Geostatistical Methods for Modelling Non-stationary Patterns in Disease Risk

Bedilu A. Ejigu[1,*], Eshetu Wencheko[1], Paula Moraga[2], Emanuele Giorgi[3]

Abstract

One of the tenets of geostatistical modelling is that close things in space are more similar than distant things, a principle also known as "the first law of geography". However, this may be questionable when unmeasured covariates affect, not only the mean of the underlying process, but also its covariance structure. In this paper we go beyond the assumption of stationarity and propose a novel modelling approach which we justify in the context of disease mapping. More specifically, our goal is to incorporate spatially referenced risk factors into the covariance function in order to model non-stationary patterns in the health outcome under investigation. Through a simulation study, we show that ignoring such non-stationary effects can lead to invalid inferences, yielding prediction intervals whose coverage is well below the nominal confidence level. We then illustrate two applications of the developed methodology for modelling anaemia in Ethiopia and Loa loa risk in West Africa. Our results indicate that the non-stationary models give a better fit than standard geostatistical models yielding a lower value for the Akaike information criterion. In the last section, we conclude by discussing further extensions of the new methods.

Keywords: Disease mapping, Gaussian process, Model-based geostatistics, Stationarity.

∗ Corresponding author:
E-mail: bedilu.alamirie@aau.edu.et (B.A Ejigu)
[1] Department of Statistics, Addis Ababa University, Ethiopia.
[2] Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK
[3] CHICAS, Lancaster Medical School, Lancaster University, UK.

1. Introduction

Model-based geostatistics (MBG) (Diggle et al., 1998) is a branch of spatial statistics that provides tools for spatially continuous inference. More specifically, using data $(y_1, \ldots, y_n)$ collected over a spatially discrete set of locations $X = \{x_1, \ldots, x_n\}$ within a region of interest A, MBG allows to make predictive inference on a spatially continuous $S(x)$ based on a principled likelihood-based paradigm. By making use of the first law of geography, whereby "close things are more related than distant things", geostatistical models aim to borrow strength of information across space in order to infer values of $S(x)$ at any location x within A. MBG has thus been increasingly used in low-resource settings where, due to the absence of disease registries, household surveys provide the main source of information for monitoring the burden of infectious diseases. MBG applications in epidemiological studies conducted in developing countries include mapping of malaria (Arab et al. (2014); Huang et al. (2011); Mokuolu and Adegboye

(2014); Nkurunziza et al. (2010); Texier et al. (2013); Jackson et al. (2010)), *Loa loa* (Thomson et al. (2004)), stunting (Amoah et al., 2017), anaemia (Amerson et al. (2017); Cook et al. (2006); Ejigu et al. (2018)), lymphatic filariasis (Moraga et al., 2015), visceral leishmaniasis (Galgamuwa et al. (2018); Yazdanpanah and M. (2013)) and river-blindness (Zoure et al., 2014), to name a few.

Let us consider a real-valued outcome Y_i taken at location x_i and a vector of covariates $d(x_i)$. In epidemiolgy, examples of real-valued health outcomes Y_i include antibodies concentration used in serological studies, antroprometric measures (e.g. height and weight) used to monitor childhood growth and measures of red blood cells loss (e.g. heamoglobin concentration) to which many infections, including malaria and various neglected tropical diseases, contribute. The standard linear geostatistical model for Y_i takes the form

$$Y_i = d(x_i)^\top \beta + S(x_i) + Z_i. \tag{1}$$

In the equation above, $\{S(x) : x \in \mathbf{R}^2\}$ is a stationary and isotropic Gaussian process with zero mean, variance σ^2 and correlation function $\text{Corr}\{S(x), S(x')\} = \rho(x, x')$. An important consequence of the assumption of stationarity and isotropy is that the correlation function of $S(x)$ is purely a function of the Euclidean distance between x and x', hence we write $\rho(x, x') = \rho(\|x - x'\|)$. Finally, the Z_i are assumed to be i.i.d. Gaussian variables with mean zero and variance τ^2.

The spatial covariance function $\rho(\cdot, \cdot)$ is traditionally assumed to belong to a parametric class of stationary functions (Diggle and Ribeiro (2007); Cressie (1993); Banerjee et al. (2015)). However, many efforts have been made to go beyond this, often questionable, assumption of stationarity. The seminal paper by Sampson and Guttorp (1992) introduced an approach through space deformation. The underlying idea is to transform the geographic region A into a new region G, such that sationarity and isotropy hold on G. Other approaches used to obtain a non-stationary covariance function are based on kernel convolution methods Higdon et al. (1999); Paciorek and Schervish (2006). Based on this approach, a stochastic process $Y(\cdot)$ is constructed by convolving a white noise process $W(\cdot)$ with a smoothing kernel $k(\cdot)$ to give

$$Y(x) = \int_A k_x(u) W(u) du.$$

Since the covariance function depends on the choice of the kernel function, a nonstationary covariance function can be constructed by defining a non-stationary kernel. To overcome the arbitrary choice of the kernel function, the weighted stationary process approach introduced by Fuentes (2001) which is an alternative approach of Higdon et al. (1999) varies the stationary process but not the kernel.

A special case non-stationary spatial processes is given by anisotropic processes, which arise in the context of directional effects in the covariance structure. For example, wind direction plays an important role in the spread of air pollutants as illustrated by Vianna Neto et al. (2014) who propose the inclusion of wind direction in a non-stationary Matérn process. The study by Schmidt *et al* Schmidt and Guttorp (2011) demonstrate the use of covariate information to model the spatial correlation between observations and develop Bayesian methods of inference based on projection models. Ver Hoef *et al.* also propose spatial models whose covariance structures incorporate flow and stream distance through the use of spatial moving averages to analyse stream networks Ver Hoef et al. (2006).

In this paper, we propose a novel geostatistical approach that allows to overcome the limits inherent to the assumptions of stationarity and isotropy for handling spatial dependence in health

outcomes. More specifically, we consider the case of non-stationary patterns that arise when unobserved risk factors may affect both the mean and the covariance structures of the underlying spatial process. To pursue this objective, our concern will be to identify parsimonious non-stationary geostatistical models that are empirically identifiable and supported by the data.

The paper is organized as follows. In Section 2 we describe the proposed modelling approach. Section 3 presents a simulation study to quantify the effects on spatial prediction when ignoring non-stationary patterns as defined in Section 2. Section 4 illustrates two applications to anaemia and Loa loa mapping in Africa. Finally, Section 5 is a discussion on further methodological extensions.

