Doctoral theses at NTNU, 2023:184

Umut Altay

Geostatistical Analysis of DHS Data: Accounting for Random Displacement of Survey Locations

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of Science and Technology

Umut Altay

Geostatistical Analysis of DHS Data: Accounting for Random Displacement of Survey Locations

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-7072-7 (printed ver.)
ISBN 978-82-326-7071-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:184

Printed by NTNU Grafisk senter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The research was funded by and carried out at the Department of Mathematical Sciences during the years 2020–2023.

First of all, I am very grateful to my supervisors Geir-Arne Fuglstad and Andrea Riebler for their great guidance and support, and also for sharing their broad knowledge and expertise with me.

I would like to thank John Paige for the close cooperation, interesting discussions and his friendhip.

I am grateful for the excellent working environment and conditions provided by the Department of Mathematical Sciences, administrative and technical staff throughout my PhD.

Thanks a lot to all my friends. It has been great three years with you. Thanks for lots of fun in the office, great road trip to the north and countless nice and happy moments and memories.

Finally, I would like to thank my family for their support.

Umut Altay
Trondheim, March 2023

Contents

1 Introduction

1.1 Measuring demographic and health indicators in low- and middle-income countries

The United Nations (UN) declared 17 sustainable development goals (SDGs) in 2015 towards ending poverty and improving socio-economical well-being of the world population (UN, 2015). Figure 1 is a graphical illustration of the main target areas. Each target area further breaks down into multiple clearly defined sub-targets. The progress on each goal is closely monitored by UN in collaboration with the member countries. The progress on each goal is measured by a relevant set of indicators (UN, 2015).

This can be a very demanding task in low- and middle-income countries (LMICs) due to the complete or partial lack of registries collecting data such as the health incidences, birth and death records as well as demographic information. Additionally, various challenges such as the lack of digital infrastructure, insufficient resources and competing priorities make it hard to keep record of even the small amount of existing information (Tervonen et al., 2017). An efficient way to fill this gap is to collect more data by directly reaching out to people by using household surveys.

The Demographic and Health Surveys Program (DHS) (`https://dhsprogram.com`) contributes significantly to the data collection. DHS was established by the U.S. Agency for International Development (USAID) in 1984. Since then it has been conducting household surveys in more than 90 LMICs. The surveys achieve high response rates that are over 90 percent in most cases (ICF International, 2023). DHS surveys are the key source of data for approximately 30 different SDGs indicators (United Nations, 2023). Figure 2 shows an interview with a household member during a DHS household survey.

Countries plan and conduct administrative tasks such as political decisions and budget allocations both on the national and subnational levels. Accordingly, obtaining estimates representing both the national and sub-national areas is important. Although the household surveys are helpful for obtaining national and sub-national estimates of the demographic and health (DH) indicators such as the vaccination rate (Local Burden of Disease Vaccine Coverage Collaborators, 2021) and neonatal and under five mortality (Wakefield et al., 2019; Golding et al., 2017), the data is still extremely sparse for in depth analysis in high resolution (e.g. 5 by 5 kilometers) across the country. Sparsity of the observations necessitates borrowing strength during estimation by incorporating data from neighbouring areas.

Figure 1: The main target areas for achieving the SDGs declared by the UN. (United Nations, 2023)

In this thesis we use Bayesian geostatistical models as described in Sections 1.3.1 and 1.3.2. Attention to the spatial dependency structure allows incorporating small scale variability into the modelling. Data can be flexibly modelled via different likelihood models such as binomial, Poisson, Gaussian, and the interactions between the geographical covariates can be captured succesfully even in the very fine scale.

The research in this thesis uses data from 2014 Kenya (KDHS2014) and 2018 Nigeria (NDHS2018) DHS surveys to analyse the proportion of contraceptive use among women aged 15-49 in Kenya and prevalence of completion of secondary education among 20-39 year old women in Nigeria, respectively. High resolution estimation and prediction through the use of the developed approach in this thesis contributes directly for monitoring the current states of SDG sub-targets 3.7 and 4.1 and constructing future projections for them. SDG sub-target 3.7 aims to ensure universal access to sexual and reproductive health-care services, including for family planning, information and education, and the integration of reproductive health into national strategies and programmes, by 2030 (UN,

Figure 2: An interview from a DHS household survey (The Demographic and Health Surveys Program, 2023).

2015). SDG sub-target 4.1 focuses on ensuring that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes, by 2030 (UN, 2015).

1.2 DHS data collection and GPS coordinate displacement

DHS Surveys are commonly conducted every five years based on a stratified two-stage cluster design. Each DHS household cluster center is located within an urban or rural part of a corresponding first level administrative area. The survey design is stratified with respect to the administrative regions, crossed by the urbanization classification (ICF International, 2012). The first stage is the random selection of a predetermined number of household clusters within each stratum. The random selection of these enumeration areas are implemented with probabilities proportional to the number of households in the corresponding clusters. Censuses that are conducted at most every ten years are the basis for the sampling design. Censuses are rare in time and do not provide detailed information. The second stage is the random selection of a set of households from the corresponding cluster (ICF International, 2012). Each cluster is geographically referenced by the GPS coordinates of the household cluster centroids.

DHS data sets are semi-public. Access needs to be requested through an

easy application process. Detailed information about the application procedure can be found in `https://dhsprogram.com/data/new-user-registration.cfm`. The application requires explaining why the data is needed and how it will be processed, within context of a brief research plan. The survey data sets and the corresponding GPS data are seperate data sources. The survey data alone can be used to obtain national and subnational level survey statistics, but it does not provide finer scale information. If the aim is to conduct high resolution analyses by geostatistical modelling, as it is in this thesis, then having access to the GPS coordinates is crucial.

DHS data is published at the cluster level in order to make individual responses indistinguishable. The GPS coordinates of the cluster centers are randomly displaced to further protect the privacy of the respondents. This is implemented by shifting the true location towards a random angle and up to a known maximum distance. Such local displacements are referred to as jittering throughout the thesis. The maximum jittering distances depend on the stratum that the corresponding cluster center is located within. The default DHS jittering distance for the locations that are within an urban stratum is 2 kilometers. On the other hand, 99% of the rural locations are displaced up to 5 kilometers and the remaining 1% is displaced up to 10 kilometers (Burgert et al., 2013). The jittering has a potential to create issues that should be both methodologically and computationally dealt with.

1.3 Geostatistical analysis for demographic and health indicators

1.3.1 Standard geostatistical model

DHS data can be modelled by a standard geostatistical model containing an intercept, a set of spatial covariates and a spatial random effect, as long as the positional error in the cluster center coordinates are not taken into account. A typical DHS survey contains data from a set of small household groups (clusters) indexed by cluster IDs $C = 1, ..., c$, and referenced by the jittered GPS coordinates of the cluster centers $\boldsymbol{s}_c$, across the corresponding spatial domain (country) $\boldsymbol{s}_c \in \mathcal{D}$. Accordingly, the geostatistical model takes the form:

$$y_c \mid \eta(\boldsymbol{s}_c), \boldsymbol{\phi} \sim \pi(y_c \mid \eta(\boldsymbol{s}_c), \boldsymbol{\phi})$$
$$\eta(\boldsymbol{s}_c) = \boldsymbol{x}(\boldsymbol{s}_c)^{\mathrm{T}}\boldsymbol{\beta} + u(\boldsymbol{s}_c)$$

where y_c is the value of the response variable at cluster c, $\eta(\boldsymbol{s}_c)$ is the linear predictor at the jittered location $\boldsymbol{s}_c$ of cluster c, $\boldsymbol{\phi}$ is a vector of parameters

belonging to the corresponding likelihood family $\pi(y_c \mid \eta(\boldsymbol{s}_c), \boldsymbol{\phi})$. The vector $\boldsymbol{\beta}$ contains the covariate effect sizes, and $\boldsymbol{x}$ is a spatially varying vector of covariates.

The spatial random effect $u(\cdot)$ is a Gaussian random field (GRF) with a mean function $\mu = \mu(\cdot)$ and a Matérn covariance function:

$$\mathrm{C}_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2; \sigma_{\mathrm{S}}^2, \rho_{\mathrm{S}}) = \sigma_{\mathrm{S}}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu} \frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\rho_{\mathrm{s}}} \right)^{\nu} \mathrm{K}_\nu \left(\sqrt{8\nu} \frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\rho_{\mathrm{s}}} \right)$$

$\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{D}$, where, σ_{S}^2 is the marginal variance, and ρ_{S} is the spatial range, with a smoothness parameter that is fixed to $\nu = 1$. The smoothness parameter is fixed to 1 for computational simplicity.

1.3.2 Bayesian setup

We include prior information via Bayesian approach to stabilize the inference and allow a flexible framework.

The models in this thesis contain the spatial random effect as a GRF. Having a GRF in the model with Matérn covariance function can cause a ridge in the likelihood for the range and the marginal variance, which might reflect itself as overfitting by estimating spurious spatial trends (Fuglstad et al., 2019). Simpson et al. (2017) suggests overcoming the overfitting problem when having GRF in the model, by assigning priors that shrink the components such as GRF, towards their base models. Therefore, as proposed by Fuglstad et al. (2019), we use penalised complexity (PC) priors on the parameters of the Matérn GRF, spatial range (ρ_{S}) and the marginal variance (σ_{S}^2). The PC-priors are formulated via the hyperparameters (*a priori* median of range) and σ_0 (*a priori* marginal variance) as $\mathrm{P}(\sigma_{\mathrm{S}} > 1) = 0.05$ and $\mathrm{P}(\rho > \rho_0) = 0.50$, respectively. PC priors penalise complexity, the distance from the base model, by shrinking the range towards infinity and the marginal variance towards zero (Fuglstad et al., 2019).

The unobserved true locations s_c^* are assigned a uniform prior $s_c^* \sim \mathcal{U}(\mathcal{D})$. This means that a true location can be anywhere within the jittering radius of the corresponding DHS cluster center. Using urbanicity was not preferrable since urbanicity was not very reliable. The DHS data includes urbanicity information for the corresponding clusters, but urbanicity was not available across the countries as high resolution raster maps (e.g. in 5 by 5 kilometers resolution). Using more informative priors may also cause issues. Choice of priors for the unobserved true locations might be an interesting potential topic for the future research.

The covariate effect sizes in the model are assigned uninformative Gaussian

prior ($\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 25\mathbf{I}_p)$).

1.3.3 Recent work in the field

There is much interest in mapping DHS indicators on a subnational level, see for example Gething et al. (2016), Bosco et al. (2017), Steele et al. (2017), Davey and Deribe (2017), Golding et al. (2017), Osgood-Zimmerman et al. (2018), Utazi et al. (2018), Graetz et al. (2018), Utazi et al. (2019b), Ganyani et al. (2019), Utazi et al. (2019a), Mosser et al. (2019), Leasure et al. (2020), Wakefield et al. (2020), Utazi and Tatem (2021), Local Burden of Disease Vaccine Coverage Collaborators (2021), Nguyen et al. (2022), Jasper et al. (2022), Woods et al. (2022), Wilson and Wakefield (2022), Utazi et al. (2022), Utazi et al. (2023). Following are two examples for such studies in more detail.

Golding et al. (2017) combined DHS household surveys with various other data sources to estimate neonatal and under five mortality, which stand for the probability of death before the age of five, and within the first month of life, per 1000 livebirths, respectively. They implemented Bayesian geostatistical models to obtain estimates of the two indicators in 5 by 5 kilometers resolution, across 46 countries in Africa.

Local Burden of Disease Vaccine Coverage Collaborators (2021) obtained individual MCV1 vaccination status data across 101 low- and middle-income countries by compiling 354 household surveys including the DHS surveys from the Institute for Health Metrics and Evaluation (IHME) Global Health Data Exchange (GHDx) database (`http://ghdx.healthdata.org`). They fitted a geostatistical model with correlated errors across space and time to estimate routine childhood MCV1 coverage at 5 by 5 kilometers resolution.

The geographical references of DHS household cluster centers are published as a set of randomly displaced (jittered) GPS coordinates of the true household cluster centroids. The positional error introduced by jittering has a great potential to cause the geostatistical analyses to yield adversely affected results, such as attenuation in covariate effect size estimates and poor predictive performance (Altay et al., 2022a,b). Golding et al. (2017) and Local Burden of Disease Vaccine Coverage Collaborators (2021) take the survey design into consideration in terms of the urbanicity strata, but they do not adress the positional error emerging from the jittering algorithm of DHS.

1.4 Handling positional uncertainty

1.4.1 State of the Art

The DHS jittering algorithm displaces the true cluster centers ($\boldsymbol{s}_c^*$) towards a random direction and up to a certain maximum jittering distance. The direction is decided by a random angle chosen from an interval between 0 and 2π. The maximum distance on the other hand, is decided by the stratum that the corresponding location is located within. Both the displacement angle and the distance is chosen from a uniform distribution with the corresponding lower and upper bounds. The DHS random displacement algorithm does not allow the jittered location to land in a different administrative area (A) than the one that it was initially located within. The country of interest is divided into K administrative regions. Let $A(\boldsymbol{s}) \in \{1, \ldots, K\}$ denote the administrative region of location $\boldsymbol{s}$ for $\boldsymbol{s} \in \mathcal{D}$. Then for an urban cluster c, which can be jittered up to 2 km, the jittering distribution is

$$\pi_{\mathrm{U}}(\boldsymbol{s}_c|\boldsymbol{s}_c^*) \propto \frac{\mathbb{I}(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*)) \cdot \mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 2)}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)}, \quad \boldsymbol{s}_c \in \mathcal{D},$$

where $d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)$ is the distance in kilometers between $\boldsymbol{s}_c$ and $\boldsymbol{s}_c^*$, and $\mathbb{I}$ is the indicator function. Similarly, for a rural cluster c, which can be jittered up to 5 km except for the 1 percent of clusters jittered up to 10 km, the jittering distribution is:

$$\pi_{\mathrm{R}}(\boldsymbol{s}_c|\boldsymbol{s}_c^*) \propto \frac{\mathbb{I}(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*))}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)} \left[\frac{99\mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 5)}{100} + \frac{\mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 10)}{100} \right], \boldsymbol{s}_c \in \mathcal{D}.$$

Fitting the geostatistical model based on the locations provided by DHS has a potential to adversely affect the two main components of the model, namely, the covariates and the spatial random effect. A design matrix with the covariate values that are extracted at the jittered locations might lead failure in capturing the true interactions between the raster-and distance-based covariates by causing a non standard form of measurement error to emerge (Gustafson, 2003). The geographical rasters usually have high resolution that reflects variability in the data within distances that are way smaller than the default maximum jittering distances of DHS. Ignoring jittering translates into ignoring this variation which then might cause the estimated covariate effect sizes to attenuate towards zero (Altay et al., 2022b).

Although there have been some methodological developments towards addressing the positional uncertainty issue in geostatistical analysis, in practice it is still common to ignore jittering because the few existing methods are either

computationally demanding, or not flexible enough to accomodate different types of likelihoods.

Fanshawe and Diggle (2011) presents an approach to reduce the effects of positional error in geostatistical model estimation and prediction. They approach the problem from the measurement error perspective and investigate uncertainty in the coordinates of data and prediction locations, emerging due to inaccurate measurements or storage of GPS coordinates. They use a standard geostatistical model which contains spatial covariates, a GRF and mutually independent Gaussian errors with zero mean and nugget variance. They break down the nugget variance into two independent processes with respect to two sources of variation, namely the small scale spatial variation and the measurement error. Then the model allows accomodating the measurement error as a bivariate normal distributed positional error. They obtain parameter estimates that are closer to the true values than the common way of not accounting for the error. On the other hand, the method focuses on accounting for the positional uncertainty only in the GRF and uses a Gaussian likelihood, which does not allow estimating prevalences. They also found that the computations were slow especially when the positional error is in the data locations, but later it became feasible, with composite likelihoods by Fronterrè et al. (2018).

Perez-Heydrich et al. (2013) proposes an approach that is based on creating buffer zones around DHS locations with respect to the maximum jittering distances of the corresponding urbanization strata. The method then uses the covariate values that are averaged over the buffer zone. Perez-Heydrich et al. (2016) explores this further and finds that averaging the covariates over a 5km buffer zone around DHS cluster centers can moderate the adverse effects of jittering for continuous and categorical rasters. The approach does not address the attenuation in the covariate effect sizes.

Wilson and Wakefield (2021) suggested using Integrated Nested Laplace Approximations (INLA) (Rue et al., 2009) within Markov chain Monte Carlo method (MCMC) to account for the positional uncertainty of DHS cluster centers in hierarchical geostatistical models within Bayesian framework. INLA focuses on approximate Bayesian inference on the models with latent Gaussian Markov random field (GMRF), and the corresponding R-package that provides the functions needed to implement the approach. The proposed mixed model is constructed based on the unknown true locations. The issue that the true cluster centers can be located anywhere within the jittering radius of the observed DHS locations with equal probability prevents obtaining a fixed set of coordinates. This breaks the underlying assumption that each observation depends on one linear predictor. Accordingly, the model is not suitable for fitting with INLA. The motivation behind using INLA within MCMC method in this context is to be able to obtain

a fixed set of locations by sampling the true locations via MCMC algorithm first, and then conditioned on them, fitting the model with INLA, iteratively. The method allows accounting for jittering in both GRF and the spatial covariates, but the long computation time makes it challenging to apply in practice.

