Understanding the effects of dichotomization of continuous outcomes on geostatistical inference

Irene Kyomuhangi[1], Tarekegn A. Abeku[2], Matthew J. Kirby[2,3], Gezahegn Tesfaye[2,4] and Emanuele Giorgi[1]

[1]CHICAS, Lancaster Medical School, Lancaster University, UK
[2]Malaria Consortium, UK
[3]London School of Hygiene and Tropical Medicine
[4]PATH/MACEPA

Abstract

Diagnosis is often based on the exceedance or not of continuous health indicators of a predefined cut-off value, so as to classify patients into positives and negatives for the disease under investigation. In this paper, we investigate the effects of dichotomization of spatially-referenced continuous outcome variables on geostatistical inference. Although this issue has been extensively studied in other fields, dichotomization is still a common practice in epidemiological studies. Furthermore, the effects of this practice in the context of prevalence mapping have not been fully understood. Here, we demonstrate how spatial correlation affects the loss of information due to dichotomization, how linear geostatistical models can be used to map disease prevalence and thus avoid dichotomization, and finally, how dichotomization affects our predictive inference on prevalence. To pursue these objectives, we develop a metric, based on the composite likelihood, which can be used to quantify the potential loss of information after dichotomization without requiring the fitting of Binomial geostatistical models. Through a simulation study and two applications on disease mapping in Africa, we show that, as thresholds used for dichotomization move further away from the mean of the underlying process, the performance of binomial geostatistical models deteriorates substantially. We also find that dichotomization can lead to the loss of fine scale features of disease prevalence and increased uncertainty in the parameter estimates, especially in the presence of a large noise to signal ratio. These findings strongly support the conclusions from previous studies that dichotomization should be always avoided whenever feasible.

Keywords: binary data; dichotomization; disease mapping; linear geostatistical model; model-based geostatistics; prevalence.

1 Introduction

Continuous measurements of disease indicators - e.g. concentration of antibodies in a blood sample - are used in many branches of health research to aid diagnosis and treatment of patients, as well as monitoring and surveillance of diseases in populations. Diagnosis is often based on the exceedance or not of a cut-off value by the continuous indicator, to identify positives and negatives for the disease of interest [1]. In some cases, for instance in anaemia epidemiology, multiple cut-offs are also used to further categorize patients into groups, such as mild, moderate and severe [2]. The rationale for such groupings is to aid and simplify both interpretation and presentation of the results [3–5], while in clinical settings the groupings are used for targeted treatment. As a result, statistical analysis is often carried out on the categorical outcome obtained through the discretization of the continuous measurement.

The disadvantages and loss of information yielded by this practice have been investigated in several studies. Fedorov et al. [6] showed that dichotomization of continuous outcome variables can lead to a loss of 36% in terms of the Fisher's information for the population average. As a result of this, the statistical power required to estimate regression relationships between a health outcome and risk factors is also diminished [7]. Furthermore, in cases where the relationship is non-linear or non-monotonic, dichotomization or categorization into few classes may make that undetectable [4, 5]. All these issues are further exacerbated when the choice of specific cut-offs is inconsistent or, in some cases, even arbitrary [8–10]. For example, cut-offs may vary within and across studies due to differences in the sample populations from which they are derived or due to changes in how they are defined according to clinical practice and operational policy.

In this paper, we investigate the effects of dichotomization of spatially-referenced continuous outcome variables on geostatistical inference. Model-based geostatistics (MBG) [11] is a likelihood-based paradigm that allows to carry out spatially continuous predictive inference on an outcome of interest using a spatially discrete set of data. Over the last two decades, MBG has been increasingly used to map the prevalence of infectious diseases [12], especially in low-resource settings where disease registries are non-existent or geographically incomplete. In this context, there have been global efforts to increase the use of rapid diagnostic tests for diseases such as malaria and HIV [13–16], which are typically recorded as binary by labelling the tested individuals as either positive or negative. In other cases, instead, dichotomization is first carried out on a continuous disease indicator variable and a geostatistical model is then developed on the binary outcome. For example, in Zimmerman et al. [17], a continuous score quantifying the deviation from normal growth in children is dichotomized in order map stunting prevalence; following a similar approach, Magalhaes et al. [18] fit a binomial geostatistical model to dichotomized continuous haemoglobin densities so as to map anaemia prevalence in West Africa.

The effects of dichotomization on geostatistical inference have not been fully understood and, to the best of our knowledge, this is the first study that attempts to pursue this objective in the context of prevalence mapping. More specifically, in this paper, we provide answers to the following questions: 1) How does spatial correlation affect the loss of information and the estimation of regression relationships? 2) Can dichotomization lead to substantially different and more uncertain spatial predictions in disease prevalence? 3) How can we use linear geostatistical models to map disease prevalence and thus avoid the drawbacks of dichotomization?

The structure of the paper is as follows. In section 2 we describe the geostatistical modelling framework for disease prevalence mapping, and outline the differences and links between geostatistical models based on binary and continuous outcomes. In section 3, we first explore the information loss due to dichotomization in terms of the Fisher's information for two observations. We then develop a metric which can be used to assess the loss of information for the estimation of the regression coefficients of any geostatistical model. We also carry out a simulation study to extend our investigation to the estimation of the covariance parameters and spatial predictions for prevalence. In section 4 we illustrate two applications on the mapping of anaemia and stunting prevalence in Africa. Section 5 is a concluding discussion.

In what follows, fitting of geostatistical models and geostatistical prediction have been carried out using the Monte Carlo maximum likelihood method implemented in the `PrevMap` package [19] available from the Comprehensive R Network archive (`cran.r-project.org`).

2

2 The link between geostatistical models for continuous and binary outcomes

Consider data from a cross-sectional survey where information on a continuous health outcome, Y_{ij}, is collected through examination of n_i individuals residing at location x_i for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. We then assume that conditionally on a spatial Gaussian process $S = \{S(x) : x \in \mathbb{R}^2\}$, the Y_{ij} are random Gaussian variables with mean $\mu_{ij} + S(x_i)$ and variance τ^2. From the linear properties of Gaussian distributions, we can write the model in the following compact form

$$Y_{ij} = \mu_{ij} + S(x_i) + Z_{ij}, \tag{1}$$

where Z_{ij} are i.i.d. Gaussian variables with mean zero and variance τ^2, representing unexplained individual-level variation and the mean component μ_{ij} is modeled as a linear regression taking the form

$$\mu_{ij} = \alpha + \beta^\top d(x_i) + \gamma^\top e_{ij}$$

where we distinguish between covariates, $d(x_i)$, that express properties of the locations and covariates, e_{ij}, for individual traits e.g. age and gender.

We model $S(x)$ as a stationary and isotropic Gaussian process with mean zero, variance σ^2 and correlation function $\text{Cor}\{S(x), S(x')\} = \rho(u)$ where $u = ||x - x'||$ denotes the Euclidean distance between x and x'. In the remainder of this paper, we shall define $\rho(\cdot)$ to be an exponentially decaying function with scale parameter ϕ, i.e. $\rho(u) = \exp\{-u/\phi\}$.

Based on a predefined threshold c, whose exceedance or not defines the disease status of an individual, we define the binary outcome $\tilde{Y}_{ij}$ as

$$\tilde{Y}_{ij} = \begin{cases} 1 & \text{if } Y_{ij} < c \\ 0 & \text{if } Y_{ij} \geq c \end{cases}, \tag{2}$$

with $\tilde{Y}_{ij} = 1$ indicating a positive case for the disease under investigation and $\tilde{Y}_{ij} = 0$ for a negative case.

From the model of the continuous outcome in (1), it follows that the model for $\tilde{Y}_{ij}$ is given by

$$
\begin{aligned}
P\left(\tilde{Y}_{ij} = 1 \Big| S(x_i)\right) &= P\left(Y_{ij} < c | S(x_i)\right) \\
&= P\left(\frac{Y_{ij} - \mu_{ij} - S(x_i)}{\tau} < \frac{c - \mu_{ij} - S(x_i)}{\tau} \Big| S(x_i)\right) \\
&= \Phi\left(\frac{c - \mu_{ij} - S(x_i)}{\tau}\right) = p_{ij},
\end{aligned}
\tag{3}
$$

where $\Phi(\cdot)$ is the cumulative density function of a standard Gaussian variable. Hence, the resulting model for $\tilde{Y}_{ij}$ is a Binomial geostatistical model with probit link function and linear predictor

$$\eta_{ij} = \Phi^{-1}(p_{ij}) = \tilde{\mu}_{ij} + \tilde{S}(x_i) \tag{4}$$

where $\tilde{\mu}_{ij} = -\mu_{ij}/\tau$ and $\tilde{S}(x_i) = -S(x_i)/\tau$.