2. Non-stationary geostatistical models: incorporating risk factors into the covariance function

The prevalence pattern of environmentally-mediated diseases exhibits strong spatial heterogeneity. Hence, in order to aid spatial prediction of disease risk at unobserved locations, spatially referenced risk factors are often introduced as covariates $d(x_i)$ in a geostatistical model, as indicated in (1). However, accurate spatial information on risk factors that directly affect the disease under investigation are often unavailable, making the use of *proxy variables* (i.e. variables that correlate with the unobservable risk factors but do not directly affect disease risk) unavoidable. For example, in the context of vector-borne diseases, elevation may be used as a proxy for the spatial distribution of the disease vector. However, the imperfect nature of this approach is apparent because proxy variables may be measured with error or are inadequate to fully capture the distribution of the disease vector, as this may also be affected by other spatially varying factors (e.g. temperature and relative humidity). As a result of this non-stationary effects in the spatial pattern of disease may still be present even after the inclusion of proxy variables into the model.

In this paper, our concern is to model non-stationary covariance structures with domain on a proxy variable space. In other words, two outcomes Y_i and Y_j taken at x_i and x_j, with associated scalar values e_i and e_j for a proxy variable, may be correlated because either $|e_i - e_j|$ or $\|x_i - x_j\|$ are close to zero. For example, a given disease metric taken at different locations may show similar values as a result of common environmental features (e.g. elevation, temperature, type of vegetation, etc.), regardless of how close the two observations are in space.

To formally express this concept, we replace the stationary Gaussian process $S(x_i)$ in (1) with another Gaussian process $S(x_i, e_i)$ which is a function over both space and a proxy domain given by e_i. The model for Y_i then takes the form

$$Y_i = d(x_i)^\top \beta + S(x_i, e_i) + Z_i, \quad i = 1, \cdots, n. \tag{2}$$

We point out that, in the above equation, e_i may also be used to model the mean of Y_i in addition its covariance function, by including this as one of the components of the vector $d(x_i)$. We then assume that $S(x_i, e_i)$ is a Gaussian process with mean zero and covariance function

$$\mathrm{Cov}\{S(x, e), S(x', e')\} = \sigma^2 \rho(x, x'; e, e').$$

In the remainder part of the paper we will focus our attention on separable covariance functions, i.e.

$$\rho(x, x'; e, e') = \rho_1(\|x - x'\|)\rho_2(|e - e'|). \tag{3}$$

more flexible models that allow for the interaction between x and e may also be considered but would require a significantly larger amount of data which is often unavailable in the context of disease mapping applications.

Our choice of either $\rho_1(\cdot)$ or $\rho_2(\cdot)$ in (3) is to use a Mateŕn function Mateŕn (1960), i.e.

$$\rho(u) = \frac{1}{\Gamma(\kappa)2^{\kappa-1}}(u/\phi)^{\kappa}K_{\kappa}(u/\phi), \quad \kappa > 0, u \geq 0, \tag{4}$$

where $K_{\kappa}(.)$ denotes the modified Bessel function of the third kind of order κ, and ϕ is the scale parameter which controls the rate at which the correlation gets close to zero with increasing separation distance u. However, Zhang (2004) has shown that σ^2, ϕ and κ cannot be consistently estimated under in-fill asymptotic which often results in κ being poorly estimated. In the context of disease mapping, where data are often sparsely sampled over space, this problem is exacerbated. Hence, in the remainder of the paper, we make the pragmatic choice of fixing k at $1/2$ which gives rise to an exponential correlation function, given by

$$\text{Cov}\{S(x,e), S(x',e')\} = \sigma^2 \exp\{-u_s/\phi_s\} \exp\{-u_e/\phi_e\}, \tag{5}$$

where $u_s = \|x - x'\|$ and $u_e = \|e - e'\|$.

We estimate the resulting model for Y_i by maximizing the profile likelihood for $\theta^{\top} = (\sigma^2, \phi_s, \phi_e, \nu^2)$, where $\nu^2 = \tau^2\sigma^2$. Let D be an n by p matrix of covariates, where p stands for number of explanatory variables, and n represents number of observations; and $V = \Sigma + \nu^2 I$, with the $(i, j) - th$ entry of Σ given by $\sigma^2 \exp\left\{-u_{s,ij}/\phi_s\right\} \exp\left\{-u_{e,ij}/\phi_e\right\}$. The resulting expression for the profile likelihood is

$$L_p(\theta) = \exp\left\{-\frac{1}{2}\left(n \log\{\hat{\sigma}^2(\theta)\} + \log|V(\theta)|\right)\right\}, \tag{6}$$

where

$$\hat{\sigma}^2(\theta) = \frac{1}{n}(y - D\hat{\beta}(\theta))^{\top}V^{-1}(\theta)(y - D\hat{\beta}(\theta)),$$

and

$$\hat{\beta}(\theta) = (D^{\top}V^{-1}D)^{-1}D^{\top}V^{-1}y.$$

Maximization of (6) is then carried out using a numerical optimization algorithm implemented in the `nliminb` function, available in the R software environment.

3. Simulation Study

In this section we carry out a simulation study to quantify the effects on spatial prediction when ignoring non-stationary effects as defined by the model in (2). To this end, we simulate 1,000 data-sets using the 615 sampled locations in Ethiopia, as shown in Figure 2, under the following geostatistical model

$$Y_i = \beta_0 + S(x_i, e_i) + Z_i, \tag{7}$$

where e_i corresponds the elevation in meters at location x_i. We also set $\beta_0 = 0$, $\sigma^2 = 1$, $\tau^2 = 1$, $\phi_s = 100$ and let ϕ_e vary over the set {100, 200, 500, 1000, 2000}. We recall that exponential correlation functions are used to model both the correlation based on the Euclidean distance and that based on the difference in elevation. For each of the $B = 1,000$ data-sets, we then fit the three following models.

- $\mathcal{M}_1$: the true model as specified in (7).
- $\mathcal{M}_2$: $Y_i = \beta_0 + S(x_i) + Z_i$.
- $\mathcal{M}_3$: $Y_i = \beta_0 + \beta_1 e_i + S(x_i) + Z_i$.

Note that $\mathcal{M}_2$ completely ignores the effect of elevation on Y_i, while $\mathcal{M}_3$ uses elevation to model the mean of Y_i but ignores its effects on the covariance structure of Y_i.

