{i2}^* \mid X_{i2} \end{pmatrix} \sim BVN\left(\begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}, \begin{bmatrix} \delta^2 & 0 \\ 0 & \delta^2 \end{bmatrix}\right).$$

Let's first define $X_1^* \mid X_1 = X_{i1}^* \mid X_{i1} - X_{j1}^* \mid X_{j1}$ and $X_2^* \mid X_2 = X_{i2}^* \mid X_{i2} - X_{j2}^* \mid X_{j2}$. It follows that $X_1^* \mid X_1 \sim N\left(x_{i1} - x_{j1}, 2\delta^2\right)$ and $X_2^* \mid X_2 \sim N\left(x_{i2} - x_{j2}, 2\delta^2\right)$. We can now exploit the fact that $R \sim Rice(\nu, \sigma)$ has a Rice distribution if $R = \sqrt{X^2 + Y^2}$where $X \sim N(\nu\cos\left(\theta\right), \sigma^2)$ and $Y \sim N(\nu\sin\left(\theta\right), \sigma^2)$ are statistically independent normal random variables and $\theta$ is any real number. Starting from this known fact, if we take $X = X_1^* \mid X_1$ and $Y = X_2^* \mid X_2$ and convert them to the polar coordinates solving the system

$$\begin{cases} \nu\cos\left(\theta\right) = x_{i1} - x_{j1} \\ \nu\sin\left(\theta\right) = x_{i2} - x_{j2} \end{cases}$$

we obtain $\nu = u_{ij}$, $\theta = \arctan\left(\frac{x_{i2} - x_{j2}}{x_{i1} - x_{j1}}\right)$ and we can express $X_1^* \mid X_1 \sim N(\nu\cos\left(\theta\right), 2\delta^2)$ and $X_2^* \mid X_2 \sim N(\nu\sin\left(\theta\right), 2\delta^2)$. Hence,

$$U_{ij}^* \mid U_{ij} = \sqrt{\left(X_1^* \mid X_1\right)^2 + \left(X_2^* \mid X_2\right)^2} \sim Rice\left(u_{ij}, \sqrt{2}\delta\right).$$

If we now substitute the two distributions calculated above inside (2.3) we get

$$\begin{aligned}
[V_{ij} \mid U_{ij}] &= \int \left[V_{ij} \mid U_{ij}^*\right] \left[U_{ij}^* \mid U_{ij}\right] dU_{ij}^* \\
&= \int \alpha^2 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v_{ij}}{2}\right\} \frac{1}{\sqrt{v_{ij}}} \frac{u_{ij}^*}{2\delta^2} \exp\left\{-\frac{\left(u_{ij}^*\right)^2 + u_{ij}^2}{4\delta^2}\right\} I_0\left(\frac{u_{ij}^* u_{ij}}{2\delta^2}\right) dU_{ij}^* \\
&= \int \left[2\tau^2 + 2\sigma^2 - 2\sigma^2 \rho\left(u_{ij}^*\right)\right] \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v_{ij}}{2}\right\} \frac{1}{\sqrt{v_{ij}}} \frac{u_{ij}^*}{2\delta^2} \exp\left\{-\frac{\left(u_{ij}^*\right)^2 + u_{ij}^2}{4\delta^2}\right\} I_0\left(\frac{u_{ij}^* u}{2\delta^2}\right. \\
&= \int \left\{2\tau^2 AB + 2\sigma^2 AB - 2\sigma^2 \rho\left(u_{ij}^*\right) AB\right\} dU_{ij}^* \\
&= 2\tau^2 A \int B dU_{ij}^* + 2\sigma^2 A \int B dU_{ij}^* - 2\sigma^2 A \int \rho\left(u_{ij}^*\right) B dU_{ij}^* \\
&= 2\tau^2 A + 2\sigma^2 A - 2\sigma^2 A \int \rho\left(u_{ij}^*\right) B dU_{ij}^* \\
&= 2A\left\{\tau^2 + \sigma^2 - \sigma^2 \int \rho\left(u_{ij}^*\right) B dU_{ij}^*\right\} \\
&= 2A\left\{\tau^2 + \sigma^2\left[1 - E\left\{\rho\left(U_{ij}^*\right)\right\}\right]\right\}
\end{aligned}$$