1.4.2 A novel, flexible and fast approach

This thesis presents a fast and flexible computational method that allows adjusting for jittering in the GPS coordinates of DHS household cluster centers. The method allows accounting for jittering either in the spatial random effect, or in the covariates, or both. The approach allows fitting geostatistical hierarchical models with Gaussian, binomial and Poisson likelihoods, within Bayesian framework. The adjusted geostatistical model can be constructed as follows:

$$\begin{aligned} y_c \mid \eta(\boldsymbol{s}_c), \boldsymbol{\phi} \sim \pi(y_c \mid \eta(\boldsymbol{s}_c), \boldsymbol{\phi}), \quad \boldsymbol{s}_c | \boldsymbol{s}_c^* \sim \pi(\boldsymbol{s}_c | \boldsymbol{s}_c^*), \\ \eta(\boldsymbol{s}^*) = \boldsymbol{x}(\boldsymbol{s}_c^*)^{\mathrm{T}} \boldsymbol{\beta} + u(\boldsymbol{s}_c^*), \quad \boldsymbol{s}^* \in \mathcal{D}. \end{aligned}$$

This setting allows approaching the unknown true locations as nuisance parameters and integrating them out of the joint likelihood function:

$$\begin{aligned} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot)) &= \int_{\mathbb{R}^2} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot), \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^* \\ &= \int_{\mathbb{R}^2} \pi(y_c | \eta(\boldsymbol{s}_c^*)) \pi(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^*, \end{aligned} \tag{1.1}$$

for $c = 1, \ldots, C$. The model in its new form provides the framework to adress the positional uncertainty issue in the observed coordinates and allows constructing the modelling hierarchy under Gaussian, binomial and Poisson likelihoods. The jittering distribution is known, which means that a true location ($\boldsymbol{s}_c^*$) can be located anywhere within the jittering radius of the corresponding observed location ($\boldsymbol{s}_c$). The method then constructs an integration grid in polar coordinates around each DHS cluster center and assigns weights to each point on the grid. The assigned weights are equal within the five kilometer radius of DHS cluster centers, and equal but lower outside it.

In the case that an integration point is less distant to the administrative area border than the maximum jittering distance of the corresponding urbanization strata, an additional set of sub-integration points are created around it and the ones that land across the border are assigned zero weight. Figure 3 shows the primary and secondary integration points and the corresponding integration weights, for a single cluster from the Kenya 2014 DHS household survey.

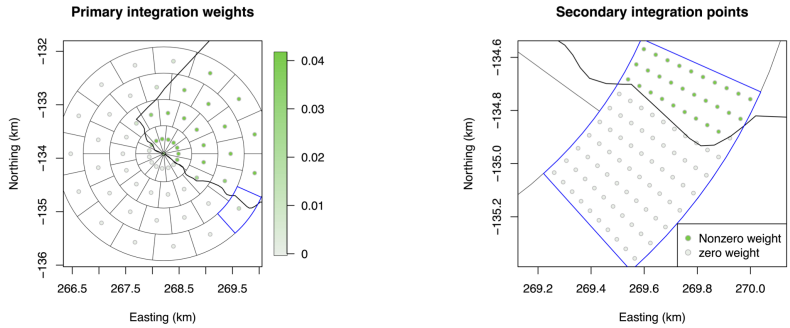


Figure 3: Illustration of primary (left) and secondary (right) integration weights for one cluster from Kenya 2014 DHS household survey.

1.5 Computationally efficient implementation

The method in this thesis requires integrating out the unknown true locations from the joint likelihood function across an integration grid via Equation (1.1). Accordingly, implementation of the developed approach relies heavily on making sure that the computations can be implemented with as low computational cost as possible. This was achieved by combining the computational benefits of the sparsity that is provided by the SPDE approach for the high spatial dimension arising from the integration grids, and the Laplace approximation and auto-differentiation feature of the Template Model Builder (TMB). Using TMB also provides flexibility in model construction by its capability of fast computation with non-Gaussian likelihoods.

1.5.1 Handling the dimensionality of the problem

A Gaussian random field with Matérn covariance is a solution to the following linear SPDE

$$(\kappa^2 - \Delta)(\tau u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \tilde{\mathcal{D}},$$

where $\kappa > 0$ and $\tau > 0$ are related to marginal variance and range, Δ is the Laplacian, $\mathcal{W}(\cdot)$ is standard Gaussian white noise, and $\tilde{\mathcal{D}} \supset \mathcal{D}$ is an extended domain to reduce boundary effects. The effective range and marginal variance are calculated from the SPDE parameters as

$$\rho_{\mathrm{S}} = \frac{\sqrt{8}}{\kappa} \quad \text{and} \quad \sigma_{\mathrm{S}}^2 = \frac{1}{4\pi\tau^2\kappa^2}.$$

Accordingly, the continuous GRF over two dimensional domain $\tilde{\mathcal{D}} \in \mathcal{R}^2$ can be approximated by the SPDE approach:

$$u(\boldsymbol{s}) = \sum_{i=1}^{m} w_i \phi_i(s), \tag{1.2}$$

where $\phi_i(\cdot)$ are pyramidal basis functions and $\boldsymbol{w} = (w_1 \ \ldots \ w_m)^{\mathrm{T}}$ are the Gaussian weights $\boldsymbol{w} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1})$ with a sparse precision matrix. Due the sparsity, the computational cost reduces to the square root of the initial computational cost, which was cubic in dimensions m. This framework makes it efficient to evaluate the continuous GRF via discretization of the continuous surface by the Delaunay triangulation mesh. In this setting, each observation point (DHS cluster center) is located on one of the non-intersecting mesh triangles.

An example two dimensional setting is illustrated in Figure 4. The upper left section of the figure shows how the values of the piece-wise linear basis functions are assigned based on a single observation point within one of the mesh triangles. If the point is located inside the triangle, the basis function is evaluated at all three vertices (nodes). Each value on the nodes is calculated as the ratio of the area of the smaller triangle that is on the opposite side, to the total area of the triangle. If the point is located on one of the three nodes, then the value of the basis function becomes one at that node and zero elsewhere. If it is located along one of the edges, the basis function values on the nodes at the two ends of the edge are calculated as the relative distances between the point and the two edges. The value is zero on the third node.

The upper right section of Figure 4 shows the pyramidal formation of the piece-wise linear basis functions. The lower left and right sections of Figure 4 show a continuous surface and its approximation by the pyramidal piece-wise basis functions, respectively (Krainski et al., 2018).

The evaluated values of basis functions form a projection matrix where each basis function is represented by one column and each observation point (cluster center) is represented by one row. Accordingly, each row contains three non-zero values. When the cluster center is located on either one of the vertices or one of the edges, the corresponding row of the projector matrix has only one value that is equal to one, or has two non-zero values, respectively (Krainski et al., 2018). Multiplying the projector matrix by the Gaussian distributed weights, the SPDE approach yields an approximation of the spatial random effect as a zero mean Gaussian distribution with a sparse precision matrix. The sparsity supports fast evaluation of GRF at any number of locations.

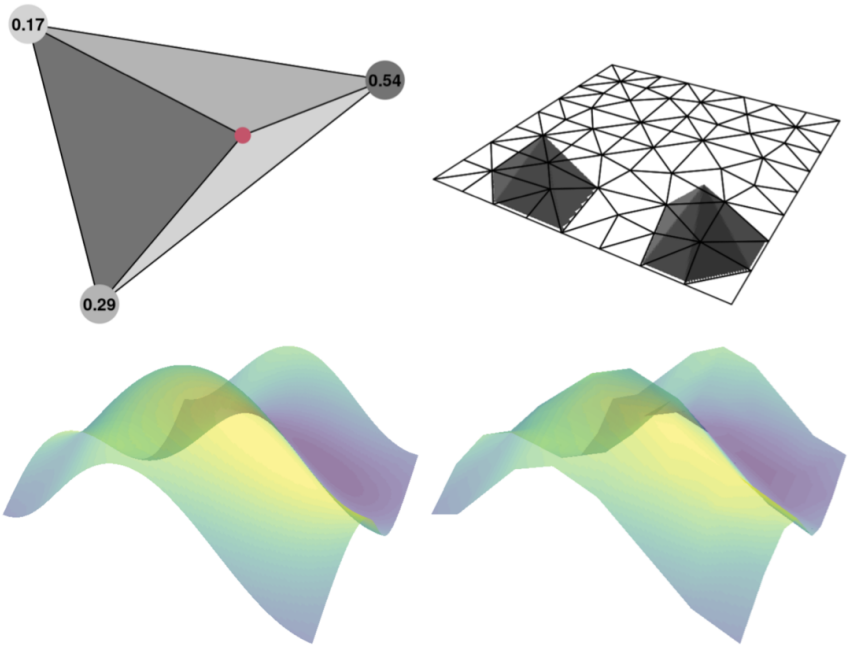


Figure 4: Approximation of a continuous surface by a set of piece-wise linear pyramidal basis functions (Krainski et al., 2018)

1.5.2 Allowing for flexible observation model

The Template Model Builder (TMB) provides fast computation by combining the auto-differentiation with the Laplace approximation (Kristensen et al., 2016). Using TMB requires the objective function, in other words the likelihood $\pi(y_c \mid \eta_c, \boldsymbol{\phi})$ of the hierarchical model, together with the whole Bayesian hierarchical structure to be defined within a C++ script. TMB approximates the likelihood in the model using the Laplace approximation constructed based on the second order Taylor series expansion (Skaug and Fournier, 2006). The TMB function "MakeADFun" can be run within R and it uses the chain rule of calculus to pre-compute the first and second order derivatives (Kristensen et al., 2016; Skaug and Fournier, 2006). The function returns the derivatives together with the objective function within a list called the core model object. Having the derivatives pre-computed provides a great computational advantage while maximizing the Laplace-approximated likelihood. The presented methodology in this thesis effectively combines all benefits of SPDE approach with the Laplace approximation

and the autodifferentiation of TMB and achieves fast computations. TMB is accessible within R via the TMB R-package.

1.6 Summary of the Papers

Paper I Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022). Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data. arXiv preprint arXiv:2202.11035v2.
In preparation for resubmission.

Paper II Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022). Jittering impacts raster- and distance-based geostatistical analyses of DHS data. arXiv preprint arXiv:2202.07442v1.
Submitted.

Paper III Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. GeoAdjust: Adjusting for positional uncertainty due to anonymisation in geostatistial analysis of DHS data.
In preparation for submission.

1.6.1 Paper I: "Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data"

This paper introduces a novel and computationally efficient approach for geostatistical analysis of DHS household surveys with positional uncertainty in the GPS coordinates of cluster centers. The method accounts for jittering in the spatial random effect and it is suitable for geostatistical analysis of Bayesian hierarchical models with binomial, Gaussian, and Poisson likelihoods. The paper provides the technical details about the implementation and includes a simulation study with binomial and Gaussian observation models, where the results of accounting and not accounting for jittering are compared under various scenarios based on different spatial range values and jittering scales. The results showed that the developed method provided more accurate parameter estimates and better predictive performance compared to the analysis where jittering is not accounted for. The improvements in the estimation and prediction got better with the increased jittering scale.

1.6.2 Paper II: "Jittering impacts raster- and distance-based geostatistical analyses of DHS data"

This paper extends the first paper, by including raster- and distance-based covariates. The improvements in the method allows accounting for jittering either in the spatial random effect, in the covariates, or in both. The simulation study analyses the prevalence of completion of secondary education among 20-39 year old women in Nigeria, based on DHS 2018 Nigeria household survey and compares the results of not adjusting for jittering, adjusting only in the covariates, and in both the covariates and the spatial random effect, under different scenarios. The results are evaluated in terms of the parameter estimation and prediction performances. The analyses in the paper showed that the estimates of the covariate effect sizes had a tendency to attenuate towards zero, when jittering was not accounted for. Accounting for jittering prevents this problem and improves the predictive performance. We also found that the improvement due to accounting for jittering largely comes from accounting for jittering in the covariates.

1.6.3 Paper III: "GeoAdjust: Adjusting for positional uncertainty in geostatistial analysis of DHS data"

The R-package GeoAdjust was constructed in order to make our methodology easily accessible for the scientific community. The package is on CRAN. The paper is targeted towards applied scientists who want to analyse DHS data in a geostatistical setting. It provides detailed step by step explanations about the package functionality and illustrates the analysis of a typical DHS dataset incorporating positional uncertainty. The results in the first and second papers can easily be reproduced using the package without complicated code. GeoAdjust offers a nice plotting function as well. To my knowledge, this is the first and only R package for flexibly analysing DHS data by accounting for the positional uncertainty in the spatial field, as well as the raster- and distance-based covariates.

References

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022a). Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data. *arXiv preprint arXiv:2202.11035v2*.

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022b). Jittering impacts

raster- and distance-based geostatistical analyses of DHS data. *arXiv preprint arXiv:2202.07442v1*.

Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., Steele, J., Hornby, G., Ruktanonchai, C., Ruktanonchai, N., et al. (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface*, 14(129):20160825.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). Geographic displacement procedure and georeferenced datarelease policy for the Demographic and Health Surveys. `https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf`. DHS Spatial Analysis Reports No. 7.

Davey, G. and Deribe, K. (2017). Precision public health: mapping child mortality in Africa. *The Lancet*, 390(10108):2126–2128.

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

Fronterrè, C., Giorgi, E., and Diggle, P. (2018). Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *Spatial Statistics*, 28:319–330.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114:445–452.

Ganyani, T., Roosa, K., Faes, C., Hens, N., and Chowell, G. (2019). Assessing the relationship between epidemic growth scaling and epidemic size: The 2014–16 ebola epidemic in west Africa. *Epidemiology & Infection*, 147.

Gething, P. W., Casey, D. C., Weiss, D. J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K. E., Dalrymple, U., Rozier, J., Rao, P. C., et al. (2016). Mapping plasmodium falciparum mortality in Africa between 1990 and 2015. *New England Journal of Medicine*, 375(25):2435–2445.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., et al. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the sustainable development goals. *The Lancet*, 390(10108):2171–2182.

Graetz, N., Friedman, J., Osgood-Zimmerman, A., Burstein, R., Biehl, M. H., Shields, C., Mosser, J. F., Casey, D. C., Deshpande, A., Earl, L., et al. (2018). Mapping local variation in educational attainment across Africa. *Nature*, 555(7694):48–53.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

ICF International (2012). Demographic and health survey sampling and household listing manual. `https://dhsprogram.com/publications/publication-dhsm4-dhs-questionnaires-and-manuals.cfm/`, last accessed on 2023-03-07.

ICF International (2023). Kenya demographic and health survey 2022 key indicators report. `https://dhsprogram.com/pubs/pdf/PR143/PR143.pdf/`, last accessed on 2023-03-07.

Jasper, P., Jochem, W. C., Lambert-Porter, E., Naeem, U., and Utazi, C. E. (2022). Mapping the prevalence of severe acute malnutrition in Papua, Indonesia by using geostatistical models. *BMC Nutrition*, 8(1):1–10.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC. `https://becarioprecario.bitbucket.io/spde-gitbook/`, last accessed on 2023-03-20.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., and Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences*, 117(39):24173–24179.

Local Burden of Disease Vaccine Coverage Collaborators (2021). Mapping routine measles vaccination in low-and middle-income countries. *Nature*, 589(7842):415–419.

Mosser, J. F., Gagne-Maynard, W., Rao, P. C., Osgood-Zimmerman, A., Fullman, N., Graetz, N., Burstein, R., Updike, R. L., Liu, P. Y., Ray, S. E., et al. (2019). Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: a spatial and temporal modelling study. *The Lancet*, 393(10183):1843–1855.

Nguyen, T. H. T., Faes, C., and Hens, N. (2022). Measles epidemic in southern vietnam: an age-stratified spatio-temporal model for infectious disease counts. *Epidemiology & Infection*, 150:e169.

Osgood-Zimmerman, A., Millear, A. I., Stubbs, R. W., Shields, C., Pickering, B. V., Earl, L., Graetz, N., Kinyoki, D. K., Ray, S. E., Bhatt, S., et al. (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature*, 555(7694):41–47.

Perez-Heydrich, C., Warren, J., Burgert, C., and Emch, M. (2013). Guidelines on the use of DHS GPS data. *ICF International, Calverton, Maryland*. `https://dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf/`, last accessed on 2023-03-20.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., and Emch, M. E. (2016). Influence of Demographic and Health Survey point displacements on raster-based analyses. *Spatial Demography*, 4(2):135–153.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Sstatistical Methodology)*, 71(2):319–392.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28.

Skaug, H. J. and Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51(2):699–709.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., De Montjoye, Y.-A., Iqbal, A. M., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127).

Tervonen, H., Bray, F., Foliaki, S., and Roder, D. (2017). Cancer registration challenges in low-and middle-income countries—the case of the Pacific islands. *European Journal of Cancer Care*, 26(1).

The Demographic and Health Surveys Program (2023). `https://www.usaid.gov/global-health/demographic-and-health-surveys-program/`, last accessed on 2023-03-12.

UN (2015). Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*.

United Nations (2023). The sustainable development goals. `https://www.un.org/en/sustainable-development-goals/`, last accessed on 2023-03-07.

Utazi, C., Thorley, J., Alegana, V., Ferrari, M., Nilsen, K., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. (2019a). A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. *Statistical Methods in Medical Research*, 28(10-11):3226–3241.

Utazi, C. E., Aheto, J. M., Wigley, A., Tejedor-Garavito, N., Bonnie, A., Nnanatu, C. C., Wagai, J., Williams, C., Setayesh, H., Tatem, A. J., et al. (2023). Mapping the distribution of zero-dose children to assess the performance of vaccine delivery strategies and their relationships with measles incidence in nigeria. *Vaccine*, 41(1):170–181.