Our predictive target for each of the three models are the observations Y_i at each of the $n = 615$ sampled locations. To compare the predictive performance of the three models, we use the bias, root-mean-square-error (RMSE) and 95% coverage probability (CP) for the predictive target. More specifically, for a given model $\mathcal{M}_k$, $k = 1, 2, 3$, these are computed by averaging over the set of observed locations, as follows

$$\begin{aligned} \text{BIAS}(\mathcal{M}_k) &= \frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}\left(\hat{Y}_i^{(j)} - Y_i^{(j)}\right) \\ \text{RMSE}(\mathcal{M}_k) &= \sqrt{\frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}\left(\hat{Y}_i^{(j)} - Y_i^{(j)}\right)^2} \\ \text{CP}(\mathcal{M}_k) &= \frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B} I\left(Y_i^{(j)} \in PI_{(j)}^{95\%}\right) \end{aligned},$$

where: $Y_i^{(j)}$ and $\hat{Y}_i^{(j)}$ are the true and estimated values of the predictive target from the j-th simulation, respectively; $I\left(Y_i^{(j)} \in PI_{(j)}^{95\%}\right)$ is an indicator function that takes value 1 if $Y_i^{(j)}$ is inside the 95% prediction interval denoted by $PI_{(j)}^{95\%}$ and 0 otherwise.

The results are reported in Table 1. We note that the largest differences in terms of bias and RMSE between $\mathcal{M}_1$, the true model (7), and the other two models are observed for value of ϕ_e smaller than 1000 meters. This can be explained by the fact that for large values of ϕ_e the impact on the overall spatial structure is less important, since $\exp\{-|e - e'|/\phi_e\}$ is closer to one, whilst in the other scenarios with a smaller ϕ_e the second factor in (3) will have a stronger impact on the overall correlation between observations.

We notice that compared to $\mathcal{M}_1$, the other two models yield a slightly larger bias and RMSE, with $\mathcal{M}_2$ generally outperforming $\mathcal{M}_3$. However, we also observe that both $\mathcal{M}_2$ and $\mathcal{M}_3$ provide unreliable prediction intervals with an actual coverage well below the 95% nominal coverage in all five scenarios. Overall, the results indicate that models that ignore non-stationary effects of the kind investigated in this paper may still provide accurate predictions but fail to reliably quantify uncertainty around these.

Table 1: Bias, root-mean-square-error (RMSE) and 95% coverage probability (CP) based on 1,000 data-sets simulated under $\mathcal{M}_1$. See the main text in Section 3 for more details.

Model	ϕ_e	BIAS	RMSE	95% CP
$\mathcal{M}_1$	100	-0.005	0.620	0.940
$\mathcal{M}_2$	100	-0.007	0.709	0.858
$\mathcal{M}_3$	100	-0.007	0.712	0.856
$\mathcal{M}_1$	200	0.001	0.602	0.948
$\mathcal{M}_2$	200	0.002	0.667	0.876
$\mathcal{M}_3$	200	0.002	0.674	0.871
$\mathcal{M}_1$	500	-0.004	0.593	0.953
$\mathcal{M}_2$	500	-0.006	0.615	0.889
$\mathcal{M}_3$	500	-0.003	0.626	0.880
$\mathcal{M}_1$	1000	-0.006	0.591	0.953
$\mathcal{M}_2$	1000	-0.006	0.577	0.897
$\mathcal{M}_3$	1000	-0.007	0.590	0.887
$\mathcal{M}_1$	2000	-0.009	0.592	0.952
$\mathcal{M}_2$	2000	-0.010	0.552	0.902
$\mathcal{M}_3$	2000	-0.014	0.566	0.889

4. Applications

4.1. Mapping Anaemia in Ethiopia

The data analysed in this section were obtained from the Demographic and Health Survey (DHS) conducted in Ethiopia in 2016. DHS are nationally representative household surveys that are generally repeated every 5 years and provide information on a range of health and population indicators, including anthropometric information. The DHS methodology is usually based on a stratified two-stage cluster design. At the first stage, enumeration areas are drawn from census files. At the second stage, for each enumeration area selected, samples of households are drawn from an updated list of households to form groups of households known as sampling clusters. The GPS location of the center of each sampling cluster is taken as the cluster location. Each child is allocated to a spatially-referenced sampling cluster. In the data-set analysed in this section, a total of 645 EAs (202 in urban areas and 443 in rural areas) were randomly selected with a probability proportional to the EA size. For more details on the DHS survey design and anaemia testing, we refer to (CSA (2016)). In this analysis, our response variable is the concentration of haemoglobin (Hb), measured in grams per decilitre (g/dl), in blood samples taken from 7,485 children aged below five years.

Figure 1 shows a histogram of Hb and a scatter plot of Hb against altitude. Although the Hb distribution shows some slight skeweness, we do not find any improvement by transforming the outcome (e.g. by taking the log), hence we model Hb on its original scale.

The number of children at a cluster ranges from 1 to 39, with an average of 14 children per cluster. A value of Hb below 11 g/dl leads a child to be diagnosed with anaemia. The mean age of children was 2.2 years with 59.31% of the children found to be anaemic. Figure 2 below presents