The functional relationships that link the parameters of the geostatistical model for $\tilde{Y}_{ij}$ with those of the model for Y_{ij} are the following

$$\begin{cases} \tilde{\alpha} = (c - \alpha)/\tau \\ \tilde{\beta} = -\beta/\tau \\ \tilde{\gamma} = -\gamma/\tau \\ \tilde{\sigma}^2 = \sigma^2/\tau^2 \end{cases}. \tag{5}$$

The above equations can thus be used to obtain the parameter estimates for a geostatistical model for $\tilde{Y}_{ij}$ by transforming the parameter estimates from the geostatistical model for Y_{ij}. Note that it is not possible, instead, to map the estimates from the model for $\tilde{Y}_{ij}$ into those for Y_{ij} because the parameter τ^2 cannot be estimated from binary data. The unstructured component Z_{ij} is in fact integrated out in (4) and, as shown in (5), all the parameters on the left hand-side are expressed as a ratio between τ and all other parameters in the model for Y_{ij}. Finally, ϕ is not included in (5), since the scale of the spatial correlation of $\tilde{S}(x)$ is the same as that of $S(x)$.

3 Quantifying the effects of dichotomization

In Section 3.1, we first study the loss of information due to dichotomization for the estimation of the mean of the process using an intercept-only model, when all other parameters are known. In Section 3.2, we carry out a simulation study to the more common case when all parameters are unknown. Here we study the effect on dichotomization on parameter estimation. In both sections, we shall restrict our attention to the scenario of a single observation per location, hence we set $n_i = 1$ for all i and drop the j subscript.

3.1 Unknown regression coefficients and known covariance parameters

3.1.1 Special case of m=2 for an intercept-only model

The objective in this section is to quantify the loss of information in terms of the expected Fisher information (EFI) with respect to $\tilde{\alpha}$, the parameter which regulates the mean level of disease prevalence. Here, we restrict our attention to the simpler case of two observations at two locations, hence $m = 2$ and $n_1 = n_2 = 1$, for an intercept-only model. As it will be shown in the applications of Section 4, this simpler scenario provides useful insights on the effects of dichotmozation which are consistently observed in the case of more than two observations. A more general scenario, however, shall also be considered in the next section.

We re-express the linear geostatistical model for a continuous outcome Y_i as

$$Y_i = \alpha + S(x_i) + Z_i, \text{ for } i = 1, \ldots, m \tag{6}$$

where $S(x_i)$ is a stationary and isotropic Gaussian process with the same properties as defined in equation (1).

Let $\Sigma_Y = \Sigma + \tau^2 I$ be the covariance matrix of the vector $Y = (Y_1, \ldots, Y_m)$, with Σ and I denoting the spatial covariance matrix with (i, j)-th entry $\sigma^2 \exp\{-\|x_i - x_j\|/\phi\}$ and an m by m identity matrix, respectively.

In order to quantify the loss of information that arises from the dichotomization of the Y_i, we first re-parametrize the linear model in (6) with respect to the prevalence parameters as defined in (5); note that $\alpha = c - \tau\tilde{\alpha}$. We then obtain the EFI for $\tilde{\alpha}$ under the linear model, given by

$$I_Y(\tilde{\alpha}) = \tau^2 \mathbb{1}^T \Sigma_Y^{-1} \mathbb{1}, \tag{7}$$

where $\mathbb{1}$ is a vector with all entries equal to 1.

In the case of the dichotomized outcome $\tilde{Y}_i$, the computation of the EFI is further complicated by the fact that the log-likelihood function is not available in closed form. More specifically, this is given by the marginal distribution of the outcome $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$, i.e.

$$\log L(\tilde{\alpha}) = \log \left(\int_{\mathbb{R}^m} f(\tilde{s}) f(\tilde{y}|\tilde{s}; \tilde{\alpha}) \, d\tilde{s} \right), \tag{8}$$

where $f(\tilde{s})$ is the density of a multivariate Gaussian distribution with mean zero and covariance matrix $\tilde{\Sigma} = \Sigma/\tau^2$, whilst

$$
\begin{aligned}
f(\tilde{y}|\tilde{s}; \tilde{\alpha}) &= \prod_{i=1}^{m} f(\tilde{y}_i|\tilde{s}_i; \tilde{\alpha}) \\
&= \exp\left\{ \sum_{i=1}^{m} \left[\tilde{y}_i \log\left(\frac{p_i}{1-p_i}\right) + (1-\tilde{y}_i)\log\{1-p_i\} \right] \right\} \\
&= \exp\{g(\tilde{y}|\tilde{s}; \tilde{\alpha})\}
\end{aligned}
\tag{9}
$$

where $\Phi^{-1}(p_i) = \tilde{\alpha} + \tilde{S}(x_i)$. To obtain the EFI for $\tilde{\alpha}$, we first take the second derivative of (8) with respect to $\tilde{\alpha}$ to give

$$
\begin{aligned}
\frac{d^2 \log L(\tilde{\alpha})}{d^2\tilde{\alpha}} &= L^{-1}(\tilde{\alpha}) \int_{\mathbb{R}^m} f(\tilde{s})f(\tilde{y}|\tilde{s}; \tilde{\alpha})\left[\left(\frac{dg(\tilde{y}|\tilde{s}; \tilde{\alpha})}{d\tilde{\alpha}}\right)^2 + \right. \\
&\left. \frac{d^2 g(\tilde{y}|\tilde{s}; \tilde{\alpha})}{d^2\tilde{\alpha}} \right] d\tilde{s} + \left(\frac{d \log L(\tilde{\alpha})}{d\tilde{\alpha}}\right)^2,
\end{aligned}
\tag{10}
$$

where

$$
\frac{d \log L(\tilde{\alpha})}{d\tilde{\alpha}} = L^{-1}(\tilde{\alpha}) \int_{\mathbb{R}^m} f(\tilde{s})f(\tilde{y}|\tilde{s}; \tilde{\alpha})\frac{dg(\tilde{y}|\tilde{s}; \tilde{\alpha})}{d\tilde{\alpha}} \, d\tilde{s}.
$$

Finally, we average over the distribution of $\tilde{Y}$

$$
I_{\tilde{Y}}(\tilde{\alpha}) = E_{\tilde{Y}}\left[-\frac{d^2 \log L(\tilde{\alpha})}{d^2\tilde{\alpha}} \right].
$$

Since the above quantity is not available in closed form we compute $I_{\tilde{Y}}(\tilde{\alpha})$ using Monte Carlo methods.

To quantify the loss of information, we then use the following metric $R(\tilde{\alpha}) = 1 - I_{\tilde{Y}}(\tilde{\alpha})/I_Y(\tilde{\alpha})$. In the special case of $S(x) = 0$ for all x, $R(\tilde{\alpha})$ reduces to

$$
R(\tilde{\alpha}) = 1 - \left[\frac{[\Phi''(\tilde{\alpha})][1 - \Phi(\tilde{\alpha})] - [\Phi'(\tilde{\alpha})]^2}{1 - \Phi(\tilde{\alpha})} - \frac{[\Phi''(\tilde{\alpha})][\Phi(\tilde{\alpha})] - [\Phi'(\tilde{\alpha})]^2}{\Phi(\tilde{\alpha})} \right]
\tag{11}
$$

where $\Phi'(\cdot)$ and $\Phi''(\cdot)$ are the first and second derivative of $\Phi(\cdot)$, respectively. Fedorov et al. [6] have shown that $I_{\tilde{Y}}(\tilde{\alpha}) \leq I_Y(\alpha)$, and that the lower limit of (11) is about 36%. Also, note that (11) is not dependent on m.