Utazi, C. E., Aheto, J. M. K., Chan, H. M. T., Tatem, A. J., and Sahu, S. K. (2022). Conditional probability and ratio-based approaches for mapping the coverage of multi-dose vaccines. *Statistics in Medicine*.

Utazi, C. E. and Tatem, A. J. (2021). Precise mapping reveals gaps in global measles vaccination coverage. *Nature*, 589:354–355.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., Cutts, F. T., and Tatem, A. J. (2019b). Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nature communications*, 10(1):1–10.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, 36(12):1583–1591.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 28(9):2614–2634.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37.

Wilson, K. and Wakefield, J. (2022). A probabilistic model for analyzing summary birth history data. *Demographic Research*, 47:291–344.

Woods, D., Cunningham, A., Utazi, C., Bondarenko, M., Shengjie, L., Rogers, G., Koper, P., Ruktanonchai, C., zu Erbach-Schoenberg, E., Tatem, A., et al.

(2022). Exploring methods for mapping seasonal population changes using mobile phone data. *Humanities and Social Sciences Communications*, 9(1):1–17.

Paper I

Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A.

In preparation

Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data

Umut Altay John Paige Andrea Riebler
Geir-Arne Fuglstad

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

Household survey data from the Demographic and Health Surveys (DHS) Program is published with GPS coordinates. However, almost all geostatistical analyses of such data ignore that the published GPS coordinates are randomly displaced (jittered). In this short report, we develop a geostatistical model that accounts for the positional uncertainty when analysing DHS surveys, and provide a fast implementation using Template Model Builder. The key focus is inference with Gaussian random fields under positional uncertainty, and our approach works for both Gaussian and non-Gaussian likelihoods. A simulation study with a binomial observation model shows that the new approach performs equally or better than the common approach of ignoring jittering, both in terms of more accurate parameter estimates and improved predictive measures. We demonstrate that the improvement would be larger under stronger jittering. An analysis of contraceptive use in Kenya shows that the approach is fast and easy to use in practice.

Geospatial analysis, Positional error, Low and Middle Income Countries, Global Health, Household Survey, Template Model Builder

1 Introduction

Demographic and health indicators are important for monitoring and evaluating progress towards achieving the United Nations' (UN's) Sustainable Development Goals (SDGs) (General Assembly of the United Nations, 2015). The DHS program has collected over 400 surveys in over 90 countries, and surveys are conducted approximately every fifth year in participating countries. DHS surveys primarily use a two-stage cluster sampling design. Georeferenced data are only available based on special permission and a known geographical displacement process is applied before releasing the GPS coordinates of the clusters. Here, DHS aims to balance the risk of disclosure of the respondents while simultaneously preserving useful information for spatial analyses (Burgert et al., 2013). Urban clusters are displaced up to 2 km, while 99% of the rural clusters are displaced up to 5 km, and the remaining 1% up to 10 km. Rural clusters are jittered more to keep the same level of disclosure risk as for urban clusters (VanWey et al., 2005). This approach can be criticized as the actual risk of disclosure is unclear. Alternative procedures exist, which include, for example, location swapping (Zhang et al., 2017), space transformations (Khoshgozaran and Shahabi, 2007) and k-anonymity (Sweeney, 2002). In the field of cyber security so-called strong protection techniques are proposed, see for example (Gahi et al., 2016).

This short report does not assess the quality of the underlying displacement process used by DHS, but proposes a novel and fast geostatistical inference approach to analyse DHS data in the presence of positional uncertainty. In a linear geostatistical model with a Gaussian likelihood, a simple approach to adjust for positional error with a known displacement distribution is to adjust the covariances between the observed locations, and assume that after marginalising out the unknown true locations, the joint distribution is still a Gaussian distribution (Cressie and Kornak, 2003). However, such approaches do not easily generalize to generalized linear models. Fanshawe and Diggle (2011) describe how to account for positional uncertainty in a hierarchical geostatistical model fitted through maximum likelihood estimation for the parameters, but found computational times to be prohibitively slow. Later work demonstrates that inference can be made faster through a composite likelihood approach in the case of a linear geostatistical model with a Gaussian likelihood (Fronterrè et al., 2018).

Recently, Wilson and Wakefield (2021) proposed a Bayesian approach for generalized linear geostatistical models in the context of DHS surveys. Each iteration in their method is composed of two parts. First, Markov chain Monte Carlo (MCMC) is used to sample the true locations, then the Integrated Nested Laplace Approximations (INLA) method (Rue et al., 2009) is applied for inference conditional on the true locations. This gives an INLA within MCMC approach

(Gómez-Rubio and Rue, 2018), which for each simulation scenario took around 52 hours to run 1,000 iterations on 398 locations.

Warren et al. (2016a) proposed a "regression calibration (RC)" method for distance-based analyses that accounts for jittering of DHS clusters by trying to estimate the true distance covariates. They found that the proposed method outperformed the naive method in almost all location and spatial density settings. In another study, Warren et al. (2016b) adressed the issue of incorrectly assigning areas to the DHS clusters, when clusters are jittered out of the corresponding true polygons. They proposed a maximum probability covariate (MPC) selection method which allows selecting the most probable covariates. They recommend using MPC to maximize the selection probability of the correct covariates. As a different approach, Gething et al. (2013) considered the impact of jittering of DHS clusters from the perspective of spatial interpolation surfaces. They proposed a geostatistical framework for creating interpolated surfaces based on DHS data.

In this short report, we present a novel approach to fit generalized linear geostatistical models that accounts for positional uncertainty in the provided GPS coordinates of the data locations. The key focus is to address the issue of inference under positional uncertantity when modelling spatial variation using Gaussian random fields (GRFs). Computation time is a key concern and the method needs to be accessible to analysts without requiring them to write complex code. We use a quadrature to integrate out the unknown true locations so that the likelihood of the observation conditional on the latent model is a mixture distribution. The random effects are then integrated out using the Laplace approximation and automatic differentiation with Template Model Builder `TMB`, which supports complex, nonlinear latent models with non-Gaussian responses (Kristensen et al., 2016). Computationally efficient inference is ensured by using the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011), which allows the spatial field to be evaluated at any location quickly. We investigate the performance of the new approach compared to standard practice of ignoring jittering in a simulation study focusing on the stability of random effect estimates.

Increasing populations have a potential to create a huge future demand for the limited resources on food in low- and middle income countries (Le Mouël and Forslund, 2017; Alexandratos and Bruinsma, 2012). In order to support family planning policies, as a source of useful insight, we analyse the proportion of contraceptive use among women aged 15-49 based on data from the 2014 Kenya Demographic and Health Survey (KDHS2014) (National Bureau of Statistics-Kenya and ICF International, 2015). This requires a geostatistical model that can handle a binomial observation model while accounting for jittering.

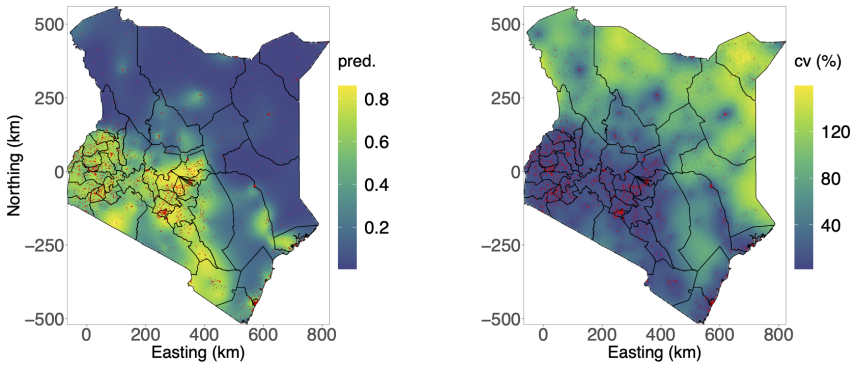


Figure 1: Predicted posterior expectations ("pred.") for the probabilities of using any contraceptive method (left) and the corresponding coefficients of variation (CV) (right) with the new approach. The red points indicate the (jittered) locations of the C =1,583 clusters in Kenya.

In Section 2, we describe the KDHS2014 data set and outline the model structure. Section 3 details the proposed method for approximate inference. Section 4 presents the simulation study and the analysis of contraception use in Kenya. We end the paper with discussion in Section 5. Supplementary results are found in the Supplementary Materials, and all R and C++ code is available in the Github repository https://github.com/umut-altay/Supplementary.git. The repository includes a data statement which outlines the application procedure to download the contraception data from DHS.

2 Data and Model Structure

Kenya consists of 47 counties, where every location in Kenya is classified as "urban" or "rural". KDHS2014 contains 1,594 observed clusters, where the true GPS coordinates have been jittered by the standard DHS procedure restricted so that each GPS location cannot be displaced outside its original county. After eliminating clusters whose coordinates did not match with their designated county or had invalid GPS coordinates, C =1,583 clusters remained. Figure 1 shows the geography together with the estimates which are obtained at the end of Section 4 from the model that we construct to account for jittering in the observation locations. Similar figures are available in Section 4 of Supplementary Materials, for the standard model that does not account for jittering.

In total, there were 31,079 interviewed women and 17,500 among them reported having used a contraceptive method in 2014. For clusters $c = 1, \ldots, C$, let n_c denote the number of interviewed women aged 15–49, y_c the number of those women who have used contraceptive methods, and $\boldsymbol{s}_c \in \mathbb{R}^2$ the jittered spatial location. The unknown true location is denoted as $\boldsymbol{s}_c^* \in \mathbb{R}^2$. We model the probability of contraception use at location $\boldsymbol{s}^*$ as

$$\text{logit}(r(\boldsymbol{s}^*)) = \mu + u(\boldsymbol{s}^*), \quad \boldsymbol{s}^* \in \mathbb{R}^2,$$

where μ is an intercept and $u(\cdot)$ is a GRF with a Matérn covariance function with marginal variance σ_{S}^2, range ρ_{S}, and fixed smoothness $\nu = 1$. We observe n_c individuals exposed to this risk and use $y_c | r(\boldsymbol{s}_c^*) \sim \text{Binomial}(n_c, r(\boldsymbol{s}_c^*))$ independently for $c = 1, \ldots, C$.

Under the DHS jittering scheme, for an urban cluster c with a maximum jittering distance of 2 km, the jittering distribution is

$$\pi_{\text{U}}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \propto \frac{\mathbb{I}(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*)) \cdot \mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 2)}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)}, \quad \boldsymbol{s}_c \in \mathbb{R}^2, \tag{2.1}$$

where, $d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)$ is the distance between $\boldsymbol{s}_c^*$ and $\boldsymbol{s}_c$, $\mathbb{I}$ is an indicator function, and clusters are jittered independently. Similarly, for a rural cluster c, with a maximum jittering distance of 5 km (and the 1 percent of clusters with a maximum jittering distance of 10 km), the jittering distribution is

$$\pi_{\text{R}}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \propto \frac{\mathbb{I}(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*))}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)} \left[\frac{99\mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 5)}{100} + \frac{\mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 10)}{100} \right], \boldsymbol{s}_c \in \mathbb{R}^2. \tag{2.2}$$

The binomial observation model is combined with the location likelihoods in Equations (2.1) and (2.2) to give the complete observation model. The underlying latent model is

$$\mu \sim \mathcal{N}(0, 1000), \quad (u(\boldsymbol{s}_1^*) \ \ldots \ u(\boldsymbol{s}_c^*))^{\text{T}} | \sigma_{\text{S}}^2, \rho_{\text{S}}, \boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_C^* \sim \mathcal{N}(\boldsymbol{0}, \Sigma),$$

where the covariance matrix Σ is a function of the unknown true locations $\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_C^*$ and the parameters σ_{S}^2 and ρ_{S}. We use the penalised complexity (PC) prior for Matérn GRFs (Fuglstad et al., 2019) for σ_{S}^2 and ρ_{S} with $\text{P}(\sigma_{\text{S}} > 1) = 0.05$ and $\text{P}(\rho > \rho_0) = 0.50$, i.e., ρ_0 is the *a priori* median range. We use uniform priors for $\boldsymbol{s}_c^*$, this effectively implies that all locations $\boldsymbol{s}_c^*$ such that $||\boldsymbol{s}_c - \boldsymbol{s}_c^*|| < 10$ are considered equally likely for $c = 1, \ldots, C$.

3 Approximating the Posterior Under Positional Uncertainty

The SPDE model decomposes the spatial effect into a linear combination of compactly supported basis functions, $u(\boldsymbol{s}) = \sum_{i=1}^{K} w_i \phi(s)$, for basis function $\phi_1, \ldots, \phi_K$, and where the K-vector of basis weights $\boldsymbol{w} = (w_1 \ \ldots \ w_K)^{\mathrm{T}}$ follows a multivariate Gaussian distribution with zero mean and with precision matrix set so as to approximate a Matérn covariance structure. This results in a highly sparse precision matrix for the basis weights, and causes the likelihood evaluation to require only $O(K^{3/2}n)$ operations for K basis elements and n observations (Lindgren et al., 2011).

We treat the unknown true locations as nuisance parameters, integrating them out of the likelihood and posterior. Letting $\boldsymbol{\beta} = (\mu)^{\mathrm{T}}$ be the vector of fixed effect coefficients for the linear predictor $\eta(\cdot)$, the full likelihood can be factorized into a product of likelihoods for individual observations: $\pi(\boldsymbol{y}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n | \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}_L) = \prod_{i=1}^{n} \pi(y_i, \boldsymbol{s}_i | \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}_L)$, where θ_L are the parameters of the likelihood. The likelihood for an individual observation can then be calculated by integrating over the distribution of its possible true spatial locations:

$$\begin{aligned}\pi(y_i, \boldsymbol{s}_i | \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}_L) &= \int_{\mathbb{R}^2} \pi(y_i, \boldsymbol{s}_i | \boldsymbol{s}_i^*, \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}_L) \pi(\boldsymbol{s}_i^* | \boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\theta}_L) \, \mathrm{d}\boldsymbol{s}_i^* \\ &= \int_{\mathbb{R}^2} \pi(y_i | \eta(\boldsymbol{s}_i^*), \boldsymbol{\theta}_L) \pi(\boldsymbol{s}_i | \boldsymbol{s}_i^*) \pi(\boldsymbol{s}_i^*) \, \mathrm{d}\boldsymbol{s}_i^*. \end{aligned} \tag{3.1}$$

Since the integral in (3.1) is two-dimensional, it can be well-approximated for each i via quadrature. We will integrate by selecting a single integration point at $\boldsymbol{s}_i$, and then building more 'rings' of points around $\boldsymbol{s}_i$. Let m_{ij} denote the number of integration points for observation i in ring j. Each numerical integration point, given by $\boldsymbol{s}_{ijk}^*$ for observation i, ring j, and index $k = 1, \ldots, m_{ij}$ has an associated integration weight given by λ_{ijk}. If we assume there are J^i rings in total (counting $\boldsymbol{s}_{i11}^* = \boldsymbol{s}_i$ as the first ring), then we can approximate the integral in (3.1) numerically as follows:

$$\begin{aligned}\int_{\mathbb{R}^2} \pi(y_i | \eta(\boldsymbol{s}_i^*), \boldsymbol{\theta}_L) \pi(\boldsymbol{s}_i | \boldsymbol{s}_i^*) \pi(\boldsymbol{s}_i^*) \, \mathrm{d}\boldsymbol{s}_i^* &= \int \pi(y_i | \eta(\boldsymbol{s}_i^*), \boldsymbol{\theta}_L) \, \mathrm{d}\left[\pi(\boldsymbol{s}_i | \boldsymbol{s}_i^*) \pi(\boldsymbol{s}_i^*)\right] \\ &\approx \sum_{j=1}^{J^i} \sum_{k=1}^{m_{ij}} \lambda_{ijk} \pi(y_i | \eta(\boldsymbol{s}_{ijk}^*), \boldsymbol{\theta}_L), \end{aligned} \tag{3.2}$$

where $\lambda_{ijk} \propto \int_{A_{ijk}} \pi(\boldsymbol{s}_i | \boldsymbol{s}_i^*) \pi(\boldsymbol{s}_i^*) \, \mathrm{d}\boldsymbol{s}_i^*$, and A_{ijk} is the area associated with integration point $\boldsymbol{s}_{ijk}^*$, and is defined in the Supplement in Section 6. We will take

$m_{i1} = 1$, and $m_{ij} = 15$ for all other $j > 1$ so that there are $1 + 15(J^i - 1)$ integration points in total. We may assume $\sum_{ij} \sum_k \lambda_{ijk} = 1$ for each i, since the scaling of these weights cancels in the posterior. Hence, if $\pi(\boldsymbol{s}^*_{ijk})$ is constant over the support of $\pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)$, then $\lambda_{ijk} \propto \int_{A_{ijk}} \pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)\, \mathrm{d}\boldsymbol{s}^*_i$. If, however, it is also known that observation i lies in spatial region $R[i]$, and $\pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)$ has any mass outside of $R[i]$, then the weights are: $\lambda_{ijk} \propto \int_{A_{ijk} \cap R[i]} \pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)\, \mathrm{d}\boldsymbol{s}^*_i$.