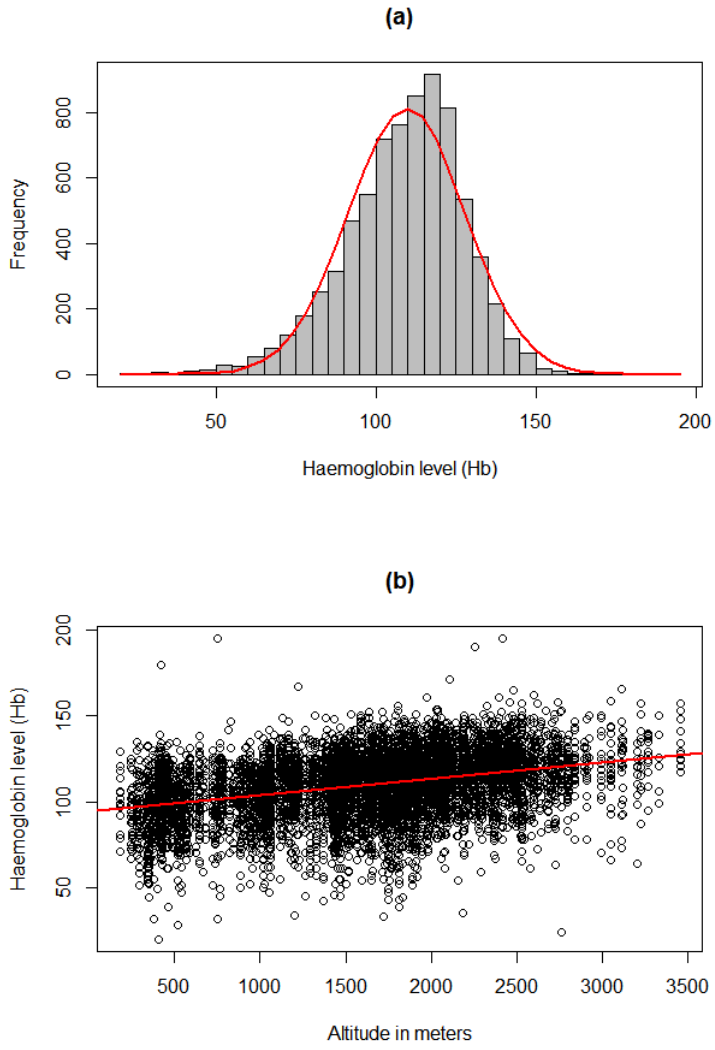

Figure 1: Histogram of haemoglobin concentration (upper panel) and scatter plot for haemoglobin against altitude (lower panel). The red curve in upper panel is the density function of a Gaussian distribution with mean and variance estimated from the raw data. The red line in the lower panel corresponds to a least squares fit.

the prevalence of anaemia at each of the sampled locations (upper panel), and the altitude raster for Ethiopia (lower panel).

Figure 3 shows the empirical semivariogram and correlogram plots for Hb based on the Euclidean distance and the altitude difference. The two variograms show evidence of correlation in the data in both domains considered.

Let Y_{ij} and d_{ij} be the Hb concentration and a vector of covariates for the j-th child at location x_i, respectively. Also let e_i be the altitude at location x_i. We then fit the three following models.

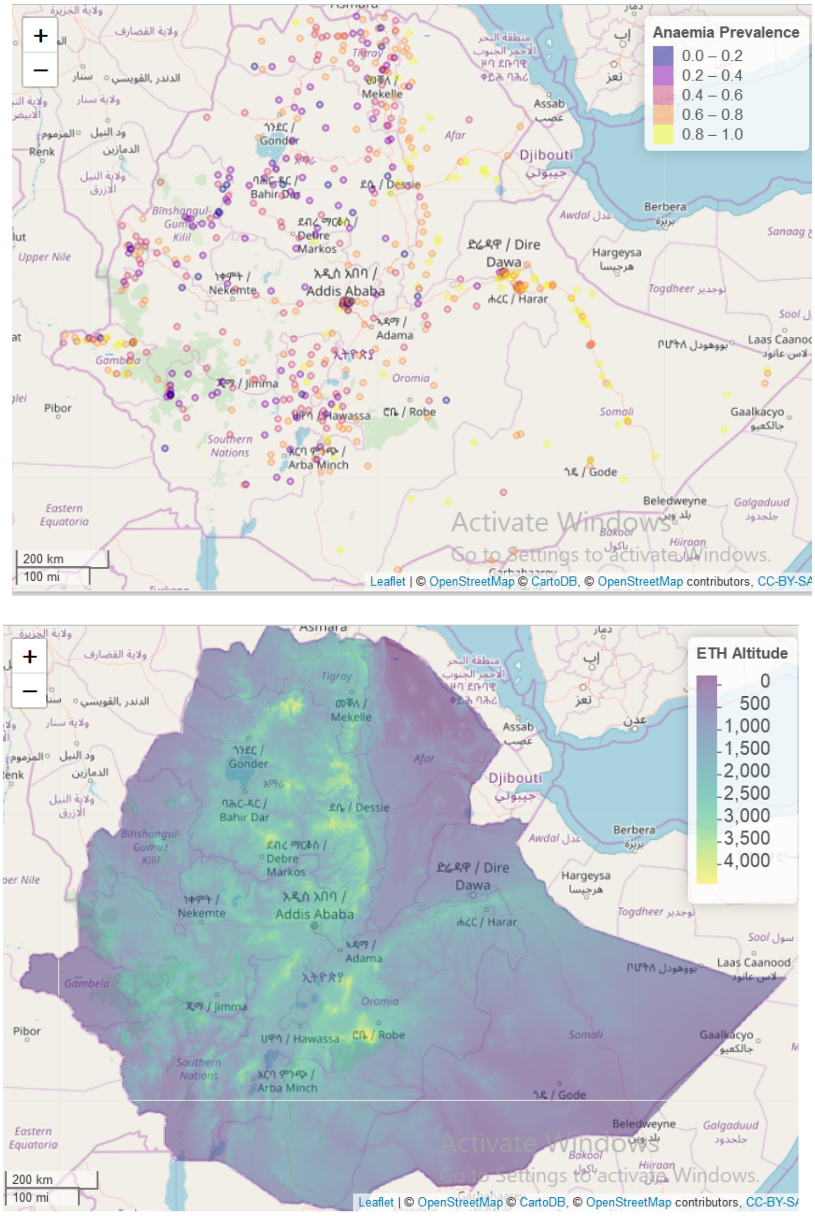

Figure 2: Anaemia prevalence among under-five children by survey clusters (upper panel), and altitude (lower panel)

- Model A, $Y_{ij} = \beta_0 + \beta^\top d_{ij} + \gamma e_i + S(x_i) + Z_{ij}$;
- Model B, $Y_{ij} = \beta_0 + \beta^\top d_{ij} + S(x_i, e_i) + Z_{ij}$;
- Model C, $Y_{ij} = \beta_0 + \beta^\top d_{ij} + \gamma e_i + S(x_i, e_i) + Z_{ij}$.