with $A = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v_{ij}}{2}\right\} \frac{1}{\sqrt{v_{ij}}}$ the density of a $\chi^2_{(1)}$, $B = \frac{u_{ij}^*}{2\delta^2} \exp\left\{-\frac{\left(u_{ij}^*\right)^2 + u_{ij}^2}{4\delta^2}\right\} I_0\left(\frac{u_{ij}^* u_{ij}}{2\delta^2}\right)$ the density of a $Rice\left(u_{ij}, \sqrt{2}\delta\right)$ and $\int B dU_{ij}^* = 1$ (since I am integrating over the support of $B$ $(0, +\infty)$). Since the closed form of $E\left[\rho\left(U_{ij}^*\right)\right]$ depends on the specific correlation function used, it will be calculated by quadrature. If we take the expectation both sides we have

$$\begin{aligned}
\frac{1}{2} E\left[V_{ij} \mid U_{ij}\right] &= E\left[A\right]\left\{\tau^2 + \sigma^2\left[1 - E\left\{\rho\left(U_{ij}^*\right)\right\}\right]\right\} \\
&= \tau^2 + \sigma^2\left\{1 - E\left[\rho\left(U_{ij}^*\right)\right]\right\}.
\end{aligned}$$

### A.1.2 R code

The R code for the thesis is all provide in an R library called geomask. It can be downloaded at the following website: https://github.com/claudiofronterre/geomask. It contains all the functions needed to reproduce the results contained in this thesis.

# Bibliography


Arbia, G., Espa, G. and Giuliani, D. (2015) Dirty spatial econometrics. *Ann. Reg. Sci.* **56**(1), 177–189.

Arbia, G., Espa, G., Giuliani, D. and Dickson, M. M. (2017) Effects of missing data and locational errors on spatial concentration measures based on ripleys k-function. *Spatial Economic Analysis* **12**(2-3), 326–346.

Armstrong, M. P., Rushton, G. and Zimmerman, D. L. (1999) Geographically masking health data to preserve confidentiality. *Stat. Med.* **18**(5), 497–525.

Bevilacqua, M. and Gaetan, C. (2015) Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Stat. Comput.* **25**(5), 877–892.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012) Estimating space and Space-Time covariance functions for large data sets: A weighted composite likelihood approach. *J. Am. Stat. Assoc.* **107**(497), 268–280.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E. and Freudenheim, J. L. (2003) Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* **14**(4), 408–412.

Burgert, C. R., Colston, J., Roy, T. and Zachary, B. (2013) Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys .

Caragea, P. and Smith, R. L. (2006) Approximate likelihoods for spatial processes. *Preprint* .

Caragea, P. C. and Smith, R. L. (2007) Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivar. Anal.* **98**(7), 1417–1440.

Cassa, C. A., Grannis, S. J., Overhage, J. M. and Mandl, K. D. (2006) A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *J. Am. Med. Inform. Assoc.* **13**(2), 160–165.

Cayo, M. R. and Talbot, T. O. (2003) Positional error in automated geocoding of residential addresses. *Int. J. Health Geogr.* **2**(1), 10.

Cressie, N. (1993) *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, N. and Kornak, J. (2003) Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Stat. Sci.* pp. 436–456.

Curriero, F. C. and Lele, S. (1999) A composite likelihood approach to semivariogram estimation. *J. Agric. Biol. Environ. Stat.* **4**(1), 9–28.

Dearwent, S. M., Jacobs, R. R. and Halbert, J. B. (2001) Locational uncertainty in georeferencing public health datasets. *J. Expo. Anal. Environ. Epidemiol.* **11**(4), 329–334.

Devillers, R. and Jeansoulin, R. (2006) *Fundamentals of spatial data quality*. ISTE Publishing Company.

Diggle, P. and Ribeiro, P. J. (2007) *Model-based Geostatistics*. Springer.

Diggle, P. J. (1993) Point process modelling in environmental epidemiology. *Statistics for the Environment* pp. 89–110.

Diggle, P. J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. CRC Press.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics. *J. R. Stat. Soc. Ser. C Appl. Stat.* **47**(3), 299–350.