If observation i is within jittering distance of the boundary of $R[i]$, then its integration weights must be adjusted accordingly. For the ijk-th integration region, we approximate $\lambda_{ijk} \propto \int_{A_{ijk} \cap R[i]} \pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)\, \mathrm{d}\boldsymbol{s}^*_i$ numerically by subdividing A_{ijk} into a 10×10 grid of 'secondary' integration regions, each with an associated secondary integration point at the center of mass of $\pi(\boldsymbol{s}_i|\boldsymbol{s}^*_i)$ on that secondary integration region. We calculate the center of mass radius by shrinking the midpoint radial coordinate of the secondary integration regions within the subregions by an equivalent factor as in Equation 2 of Section 6 in the Supplementary Material, except replacing the subregion boundary angles $a_{ij2} - a_{ij1}$ (defined in the Supplement in Section 6) with $(a_{ij2} - a_{ij1})/10$. We then scale λ_{ijk} depending on the proportion of associated subintegration points in $R[i]$. This is equivalent to assuming that all secondary integration points associated with a given integration region have approximately equal weight. This adjustment to the weights, as well as the integration regions and points for an urban cluster in Nairobi, are depicted in Figure 2. Technical details regarding the generation of the integration points, weights, and regions, including derivations, are given in Section 6 of the Supplementary Material.

We implement the above model in C++ using TMB, which integrates out $\boldsymbol{w}$, and uses autodifferentiation to maximize and takes a Laplace approximation of the posterior. As a result, the proposed method has the computational advantages of both the SPDE model and of its implementation in TMB. If M_{P} and M_{S} are respectively the average number of primary integration points per observation and the number of secondary integration points per primary integration point, then our method still only requires $O(M_{\mathrm{P}} M_{\mathrm{S}} n K^{3/2}) = O(nK^{3/2})$ computational operations per likelihood evaluation. The autodifferentiation of TMB also helps to reduce the number of operations required for optimizing the approximated posterior.

The integration weights before correction for boundary effects, the radial displacement of the integration points, and the number of points per integration ring are given in Table 5 in Section 6 of the Supplementary Material.

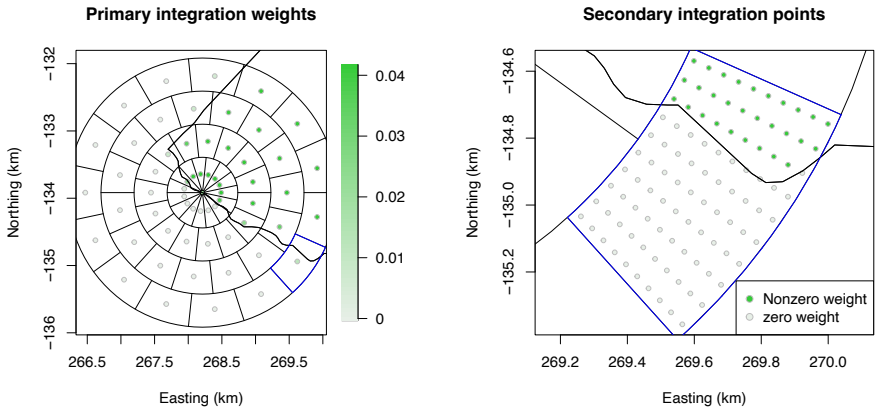


Figure 2: Primary integration points and their weights for a cluster in Nairobi (left), and an integration region with associated secondary integration points (right). The integration region is outlined in blue, and Nairobi is outlined in black in both plots.

4 Simulation Study and Analysis of Contraception Use in Kenya

We evaluate the gain when accounting for jittering through a simulation study where data is generated according to the model described in Section 2. The GRF is simulated using marginal variance $\sigma^2_{\text{S}} = 1$, spatial ranges $\rho_{\text{S}} \in \{160, 340\}$ (km), and smoothness $\nu = 1$. These ranges correspond to approximately 1/5 and 2/5 of the extent of Kenya in West-East-direction. We fix the true coordinates to match the $C = 1,583$ clusters with reliable location information in KDHS2014, and set the intercept $\mu = 0$, which corresponds to 50% contraception use. This is motivated by the fact that contraception use in Kenya has strong spatial variation, but with a national level around 58% (National Bureau of Statistics-Kenya and ICF International, 2015). Datasets are generated by simulating y_c at location $\boldsymbol{s}^*_c$ from a binomial distribution where the success probability is $r(\boldsymbol{s}^*_c)$ and the number of trials $n_c = 100$ for $c = 1, \ldots, C$. Section 3 in the Supplementary Material presents the corresponding study with a Gaussian observation model.

For each of the two ranges, we simulate the GRF and responses repeatedly to give 50 datasets. To each of these datasets we apply two jittering strategies: 1) standard DHS jittering, and 2) DHS jittering with maximum distances multiplied

with 4 (termed 4 $\times$ DHS jittering). This gives 200 datasets for the four combinations of ranges (160 km and 340 km) and jittering options. For each dataset we fit a standard spatial model that assumes locations are correct (Model-S) and the new model that accounts for positional uncertainty (Model-J). For the model specification in Section 2, we set the *a priori* median of range ρ_0 equal to true range. After fitting the model, we compute the continuous rank probability score (CRPS) and the logarithmic score (log-score) (Gneiting and Raftery, 2007) for 1,000 evenly distributed prediction locations (shown in Figure 1 in the Supplementary Material).

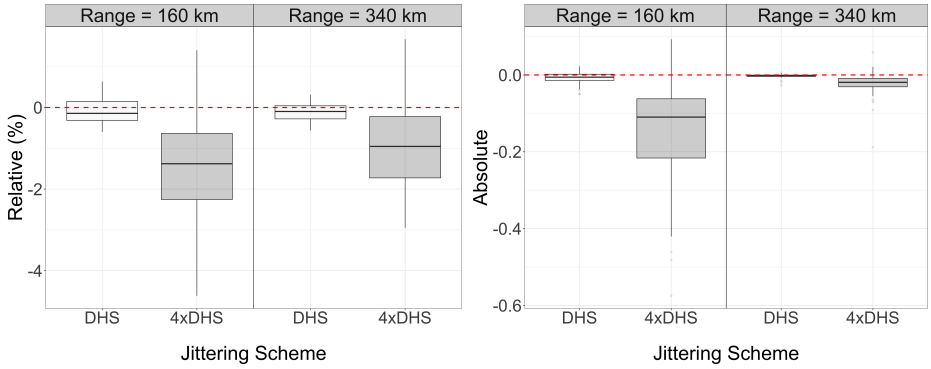
Posterior inference is approximately Bayesian using TMB, and parameter estimates are computed using posterior medians. Table 1 shows that there is less bias in the parameter estimates when using Model-J than Model-S. The difference between the two approaches becomes larger for 4 $\times$ DHS jittering than standard DHS jittering. The positional uncertainty in Model-J gives larger 95% credible intervals (CIs) for the parameters compared to the Model-S, and the difference is larger for more jittering.

Figure 3a shows a minor improvement in relative difference in CRPS for the prediction locations with Model-J compared to Model-S under standard DHS jittering. For 4 $\times$ DHS jittering, there is a clear improvement. Figure 3b shows similar behavior for the log-score, but with a less clear difference with $\rho_S = 340$ km and 4 $\times$ DHS jittering. There were only minor differences in the average coverage of the predictive distributions as shown in Table 3 in the Supplementary Material. A corresponding simulation study with a Gaussian observation model in Section 3 in the Supplementary Materials leads to similar conclusions, and demonstrates that a nugget variance is overestimated when jittering is not accounted for.

We apply the new approach to the contraception use dataset described in Section 2. Model-S and Model-J were estimated in 21 seconds and 8 minutes, respectively. For the real data analysis we again place a PC prior on the spatial range parameter, setting the median spatial range to $\rho_0 = 160$ km. The estimated contraception use probabilities and coefficients of variation for Model-J are shown in Figure 1. The map shows contraception use is high in the southwest direction and low in northeast. On average CVs are 2.7% higher for Model-J relative to Model-S and point estimates are nearly indistinguishable; see Section 5 of the Supplementary Materials for more details and figures for Model-S.

Table 1: Average biases and average 95% CI lengths of parameter estimates under Model-J. We use absolute bias for μ and relative bias for ρ_S and σ^2_S. The corresponding values using Model-S are shown in parentheses.

		DHS jittering		4xDHS jittering	
Parameter	**Truth**	**Bias**	**CI length**	**Bias**	**CI length**
Short range					
μ	0	-0.03 *(-0.03)*	0.79 *(0.77)*	-0.03 *(-0.03)*	0.84 *(0.69)*
ρ_S	160	-3% *(-6%)*	69 *(65)*	7% *(-13%)*	82 *(59)*
σ^2_S	1	-2% *(-2%)*	0.34 *(0.33)*	-4% *(-7%)*	0.37 *(0.30)*
Long range					
μ	0	-0.04 *(-0.04)*	1.24 *(1.23)*	-0.10 *(-0.10)*	1.27 *(1.18)*
ρ_S	340	-7% *(-9%)*	203 *(200)*	-3% *(-10%)*	221 *(199)*
σ^2_S	1	-7% *(-8%)*	0.51 *(0.50)*	-9% *(-11%)*	0.52 *(0.48)*



(a) Pair-wise relative differences $\left(100 \times \frac{CRPS_J - CRPS_S}{CRPS_S}\right)$ in CRPS

(b) Pair-wise absolute differences in Log Scores

Figure 3: Pair-wise differences in CRPS and Log Scores that are obtained from Model-J and Model-S for binomial observation model.

5 Discussions and Conclusions

Our simulation study suggests that accounting for the presence of jittering, or positional uncertainty, in the geostatistical analysis of DHS data on contraception use leads to more accurate parameter estimates than a standard geostatistical analysis. The improvement becomes more pronounced if more jittering is applied than DHS applies by default. Further, we see slight improvement in predictive quality under standard DHS jittering, and this improvement becomes clearer for higher amounts of jittering. Our novel approach represents a major improvement over existing inference approaches that are suitable for binomial observation models such as INLA within MCMC, where computation time is measured in days (Wilson and Wakefield, 2021). The computation time of the new approach is measured in minutes as compared to days for INLA within MCMC.

In the simulation study we encountered numerical issues when fitting a small number of the simulations. These occurred when the amount of jittering was large compared to the spatial range. In the case of range 160 km and 4×DHS jittering, 2 out of 50 model runs crashed. Though, this amount of jittering is large compared to what is used in practice by DHS, but there is a need for future investigation into methods that are more stable for higher amounts of positional uncertainty. The focus of this paper is to present a fast geostatistical model that accounts for jittering during inference. Our approach supports generalized linear geostatistical models with a wide variety of non-Gaussian observation models due to its implementation in TMB. It also is applicable in the context of other known jittering distributions, such as in cases where the administrative area of a cluster is known, but the exact location within the area is not. One limitation, however, is that the computational efficiency will decrease when large displacements of coordinates are possible relative to the size of the domain of interest. This is due to decreasing sparsity in the precision matrix induced by jittering distributions overlapping with more spatial basis functions. An interesting potential direction of future research would be to model positional uncertainty when including spatially varying covariates. Furthermore, it would be interesting to investigate the accuracy of the approach presented in this paper to other jittering strategies such as swapping and truncating can also be applied (Burgert et al., 2013).

References

Alexandratos, N. and Bruinsma, J. (2012). World agriculture towards 2030/2050: the 2012 revision.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). Geographic displacement procedure and georeferenced datarelease policy for the Demographic and Health Surveys. `https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf`. DHS Spatial Analysis Reports No. 7.

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

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

Fronterrè, C., Giorgi, E., and Diggle, P. (2018). Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *Spatial Statistics*, 28:319–330.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114:445–452.

Gahi, Y., Guennoun, M., and Mouftah, H. T. (2016). Big data analytics: Security and privacy challenges. In *2016 IEEE Symposium on Computers and Communication (ISCC)*, pages 952–957. IEEE.

General Assembly of the United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1.

Gething, P., Tatem, A., Bird, T., and Burgert-Brucker, C. R. (2013). Creating spatial interpolation surfaces with DHS data. `https://dhsprogram.com/pubs/pdf/SAR11/SAR11.pdf`. DHS Spatial Analysis Reports No. 11.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gómez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051.

Khoshgozaran, A. and Shahabi, C. (2007). Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy. In *International symposium on spatial and temporal databases*, pages 239–257. Springer.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5):1–21.

Le Mouël, C. and Forslund, A. (2017). How can we feed the world in 2050? a review of the responses from global scenario studies. *European Review of Agricultural Economics*, 44(4):541–591.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73:423–498.

National Bureau of Statistics-Kenya and ICF International (2015). 2014 KDHS key findings. `https://www.dhsprogram.com/pubs/pdf/sr227/sr227.pdf`.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Sstatistical Methodology)*, 71(2):319–392.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

VanWey, L. K., Rindfuss, R. R., Gutmann, M. P., Entwisle, B., and Balk, D. L. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences*, 102(43):15337–15342.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016a). Influence of demographic and health survey point displacements on distance-based analyses. *Spatial Demography*, 4(2):155–173.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016b). Influence of demographic and health survey point displacements on point-in-polygon analyses. *Spatial Demography*, 4(2):117–133.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100421.

Zhang, S., Freundschuh, S. M., Lenzer, K., and Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1):22–34.

Supplementary materials: Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data

Umut Altay John Paige Andrea Riebler
Geir-Arne Fuglstad

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

1 Introduction

This document consists of the supplementary results and materials for our paper titled "Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data". We used (jittered) 2014 Kenya Demographic and Health Survey (KDHS2014) clusters for our study. Figure 1 shows them together with the prediction locations. The rest of the document is structured as follows:

Section 2 presents the supplementary figures of continuous rank probability score (CRPS) and log-score that are obtained from the simulations with the binomial observation model.

Section 3 consists of figures and tables of CRPS and log-score that are obtained from the simulations with the Gaussian observation model. Section 4 presents the tables of coverage values that are obtained from the simulations with both the binomial and Gaussian observation models. Average computation (model estimation) times that are measured for Model-J during the simulation study under different scenarios are also shared in this section. Section 5 shows the results of additional predictions that are done using the binomial model on KDHS2014 contraceptive usage data. Section 6 explains how numerical integrations are conducted in our approach.

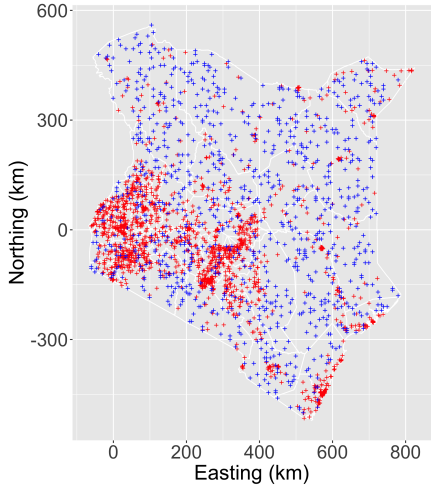


Figure 1: Locations that are used for the study, within Kenya. Jittered locations of the C = 1,583 clusters are indicated by red. Prediction locations are indicated by blue.

2 Supplementary Results for Binomial Likelihood

This section presents the supplementary results for the simulation study with the binomial observation model. Figure 2 shows the box-plots of CRPS and log-score values that are obtained from Model-S and Model-J for the scenarios combining ranges $\rho_S \in \{160, 340\}$ (km) with jittering schemes (DHS and 4xDHS). Smaller CRPS and log-scores indicate better predictions. Figure 2 shows that Model-J tends to achieve smaller prediction scores and to make better predictions than Model-S as the jittering gets larger. Both models react to the increasing spatial range by providing better predictions.

3 Simulation Study for Gaussian Likelihood

This section presents the results of the simulation study with the Gaussian observation model. Figure 3 shows the box plots of pair-wise relative differences in CRPS and the absolute differences in log-score for the prediction locations, with Model-J compared to Model-S. Figure 4 shows the box-plots of CRPS and log-score values for the scenarios combining ranges $\rho_S \in \{160, 340\}$ (km) and

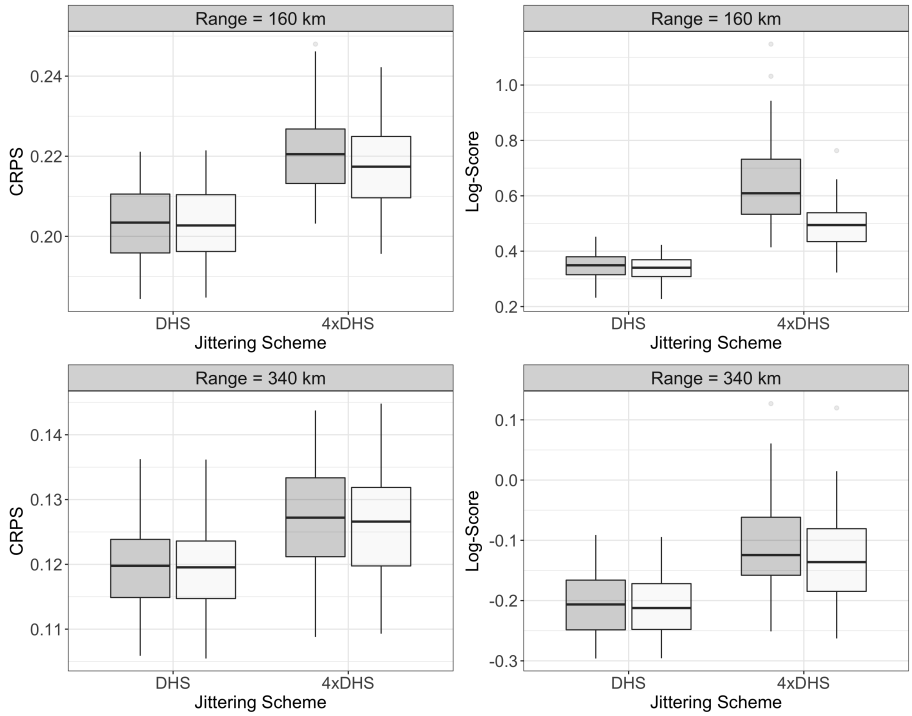
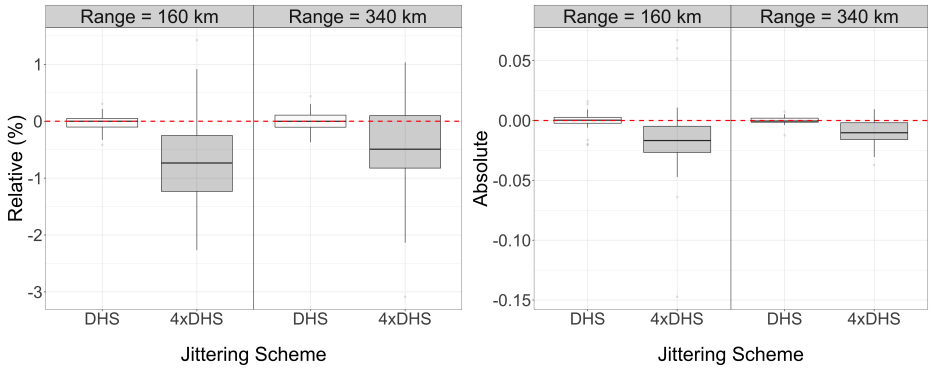


Figure 2: Box-plots of CRPS and log-score values that are obtained from Model-S (boxes with the darker color) and Model-J (boxes with the lighter color) at 1000 prediction locations, for the simulations with the binomial observation model.