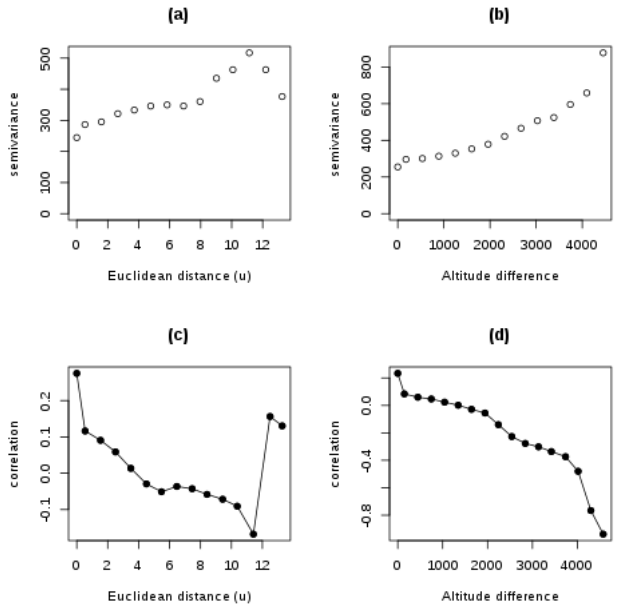

Figure 3: Empirical semivariogram (upper panel) based on Euclidean distance (left) and altitude difference (right panel), and correlogram plots (lower panel)

In the equations above d_{ij} includes both individual- and EA-level covariates; $S(x_i)$ and $S(x_i, e_i)$ are Gaussian processes as defined in (1) and (2), respectively. The Z_{ij} are i.i.d. zero-mean Gaussian variable with variance ω^2. More specifically, Model A is a standard geostistical model that takes account of the altitude effect on the mean of Hb; Model B incorporates the effect of altitude in the covariance structure only, while Model C also in the mean of the outcome.

Table 2 presents the results of the fitted models to the data. The point and interval estimates of the explanatory variables in common to all models are all comparable. We note that the 95% confidence intervals for covariance parameters are narrower in Model A than in Model B and C. However, Model C has the lowest AIC and we thus consider this to be the best model among the three considered.

Figure 4 shows the impact for three different values of altitude difference on the spatial correlation based on the Euclidean distance and difference between these clearly indicate that non-stationary effects play an important role in the stochastic variation of the data.

4.2. Loa loa prevalence mapping in West Africa

We now analyse data collected from 197 villages in Cameroon and southern Nigeria on *Loa loa*. *Loa loa* (also known as African eye worm) is an infectious disease caused by the filarial nematode (roundworm) *Loa loa*. Although *Loa loa* is not a life threatening disease, it has lately increasingly become of public health concern since individuals who are highly co-infected with Loa loa and lymphatic falariasis, are at risk of developing serious adverse events, such as encephalopathy, which can lead to permanent brain damage or even death.

Table 2: Maximum likelihood estimates and 95% confidence intervals for the parameters of Models A, B and C as specified in Section 4.1.

Term	**Model A**		**Model B**		**Model C**	
	Estimates	**95%CI**	**Estimates**	**95%CI**	**Estimates**	**95%CI**
β_0	92.777	(87.577, 97.977)	100.216	(91.190, 109.242)	91.365	(85.267, 97.463)
Age	0.304	(0.282, 0.326)	0.305	(0.283, 0.327)	0.305	(0.283, 0.327)
Altitude	0.006	(0.0058, 0.0062)			0.007	(0.003, 0.011)
Residence(ref:Rural)						
Urban	1.613	(-0.116, 3.342)	1.607	(-0.130, 3.344)	1.563	(-0.185, 3.311)
Gender						
Male	-0.64	(-1.324, 0.044)	-0.653	(-1.337, 0.031)	-0.656	(-1.338 , 0.026)
Wealth index (ref. Middle)						
Poor	-2.262	(-3.487, -1.037)	-2.201	(-3.416, -0.986)	-2.223	(-3.440, -1.006)
poorest	-3.723	(-4.954, -2.492)	-3.492	(-4.729, -2.255)	-3.529	(-4.768, -2.290)
rich	-1.336	(-2.669, -0.003)	-1.407	(-2.740, -0.074)	-1.417	(-2.750 , -0.084)
richest	2.448	(0.729, 4.167)	2.266	(0.547, 3.985)	2.267	(0.544, 3.990)
Education (ref. Higher)						
no education	-2.759	(-4.964, -0.554)	-2.554	(-4.763, -0.345)	-2.624	(-4.833, -0.415)
Primary	-1.825	(-4.003, 0.353)	-1.672	(-3.850, 0.506)	-1.731	(-3.909, 0.447)
Secondary	-0.57	(-2.924, 1.784)	-0.422	(-2.788, 1.944)	-0.443	(-2.809, 1.923)
σ^2	62.977	(62.485, 63.473)	104.225	(54.692, 198.620)	54.693	(35.189, 85.007)
ϕ_s	1.351	(0.885, 2.063)	3.606	(1.795, 7.245)	1.606	(0.896,2.880)
ϕ_e	-	-	3389.188	(1185.341, 9690.546)	2256.45	(770.830, 6605.302)
ω^2	221.22	(219.063,223.398)	220.731	(115.828,420.643)	220.49	(141.862,342.699)
AIC	48275.069		40817.836		**40809.321**	

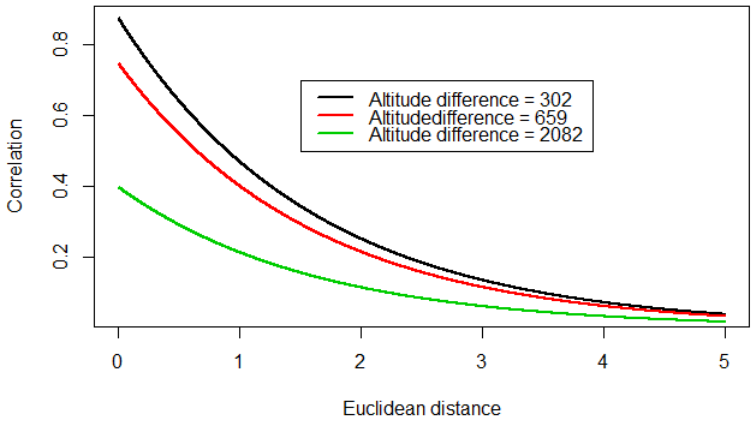

Figure 4: The black line corresponds to the function $exp(-u/\phi_s) * exp(-302/\phi_e)$, the red line corresponds to $exp(-u/\phi_s) \times exp(-659/\phi_e)$ and the green line corresponds to $exp(-u/\phi_s) \times exp(-2082/\phi_e)$, where $exp(\cdot)$ is an exponential correlation function, (302,659,2082) are first, second and third quartiles of altitude, and ϕ_s, ϕ_e are parameter estimates from Model C in Table 2.

The analysed dataset is freely available from the R package PrevMap Giorgi and Diggle (2017). Previous studies by Thomson et al. (2004) and Diggle et al. (2007) have reported a significant association with the maximum normalized difference vegetation index (MNDVI), a spatial indicator of live green vegetation. NDVI quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). The measured NDVI value ranges from -1 to +1, with negative values indicating the presence of water bodies and positive values corresponding to dense green leaves.