To compute the integrals which define $-d^2 \log L(\tilde{\alpha})/d^2\tilde{\alpha}$, we use a quadrature apporoach based on Quasi Monte Carlo methods. Finally, for the computation of the expectation in $I_{\tilde{Y}}(\tilde{\alpha})$, we use 10,000 samples and vary the spatial correlation between the two observations over $\rho \in \{i/10; i = 1, \ldots, 7\}$. Figure 1 shows different curves of $R(\tilde{\alpha})$, as a percentage, by setting $\sigma^2 = 1$ and letting τ^2 vary over the set $\{0.5, 1, 2\}$. Notice that these curves are symmetric with respect to 0, although they are shown only for positive values of $\tilde{\alpha}$. Across all three panels of Figure 1, we observe that increasing values of ρ lead to a reduction in the loss of information, although such reduction becomes smaller when the data are more noisy, i.e. when τ^2 also increases. Most notably, the largest loss of information is observed for values of prevalence close to 100% and 0% corresponding to large positive and negative values for $\tilde{\alpha}$, respectively. The variance τ^2 of the unstructured residuals Z_i also plays a very important role as shown by the dramatic increase in $R(\tilde{\alpha})$ for $\tau^2 = 2$, with all curves placed above $R(\tilde{\alpha}) = 0.65$.
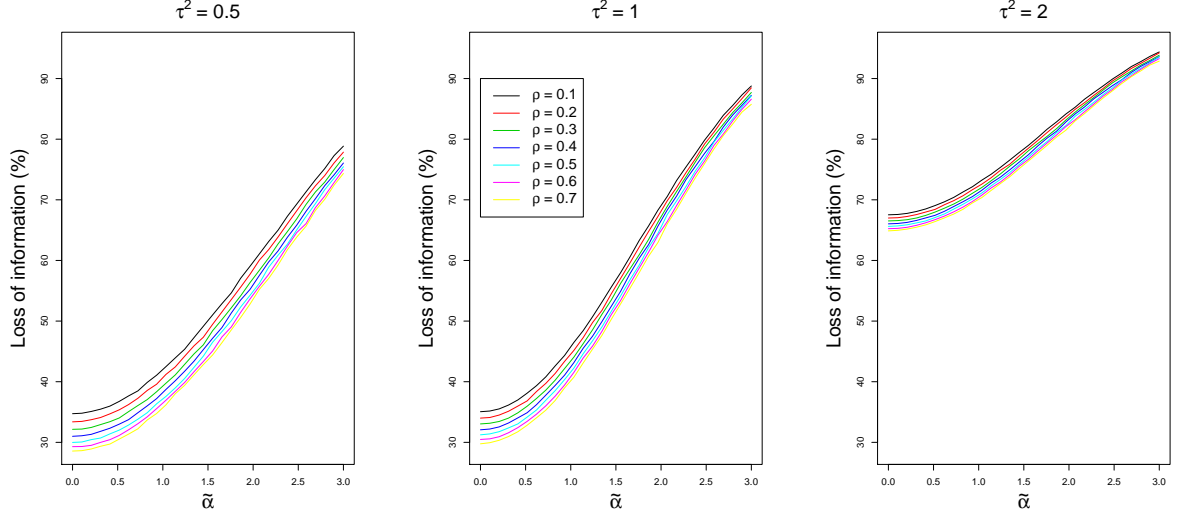
Figure 1: Curves for $R(\tilde{\alpha})$, shown as a percentage, by fixing $\sigma^2 = 1$ and varying $\tau^2 \in \{0.5, 1, 2\}$ and the spatial correlation between two observations $\rho \in \{i/10; i = 1, \ldots, 7\}$.

3.1.2   General case m > 2

For the general case of more than two locations (i.e. $m > 2$), the effects of dichotomization will also be dependent on the spatial arrangement of the sampled locations. In this section, we develop a metric that allows to quantify the potential loss of information due to dichotomization of continuous outcomes with respect the estimation of the regression coefficients $\theta = (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ of any geostatistical model as defined by (1).

In order to lower the computational burden, we first approximate the likelihood function of both the continuous and dichotomized data using a composite likelihood approach [20]. More specifically, we consider all possible pairs of $\tilde{Y}_{ij}$ and treat each of these as independent bivariate distributions. Let $\tilde{Y}_h$ and $\tilde{Y}_k$ denote the vectors of binary outcomes associated with locations x_h and x_k and which are obtained through dichotomization of Y_h and Y_k, respectively. We then write

$$L(\theta) \approx L_{CL}(\theta) = \prod_{h=1}^{m-1} \prod_{k=h+1}^{m} f\left(\tilde{y}_h, \tilde{y}_k; \tilde{\theta}\right). \tag{12}$$

In the above equation the bivariate probability functions $f(\tilde{y}_h, \tilde{y}_k; \tilde{\theta})$ are expressed by the following integral in two dimensions

$$f(\tilde{y}_h, \tilde{y}_k; \tilde{\theta}) = \int_{R^2} f(\tilde{s}) f\left(\tilde{y}_h, \tilde{y}_k | \tilde{s}; \tilde{\theta}\right) d\tilde{s}$$

where $f\left(\tilde{y}_h, \tilde{y}_k | \tilde{s}; \tilde{\theta}\right)$ consists of a product of $n_h + n_k$ probability functions for the binary observations in y_h and y_k.

Let $\hat{\theta}_{LM}$ denote the maximum likelihood estimates of θ obtained from the linear model using the system of equations in (5). In order to understand how more or less dispersed the composite likelihood becomes after dichotmization, we proceed as follows. We first compute the second derivative of the composite log-likelihood at $\hat{\theta}_{LM}$, i.e.

$$H_{\tilde{Y}}(\hat{\theta}_{LM}) = \left[\frac{\partial^2 \log L_{CL}}{\partial \theta_i \partial \theta_j}\right]_{\theta = \hat{\theta}_{LM}}.$$

6

For a continuous outcome Y, we have

$$H_Y(\theta) = -\tau^2 \sum_{h=1}^{m-1} \sum_{k=h+1}^{m} D_h^\top \Sigma_{hk}^{-1} D_k$$

where Σ_{hk} is covariance matrix between Y_h and Y_k, and D_h and D_k are the design matrices associated with locations x_h and x_k, respectively. To quantify the change in the dispersion of the composite likelihood around $\hat{\theta}_{LM}$, we finally compute

$$
\begin{aligned}
CLD(\hat{\theta}_{LM}) &= \log\{\det[-H_{\tilde{Y}}^{-1}(\hat{\theta}_{LM})]\} - \log\{\det[-H_Y^{-1}(\hat{\theta}_{LM})]\} \\
&= \log\{\det[-H_Y(\hat{\theta}_{LM})]\} - \log\{\det[-H_{\tilde{Y}}(\hat{\theta}_{LM})]\}
\end{aligned}
$$

where $\det(A)$ denotes the determinant of a square matrix A. Large values of $CLD(\hat{\theta}_{LM})$ indicate a more dispersed composite likelihood around $\hat{\theta}_{LM}$ for the binary data $\tilde{Y}$, which we interpret as evidence of loss of information due to dichotmization. Also, note that computation of CLD can be carried out without fitting any Binomial geostatatistical model.

In the applications of Sections 4.1 and 4.2, we compute the CLD metric, using by plugging-in the maximum likelihood estimates for the covariance parameters from the linear geostatistical models.

3.2   Simulation study

The objective of this simulation is to quantify the effects of dichotomization on parameter estimation and spatial predictions of prevalence. To this end we consider the linear model for Y_i as specified in (6) and its dichotomization using a threshold c to give $\tilde{Y}_i = 1$ if $Y_i < c$ and 0 otherwise.

In the simulation, we set $\alpha = 0$ and $\sigma^2 = 1$. We consider several scenarios obtained through all possible combinations of values for the model parameters defined in Table 1. For a given scenario, we simulate 1,000 data-sets of both continuous and dichotomized outcomes and fit their corresponding geostatistical models. We use a regular grid covering the unit square with spacing of $1/14$ to give a sample size of $m = 225$. For each of the fitted models, we extract the estimates for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and ϕ, and predict prevalence $p_i = \Phi\left(\tilde{\alpha} + \tilde{S}(x_i)\right)$ at each of the grid points. We summarise the results using the bias and mean square error (MSE) and, for the prevalence predictions, average these two indices over the grid locations.