(a) Relative differences ($100 * \frac{CRPS_J - CRPS_S}{CRPS_S}$) in CRPS

(b) Absolute differences in Log Scores

Figure 3: Box plots of pair-wise differences of the prediction scores that are obtained from Model-S and Model-J at 1000 prediction locations, for the simulations with the Gaussian observation model.

jittering schemes (DHS and 4xDHS). Table 1 presents the average biases and average CI lengths of parameter estimates.

4 Model Estimation Times and Coverage

Table 2 shows the average model estimation times (in minutes) obtained by running Model-J on different simulation scenarios. Table 3 shows the coverage values obtained from each scenario, using both Model-S and Model-J.

5 Additional KDHS2014 Contraceptive Usage Results

Figure 5 shows the predicted posterior expectations for the probabilities of using any contraceptive method and the corresponding coefficients of variation (CV) for KDHS2014 contraceptive usage data with Model-S. Similar figures for Model-J are shared in Section 2 of "Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data" paper. Figure 6 shows the comparison of the predicted posterior expectations for the probabilities of using

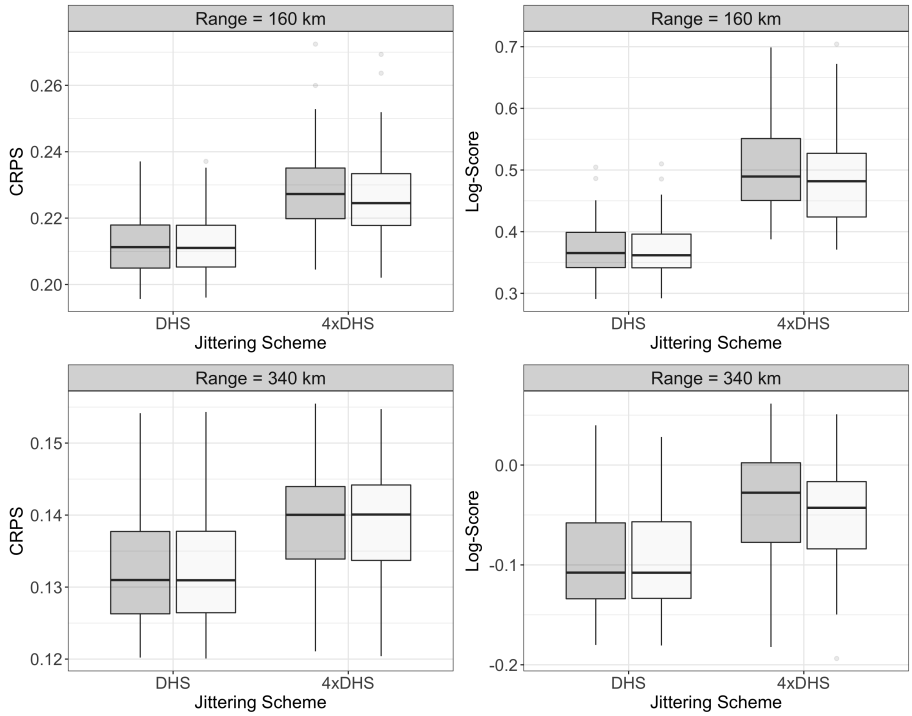


Figure 4: Box-plots of CRPS and log-scores that are obtained from Model-S (boxes with the darker color) and Model-J (boxes with the lighter color) at 1000 prediction locations, for the simulations with the Gaussian observation model.

Table 1: Average biases and average 95% CI lengths of parameter estimates. We use absolute bias for μ and relative bias for ρ_S, σ_N^2 and σ_S^2. The corresponding values using Model-S are shown in parantheses.

		DHS jittering		**4xDHS jittering**	
Parameter	**Truth**	**Bias**	**CI length**	**Bias**	**CI length**
Short range					
μ	0	-0.03 *(-0.03)*	0.80 *(0.79)*	-0.04 *(-0.04)*	0.86 *(0.84)*
ρ_S	160	-0.6% *(-1%)*	77 *(76)*	8% *(7%)*	89 *(89)*
σ_N^2	0.1	6% *(8%)*	0.01 *(0.01)*	11% *(38%)*	0.02 *(0.02)*
σ_S^2	1	-3% *(-3%)*	0.35 *(0.35)*	-3% *(-4%)*	0.37 *(0.37)*
Long range					
μ	0	-0.04 *(-0.04)*	1.25 *(1.25)*	-0.04 *(-0.04)*	1.23 *(1.23)*
ρ_S	340	-7% *(-7%)*	216 *(215)*	-7% *(-6%)*	220 *(222)*
σ_N^2	0.1	0.7% *(1%)*	0.01 *(0.01)*	2% *(11%)*	0.01 *(0.01)*
σ_S^2	1	-8% *(-8%)*	0.51 *(0.51)*	-10% *(-10%)*	0.50 *(0.50)*

Table 2: Average model estimation times with Model-J (in minutes) during the simulation study .

	Short range		**Long range**	
Simulations	**DHS jittering**	**4xDHS jittering**	**DHS jittering**	**4xDHS jittering**
Binomial	4.61	9.58	4.60	8.32
Gaussian	4.04	7.32	3.34	6.02

Table 3: Coverage values of Model-J. The corresponding values using Model-S are shown in the parantheses.

Simulations	**Range**	**DHS jittering**	**4xDHS jittering**
Gaussian	Short	0.92 *(0.92)*	0.89 *(0.89)*
	Long	0.93 *(0.93)*	0.92 *(0.92)*
Binomial	Short	0.91 *(0.91)*	0.87 *(0.88)*
	Long	0.93 *(0.93)*	0.91 *(0.90)*

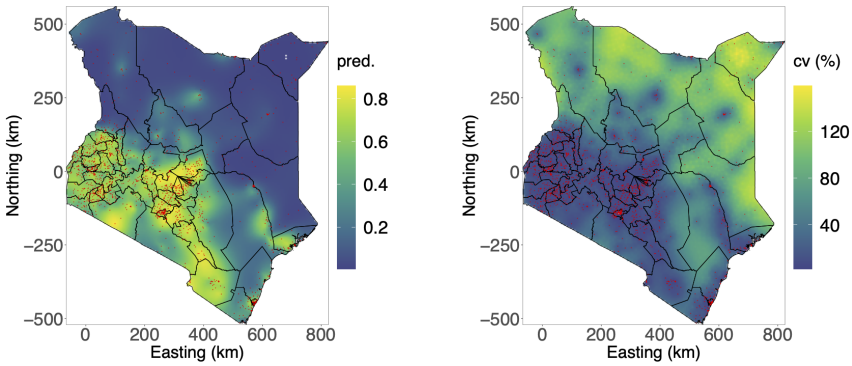


Figure 5: Predicted posterior expectations ("pred.") for the probabilities of using any contraceptive method (left) and the corresponding coefficients of variation (CV) (right) for Model-S. The red points indicate the (jittered) locations of the $C = 1,583$ clusters in Kenya.

Table 4: Parameter estimates and corresponding 95% intervals

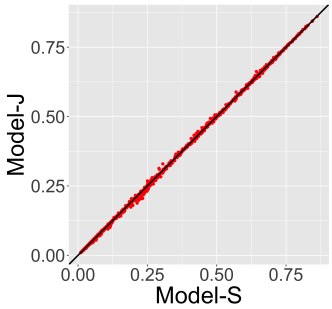
	Median	**Lower**	**Upper**	**Length**
β_0	-1.78*(-1.76)*	-2.61*(-2.60)*	-0.97*(-0.93)*	1.64*(1.66)*
ρ	183*(188)*	143*(147)*	233*(241)*	90*(94)*
σ^2_{SF}	1.74*(1.72)*	1.43*(1.42)*	2.11*(2.09)*	0.68*(0.67)*

any contraceptive method and the coefficient of variations that are obtained from Model-S and Model-J, by using KDHS2014 contraceptive usage data. Coefficient of variation values are slightly higher for Model-J compared to Model-S, while the predicted posterior expectations from both models are very similar to each other, as it is also mentioned in Section 4 of the main manuscript.

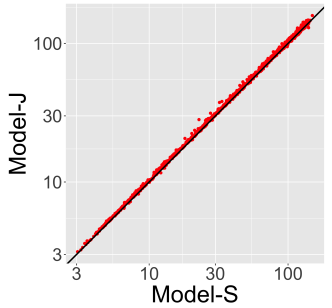
Table 4 shows the parameter estimates and corresponding 95% intervals for KDHS2014 contraceptive usage data with Model-J. The corresponding values using Model-S are shown in parantheses.

6 Technical Derivation of Numerical Integration Procedure

If we take integration points in each ring to be angularly equidistant, and represent the area associated with the ijk-th integration point (for observation i,



(a) Scatter plot of predicted posterior expectations for the probabilities of using any contraceptive method with Model-S versus Model-J



(b) Scatter plot for the coefficient of variation values of the predictions (in log scale) with Model-S versus Model-J

Figure 6: Scatter plots of the predictions and corresponding uncertainty

integration ring j, and the k-th integration point in the ring) as,

$$A_{ijk} = \{\boldsymbol{s}_i + (r\cos a,\ r\sin a)^T : r_{i(j-1)} \leq r < r_{ij},\ a_{ij(k-1)} \leq a < a_{ijk}\},$$

where r_{i0} is taken to be 0 for all i, and r_{iJ^i} is L_i, then the weights depend on the probability mass of the jittering distribution in each A_{ijk}. We take the integration area boundaries as equispaced,

$$a_{ijk} = \begin{cases} \frac{2\pi}{m_{ij}}(k-1) + \frac{\pi}{m_{ij}}, & j \bmod 2 = 1,\ j \geq 5 \\ \frac{2\pi}{m_{ij}}(k-1), & \text{otherwise,} \end{cases}$$

where $\frac{\pi}{m_{ij}}$ intersperses the integration points for every other ring based on m_{ij}, the number of integration points for observation i and ring j. Now that each a_{ijk} has been specified for $i = 1, \ldots, n$, $j = 1, \ldots, J^i$ given J^i the number of integration rings for observation i, and $k = 1, \ldots, m_{ij}$, the probability mass of the jittering distribution in A_{ijk} and therefore the integration point weights depend only on the choice of the radii r_{ij}. Since the jittering density distribution in Equation 1 in the main manuscript is radially symmetric, interspersing the points along each ring does not influence the integration weights. Our choice of the r_{ij} will depend on whether the observed cluster is urban or rural.

For urban clusters, the jittering process density is continuous on the support of the density, unlike for the rural clusters. We choose the radii, r_{ij}, for any fixed urban observation i so that the integration weights are equal for each of the integration points. If the prior density $\pi(\boldsymbol{s}_i^*)$ is constant over the support of

$\pi(\boldsymbol{s}_i|\boldsymbol{s}_i^*)$, then $\pi(\boldsymbol{s}_i|\boldsymbol{s}_i^*)$ being uniform when represented in radial coordinates on $[0, 2\pi]\times[0, L_i]$ implies setting $r_{ij} = \frac{L_i \sum_{j'=0}^{j} m_{ij'}}{\sum_{j'=1}^{J^i} m_{ij'}}$ results in equal urban integration weights in Equation 4 in the main manuscript, with,

$$\lambda_{ijk} \propto \frac{r_{ij} - r_{i(j-1)}}{L_i} \frac{a_{ijk} - a_{ij(k-1)}}{2\pi} \pi(\boldsymbol{s}_{ijk}^*),$$

so that $\lambda_{ijk} \propto \frac{r_{ij}-r_{i(j-1)}}{L_i} \frac{a_{ijk}-a_{ij(k-1)}}{2\pi}$ if $\pi(\boldsymbol{s}_{ijk}^*)$ is constant.

If $\boldsymbol{s}_i$ is rural, there is a discontinuity in $\pi(\boldsymbol{s}_i|\boldsymbol{s}_i^*)$ where $||\boldsymbol{s}_i - \boldsymbol{s}_i^*|| = L_i'$ for discontinuity radius L_i' due to the fact that there is a 0.01 probability of rural points having a larger maximum jittering distance. We therefore define 'inner' and 'outer' rings with $J^i = J^i_{\text{inner}} + J^i_{\text{outer}}$, where the inner rings and outer rings are inside and outside of the discontinuity radius respectively. For rural DHS spatial locations, $r_{iJ^i_{\text{inner}}} = L_i' = 5$ and $r_{iJ^i} = L_i = 10$. We choose the inner and outer ring radii so that the integration points in the inner and outer rings have equal weights respectively, so that:

$$r_{ij} = \begin{cases} \frac{\sum_{j'=1}^{j} m_{ij'}}{\sum_{j'=1}^{J^i_{\text{inner}}} m_{ij'}} L_i', & 1 \le j \le J^i_{\text{inner}} \\ L_i' + \frac{\sum_{j'=J^i_{\text{inner}}+1}^{j} m_{ij'}}{\sum_{j'=J^i_{\text{inner}}+1}^{J^i} m_{ij'}} (L_i - L_i'), & J^i_{\text{inner}} < j \le J^i. \end{cases}$$

These ring radii result in the following rural integration weights:

$$\lambda_{ijk} \propto \begin{cases} \frac{r_{ij}-r_{i(j-1)}}{m_{ij}} \left(\frac{99}{100} \frac{1}{r_{J^i_{\text{inner}}}} + \frac{1}{100} \frac{1}{L_i} \right), & 1 \le j \le J^i_{\text{inner}} \\ \frac{r_{ij}-r_{i(j-1)}}{m_{ij}} \frac{1}{100} \frac{1}{L_i}, & J^i_{\text{inner}} < j \le J^i. \end{cases} \tag{6.1}$$

The $\frac{99}{100}$ and $\frac{1}{100}$ factors in the above expressions are due to rural clusters having a probability of $\frac{1}{100}$ of being displaced by up to 10 km. We set $J^i = 5$ for urban points, and $J^i_{\text{inner}} = 5$ and $J^i_{\text{outer}} = 5$ for rural points. Although the rural outer ring weights are much smaller than the inner weights to the point where leaving them out and renormalizing the weights would likely not influence the predictions, and would improve computation times, we choose to include them for greater precision.

We set each integration point $\boldsymbol{s}_{ijk}^*$ to be the center of mass of $\pi(\boldsymbol{s}_i|\boldsymbol{s}_i^*)$ within the associated ijk-th integration area A_{ijk}, with $\boldsymbol{s}_{ijk}^* = (r_{ij}^* \cos((a_{ijk} + a_{ijk})/2), r_{ij}^* \sin((a_{ijk} + a_{ijk})/2))^T$, and where,

$$r_{ij}^* = \frac{r_{i(j-1)} + r_{ij}}{2} \frac{\sqrt{2(1 - \cos(a_{ij2} - a_{ij1}))}}{a_{ij2} - a_{ij1}}, \tag{6.2}$$

for $j > 1$ (if $j = 1$, then $r^*_{ij} = 0$). A derivation of (6.2) is given in more detail below.