The scatter plot of Figure 5 (lower panel) shows a clear linear relationship between the empirical logit-transformed prevalence against MNDVI.

Figure 6 below presents the semivariogram (upper panel) and correlogram (lower panel) plot based on the Euclidean distance and MNVDI difference. Each of the four plots suggest the presence of unexplained correlation in the data which we model using the three following models.

- Model A, $Y_i = \beta_0 + S(x_i) + Z_i$;
- Model B, $Y_i = \beta_0 + \beta_1 e_i + S(x_i) + Z_i$;
- Model C, $Y_i = \beta_0 + S(x_i, e_i) + Z_i$;
- Model D, $Y_i = \beta_0 + \beta_1 e_i + SE(x_i, e_i) + Z_i$.

In the equations above, e_i denotes the MNDVI at location x_i.

The parameter estimate and the 95% confidence intervals are reported in Table 3. We note that after incorporating MNDVI into the covariance structure, this leads to an increase in the estimate of the scale parameter. Also, as in the application from the previous section, the model

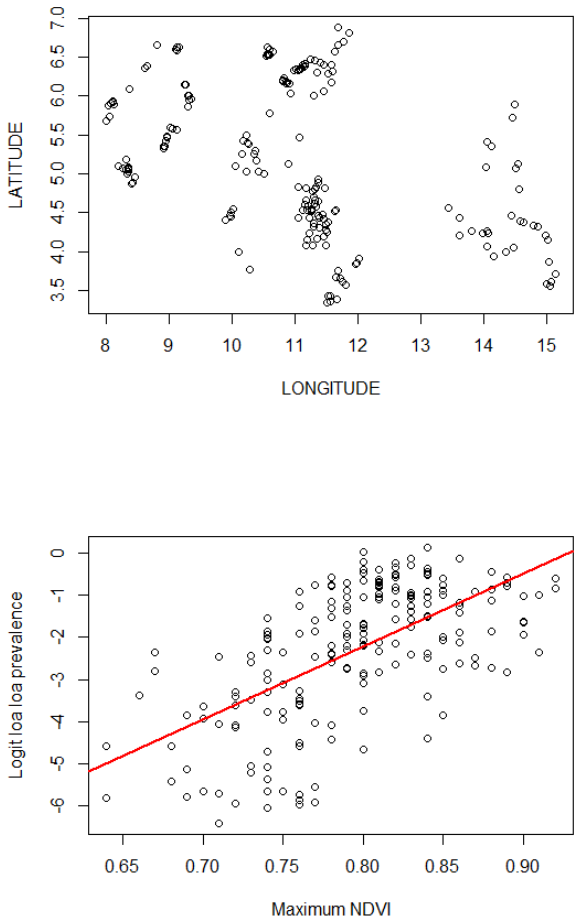

Figure 5: Plot of the locations of sampled villages across Cameroon and Nigieria (upper panel) and scatter plot of the empirical logit-transformed *Loa loa* prevalence against the maximum normalized difference vegetation index (MNDVI) (lower panel). The red line in the lower panel has been obtained via a least squares fit.

with the lowest AIC is the one that uses MNDVI to model both the mean and covariance of the data, i.e. Model D (Table 3). We observe that the introduction of MNDVI as a covariate in Model D leads to a reduction both in the variance σ^2 and the scale parameters (ϕ_s, ϕ_e) of the Gaussian process.

We use the formulation provided by Figure 7 in order to assess the impact of MNDVI on the fitted correlation structure. Figure 7 shows how the correlation on the Euclidean distance domain changes for different values of MNDVI difference. Although the non-stationary effects due to MNDVI are less strong than those observed for altitude in the previous application (Section 4.1), these are, however, non-negligible.

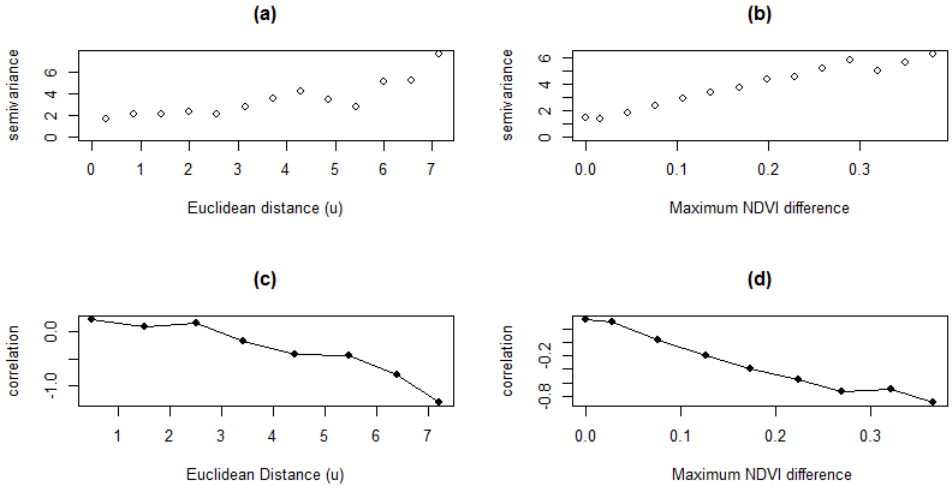

Figure 6: Empirical semivariogram (upper panel) based on separation distance (left) and max NDVI difference (right panel), and respective correlogram plots (lower panel)

Table 3: Maximum likelihood estimates and 95% confidence intervals of the Models A to D as specified in Section 4.2.

Parameter	Model A	Model B	Model C	Model D
β_0	-2.299	-9.580	-2.488	-10.189
	(-3.371, -1.227)	(-12.667, -6.493)	(-4.507, -0.469)	(-14.509, -5.869)
β_1	-	9.133	-	9.873
	-	(5.376, 12.890)	-	(4.546, 15.200)
σ^2	2.451	1.522	3.671	1.443
	(1.866, 3.218)	(1.036, 2.235)	(0.812, 16.604)	(0.827,2.517)
ϕ_s	0.844	0.616	2.112	0.633
	(0.321, 2.218)	(0.188, 2.020)	(0.403, 11.065)	(0.300, 1.336)
ϕ_e	-	-	0.931	0.555
	-	-	(0.113, 7.671)	(0.068, 4.528)
τ^2	0.369	0.368	0.347	0.331
	(0.037, 3.663)	(0.038, 3.554)	(0.076, 1.588)	(0.165,0.665)
AIC	188.681	167.127	-13.093	-19.897

5. Discussion

We have developed a geostatstical modelling approach to model non-stationary effects when covariates also affect the covariance structure of the data. Our simulation study has shown that ignoring such effects by using standard stationary geostatsitical models can lead to invalid predictive inferences on the outcome of interest, with a significantly smaller actual coverage than the nominal level. Our two applications in the context of disease mapping have shown that the proposed non-stationary models give better fits to the data.