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M. and Molyneux, D. H. (2007) Spatial modelling and the prediction of loa loa risk: decision making under uncertainty. *Ann. Trop. Med. Parasitol.* **101**(6), 499–509.

Fanshawe, T. R. and Diggle, P. J. (2011) Spatial prediction in the presence of positional error. *Environmetrics* **22**(2), 109–122.

Gabrosek, J. and Cressie, N. (2002) The effect on attribute prediction of location uncertainty in spatial data. *Geogr. Anal.* **34**(3), 262–285.

Gething, P., Tatem, A., Bird, T. and Burgert-Brucker, C. R. (2015) Creating spatial interpolation surfaces with DHS data DHS spatial analysis reports no. 11. *Rockville, Maryland: ICF* .

Goldberg, D. W. and Cockburn, M. G. (2012) The effect of administrative boundaries and geocoding error on cancer rates in california. *Spat. Spatiotemporal Epidemiol.* **3**(1), 39–54.

Gotway, C. A. and Young, L. J. (2002) Combining incompatible spatial data. *J. Am. Stat. Assoc.* **97**(458), 632–648.

Grosh, E.*Munoz, M. and Juan (1996) A manual for planning and implementing the living standards measurement study survey. Technical Report LSM126, The World Bank.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L. and Miller, W. C. (2010) Mapping health data: improved privacy protection with donut method geomasking. *Am. J. Epidemiol.* **172**(9), 1062–1069.

Hjort, N. L., Omre, H., Frisén, M., Godtliebsen, F., Jon Helgeland, Møller, J., Eva B. Vedel Jensen, Rudemo, M. and Stryhn, H. (1994) Topics in spatial statistics [with discussion, comments and rejoinder]. *Scand. Stat. Theory Appl.* **21**(4), 289–357.

Jacquez, G. M. (2012) A research agenda: does geocoding positional error matter in health GIS studies? *Spat. Spatiotemporal Epidemiol.* **3**(1), 7–16.

Jacquez, G. M. and Waller, L. A. (2000) The effect of uncertain locations on disease cluster statistics. *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing* pp. 53–64.

Kravets, N. and Hadden, W. C. (2007) The accuracy of address coding and the effects of coding errors. *Health Place* **13**(1), 293–298.

Malizia, N. (2013) The effect of data inaccuracy on tests of Space-Time interaction. *Trans. GIS* **17**(3), 426–451.

Matérn, B. (1960) *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations.* Statens skogsforskningsinstitut.

Matern, B. (1986) Spatial variation, number 36 in lectures notes in statistics .

Mateu, J., Porcu, E., Christakos, G. and Bevilacqua, M. (2007) Fitting negative spatial covariances to geothermal field temperatures in nea kessani (greece). *Environmetrics* **18**(7), 759–773.

Mazumdar, S., Rushton, G., Smith, B. J., Zimmerman, D. L. and Donham, K. J. (2008) Geocoding accuracy and the recovery of relationships between environmental exposures and health. *Int. J. Health Geogr.* **7**, 13.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C., Moser, W. K., Lister, A. J., King, S. L., LaPoint, E. B., Coulston, J. W., Smith, W. B. and Reams, G. A. (2005) Estimating and circumventing the effects of perturbing and swapping inventory plot locations. *J. For.* **103**(6), 275–279.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Stat. Soc. Ser. A* **135**(3), 370–384.

Neyman, J. and Scott, E. L. (1958) Statistical approach to problems of cosmology. *J. R. Stat. Soc. Series B Stat. Methodol.* **20**(1), 1–43.

Ripley, B. D. (1976) The Second-Order analysis of stationary point processes. *J. Appl. Probab.* **13**(2), 255–266.

Ripley, B. D. (1977) Modelling spatial patterns. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**(2), 172–212.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M. and Zimmerman, D. L. (2006) Geocoding in cancer research: a review. *Am. J. Prev. Med.* **30**(2 Suppl), S16–24.

Stanton, M. C. and Diggle, P. J. (2013) Geostatistical analysis of binomial data: generalised linear or transformed gaussian modelling? *Environmetrics* **24**(3), 158–171.

Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Series B Stat. Methodol.* **66**(2), 275–296.