If observation i is urban, we can calculate the expectation of the horizontal coordinate, say x_{ij1}, for the center of mass of the first integration area in ring j, and assuming $a_{ij0} = 0$ and $a_{ij1} = 2\pi/m_{ij}$, as follows:

$$\begin{aligned}
E[x_{ij1}] &= \int_{r_{i(j-1)}}^{r_{ij}} \int_0^{2\pi/m_{ij}} rx \frac{C_{ij}}{2\pi D_i r} \, \mathrm{d}a \, \mathrm{d}r \\
&= \int_{r_{i(j-1)}}^{r_{ij}} \int_0^{2\pi/m_{ij}} r\cos(a) \frac{C_{ij}}{2\pi D_i} \, \mathrm{d}a \, \mathrm{d}r \\
&= \int_{r_{i(j-1)}}^{r_{ij}} r \frac{C_{ij}}{2\pi D_i} (\sin(2\pi/m_{ij}) - \sin(0)) \, \mathrm{d}r \\
&= \frac{C_{ij}}{2\pi D_i} \sin(2\pi/m_{ij})(r_{ij}^2 - r_{i(j-1)}^2) \\
&= \frac{\sin(2\pi/m_{ij})}{2\pi/m_{ij}} \frac{r_{ij} + r_{i(j-1)}}{2},
\end{aligned} \tag{6.3}$$

where $C_{ij} = \frac{2\pi D_i}{(r_{ij} - r_{i(j-1)})(a_{ijk} - a_{ij(k-1)})}$. Similar reasoning yields the following expectation for $E[y_{ij1}]$, where y_{ij1} is the vertical coordinate of the center of mass of the first integration area for observation i in ring j:

$$E[y_{ij1}] = \frac{1 - \cos(2\pi/m_{ij})}{2\pi/m_{ij}} \frac{r_{ij} + r_{i(j-1)}}{2}. \tag{6.4}$$

We can combine the above two expectations to get the radial displacement of the center of mass of integration area A_{ijk}, relative to $\boldsymbol{s}_i$:

$$\begin{aligned}
r^*_{ij} &= \sqrt{E[x_{ij1}]^2 + E[y_{ij1}]^2} \\
&= \sqrt{\frac{\sin(2\pi/m_{ij})^2}{4\pi^2/m_{ij}^2} \frac{(r_{ij} + r_{i(j-1)})^2}{4} + \frac{(1 - \cos(2\pi/m_{ij}))^2}{4\pi^2/m_{ij}^2} \frac{(r_{ij} + r_{i(j-1)})^2}{4}} \\
&= \frac{r_{ij} + r_{i(j-1)}}{2} \frac{m_{ij}}{2\pi} \sqrt{\sin(2\pi/m_{ij})^2 + (1 - \cos(2\pi/m_{ij}))^2} \\
&= \frac{r_{ij} + r_{i(j-1)}}{2} \frac{m_{ij}}{2\pi} \sqrt{2(1 - \cos(2\pi/m_{ij}))}.
\end{aligned} \tag{6.5}$$

Due to the radial symmetry of the jittering distribution under a flat prior $\pi(\boldsymbol{s}^*_i)$, we obtain $\sqrt{E[x_{ijk}]^2 + E[y_{ijk}]^2} = \sqrt{E[x_{ij1}]^2 + E[y_{ij1}]^2}$ for all $1 \le k \le m_{ij}$.

If observation i is rural, we must use the rural jittering density taking the form,

$$\pi(r) = \begin{cases} \frac{99}{100}\frac{C_{ij}}{2\pi D_i r} + \frac{1}{100}\frac{C'_{ij}}{2\pi D'_i r}, & 0 < r \leq L'_i \\ \frac{1}{100}\frac{C'_{ij}}{2\pi D'_i r}, & L'_i < r \leq L_i \\ 0, & \text{otherwise,} \end{cases}$$

for $C_{ij} = \frac{2\pi D_i}{(r_{ij} - r_{i(j-1)})(a_{ijk} - a_{ij(k-1)})}$ and $C'_{ij} = \frac{2\pi D'_i}{(r_{ij} - r_{i(j-1)})(a_{ijk} - a_{ij(k-1)})}$. We can then calculate the expected horizontal coordinate of the integration area with respect to the rural jittering density in the same way as for the urban density:

$$E[x_{ij1}] = \int_{r_{i(j-1)}}^{r_{ij}} \int_0^{2\pi/m_{ij}} rx \left(\frac{99}{100}\frac{C'_{ij}}{2\pi D'_i r} + \frac{1}{100}\frac{C_{ij}}{2\pi D_i r} \right) \, \mathrm{d}a \, \mathrm{d}r.$$

Since (6.5) does not depend on D_i, we reach the same result for rural as for urban integration points for 'inner' integration area A_{ijk}:

$$\begin{aligned} E[x_{ij1}] &= \frac{99}{100} \int_{r_{i(j-1)}}^{r_{ij}} \int_0^{2\pi/m_{ij}} rx \frac{C'_{ij}}{2\pi D'_i r} \, \mathrm{d}a \, \mathrm{d}r + \\ &\quad \frac{1}{100} \int_{r_{i(j-1)}}^{r_{ij}} \int_0^{2\pi/m_{ij}} rx \frac{C_{ij}}{2\pi D_i r} \, \mathrm{d}a \, \mathrm{d}r \\ &= \frac{99}{100} \frac{r_{ij} + r_{i(j-1)}}{2} \frac{m_{ij}}{2\pi} \sqrt{2(1 - \cos(2\pi/m_{ij}))} + \\ &\quad \frac{1}{100} \frac{r_{ij} + r_{i(j-1)}}{2} \frac{m_{ij}}{2\pi} \sqrt{2(1 - \cos(2\pi/m_{ij}))} \\ &= \frac{r_{ij} + r_{i(j-1)}}{2} \frac{m_{ij}}{2\pi} \sqrt{2(1 - \cos(2\pi/m_{ij}))}. \end{aligned}$$

Similar lines of reasoning show that the above expression for $E[x_{ij1}]$ holds even for 'outer' integration areas, and that (6.4) and (6.5) also hold for rural integration areas (both inner and outer).

Table 5 gives the radial displacement, number of integration points, and integration weights (uncorrected for potential administrative boundary effects) as a function of j, the ring index.

Table 5: For each numerical integration ring, the displacement, number, and weights of the individual integration points. Weights here have not been corrected for edge effects, and have been normalized to sum to 1. Displacements are scaled to match the DHS jittering distribution.

	Ring Number	Displacement (km)	Number of Points	Integration Weights
Urban	1	0.00	1	0.0164
	2	0.28	15	0.0164
	3	0.76	15	0.0164
	4	1.25	15	0.0164
	5	1.74	15	0.0164
Rural	1	0.00	1	0.0163
	2	0.69	15	0.0163
	3	1.91	15	0.0163
	4	3.13	15	0.0163
	5	4.35	15	0.0163
	6	5.46	15	0.0001
	7	6.45	15	0.0001
	8	7.45	15	0.0001
	9	8.44	15	0.0001
	10	9.43	15	0.0001

Paper II

Jittering impacts raster- and distance-based geostatistical analyses of DHS data

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A.

Submitted to Statistical Modelling Journal on 11.11.2022

Jittering Impacts Raster- and Distance-based Geostatistical Analyses of DHS Data

Umut Altay John Paige Andrea Riebler
Geir-Arne Fuglstad

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

Fine-scale covariate rasters are routinely used in geostatistical models for mapping demographic and health indicators based on household surveys from the Demographic and Health Surveys (DHS) program. However, the observations in these surveys have GPS coordinates that are jittered for privacy purposes. We demonstrate the need to account for this jittering when analysing DHS data and propose a computationally efficient approach. We analyse the prevalence of completion of secondary education among 20-39 year old women in Nigeria in 2018 based on the 2018 DHS survey in Nigera, and demonstrate substantial changes in the estimates of range and fixed effects compared to ignoring jittering. Then based on a simulation study that mimics the dataset, we demonstrate that accounting for jittering reduces attenuation in the estimated coefficients and improves predictions.

Keywords: Jittering, DHS surveys, Demographic and health indicators, geostatistical analysis, Template Model Builder (TMB).

1 Introduction

Fine-scale spatial estimation of demographic and health indicators has become commonplace (Burstein et al., 2019; Utazi et al., 2019; Local Burden of Disease Vaccine Coverage Collaborators, 2021). This paper is focused on prevalences, which include many important indicators such as completion of secondary education, neonatal mortality, and vaccination coverage (Fuglstad et al., 2021). For low- and middle-income countries (LMICs), the household surveys conducted by the Demographic and Health Surveys (DHS) Program are a crucial data source. Geographic information in DHS data is given through GPS coordinates, which describe centres of clusters of households. However, cluster centres are randomly displaced by up to 10 km before being published in order to protect participants' privacy (Burgert et al., 2013). We refer to the small random displacements as *jittering* of the GPS coordinates.

In global health, it is common practice to ignore jittering and estimate risk using a standard geostatistical model with a binomial likelihood. The latent spatial variation in risk is modelled as the combination of raster- and distance-based covariates and a Gaussian random field (GRF). However, covariate values extracted from rasters can vary widely on the distance scale of jittering. Using the covariate value at the jittered location instead of the original location induces a

non-standard form of measurement error (Gustafson, 2003). This may in turn lead to attenuation of effect estimates and errors in uncertainty. Furthermore, not accounting for the positional uncertainty for the GRF, artificially reduces estimated spatial dependency and may reduce predictive power as well (Cressie and Kornak, 2003; Fanshawe and Diggle, 2011; Fronterrè et al., 2018).

To address uncertainty in covariates, Perez-Heydrich et al. (2013, 2016) suggested 1) to use regression calibration in the context of distance-based covariates (Warren et al., 2016), and 2) to average spatial covariates within a 5 km buffer zone for continuous and categorical rasters. However, this approach does not address the issue of attenuation of associations.

Fanshawe and Diggle (2011) proposed a Bayesian approach to account for positional uncertainty for the GRF, but did not propagate uncertainty in the covariates, and only used Gaussian likelihoods that are not applicable to prevalences. The approach was also computationally expensive, but Fronterrè et al. (2018) made the approach computationally efficient and demonstrated its applicability to analyse malnutrition based on DHS data.

Recently, Wilson and Wakefield (2021) formulated a full geostatistical model for DHS data that includes an observation model for the jittered GPS coordinates, and estimated the model with integrated nested Laplace approximations (INLA) (Rue et al., 2009) within Markov chain Monte Carlo (MCMC) (Gómez-Rubio and Rue, 2018). Their approach addresses the effect of positional uncertainty on both the spatial covariates and the GRF, but was computationally expensive with 1000 MCMC iterations requiring 52 hours in their simulation study.

Altay et al. (2022) proposed a similar model as Wilson and Wakefield (2021), but used a more efficient inference scheme with computation time being measured in minutes instead of hours. Their approach was made possible through an approximation of the likelihood, the SPDE approach (Lindgren et al., 2011), and Laplace approximations through template model builder (TMB) (Kristensen et al., 2016).

The simulation study in Altay et al. (2022) revealed that small spatial ranges for the GRF or larger jittering than the DHS scheme were required to see substantial improvements with the new approach over ignoring jittering. However, Altay et al. (2022) focused on the impact of jittering on the GRF, and lacked raster- and distance-based covariates.

Such covariates are far more variable at small spatial scales than a smoothly varying GRF. The aim of this paper is to extend the approach in Altay et al. (2022) to a full generalized geostatistical model for prevalence, and to demonstrate that ignoring jittering can lead to attenuation of associations and reduced predictive power when analysing DHS data. We show this via a spatial analysis of the prevalence of secondary education completion among women aged 20–39 in 2018 based on the 2018 Nigeria DHS (NDHS2018) (National Population Commission - NPC and ICF, 2019).

The new approach, which adjusts for jittering, and the standard approach, which ignores jittering, cannot be compared with cross-validation since the true coordinates of the clusters are not known. Therefore, we construct a simulation study that mimics the NDHS2018 dataset to perform the comparison in terms of their ability to estimate parameters and to predict risk at unobserved locations. We use bias and root mean square error (RMSE) to assess parameter estimation, and RMSE and continuous rank probability score (CRPS) (Gneiting and Raftery, 2007) to assess predictive ability.

In Section 2, we describe the new approach that adjusts for jittering, and discuss its implementation. In Section 3, we demonstrate the differences between adjusting and not adjusting for jittering when analysing the prevalence of completion of secondary education among women in

Nigeria.

Then in Section 4, we evaluate parameter estimation and prediction through the simulation study that mimics the data in the application. The paper ends with discussion and conclusions in Section 5. The code used for the paper can be found in the GitHub repository `https://github.com/umut-altay/GeoAdjust`.

2 Adjusting for jittering in a geostatistical model

2.1 Notation for DHS data

For a given country and DHS household survey, C clusters are visited. These clusters constitute small geographic areas and are collections of households. A total of n_c people at risk are observed and $y_c \leq n_c$ individuals have positive outcomes for clusters $c = 1, 2, \ldots, C,$. The reported GPS coordinates of the cluster centres are $\boldsymbol{s}_c \in \mathbb{R}^2$, $c = 1, \ldots, C$. These locations are not the true GPS coordinates, but the jittered GPS coordinates. Additionally, the urban/rural designation is known for each visited cluster.

2.2 Geostatistical model

2.2.1 Model for spatial variation in risk

We envision a spatially varying risk, $r(\cdot)$, for the country of interest $\mathcal{D} \subset \mathbb{R}^2$ modelled through

$$r(\boldsymbol{s}) = \text{logit}^{-1}(\eta(\boldsymbol{s})) = \text{logit}^{-1}(\boldsymbol{x}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + u(\boldsymbol{s})), \quad \boldsymbol{s} \in \mathcal{D},$$

where $\boldsymbol{x}(\cdot)$ is a p-dimensional vector of covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of covariate effect sizes, and $u(\cdot)$ is a Matérn GRF. The Matérn covariance function is parametrized as

$$\mathrm{C}_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2; \sigma^2_{\mathrm{S}}, \rho_{\mathrm{S}}) = \sigma^2_{\mathrm{S}} \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu} \frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\rho_{\mathrm{s}}} \right)^{\nu} \mathrm{K}_\nu \left(\sqrt{8\nu} \frac{||\boldsymbol{s}_1 - \boldsymbol{s}_2||}{\rho_{\mathrm{s}}} \right), \quad \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{D},$$

where σ^2_{S} is the marginal variance, ρ_{S} is the spatial range, and the smoothness is fixed to $\nu = 1$.

2.2.2 Unadjusted model

When jittering is ignored, the reported cluster locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_C$ are treated as the true locations. This gives the unadjusted observation model:

$$\begin{aligned} y_c | r_c, n_c &\sim \text{Binomial}(n_c, r_c), \\ r_c &= r(\boldsymbol{s}_c) = \text{logit}^{-1}(\eta(\boldsymbol{s}_c)), \end{aligned} \tag{1}$$

where r_c is the risk in cluster c, for $c = 1, \ldots, C$.

2.2.3 Adjusted model

Let $\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_C^* \in \mathcal{D}$ denote the true locations corresponding to the jittered locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_C$. The adjusted observation model is,

$$\begin{aligned} y_c | r_c, n_c &\sim \text{Binomial}(n_c, r_c), \quad \boldsymbol{s}_c | \boldsymbol{s}_c^* \sim \pi_{\text{Urb}[c]}(\boldsymbol{s}_c | \boldsymbol{s}_c^*), \\ r_c | \boldsymbol{s}_c^* &= r(\boldsymbol{s}_c^*) = \text{logit}^{-1}(\eta(\boldsymbol{s}_c^*)), \end{aligned} \tag{2}$$

where r_c is the risk in cluster c, and $\text{Urb}[c] \in \{\text{U}, \text{R}\}$ corresponds to the cluster's urban (U) or rural (R) designation, for $c = 1, \ldots, C$. In this observation model, both y_c and $\boldsymbol{s}_c$ are treated as observed quantities. The unobserved true locations s_c^* are treated as random quantities and assigned a uniform prior $s_c^* \sim \mathcal{U}(\mathcal{D})$. This implies that we treat all locations $\boldsymbol{s}_c^*$ within the maximum jittering distance from $\boldsymbol{s}_c$ as equally likely *a priori*.

The jittering distributions π_{U} and π_{R} follow from the (known) DHS jittering scheme. The country of interest is divided into K administrative regions. Let $A(\boldsymbol{s}) \in \{1, \ldots, K\}$ denote the administrative region of location $\boldsymbol{s}$ for $\boldsymbol{s} \in \mathcal{D}$. Then for an urban cluster c, which can be jittered up to 2 km, the jittering distribution is

$$\pi_{\text{U}}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \propto \frac{\mathbb{I}_(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*)) \cdot \mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 2)}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)}, \quad \boldsymbol{s}_c \in \mathcal{D},$$

where $d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)$ is the distance in kilometers between $\boldsymbol{s}_c$ and $\boldsymbol{s}_c^*$, and $\mathbb{I}$ is the indicator function. Similarly, for a rural cluster c, which can be jittered up to 5 km except for the 1 percent of clusters jittered up to 10 km, the jittering distribution is:

$$\pi_{\text{R}}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \propto \frac{\mathbb{I}(A(\boldsymbol{s}_c) = A(\boldsymbol{s}_c^*))}{d(\boldsymbol{s}_c, \boldsymbol{s}_c^*)} \left[\frac{99 \mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 5)}{100} + \frac{\mathbb{I}(d(\boldsymbol{s}_c, \boldsymbol{s}_c^*) < 10)}{100} \right], \quad \boldsymbol{s}_c \in \mathcal{D}.$$

2.2.4 Priors

We assume linear covariate associations, and use the prior $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, 25\mathbf{I}_p)$. The range, ρ_{S}, and marginal variance, σ_{S}^2, of the Matérn GRF is assigned a penalised complexity (PC) prior (Fuglstad et al., 2019). This requires selecting two hyperparameters: the *a priori* median of range R_0, and the *a priori* median of marginal variance V_0.