We also notice that the best models in the two applications incorporated the environmental factors - i.e. altitude for the analysis on anaemia (Section 4.1) and the maximum normalized difference vegetation index (MNDV) (Section 4.2) for the Loa loa analysis - both in the mean

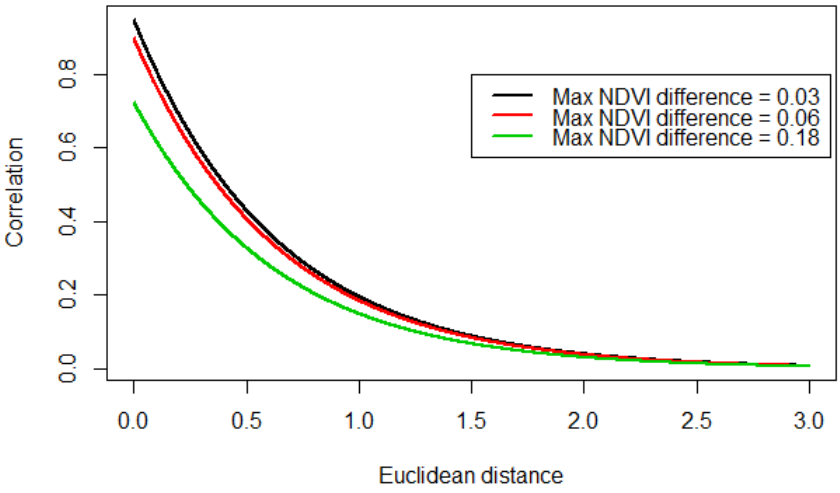

Figure 7: The black line corresponds to the function $exp(-u/\phi_s)*exp(-0.03/\phi_e)$, the red line corresponds to $exp(-u/\phi_s)*exp(-0.06/\phi_e)$ and green line corresponds to $exp(-u/\phi_s) * exp(-0.18/\phi_e)$, where $exp(\cdot)$ is an exponential correlation function, (0.03,0.06,0.18) are first, second and third quartiles of maximum NDVI, and ϕ_s, ϕ_e are parameter estimates from Model D.

and the covariance function of the model. More specifically, the introduction of environmental factors in the mean component of the outcome leads to a reduction of the variance and scale parameters of the non-stationary correlation function in both applications. This suggests that these two contributions from disease risk factors into the spatial variation of health outcomes are equally important and neither should be ignored. Our recommendation is thus to assess both the regression relationship of health outcomes with risk factors and how the latter affects the covariance structure of the former using the proposed approach in this paper.

Future extensions of the presented methodology should focus on the incorporation of multiple covariates $e_1, \ldots, e_m$ into the covariance functions to re-write (3) as

$$\rho(x, x', e_1, e_1', \ldots, e_m, e_m') = \rho(\|x - x'\|) \prod_{j=1}^{m} \rho(|e_j - e_j'|). \tag{8}$$

This approach could be especially useful to model unexplained correlation in the data that is not spatially structured or that occurs at a small spatial scale. For example, some of the e_j (8) might include individual traits (e.g. ethnicity, employment, education, etc.) and household characteristics (e.g. socio-economic status, material of the house, etc.). However, robust methods for model selections should also be developed for such complex models in order to avoid over-fitting of the data.

One of the limitations of our modelling approach is the assumption of separable correlation functions $\rho_1(\cdot)$ and $\rho_2(\cdot)$, as given by (3). Classes of non-separable correlation functions could then be used to relax this assumption; see, for example, Gneiting (2002) in the context of

space-time covariance functions. However, this would require a larger amount of data than that available in the two applications presented in this paper. To show this, we have also fitted the following non-separable model belonging to the Gneiting (2002) family of correlations to the Loa loa data

$$\text{Cov}\{S(x,e),S(x',e')\} = \sigma^2 \frac{1}{(1+u_e/\phi_e)^{\gamma+1}} \exp\left\{-\frac{u_s/\phi_s}{(1+u_e/\phi_e)^{\delta/2}}\right\}, \gamma > 0, \delta \in [0,1] \qquad (9)$$

where δ is the parameter regulating the strength of the interaction between x and e, with $\delta = 1$ being the case of strongest interaction and $\delta = 0$ yielding a separable correlation function with no interaction. Figure 8 shows the profile deviance for the parameter δ, given by

$$D(\delta) = -2\left(l_p(\delta) - l_p(\hat{\delta})\right)$$

where $l_p(\delta)$ is the profile log-likelihood for δ and $\hat{\delta}$ is its maximum likelihood estimate. Note that the whole of $D(\delta)$ lies below the quantile 0.95 of a chi-square distribution with one degree of freedom, which is approximately 3.84. This indicates that the data carry very little information about the interaction parameter δ.

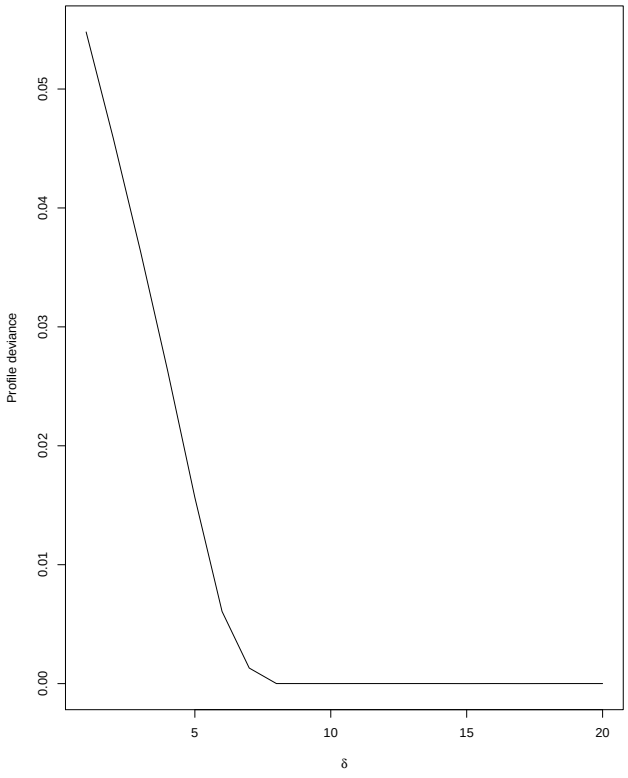

Figure 8: Profile deviance (solid line) for the parameter of interaction using the Gneiting (2002) family given by (9).