Tables 2 and 3 report the results for the model parameters and spatial predictions for prevalence, respectively. Overall, bias and MSE for $\tilde{\alpha}$ and ϕ are consistently smaller in the model fitted to the continuous data (C) than for that fitted to the binary data (B). In the case of $\tilde{\sigma}^2$, instead, the performance of both models is strongly affected by the scale of the spatial correlation ϕ: for $\phi = 0.1$ the model B outperforms model C in terms of bias and MSE, whilst the opposite is observed for $\phi = 0.2$. A possible explanation for this may be due to the fact that in the linear geostatistical model, higher spatial correlation helps to better separate the individual contributions of the signal component $S(x_i)$ and the noise component Z_i to the total variation in the outcome Y_i, thus improving the estimation of $\tilde{\sigma}^2 = \sigma^2/\tau^2$. In the case of the binary data, instead, $\tilde{S}(x_i)$ is the only source of over-dispersion and, as a result of this, a higher spatial correlation leads to a larger number of concordant binary outcomes and, therefore, to a poorer estimate of the variance of $\tilde{S}(x_i)$. Also, we notice that the estimation of $\tilde{\sigma}^2$ and ϕ does not appear to be affected by the threshold c, unlike $\tilde{\alpha}$. Finally, the results for the spatial predictions of prevalence show that the performance of model C is unaffected by changes in c and τ^2, while for $\phi = 0.2$ the predictions have slightly lower MSE than for $\phi = 0.1$. Model B, instead, delivers predictions with higher bias for increasing c which can be partly explained by the positive increase in the bias in the estimates of $\tilde{\alpha}$ for increasing c.

7

We have also conducted further simulations under the same scenarios defined in Table 1 but for a larger sample size $m = 450$, by placing additional points on a regular grid adjacent to the unit square so as to cover the rectangle $[0, 2] \times [0, 1]$. The results, reported in the Appendix (Tables A2 and A2), lead to the same conclusions drawn for $m = 225$.

Table 1: True values for τ^2, ϕ and c used in the simulation study.

	Symbol	Variations
Variance of the nugget effect	τ^2	0.5, 1, 2
True scale of spatial correlation	ϕ	0.1, 0.2
Cut-off	c	0, 0.2, 0.4

Table 2: Bias and mean square error (in brackets) for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and the estimate $\hat{\phi}$ obtained from the geostatistical models fitted the binary (B) and continuous (C) outcomes.

Parameter	τ^2	ϕ	c=0 B	c=0 C	c=0.2 B	c=0.2 C	c=0.4 B	c=0.4 C
$\tilde{\alpha}$	0.5	0.1	0.009 (0.168)	0.005 (0.113)	0.063 (0.312)	0.017 (0.115)	0.157 (0.263)	0.042 (0.139)
	1	0.1	-0.009 (0.126)	-0.008 (0.068)	0.060 (0.409)	0.004 (0.064)	0.153 (0.119)	0.041 (0.080)
	2	0.1	-0.006 (0.079)	-0.003 (0.036)	0.077 (0.493)	0.022 (0.040)	0.148 (0.169)	0.050 (0.054)
	0.5	0.2	-0.025 (0.624)	-0.013 (0.296)	0.156 (0.648)	0.048 (0.282)	0.238 (0.722)	0.028 (0.303)
	1	0.2	-0.018 (0.332)	-0.007 (0.150)	0.093 (0.585)	0.008 (0.137)	0.215 (0.323)	0.025 (0.145)
	2	0.2	-0.007 (0.185)	-0.004 (0.080)	0.106 (0.566)	0.023 (0.080)	0.164 (0.272)	0.022 (0.088)
$\tilde{\sigma}^2$	0.5	0.1	-0.011 (1.296)	0.788 (5.843)	-0.033 (1.324)	0.604 (5.170)	0.022 (1.594)	0.734 (5.741)
	1	0.1	0.234 (0.787)	0.822 (5.953)	0.190 (0.688)	0.750 (5.255)	0.224 (0.701)	0.741 (5.942)
	2	0.1	0.211 (0.326)	0.672 (5.047)	0.217 (0.364)	0.600 (4.961)	0.204 (0.324)	0.653 (4.833)
	0.5	0.2	1.641 (8.894)	0.527 (3.091)	1.784 (12.672)	0.574 (3.268)	1.566 (8.464)	0.515 (3.401)
	1	0.2	1.162 (3.755)	0.399 (1.712)	1.064 (3.243)	0.372 (1.619)	1.048 (2.949)	0.365 (1.671)
	2	0.2	0.548 (0.871)	0.254 (1.104)	0.575 (1.074)	0.304 (1.829)	0.534 (0.910)	0.341 (1.707)
$\hat{\phi}$	0.5	0.1	0.088 (0.016)	0.007 (0.002)	0.084 (0.015)	0.009 (0.002)	0.085 (0.015)	0.007 (0.002)
	1	0.1	0.071 (0.014)	0.004 (0.003)	0.072 (0.014)	0.006 (0.004)	0.074 (0.015)	0.006 (0.003)
	2	0.1	0.060 (0.017)	0.010 (0.007)	0.063 (0.022)	0.009 (0.007)	0.056 (0.014)	0.010 (0.006)
	0.5	0.2	0.076 (0.029)	-0.017 (0.011)	0.083 (0.034)	-0.020 (0.008)	0.076 (0.030)	-0.021 (0.009)
	1	0.2	0.068 (0.030)	-0.014 (0.016)	0.058 (0.027)	-0.028 (0.010)	0.061 (0.034)	-0.023 (0.013)
	2	0.2	0.032 (0.025)	-0.024 (0.019)	0.037 (0.031)	-0.024 (0.016)	0.028 (0.019)	-0.032 (0.015)

Table 3: Bias and mean square error (in brackets), averaged over a 1/14 by 14 regular grid covering the unit square (hence, $m = 225$), for the spatial predictions of prevalence obtained from the geostatistical models fitted to the binary (B) and continuous (C) outcomes.

τ^2	ϕ	c=0 B	c=0 C	c=0.2 B	c=0.2 C	c=0.4 B	c=0.4 C
0.5	0.1	0.001 (0.060)	0.001 (0.039)	0.018 (0.059)	0.001 (0.038)	0.034 (0.058)	0.001 (0.036)
1	0.1	-0.001 (0.051)	0.001 (0.038)	0.018 (0.051)	0.001 (0.038)	0.038 (0.050)	0.001 (0.036)
2	0.1	-0.001 (0.040)	0.001 (0.033)	0.020 (0.040)	-0.001 (0.033)	0.038 (0.040)	-0.002 (0.032)
0.5	0.2	-0.001 (0.042)	0.001 (0.030)	0.013 (0.042)	-0.001 (0.030)	0.025 (0.041)	-0.001 (0.029)
1	0.2	0.001 (0.037)	0.001 (0.028)	0.014 (0.037)	-0.001 (0.028)	0.030 (0.036)	-0.001 (0.027)
2	0.2	-0.001 (0.031)	0.001 (0.024)	0.019 (0.031)	-0.001 (0.024)	0.034 (0.031)	-0.002 (0.024)

4   Applications

4.1   Mapping anaemia prevalence in Ethiopia

In this section, we analyse data collected from the Beyond Garki project [1]. This project consisted of cross-sectional surveys which were conducted in selected study sites in Ethiopia

---
[1] www.malariaconsortium.org/beyondgarki

and Uganda to monitor changes in malaria risk in the context of interventions that had been implemented. The study sites were defined as a 'health centre and the catchment population in selected villages around it'. Here we subset the data for the Hembecho site, in Ethiopia, collected during the 2012 survey, where a random sample of households were selected from a list of enumerated households from all villages within a radius of 2 to 6 kilometers of the health facility. Among the data obtained in this survey were continuous measurements of haemoglobin density (g/dL), taken from blood samples of individuals living in the households. These measurements were then used to determine the anaemia status of individuals. Further details of the study design and data collection can be found in Abeku et al. [21].

In this analysis, the objective is to identify areas where the anaemia prevalence is highly likely to exceed a 20% threshold for 20 year old women. Hence, we map and compare exceedance probabilities from the geostatistical models for the continuous and binary outcomes. The chosen threshold for anaemia prevalence has clinical, operational and public health significance for policy decisions, with the World Health Organisation (WHO) classifying 20% anaemia prevalence as 'moderate public health significance'[2]. Finally, the rationale for carrying out predictions for 20 years old women is that one of the key WHO Global Nutrition Targets for 2025 is a 50% reduction of anaemia in women of reproductive age [2, 22], which is defined as 15-49 years [23].