2.3 Implementation

2.3.1 Inference scheme

The observation model in Equation (2) can be written as,

$$\begin{aligned} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot)) &= \int_{\mathbb{R}^2} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot), \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^* \\ &= \int_{\mathbb{R}^2} \pi(y_c | \eta(\boldsymbol{s}_c^*)) \pi_{\text{Urb}[c]}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^*, \end{aligned} \tag{3}$$

for $c = 1, \ldots, C$. This formulation suggests that we can avoid sampling the true locations with an MCMC approach. Let $\boldsymbol{\theta} = (\log(\sigma_{\text{S}}^2), \log(\rho_{\text{S}}))$. We propose an empirical Bayes approach:

- **Step 1:** Calculate the maximum a posteriori (MAP) estimate, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ using $\pi(\boldsymbol{\theta}|y_1, \ldots, y_C, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_C)$.
- **Step 2:** Extract inference about $\boldsymbol{\beta}$ from $\pi(\boldsymbol{\beta}|y_1, \ldots, y_C, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_C, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$.
- **Step 3:** Estimate risk $r(\boldsymbol{s})$ at location $\boldsymbol{s}$ using $\pi(r(\boldsymbol{s})|y_1, \ldots, y_C, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_C, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}})$.

Two key components are combined for rapid inference: the SPDE approach to approximate the Matérn GRF (Lindgren et al., 2011), and TMB for empirical Bayesian inference (Kristensen et al., 2016).

2.3.2 SPDE approach

For each cluster c, the true location $\boldsymbol{s}_c^*$ is not known, and the observation model in Equation (3) involves the spatial field $u(\cdot)$ at all locations that are compatible with the jittered location $\boldsymbol{s}_c$. If we replace the integral in Equation (3) by a integration scheme using N_{Int} integration points, we need to evaluate the spatial field at $C \cdot N_{\text{Int}}$ locations. A standard implementation of the Matérn model would result in a dense $C \cdot N_{\text{Int}} \times C \cdot N_{\text{Int}}$ matrix and make computations infeasible even for a few locations.

To overcome these issues, the stochastic partial differential equations (SPDE) approach (Lindgren et al., 2011) provides an approximation to the Matérn GRF that results in a sparse precision matrix. First, the area of interest is triangulated with a triangulation consisting of m nodes. Then the GRF $u(\cdot)$ is approximated by

$$\tilde{u}(\boldsymbol{s}) = \sum_{i=1}^{m} w_i \phi_i(s), \tag{4}$$

where $\phi_i(\cdot)$ are pyramidal basis functions and $\boldsymbol{w} = (w_1 \ \ldots \ w_m)^{\mathrm{T}}$ are weights for the basis functions. The SPDE approach results in a distribution $\boldsymbol{w} \sim \mathcal{N}_m(\boldsymbol{0}, \mathbf{Q}(\boldsymbol{\theta})^{-1})$, where $\mathbf{Q}(\boldsymbol{\theta})$ is sparse.

From Equation (4), the value at any location is a linear transformation $\tilde{u}(\boldsymbol{s}) = \boldsymbol{a}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{w}$, $\boldsymbol{s} \in \mathcal{D}$, where $\boldsymbol{a}(\boldsymbol{s}) \in \mathbb{R}^m$ is sparse with at most three nonzero elements depending on the location $\boldsymbol{s}$. This means that the spatial field can be evaluated at a large number of locations quickly.

The SPDE is given by

$$(\kappa^2 - \Delta)(\tau u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \tilde{\mathcal{D}},$$

where $\kappa > 0$ and $\tau > 0$ are related to marginal variance and range, Δ is the Laplacian, $\mathcal{W}(\cdot)$ is standard Gaussian white noise, and $\tilde{\mathcal{D}} \supset \mathcal{D}$ is an extended domain to reduce boundary effects. We use Neumann boundary conditions to make the problem well defined, and following Lindgren et al. (2011), the effective range and marginal variance are calculated from the SPDE parameters as

$$\rho_{\text{S}} = \frac{\sqrt{8}}{\kappa} \quad \text{and} \quad \sigma_{\text{S}}^2 = \frac{1}{4\pi\tau^2\kappa^2}.$$

2.3.3 Template Model Builder

We implement the empirical Bayesian inference scheme by employing the built-in auto-differentiation and Laplace approximations of the Template Model Builder (TMB) R package. TMB allows us

Table 1: The three approaches considered.

Approach	Description
UnAdj	Ignore jittering for covariates and spatial effect.
CovAdj	Ignore jittering for the spatial effect.
FullAdj	Fully adjust for jittering.

to approximate the likelihood in Equation (3) through the integration scheme

$$\pi(y_c, \boldsymbol{s}_c|\eta(\cdot)) \propto \sum_{k=1}^{K} \alpha_k \pi(y_c|\eta(\boldsymbol{s}^*_{c,k}))\pi_{\text{Urb}[c]}(\boldsymbol{s}_c|\boldsymbol{s}^*_{c,k})\pi(\boldsymbol{s}^*_{c,k}), \tag{5}$$

where $\alpha_1, \ldots, \alpha_K$ are integration weights. More details are available in Altay et al. (2022). Critically, the integration scheme in Equation (5) involves

$$\eta(\boldsymbol{s}^*_{c,k}) = \boldsymbol{x}(\boldsymbol{s}^*_{c,k})^{\mathrm{T}}\boldsymbol{\beta} + u(\boldsymbol{s}^*_{c,k}), \quad k = 1, \ldots, K, \quad c = 1, \ldots, C.$$

Based on the known jittering distribution, we construct the integration scheme with rings of integration points around each cluster center. The observed cluster center is the first integration point, and we use 5 and 10 rings for the clusters that are located within urban and rural administrative areas, respectively. Each ring consists of 15 angularly equidistant integration points.

Through this paper we consider the three approaches shown in Table 1. FullAdj should be used if possible, but Altay et al. (2022) found that accounting for jittering in the spatial effect resulted in: 1) some numerical instability when the positional error grew large enough compared to the spatial range, and 2) no major changes except under wider jittering distributions than in the DHS scheme. Thus if CovAdj is more numerically stable or faster than FullAdj, CovAdj may be preferred. UnAdj ignores jittering and is included as a baseline for comparisons.

3 Analysis of completion of secondary education

Our outcome of interest is completion of secondary education, which is as an indicator of social well-being and life outcome (Lewin, 2008). Rates vary strongly between women and men, but also between urban and rural areas. According to (UNESCO, 2019), only 1% of the poorest girls in low income countries will complete secondary education. If a girl completes secondary education, the risk of HIV infection is reduced by about 50% (UNAIDS, 2022).

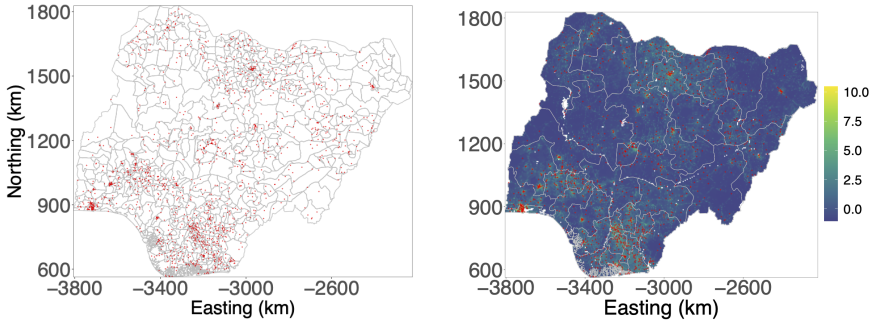
In this section, we model the prevalence of completion of secondary education among 20-39 year old women in Nigeria in 2018 based on the NDHS2018. Our analysis has two aims. First, to map the spatial variation in the risk of completion of secondary education for women aged 20–39 years in 2018 for 5 km × 5 km pixels and for the admin1 areas. Second, to determine the associations between the spatial variation in risk and a set of explanatory spatial covariates.

Nigeria is an LMIC with a population of more than 200 million. The first administrative level (admin1) consists of 37 admin1 areas, which are the 36 states and the federal capital territory, shown in Figure 1(b). The second administrative level (admin2) is nested within the admin1 areas and consists of the 774 local government areas (LGAs) shown in Figure 1(a). We use the national boundary, admin1 boundaries and admin2 boundaries specified by GADM version 4.0 (GADM, 2021).

Table 2: Summary of covariate rasters providing name, description and figure. CityA, Elev, DistW and UrbR are transformed, while UrbR is not.

Name	Description	Figure
PopD	Population count (250 m × 250 m)	1(b)
CityA	Travel time in minutes (1 km × 1 km)	2(a)
Elev	Elevation in meters (1 km × 1 km)	2(b)
DistW	Distance to nearest river or lake in degrees (1 km × 1 km)	2(c)
UrbR	Urbanicity ratio (250 m × 250 m)	2(d)

The NDHS2018 has $C = 1{,}380$ clusters with jittered GPS coordinates available under the same jittering distribution as in Section 2.2. For all clusters, the jittering was restricted to stay within the correct admin2 area. In total, 25,287 women aged 20–39 years were interviewed and 12,911 of these had completed secondary education. We use the notation n_c individuals-at-risk, y_c successes, and jittered GPS coordinate $\boldsymbol{s}_c$ for $c = 1, \ldots, C$.



(a) Admin2 areas.

(b) Population count and admin1 areas.

Figure 1: Maps of Nigeria with a) admin2 areas, and b) admin1 areas and $\log(1 + x)$-transformed and scaled 100 m × 100 m population count raster. The red dots are the jittered locations of the 1,380 clusters.

We expect the prevalence of completion of secondary education to be closely related to the access to educational resources, such as technological infrastructure, schools and teachers. We consider five spatial covariates: population count (PopD) (World Pop, 2022), travel time to nearest city (CityA) (Weiss et al., 2018), elevation (Elev) (National Oceanic and Atmospheric Administration, 2022), distance to nearest river or lake (DistW) (Natural Earth, 2012), and urbanicity ratio (Pesaresi et al., 2016). For UrbR, we use the original covariate, and for the other four covaraties, we use a $\log(1 + x)$-transformation and then center and standardize the covariate rasters accross the pixels. The information about the covariate rasters and figures is summarized in Table 2.

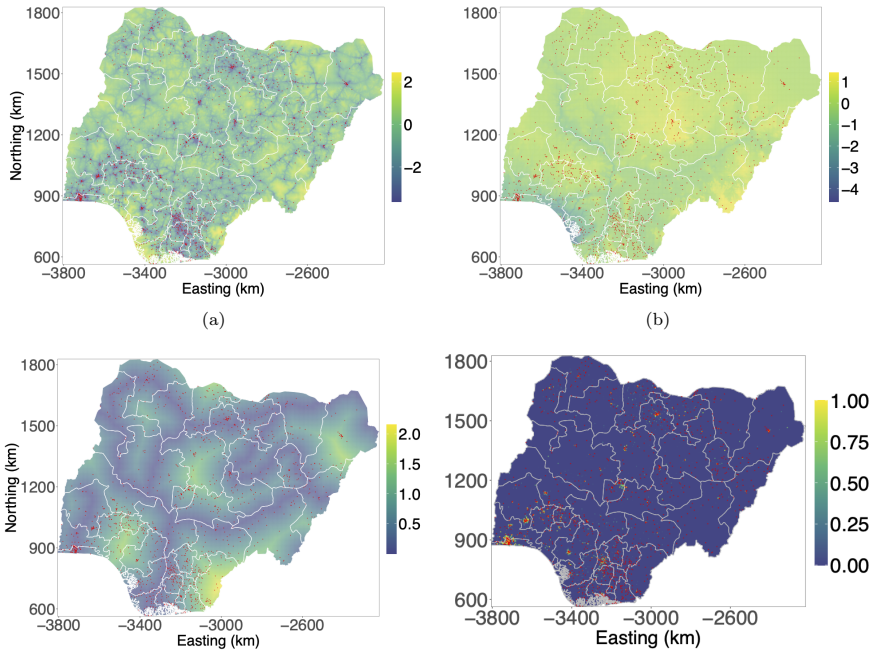


Figure 2: Covariate rasters for Nigeria: a) travel time to the nearest city center, b) elevation, c) minimum distance to nearest river or lake, and d) urbanicity ratio. The red dots indicate the (jittered) locations of $C = 1,380$ clusters.

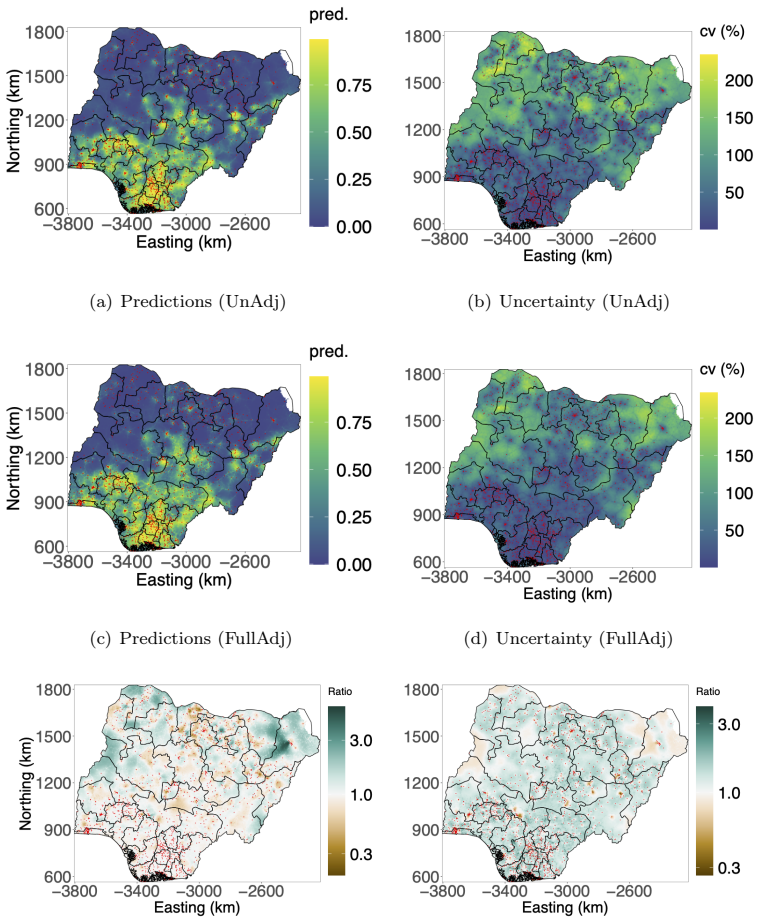
We fit the models UnAdj, CovAdj and FullAdj described in Section 2 using an intercept and the five covariates described above, and the PC prior on the Matérn GRF specified by $\mathrm{P}(\sigma_S > 1) = 0.05$ and $\mathrm{P}(\rho_S > R_0) = 0.50$, where $R_0 = 160\,\mathrm{km}$ is *a priori* median range. Inference is performed as described in Section 2.3, and Table 3 shows the estimated parameters and their corresponding credible interval lengths (except for ρ_S and σ_S^2, which are fixed to their MAP estimates). The parameters of the GRF, ρ_S and σ_S^2, are estimated lower for UnAdj than CovAdj and FullAdj.

This suggests that the noisy covariates and the positional uncertainty for the GRF are interpreted as reduced spatial dependency and reduced spatial signal strength. For PopD and UrbR, there is a strong attenuation when jittering is ignored. The credible intervals for the coefficient of UrbR, β_{UrbR}, suggest that β_{UrbR} is not significant at the 95% level for UnAdj, whereas β_{UrbR} is clearly significant for CovAdj and FullAdj. This suggest that not accounting for jittering can lead to misleading conclusions. Lastly, CovAdj and FullAdj give similar results, which indicates that accounting for the jittering in the covariates is more important than accounting for jittering for the spatial field.

Table 3: Parameter estimates and the corresponding 95% credible interval lengths in parentheses. Uncertainty is not computed for ρ_S and σ_S^2.

					Parameter			
Model	ρ_S	σ_S^2	μ	β_{DistW}	β_{CityA}	β_{Elev}	β_{PopD}	β_{UrbR}
UnAdj	68.60	2.00	-2.03 *(0.98)*	0.96*(1.35)*	-0.39 *(0.15)*	-0.20 *(0.77)*	0.19 *(0.11)*	-0.22 *(0.49)*
CovAdj	111.62	1.75	-1.97 *(1.14)*	0.62 *(1.39)*	-0.41 *(0.19)*	-0.01 *(0.71)*	0.40 *(0.16)*	-1.53 *(0.85)*
FullAdj	114.26	1.74	-1.98 *(1.11)*	0.64 *(1.39)*	-0.41 *(0.18)*	-0.01 *(0.71)*	0.40 *(0.16)*	-1.54 *(0.84)*

Figures 3(a) and 3(c) shows the pixel maps of predicted risk from UnAdj and FullAdj, respectively. Figures 3(b) and 3(d) show the corresponding coefficient of variation (CV) in percent. Figure 3(e) shows that some areas such as Borno (in the north-east) have up to three times the risk under the UnAdj approach as under FullAdj. And Figure 3(f) makes it clear that UnAdj tends to have higher uncertainty than FullAdj.