6. Conclusion

Our proposed model-based geostatistical framework provides a parsimonious approach for handling nonstationarity patterns in disease risk through the introduction of risk factors into the covariance structure of the model. Our study has shown that this approach is empirically feasible even with moderate sized data-sets. Ignoring non-sationarity using standard geostatistical models can deliver invalid predictive inferences on the health outcome of interest. Future research should be focused on developing model selections criteria in order to identify the optimal set of variables that should be used to model non-stationarity.

References

Amerson, R., Miller, L., M., G., Ramsey, K., Bake, J., 2017. Assessment of anemia levels in infants and children in high altitude peru. Global Journal of Health Science 9, 87–95.

Amoah, B., Giorgi, E., Heyes, D., Burren, S., Diggle, P., 2017. Geostatistical modelling of the association between malaria and child growth in africa. International Journal of Health Geographics 17.

Arab, A., Jackson, M., Kongoli, C., 2014. Modelling the effects of weather and climate on malaria distributions in west africa. Malaria Journal 13.

Banerjee, S., Carlin, B., Gelfand, A., 2015. Hierarchical modeling and analysis for spatial data. Taylor & Francis, CRC Press.

Cook, J., Boy, E., Flowers, C., Daroca, D., 2006. The influence of high-altitude living on body iron. Blood 106, 1441–1446.

Cressie, N., 1993. Statistics for Spatial Data. Wily, New York.

CSA, 2016. Ethiopian Demographic and Health Survey 2016. Central Statistical Agency Ethiopia. Addis Ababa, Ethiopia and Calverton, Maryland, USA.

Diggle, P., Ribeiro, P., 2007. Model-based Geostatistics. Springer-Verlag, New York.

Diggle, P., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. Journal of the Royal Statistical Society. Series C (Applied Statistics) 47, 299–350.

Diggle, P., Thomson, M., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J.and Remme, H., Boussinesq, M., Molyneux, D., 2007. Spatial modelling and prediction of loa loa risk: Decision making under uncertainty. Annals of Tropical Medicine and Parasitology 101, 499–509.

Ejigu, B., Wencheko, E., Berhane, K., 2018. Spatial pattern and determinants of anaemia in ethiopia. PLoS ONE 13, e0197171.

Fuentes, M., 2001. A high frequency kriging approach for non-stationary environmental processes. Environmetrics 12, 469–483.

Galgamuwa, L., Dharmaratne, S., Iddawela, D., 2018. Leishmaniasis in sri lanka: Spatial distribution and seasonal variations from 2009 to 2016. Parasites & Vectors 11.

Giorgi, E., Diggle, P., 2017. Prevmap: an r package for prevalence mapping. Journal of Statistical Software 78, 1–29.

Gneiting, T., 2002. Nonseparable, stationary covariance functions for space–time data. Journal of the American Statistical Association 97, 590–600. doi:10.1198/016214502760047113.

Higdon, D., Swall, J., Kern, J., 1999. A high frequency kriging approach for non-stationary environmental processes. ayesian Statistics 6.

Huang, F., Zhou, S., Zhang, S., Zhang, H., Li, W., 2011. Meteorological factors-based spatio-temporal mapping and predicting malaria in central china. Am. J. Trop.Med. Hyg. 85, 560 – 567.

Jackson, C., Laura, J., Cathy, F., Abigail, C., Kimberly, F., 2010. Modelling the effect of climate change on prevalence of malaria in western africa. Statistica Neerlandica 64, 388–400.

Mateŕn, B., 1960. Spatial Variation:Technical report. Statens Skogsforsningsinstitut. Stockholm.

Mokuolu, A., Adegboye, D., 2014. The impact of environmental factors on malaria prevalence in a peri-urban community. International Journal of Public Health Science 3, 173–178.

Moraga, P., Cano, J., Baggaley, R.F., Gyapong, J.O., Njenga, S., Nikolay, B., Davies, E., Rebollo, M.P., Pullan, R.L., Bockarie, M.J., Hollingsworth, D., Gambhir, M., Brooker, S.J., 2015. Modelling the distribution and transmission intensity of lymphatic filariasis in sub-saharan africa prior to scaling up interventions: integrated use of geostatistical and mathematical modelling. Parasites & Vectors 8, 560. doi:10.1186/s13071-015-1166-x.

Nkurunziza, H., Gebhardt, A., Pilz, J., 2010. Bayesian modelling of the effect of climate on malaria in burundi. Malaria Journal 9.

Paciorek, C., Schervish, M., 2006. Spatial modeling using a new class of nonstationary covariance functions. Environmetrics 17, 483–506.

Sampson, P., Guttorp, P., 1992. Nonparametric estimation of non-stationary spatial covariance structure. Journal of the American Statistical Association 87, 108–119.

Schmidt, A., Guttorp, P., 2011. Considering covariates in the covariance structure of spatial processes. Environmetrics 22, 487–500.

Texier, G., Machault, V., Barragti, M., Boutin, J., 2013. Environmental determinant of malaria cases among travellers. Malaria Journal 12.

Thomson, M., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J., Molyneux, D., Boussinesq, M., 2004. Mapping the distribution of loa boa in cameroon in support of the african programme for onchocerciasis control. Filaria Journal 3.

Ver Hoef, J., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environmental and Ecological Statistics 13, 449–464.

Vianna Neto, J., Schmidt, A., Guttorp, P., 2014. Accounting for spatially varying directional effects in spatial covariance structures. Journal of the Royal Statistical Society: Series C 63, 103–122.

Yazdanpanah, H., M., R., 2013. Analysis of spatial distribution of leishmaniasis and its relationship with climatic parameters (case study: Ilam province). Bull. Env. Pharmacol. Life Sci. 2, 80–86.

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

Zoure, H., Noma, M., Tekle, A., Amazigo, U., Diggle, P., Giorgi, E., Remme, J., 2014. The geographic distribution of onchocerciasis in the 20 participating countries of the african programme for onchocerciasis control:(2) pre-control endemicity levels and estimated number infected. Parasites & vectors 7.