The data-set contains information on 1712 individuals distributed over 457 households, with an average of 3.7 individual in each household. The continuous outcome variable, Y_{ij}, is the log-transformed haemoglobin density for the j-th individual at the i-th household. To account for the non-linear relationship between the log-transformed anaemia density and age, as shown in Figure 2, we use a linear spline with knots at 15 and 30 years. Our proposed linear model for Y_{ij} is thus expressed as

$$Y_{ij} = \alpha + \sum_{h=1}^{3} \beta_h b_h(a_{ij}) + \beta_4 d_{ij} + S(x_i) + Z_{ij}, \tag{13}$$

where: a_{ij} is the age, in years, of an individual; d_{ij} is a binary indicator of the sex of an individual, with "female" as reference category; $b_h(\cdot)$ are the base functions of the linear spline defined as $b_1(a) = a$, $b_2(a) = max\{0, a - 15\}$ and $b_3(a) = max\{0, a - 30\}$.
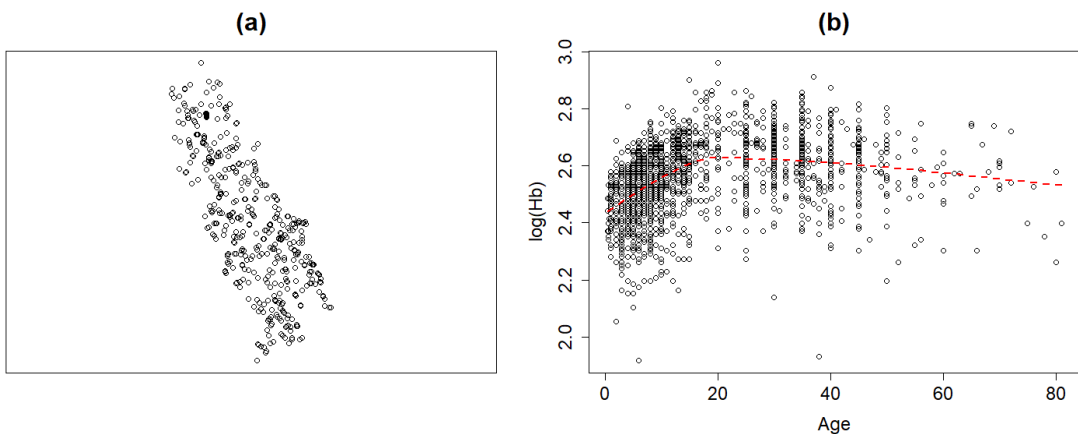
Figure 2: (a) locations of the households in the survey; (b) scatter plot of the log-transformed haemoglobin density against age, in years. The dashed red line in the in panel (b) is a least square fit of the linear spline defined in the main text of Section 4.1.

Dichotomization of Y_{ij} results in the binary outcome variable Y_{ij}^* indicating anaemia status, where $Y_{ij}^* = 1$ denoting a positive case and $Y_{ij}^* = 0$ a negative case for severe anaemia. In order to classify an individual as positives or negatives, thresholds for severe anaemia have

9

been applied using individual-level information on age, sex and pregnancy status as defined in Table 4. As result of this, we then modify equation (3) as

$$p_{ij} = \Phi\left(\frac{c_{ij} - \mu_{ij} - S(x_i)}{\tau}\right)$$

where c_{ij} is the logarithm of the threshold values which must now be incorporated as an additional covariate into the linear predictor, i.e.

$$\Phi^{-1}(p_{ij}) = \tilde{\alpha} + \sum_{h=1}^{3} \tilde{\beta}_h b_h(a_{ij}) + \tilde{\beta}_4 d_{ij} + \tilde{\beta}_5 c_{ij} + \tilde{S}(x_i)$$

where $\tilde{\beta}_5 = 1/\tau$.

**Table 4:** Thresholds of haemoglobin densities (g/dL) for anaemia diagnosis [2].

Age or Sex group	Mild	Anaemia Moderate	Severe
Children (Age 6-59 months)	10.0-10.9	7.0-9.9	< 7.0
Children (Age 5-11 yrs)	11.0-11.4	8.0-10.9	< 8.0
Children (Age 12-15yrs)	11.0-11.9	8.0-10.9	< 8.0
Pregnant women (Age > 15yrs)	10.0-10.9	7.0-9.9	< 7.0
Non-pregnant women (Age > 15 yrs)	11.0-11.9	8.0-10.9	< 8.0
Men (Age > 15 yrs)	11.0-12.9	8.0-10.9	< 8.0

Table 5 reports the maximum likelihood estimates and 95% confidence intervals of the model parameters for the binary and continuous outcomes. The linear geostatistical model gives an estimate for τ^2, the variance of Z_{ij}, of about 1.133×10^{-2} (95% CI: 1.050×10^{-2}, 1.222×10^{-2}) and for σ^2 of about 1.558×10^{-3} (95% CI: 0.954×10^{-3}, 2.422×10^{-3}), yielding an estimated noise to signal ratio τ^2/σ^2 of about 7.3. The CLD metric (Section 3.1.2) is 771.235 which indicates a larger dispersion of the composite likelihood for the binary data than for the continuous data. We observe that the estimates of the parameters are all comparable except for $\tilde{\sigma}^2$ $(= \sigma^2/\tau^2)$, as indicated by the non-overlapping confidence intervals from the two models. More importantly, we observe that the confidence intervals for the regressions coefficients are narrower for the linear model.

Figure 3 shows the resulting anaemia prevalence predictions for 20 year old women from the two models. While the overall pattern of predicted prevalence is similar between the models (see Figures 3(a) and 3(b)), there are non-negligible differences ranging from -8.23% to 7.85% prevalence (Figure 3(c)). Similarly, the maps of the exceedance probability qualitatively show similar spatial patterns (figures 3(d) and 3(e)). However, we identify small areas where the differences range from -38.6% to 31.20% (Figure 3(f)).

Table 5: Maximum likelihood estimates with associated 95% confidence intervals (CI) for the geostatistical models fitted to the anaemia data.

Term	Binomial model Estimate	95% CI	Linear model Estimate	95% CI
$\tilde{\beta}_1$	0.079	(-0.220, 0.378)	-0.278	(-0.377, -0.179)
$\tilde{\beta}_2$	-0.066	(-0.124, -0.008)	-0.115	(-0.131, -0.100)
$\tilde{\beta}_3$	0.052	(-0.025, 0.129)	0.097	(0.072, 0.122)
$\tilde{\beta}_4$	0.071	(0.021, 0.121)	0.050	(0.031, 0.069)
$\tilde{\sigma}^2$	0.527	(0.395,0.705)	0.138	(0.082, 0.218)
ϕ	0.325	(0.201,0.396)	0.250	(0.093, 0.549)
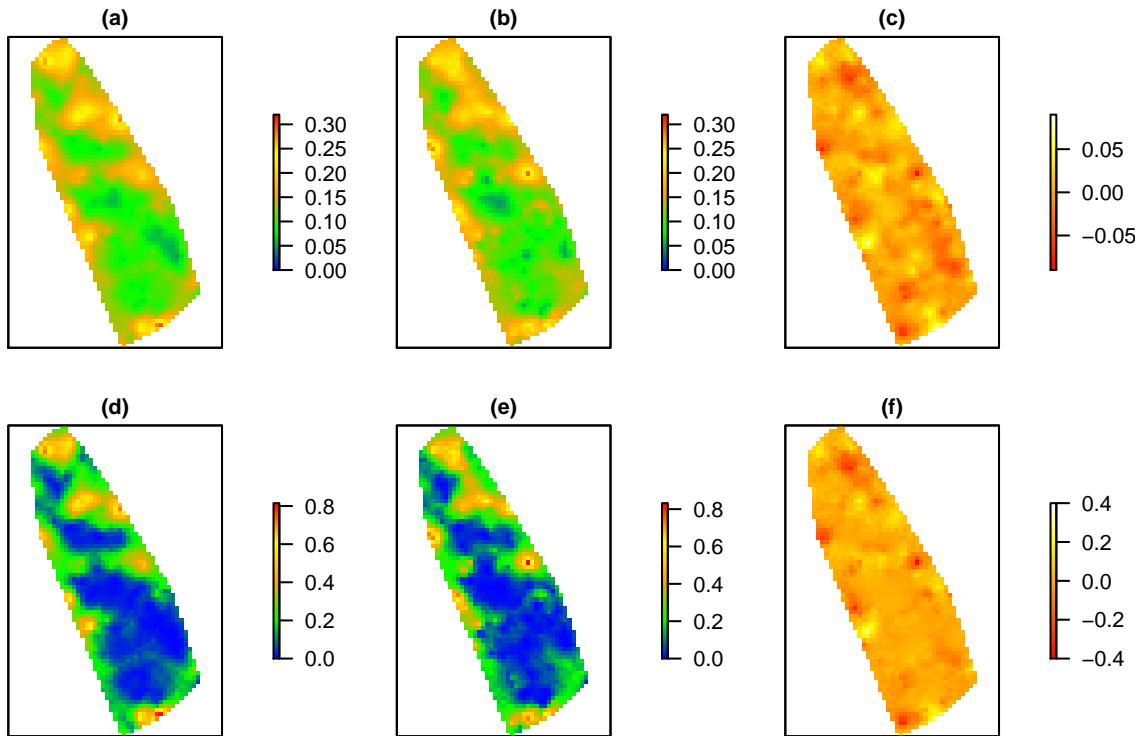
Figure 3: Predicted anaemia prevalence for a 20 year old woman. Upper panels: prevalence surfaces from the binomial model (a) and the continuous models (b), and the difference between the first and the second (c). Lower panels: exceedance probabilities for a 20% prevalence threshold obtained from the binary (d) and the continuous models (e), and their difference (f).