Thomas, M. (1949) A generalization of poisson's binomial limit for use in ecology. *Biometrika* **36**(Pt. 1-2), 18–25.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Stat. Sin.* **21**(1), 5–42.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Stat. Soc. Series B Stat. Methodol.* **50**(2), 297–312.

WHO (2013) The world health organization year 2013 progress report. Technical report, WHO.

Woodward, M. (2013) *Epidemiology: Study Design and Data Analysis, Third Edition*. CRC Press.

Zandbergen, P. A. (2007) Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health* **7**, 37.

Zandbergen, P. A. (2009) Geocoding quality and implications for spatial analysis. *Geography Compass* **3**(2), 647–680.

Zandbergen, P. A. (2014) Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for Individual-Level data. *Adv Med* **2014**, 567049.

Zimmerman, D. L. (2007) Statistical methods for incompletely and incorrectly geocoded cancer data. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, pp. 165–180. CRC Press.

Zimmerman, D. L., Li, J. and Fang, X. (2010) Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Stat. Med.* **29**(9), 1025–1036.

Zimmerman, D. L. and Sun, P. (2006) Estimating spatial intensity and variation in risk from locations subject to geocoding errors. *Iowa City: University of Iowa* .

Zinszer, K., Jauvin, C., Verma, A., Bedard, L., Allard, R., Schwartzman, K., de Montigny, L., Charland, K. and Buckeridge, D. L. (2010) Residential address errors in public health surveillance data: a description and analysis of the impact on geocoding. *Spat. Spatiotemporal Epidemiol.* **1**(2-3), 163–168.

# Claudio Fronterré

CURRICULUM VITAE

## Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: fronterre@stat.unipd.it

## Current Position

*Since November 2014; (expected completion: March 2018)*
**PhD Student in Statistical Sciences, University of Padova.**
*Thesis title: Spatial analysis of geomasked and aggregated data...*
Supervisor: Prof. Giuseppe Espa

## Research interests

- Spatial and spatio-temporal statistics
- Biostatistics
- Computational statistics

## Education

*2012 – 2014*
**Master degree in Finance**.
University of Trento, Department of Economics and Management Title of dissertation: "Spatio-temporal analysis of crime diffusion in the province of Trento"
Supervisor: Prof. Giuseppe Espa
Final mark: 110/110 cum laude

*2009 – 2012*
**Bachelor degree in Economics and Management**.
University of Trento, Department of Economics and Management
Title of dissertation: "Business dynamics in the province of Trento"
Supervisor: Prof. Giuseppe Espa
Final mark: 103/110.

## Visiting periods

*April 2016 – June 2017*
CHICAS, Medical School, University of Lancaster,
Lancaster, UK.
Supervisor: Prof. Peter J. Diggle

## Further education

*August 2015*
Geostat Summer School
University of Lancaster

*December 2013*
2nd Trento winter school in Spatial Statistics and Econometrics
University of Trento

## Work experience

*July 2017 – September 2017*
**Lancaster University**.
Research Associate in Spatial Statistics.

*September 2013 – September 2014*
**Department of Economics and Management**.
Tutor for the first year course Data analysis and Statistics.

*January 2014 – March 2014*
**Research group eCrime**.
Intern for the European project "eSecurity - ICT for knowledge-based and predictive urban security".

## Awards and Scholarship

*2014*
Master Merit Award : money prize awarded by evaluating the overall results achieved during the university career.

*2012*
Bachelor Merit Award: money prize awarded by evaluating the overall results achieved during the university career.

## Computer skills

- R, Matlab
- GRASS, QGIS
- Latex

## Language skills

Italian: native; English: fluent.

## Publications

**Articles in journals**
Fronterré, C., Giorgi, E., Diggle, P.J. (2017). Geostatistical inference in the presence of geomasking:

a composite-likelihood approach. Submitted to *Spatial Statistics*

## References

**Prof. Peter J. Diggle**
CHICAS, Medical School, Lancaster University
Lancaster, LA1 4YB, UK
Phone: 1524-593957
e-mail: p.diggle@lancaster.ac.uk

**Prof. Giuseppe Espa**
Department of Economics and Management, University of Trento
Vi Inama 5 - 38122, Trento
Phone: +390461282157
e-mail: giuseppe.espa@unitn.it