4.2 Mapping stunting prevalence in Ghana

The data analysed in this section are from the 2014 Demographic and Health Survey[2] (DHS) conducted in Ghana. DHS are nationally representative household surveys conducted about every 5 years, and provide data on health and population indicators for monitoring and impact evaluation across Africa. The DHS surveys follow a stratified two-stage cluster design where in the first stage, enumeration areas are selected from previous population census files, followed by a second stage where, for each selected enumeration area, samples of households are sampled from updated lists of households to generate the so called sampling clusters. The GPS locations of a sampling cluster is then assigned to each of the individuals falling within that cluster.

Among the health indicators collected in this survey are anthropometric measurements, which are used to calculate the height-for-age Z-score (HAZ). HAZ are standardized scores which indicate the standard deviation from the mean of children's heights based on the WHO growth standards [24, 25] and are comparable across ages and sex. HAZ values below -2 are taken as an indication of stunted growth.

One of the key WHO Global Nutrition Targets for 2025 is a 40% reduction in the number of children under-5 who are stunted [22, 26]. Additionally, a stunting prevalence above 40% is considered a high public health significance [27]. Accordingly, we aim to map the exceedance probability of 40% stunting prevalence for a 2 year old who falls in the lowest wealth index category and whose mother has poor education.

The data include information on children under 5 years old, with a total of 2671 sampled children and 410 clusters, giving an average of 6.5 children per cluster. The continuous

---

[2] dhsprogram.com

outcome variable, Y_{ij}, is the HAZ for child j in cluster i. Figure 4 (a) shows the empirical relationship between HAZ and age in years. Using a similar approach of the previous analysis, we capture the non-linear relationship with a linear spline having knots at 1 and 2 years. Hence, the resulting linear geostatistical model is

$$Y_{ij} = \alpha + \sum_{h=1}^{3} \beta_h b_h(a_{ij}) + \beta_4 d_{ij} + \beta_5 e_{ij} + S(x_i) + Z_{ij}, \tag{14}$$

where: a_{ij} is the age of a child; the basis functions of the linear splines are $b_1(a) = a$, $b_2(a) = \max\{0, a-1\}$ and $b_2(a) = \max\{0, a-2\}$; d_{ij} is a score of maternal education, taking integer values from 1="Poorly educated" to 3="Highly educated"; e_{ij} is a wealth index of the household, taking integer values from 1="Poor" to 3="Rich".

(a) **(b)**
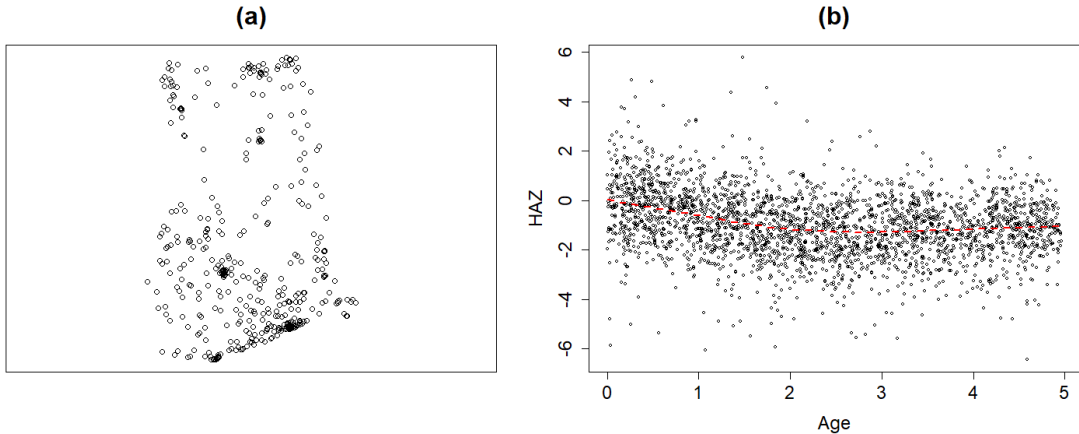
Figure 4: Figure (a) shows the spatial distribution of households included in the analysis, while (b) shows the relationship between HAZ and age. The red dashed line in panel (b) corresponds to a least square fit of the linear spline defined in (14).

The maximum likelihood estimates and associated 95% confidence intervals are shown in Table 6. We also report that the estimates for σ^2 and τ^2 from the linear geostatistical model are 0.071 (95% CI: 0.037, 0.126) and 1.396 (95% CI: 1.318, 1.477), respectively. Hence, the estimated ratio τ^2/σ^2 is about 20, indicating that the data are substantially more noisy than those analysed in the previous section. This is also reflected in the CLD metric yielding a value of 9667.012 which is substantially larger than that reported for the anaemia analysis. Following from the results of Section 3, this suggests that the effects of dichotomization on geostatistical inference will also be stronger. We observe that the estimates of the regression coefficients are concordant in sign but the size of the effects of the covariates are different as indicated by the non-overlapping confidence intervals; as in the previous section, we observe that the confidence intervals for the regression coefficients from the linear model are all narrower. The estimated $\tilde{\sigma}^2$ and ϕ are also substantially different, with the linear geostatistical model providing lower estimates and narrower confidence intervals for both parameters.

The differences in the parameter estimates are also reflected in Figure 5 which shows the predicted surfaces of stunting prevalence and the exceedance probabilities from the two models. These predictions are for a for a 2 year old who falls in the lowest wealth index category, and whose mother has poor education. Qualitatively, both models identify high and low levels of prevalence in the same areas. However, the differences in the predicted prevalence between the binomial model (Figure 5(a)) and the continuous model (Figure 5(b)), range from -9.38% to 19.98% prevalence (Figure 5(c)). Most notably, the binomial model presents much smoother maps than those from the linear model. For example, the binomial model identifies a single large hot-pot in the eastern part of Ghana, as being highly likely to

exceed 40%. The linear model, instead, shows three neighbouring hot-spots in the same area. The differences in exceedance probabilities between the two models range from -51.50% to 64.90% (Figure 5(f))

Table 6: Maximum likelihood estimates with associated 95% confidence intervals (CI) for the geostatistical models fitted to the data on childhood malnutrition.

Term	Binomial model		Linear model	
	Estimate	95% CI	Estimate	95% CI
$\tilde{\beta}_1$	0.772	(-0.002, 1.545)	0.554	(0.338, 0.771)
$\tilde{\beta}_2$	0.774	(-0.331, 1.879)	0.168	(-0.168, 0.502)
$\tilde{\beta}_3$	-1.942	(-2.463, -1.421)	-0.830	(-1.024, -0.639)
$\tilde{\beta}_4$	-0.721	(-1.174, -0.269)	-0.139	(-0.239, -0.039)
$\tilde{\beta}_5$	-0.444	(-0.666, -0.221)	-0.259	(-0.332, -0.186)
$\tilde{\sigma}^2$	0.256	(0.102, 0.528)	0.051	(0.026, 0.091)
ϕ	157.301	(59.706, 341.451)	51.899	(14.657, 136.232)
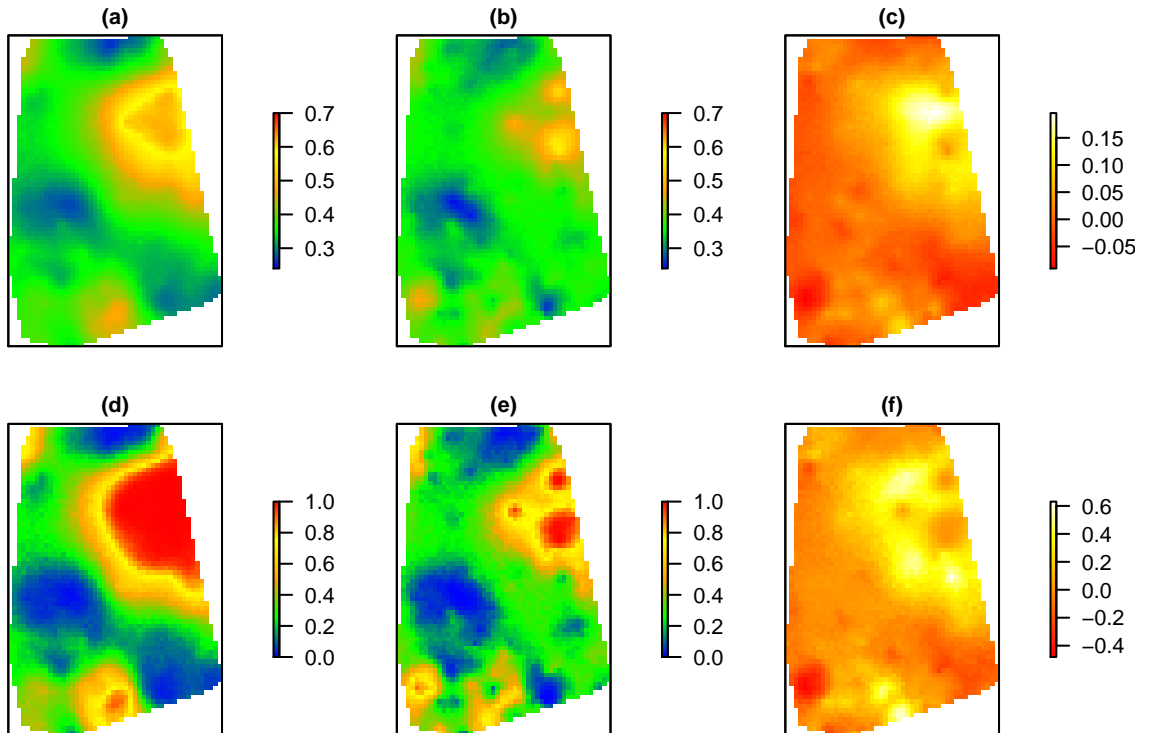
Figure 5: Predicted stunting prevalence for a 2 year old who falls in the lowest wealth index category and whose mother has poor education. Upper panels: prevalence surfaces from the binomial model (a) and the continuous models (b), and the difference between the first and the second (c). Lower panels: exceedance probabilities for a 40% prevalence threshold obtained from the binary (d) and the continuous models (e), and their difference (f).

5 Discussion

Understanding of the effects of dichotomization of continuous outcomes is especially important in medical research where cut-offs are used for diagnosis. These can be derived using different approaches: empirical approaches, where thresholds are obtained through summary

13

statistics of measurements taken from healthy individuals; clinical approaches, which utilize a specified threshold above which symptoms and complications become more frequent; prognostic approaches, where a threshold is defined based on clinical results which may be symptom-less but carry an adverse prognosis; or, finally, operational approaches, where a threshold may be based on management and/or operational guidelines [1]. In this study, we have investigated the effects of dichotomization in the context of geostatistical modelling of disease prevalence data through a simulation study and two applications, and have proposed a likelihood-based metric to quantify the potential loss informaiton arising from this practice. All of these provided evidence that dichotomization of the data can lead to substantial loss of information for both parameter estimation and spatial prediction. We found that spatial correlation may alleviate the effects of dichotomization for parameter estimation but the increase in uncertainty and bias still remained substantially larger than those of the linear model. In particular, one of the key factors that more strongly affects the loss of accuracy and precision is the distance of the threshold from the mean of the underlying process. As such distance increases, both the bias and the MSE in the estimation of the mean component and in the spatial predictions of prevalence also increase. Another important factor is the magnitude of the noise variance τ^2 relative to the signal variance σ^2. This was especially evident in the application of Section 4.1, where a τ^2 about 20 times larger than σ^2 led to the loss of the fine scale features in the spatial pattern of disease prevalence.

As shown in the application of Section 4.1, when thresholds vary across individuals, these should be accounted for in the model for binary data, an aspect that has been ignored in previous studies of anaemia mapping. Also, this could be especially problematic if some of the covariates on which the the cut-offs are based are missing (e.g. the pregnancy status of a woman). An additional problem that arises in the context of anaemia epidemiology, is that the cut-offs described in Table 4 are based on guidelines from 1992 and 2001 [2] which may be subject to amendment as scientific research or clinical practices evolve.

We have only considered the case of a Gaussian distribution for the unstructured component Z_i. Assuming a symmetric distribution for Z_i implies that, on average, misclassifications of individuals as false positives and false negatives balance out after dichotomization. However, if Z_i followed a skewed distribution, this could introduce additional bias in the geostatistical model for binary data as more individuals could be misclassified as either false positives or false negatives. Hence, we expect that under these scenarios the negative effects of dichotomization on geostatistical inference would be even stronger than those shown in this study.

It is important to note that in our study we compared the performance of binary and linear geostatistical models for cut-offs that are dependent on the scale of the continuous measurement. In other cases, the Y_i may follow a mixture distribution with a probability mass in zero. For example, malaria parasite density may exhibit this feature if a large proportion of the general population has not been infected and is thus clear of parasites. In this case dichotomization of the continuous outcome as $Y_i^* = 1$ if $Y_i > 0$ and $Y_i^* = 0$ otherwise $Y_i = 0$, would not lead to any loss of information.

A final remark relates to the computational burden of binary and linear geostatistical models. The likelihood function of the latter can be, most of the times, expressed in closed form, while the former requires numerical procedures based on analytical or Monte Carlo approximations of the likelihood function in order to be fitted. Hence, the increase in the computational burden is a further reason to avoid dichotomization of the data.

6 Conclusion

In the context of geostatistical inference, dichotomization of continuous outcomes can lead to a substantial loss of efficiency for both parameter estimation and spatial prediction. Such

loss is further compounded as cut-offs used for dichotomization are further away from the mean. In addition, dichotomization can also result in the loss of fine scale features of disease prevalence, especially in the presence of a large noise to signal ratio. The findings in this study strongly support the conclusions drawn from previous studies that, whenever feasible, dichotomization should be avoided by developing models for the continuous measurements which can then be used to estimate prevalence.

Authors' contributions

IK wrote the first draft of the manuscript. IK and EG conducted the statistical analysis. IK, EG and TAA reviewed the manuscript. TAA, MJK, and GT formed part of the Beyond Garki project team. TAA coordinated the Beyond Garki project, MJK co-designed and supervised the data collection, and GT provided technical support.

Competing interests

The authors declare that they have no competing interests

References

[1] David Coggon, David Barker, and Geoffrey Rose. *Epidemiology for the Uninitiated.* John Wiley & Sons, 2009.

[2] World Health Organisation. *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity.* Tech. rep. World Health Organisation, 2011.

[3] Giuseppe Del Priore, Peyman Zandieh, and Men-Jean Lee. "Treatment of continuous data as categoric variables in obstetrics and gynecology". In: *Obstetrics & Gynecology* 89.3 (1997), pp. 351–354.

[4] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Statistics in medicine* 25.1 (2006), pp. 127–141.

[5] Caroline Bennette and Andrew Vickers. "Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents". In: *BMC medical research methodology* 12.1 (2012), p. 21.

[6] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. "Consequences of dichotomization". In: *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 8.1 (2009), pp. 50–61.

[7] Douglas G Altman and Patrick Royston. "The cost of dichotomising continuous variables". In: *Bmj* 332.7549 (2006), p. 1080.

[8] Frank Harrell. *Problems caused by categorising continuous variables.* `http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous`. Accessed: 2019-8-8. June 2004.

[9] Petra Buettner, Claus Garbe, and Irene Guggenmoos-Holzmann. "Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma". In: *Journal of clinical epidemiology* 50.11 (1997), pp. 1201–1210.

[10] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. "On the practice of dichotomization of quantitative variables." In: *Psychological methods* 7.1 (2002), p. 19.

[11] Peter J Diggle, JA Tawn, and RA Moyeed. "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3 (1998), pp. 299–350.

[12] Peter J Diggle and Emanuele Giorgi. "Model-based geostatistics for prevalence mapping in low-resource settings". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1096–1120.

[13] Sandra Incardona, Elisa Serra-Casas, Nora Champouillon, Christian Nsanzabana, Jane Cunningham, and Iveth J González. "Global survey of malaria rapid diagnostic test (RDT) sales, procurement and lot verification practices: assessing the use of the WHO–FIND Malaria RDT Evaluation Programme (2011–2014)". In: *Malaria journal* 16.1 (2017), p. 196.

[14] Jinkou Zhao, Marcel Lama, Eline Korenromp, Patrick Aylward, Estifanos Shargie, Scott Filler, Ryuichi Komatsu, and Rifat Atun. "Adoption of rapid diagnostic tests for the diagnosis of malaria, a preliminary analysis of the Global Fund program data, 2005 to 2010". In: *PloS one* 7.8 (2012), e43549.

[15] GCM Kalla, E Voundi Voundi, R Guiadem, F Angwafor Iii, L Bélec, and F-X Mbopi-Keou. "Mass campaigns for HIV, HBV (HBsAg) and HCV screening by multiplex rapid diagnostic test in sub-Saharan Africa using mobile units: the game changer". In: *International Journal of Infectious Diseases* 79 (2019), p. 107.

[16] World Health Organization. "Consolidated Guidelines on HIV Testing Services: 5Cs: consent, confidentiality, counselling, correct results and connection 2015". In: (2015).

[17] Aaron Zimmerman, Anoushka I. Millear, Rebecca W. Stubbs, Chloe Shields, Brandon Pickering, Lucas Earl, Nicholas Graetz, Damaris K. Kinyoki, Sarah E. Ray, Samir Bhatt, et al. "Mapping child growth failure in Africa between 2000 and 2015". In: *Nature* 555 (Mar. 2018), pp. 41–47. DOI: `10.1038/nature25760`.

[18] Ricardo J Soares Magalhaes and Archie CA Clements. "Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria, and helminth infections in West Africa". In: *PLoS medicine* 8.6 (2011), e1000438.

[19] Emanuele Giorgi and Peter J Diggle. "PrevMap: an R package for prevalence mapping". In: *J Stat Softw* 78.8 (2017), pp. 1–29.

[20] Cristiano Varin, Nancy Reid, and David Firth. "An overview of composite likelihood methods". In: *Statist. Sinica* (2011), pp. 5–42.

[21] Tarekegn A Abeku, Michelle EH Helinski, Matthew J Kirby, Takele Kefyalew, Tessema Awano, Esey Batisso, Gezahegn Tesfaye, James Ssekitooleko, Sarala Nicholas, Laura Erdmanis, et al. "Monitoring changes in malaria epidemiology and effectiveness of interventions in Ethiopia and Uganda: Beyond Garki Project baseline survey". In: *Malaria journal* 14.1 (2015), p. 337.

[22] World Health Organisation. *Global nutrition targets 2025: Policy brief series*. Tech. rep. World Health Organisation, 2014.

[23] *WHO Infertility definitions and terminology.* `https://www.who.int/reproductivehealth/topics/infertility/definitions/en/`. Accessed: 2019-8-5. Oct. 2016.

[24] World Health Organization. "WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development". In: (2006).

[25] WHO Multicentre Growth Reference Study Group. "WHO Child Growth Standards based on length/height, weight and age." In: *Acta paediatrica (Oslo, Norway: 1992). Supplement* 450 (2006), p. 76.

[26] World Health Organisation. *Global nutrition targets 2025: Stunting policy brief.* Tech. rep. World Health Organisation, 2014.

[27] World Health Organization. *Nutrition Landscape Information System (NLIS) country profile indicators: interpretation guide.* Tech. rep. World Health Organisation, 2010.

A   Simulation study results for sample size m=450

Table A1: Bias and mean square error (in brackets) for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and the estimate $\hat{\phi}$ obtained from the geostatistical models fitted the binary (B) and continuous (C) outcomes. The following are results when the number of observations, $n = 450$

Parameter	τ^2	ϕ	c=0 B	c=0 C	c=0.2 B	c=0.2 C	c=0.4 B	c=0.4 C
$\tilde{\alpha}$	0.5	0.1	-0.013 (0.090)	-0.014 (0.056)	0.089 (0.216)	0.017 (0.068)	0.173 (0.180)	0.045 (0.085)
	1	0.1	-0.007 (0.06)	0.001 (0.032)	0.081 (0.325)	0.019 (0.039)	0.159 (0.054)	0.034 (0.040)
	2	0.1	-0.009 (0.031)	-0.004 (0.017)	0.073 (0.448)	0.021 (0.019)	0.120 (0.128)	0.027 (0.022)
	0.5	0.2	-0.030 (0.292)	-0.029 (0.151)	0.100 (0.382)	0.013 (0.151)	0.228 (0.417)	0.023 (0.162)
	1	0.2	-0.014 (0.169)	-0.012 (0.077)	0.093 (0.401)	0.008 (0.071)	0.175 (0.167)	0.009 (0.079)
	2	0.2	-0.006 (0.093)	-0.001 (0.040)	0.079 (0.497)	0.011 (0.041)	0.156 (0.169)	0.015 (0.040)
$\tilde{\sigma}^2$	0.5	0.1	0.603 (0.554)	0.735 (4.688)	0.624 (0.579)	0.747 (4.983)	0.570 (0.508)	0.780 (4.932)
	1	0.1	0.289 (0.137)	0.543 (3.119)	0.297 (0.136)	0.679 (3.828)	0.274 (0.122)	0.508 (3.276)
	2	0.1	0.125 (0.029)	0.417 (2.781)	0.112 (0.027)	0.462 (2.806)	0.107 (0.026)	0.446 (2.966)
	0.5	0.2	1.096 (2.059)	0.289 (1.197)	1.120 (2.086)	0.387 (1.408)	1.078 (2.015)	0.230 (1.055)
	1	0.2	0.566 (0.548)	0.131 (0.334)	0.565 (0.559)	0.133 (0.267)	0.533 (0.498)	0.105 (0.218)
	2	0.2	0.252 (0.126)	0.100 (0.519)	0.243 (0.119)	0.116 (0.471)	0.228 (0.114)	0.075 (0.157)
ϕ	0.5	0.1	0.031 (0.002)	0.001 (0.001)	0.033 (0.002)	0.002 (0.001)	0.031 (0.002)	0.001 (0.001)
	1	0.1	0.031 (0.002)	0.001 (0.002)	0.032 (0.002)	-0.002 (0.001)	0.03 (0.002)	0.001 (0.002)
	2	0.1	0.028 (0.002)	0.002 (0.002)	0.026 (0.002)	0.001 (0.002)	0.024 (0.002)	0.001 (0.003)
	0.5	0.2	0.088 (0.018)	-0.017 (0.004)	0.087 (0.017)	-0.017 (0.005)	0.088 (0.018)	-0.011 (0.005)
	1	0.2	0.095 (0.021)	-0.012 (0.007)	0.093 (0.019)	-0.014 (0.006)	0.090 (0.018)	-0.014 (0.006)
	2	0.2	0.088 (0.020)	-0.011 (0.010)	0.083 (0.018)	-0.015 (0.009)	0.078 (0.017)	-0.016 (0.009)

Table A2: Bias and mean square error (in brackets), averaged over a 1/14 by 1/14 regular grid in $[0, 2] \times [0, 1]$ (hence, $m = 450$), for the spatial predictions of prevalence obtained from the geostatistical models fitted to the binary (B) and continuous (C) outcomes.

τ^2	ϕ	c=0 B	c=0 C	c=0.2 B	c=0.2 C	c=0.4 B	c=0.4 C
0.5	0.1	0.001 (0.053)	0.001 (0.038)	0.013 (0.053)	0.001 (0.037)	0.029 (0.051)	0.001 (0.036)
1	0.1	-0.001 (0.047)	0.001 (0.037)	0.019 (0.047)	-0.001 (0.037)	0.034 (0.046)	-0.001 (0.035)
2	0.1	-0.002 (0.037)	-0.001 (0.032)	0.020 (0.037)	0.001 (0.032)	0.037 (0.036)	-0.001 (0.031)
0.5	0.2	-0.001 (0.041)	0.001 (0.029)	0.012 (0.041)	0.001 (0.029)	0.026 (0.039)	-0.001 (0.027)
1	0.2	0.001 (0.036)	0.001 (0.027)	0.017 (0.036)	-0.001 (0.027)	0.032 (0.035)	-0.001 (0.025)
2	0.200	-0.001 (0.029)	0.001 (0.022)	0.020 (0.029)	-0.001 (0.022)	0.037 (0.029)	-0.001 (0.021)