(a) Predictions (UnAdj)

(b) Uncertainty (UnAdj)

(c) Predictions (FullAdj)

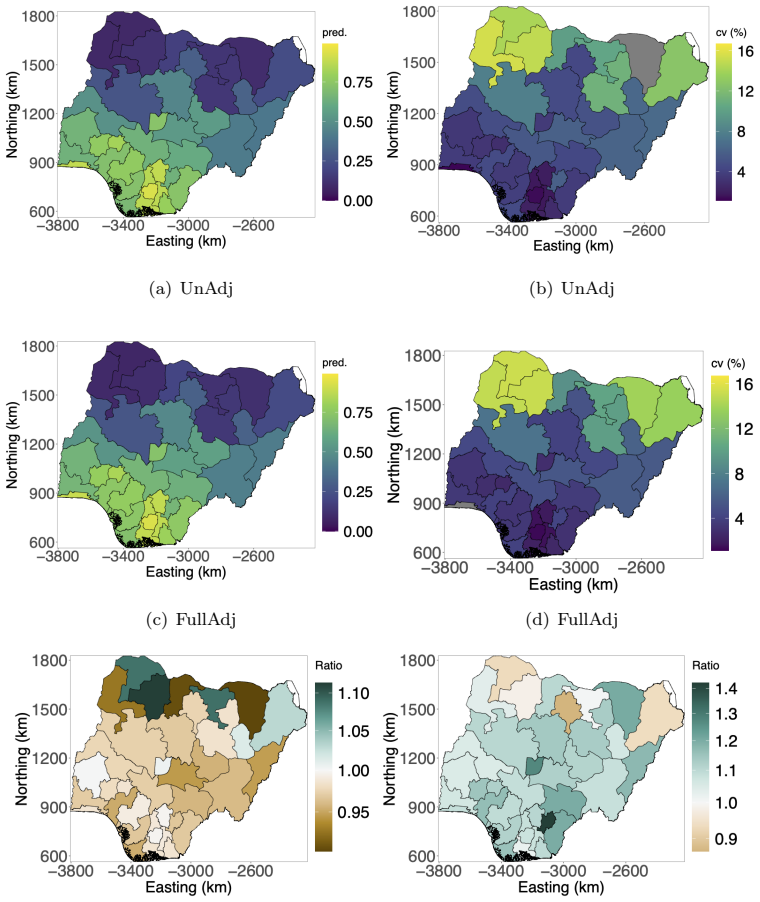
(d) Uncertainty (FullAdj)

(e) Ratio of predictions

(f) Ratio of uncertainties

Figure 3: Row 1 and 2 are predicted risk and the CVs, and row 3 shows ratios (UnAdj/FullAdj) of predictions and CVs. The red dots indicate the (jittered) locations of the 1,380 clusters.

We aggregate point level predictions with respect to population density to produce areal estimates at the 37 admin1 areas (for more on aggregating point level predictions with respect to a population, see Paige et al. 2022). Figures 4(a) and 4(c) show the predicted risk for UnAdj and CovAdj, respectively. And Figures 4(b) and 4(d) show the corresponding CVs. From Figure 4(e), we see that the point estimates vary from a factor 0.9 to 1.1, and Figure 4(f) shows that some areas differ with a factor of up to 1.4 in CV.



(a) UnAdj

(b) UnAdj

(c) FullAdj

(d) FullAdj

(e) Ratio of aggregated predictions

(f) Ratio of aggregated uncertainties

Figure 4: Rows 1 and 2 are predicted risk and CVs for UnAdj and FullAdj respectively at the admin1 level, and row 3 shows ratios (UnAdj/FullAdj) of predictions and CVs.

The ability to predict risk at unobserved locations for UnAdj, CovAdj and FullAdj cannot be compared with cross validation. If data is held-out from NDHS2018, we can only evaluate the models' abilities to predict risk at a new jittered cluster with unknown true location. Thus we aim to investigate if the differences we have seen for completion of secondary education in this section is consistent in a simulation study.

4 Simulation study

The aim of the simulation study is to compare the parameter estimation and prediction of the UnAdj, CovAdj and FullAdj approaches under the same design as the NDHS2018 survey used in Section 3. We fix the number of clusters to $C = 1{,}380$, and for each cluster c, we set its true location $\boldsymbol{s}_c^*$ and number-at-risk n_c to the observed location and number-at-risk respectively in the 2018NDHS, for $c = 1, \ldots, C$.

We assume that the true spatial risk varies as

$$r(\boldsymbol{s}) = \text{logit}^{-1}(\boldsymbol{x}(\boldsymbol{s})^{\text{T}}\boldsymbol{\beta} + u(\boldsymbol{s})), \quad \boldsymbol{s} \in \mathcal{D}, \tag{6}$$

where $\boldsymbol{x}(\cdot)$ is a 6-dimensional spatially varying vector with 1 as the first element and the covariates DistW, CityA, Elev, PopD, and UrbR as the five last elements,
$\boldsymbol{\beta} = (\mu, \beta_{\text{DistW}}, \beta_{\text{CityA}}, \beta_{\text{Elev}}, \beta_{\text{PopD}}, \beta_{\text{UrbR}})^{\text{T}}$, and $u(\cdot)$ is a Matérn GRF. We fix the spatial range ρ_{S} and the marginal variance σ_{S}^2 of the GRF to the values estimated using FullAdj in Table 1. Then we construct three scenarios based on the estimated coefficients under FullAdj in Table 1:

1. **SignalLow**: $\boldsymbol{\beta}$ is set to half the estimated value.
2. **SignalMed**: $\boldsymbol{\beta}$ is set to the estimated value.
3. **SignalHigh**: $\boldsymbol{\beta}$ is set to twice the estimated value.

For each scenario, we generate $n_{\text{sim}} = 50$ simulations by first generating the true risk surface $r(\cdot)$ using Equation (6). Then, for each cluster $c = 1, \ldots, C$, we simulate response $y_c | r(\boldsymbol{s}_c^*), n_c \sim \text{Binomial}(n_c, r(\boldsymbol{s}_c^*))$ and observed location $\boldsymbol{s}_c | \boldsymbol{s}_c^*$ according to the DHS jittering scheme. For each of these 150 datasets we apply the UnAdj, CovAdj and FullAdj approaches as in Section 3. The results for CovAdj are provided in the appendices since they exhibit the same behaviour as FullAdj.

Parameter estimation is evaluated by computing the RMSE, $\frac{1}{n_{\text{sim}}} \sum_{b=1}^{n_{\text{sim}}} (\hat{\theta} - \theta)^2$, and the Bias, $\frac{1}{n_{\text{sim}}} \sum_{b=1}^{n_{\text{sim}}} (\hat{\theta} - \theta)$, where $\hat{\theta}$ is the posterior mean (or MAP in the case of ρ_{S} and σ_{S}^2) and θ is the true value of the coefficient. Predictions are evaluated on a fixed set of 1,000 randomly selected locations within Nigeria, where we predict $\eta(\boldsymbol{s}) = \text{logit}(r(\boldsymbol{s}))$ with the posterior median. These predictions are evaluated by the average RMSE and CRPS defined by $\int_{\mathbb{R}^2} (F(x) - \mathbb{I}(y \leq x))^2 \text{d}x$, where y is the true value and $F(\cdot)$ is the predictive distribution.

Table 4 shows the Bias and RMSE for parameter estimation for UnAdj and FullAdj for the SignalMed scenario. The bias we observed in Section 3 for ρ_{S} and β_{UrbR} is consistent across simulations. RMSEs for FullAdj is lower or comparable to CovAdj for all parameters. The box plots in Figure 5 show that the differences are amplified when the strength of the signal of the spatial covariates is increased in SigHigh and reduced when the signal of the spatial covariates is decreased in SigLow.

Table 4: Bias and RMSE for parameter estimation.

	Model	Parameter ρ_S	σ_S^2	μ	β_{DistW}	β_{CityA}	β_{Elev}	β_{PopD}	β_{UrbR}
Bias	**UnAdj**	-21.96	0.01	0.37	-0.06	0.06	0.01	-0.16	0.88
	FullAdj	9.46	-0.01	0.38	-0.10	0.03	0.01	-0.10	0.44
RMSE	**UnAdj**	23.90	0.09	0.45	0.36	0.08	0.21	0.16	0.89
	FullAdj	15.09	0.11	0.45	0.36	0.05	0.17	0.10	0.47

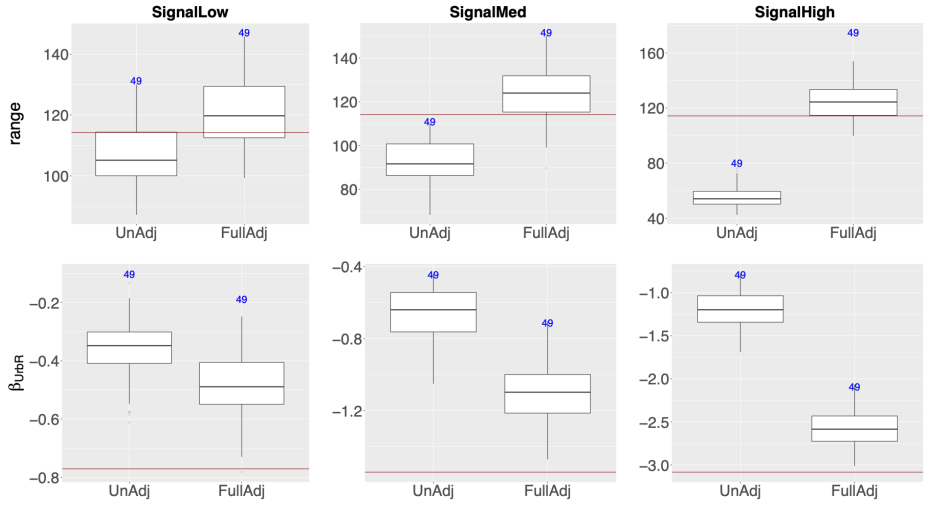


Figure 5: Box plots of estimated ρ_S and β_{UrbR} for SignalLow, SignalMed and SignalHigh. Numbers written in blue show the number of simulations that ran successfully. The horizontal red lines show the true parameter value.

Figure 6 shows the variation in RMSE and CRPS across datasets for predictions. We see that FullAdj and UnAdj perform the same in prediction for SignalLow, FullAdj is slightly better for SignalMed, and substantially better for SignalHigh. This indicates that the stronger the signal of the spatial covariates, the larger the gain from adjusting for jittering. The results for other parameters in all scenarios and for CovAdj can be found in Section B of appendices. One dataset was excluded in each scenario due to numerical issues with CovAdj or FullAdj.

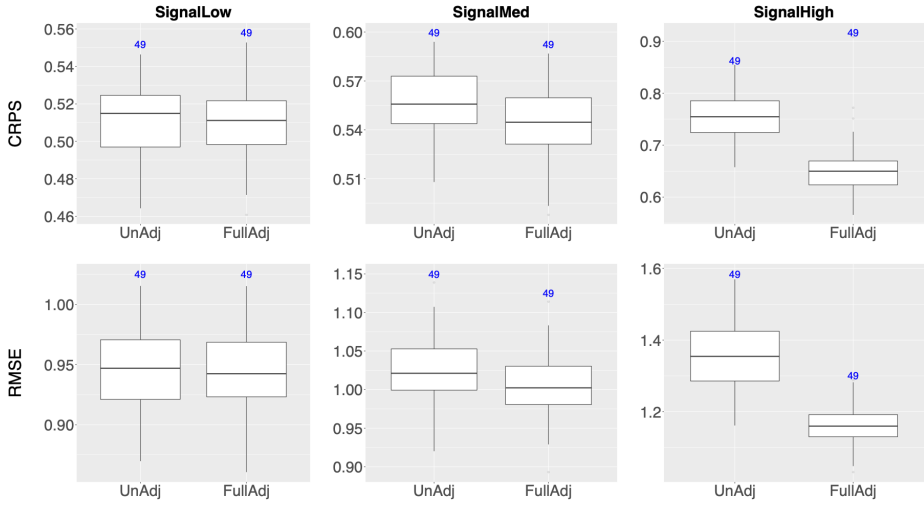


Figure 6: Box plots of CRPS and RMSE for predictions for SignalLow, SignalMed and SignalHigh.Numbers written in blue show the number of simulations that ran successfully for all models.

In influential work such as Burstein et al. (2019) and Local Burden of Disease Vaccine Coverage Collaborators (2021), covariates are resampled to a $5\,\text{km} \times 5\,\text{km}$ grid. This is similar, but slightly different than the buffer zone approach discussed in Section 1. In Section C of appendices, we show that this approach does not address the loss of predictive performance.

5 Discussion

Accounting for jittering substantially changed the parameter estimates for the geostatistical model for risk of completion of secondary education among women aged 20–39 years. The simulation study demonstrated that these differences were linked to the strength of the signal of the spatial covariates when explaining the spatial variation. For strong signals, the associations were attenuated and the predictive power was reduced.

The most important aspect of jittering in the context of geostatistical models for DHS data is to account for the resulting uncertainty in covariates extracted from rasters or extracted based on distances. This induces measurement error that may lead to attenuation in associations between covariates and the responses. Some covariates such as sanitation practices and household assets can be known exactly (Burgert-Brucker et al., 2016), but these cannot be included when the goal is prediction since fine-scale rasters are not available.

This work used uniform priors for the unknown true locations. One could expect including information about population density and urbanicity into the priors would produce more accurate inference. However, population density maps and urbanicity maps are also modelled surfaces with biases and uncertainties that are not well understood. This means that evaluation of the

sensitivity to such maps would have to be investigated, and one would need a way to evaluate whether such a model works better.

The inference scheme uses empirical Bayes. It is possible to investigate methods such as INLA, but the implementation in the R package `inla` does not allow the likelihood to depend on the latent risk at multiple locations, which is necessary due to the integration points. MCMC algorithms such as STAN (Stan Development Team, 2020) has the required flexibility, but is infeasible for thousands of spatial locations.

When analysing completion of secondary education, we found that, for urbanicity, an effect size of 0 was contained in the 95% credible interval when not adjusting for jittering, and not contained when adjusting for jittering. This suggests that not accounting for jittering when analysing DHS data is a practice that can alter conclusions about statistical significance. Since the proposed approach is fast for spatial analysis, we suggest to use the new approach for analysing DHS data to avoid the risk of misleading conclusions and reduced predictive power.

References

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022). Accounting for spatial anonymization in DHS household surveys. *arXiv preprint arXiv:2202.11035*.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). Geographic displacement procedure and georeferenced datarelease policy for the Demographic and Health Surveys. `https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf`. DHS Spatial Analysis Reports No. 7.

Burgert-Brucker, C. R., Domtamsetti, T., Marshall, A. M., and Gething, P. (2016). Guidance for use of the DHS program modeled map surfaces. Spatial Report No. 14.

Burstein, R., Henry, N. J., Collison, M. L., Marczak, L. B., Sligar, A., Watson, S., Marquez, N., Abbasalizad-Farhangi, M., Abbasi, M., Abd-Allah, F., et al. (2019). Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature*, 574(7778):353–358.

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

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

Fronterrè, C., Giorgi, E., and Diggle, P. (2018). Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *Spatial Statistics*, 28:319–330.

Fuglstad, G.-A., Li, Z. R., and Wakefield, J. (2021). The two cultures for prevalence mapping: Small area estimation and spatial statistics. *arXiv preprint arXiv:2110.09576*.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114:445–452.

GADM (2021). GADM (version 4.0). `https://gadm.org/download_country.html`. Accessed: 2022-02-20.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gómez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5):1–21.

Lewin, K. M. (2008). *Strategies for sustainable financing of secondary education in Sub-Saharan Africa*, volume 136. World Bank Publications.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73:423–498.

Local Burden of Disease Vaccine Coverage Collaborators (2021). Mapping routine measles vaccination in low-and middle-income countries. *Nature*, 589(7842):415–419.

National Oceanic and Atmospheric Administration (2022). National centers for environmental information.

National Population Commission - NPC and ICF (2019). Nigeria Demographic and Health Survey 2018 - final report. `http://dhsprogram.com/pubs/pdf/FR359/FR359.pdf`.

Natural Earth (2012). Rivers + lake centerlines.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2022). Spatial aggregation with respect to a population distribution: Impact on inference. *Spatial Statistics*. In press.

Perez-Heydrich, C., Warren, J., Burgert, C., and Emch, M. (2013). Guidelines on the use of DHS GPS data. *ICF International, Calverton, Maryland*. Spatial analysis reports no. 8.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., and Emch, M. E. (2016). Influence of Demographic and Health Survey point displacements on raster-based analyses. *Spatial demography*, 4(2):135–153.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., Syrris, V., et al. (2016). Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. *Publications Office of the European Union*, pages 1–62.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Sstatistical Methodology)*, 71(2):319–392.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

UNAIDS (2022). School saves lives: World leaders back a courageous goal, "Education Plus", to prevent new HIV infections through education and empowerment.

UNESCO (2019). Her education, our future: UNESCO fast-tracking girls' and women's education.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., Cutts, F. T., and Tatem, A. J. (2019). Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nature communications*, 10(1):1–10.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016). Influence of demographic and health survey point displacements on distance-based analyses. *Spatial Demography*, 4(2):155–173.

Weiss, D. J., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100421.

World Pop (2022). Open spatial demographic data and research.

Supplementary materials: Jittering Impacts Raster- and Distance-based Geostatistical Analyses of DHS Data

Umut Altay John Paige Andrea Riebler
Geir-Arne Fuglstad

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

1 Introduction

This document consists of the supplementary results and materials for our paper titled "Jittering Impacts Raster- and Distance-based Geostatistical Analyses of DHS Data". In the main paper, we presented the results that are obtained by fitting the models that doesn't account for jittering (UnAdj), accounts for jittering only in covariates (CovAdj) and both in covariates and the spatial random effect (FullAdj), to NDHS2018 survey data. Section 2 shows these results in comparison with the ones that are obtained from the model that accounts for jittering only in the random effect (RandAdj). Section 3 presents the supplementary results of the simulation study. Section 4 compares the results based on the smoothed covariates, to the results that are obtained from the UnAdj and FullAdj models.

2 Additional results for analysis of NDHS2018 data set

This section presents the results that are obtained by fitting UnAdj, CovAdj, RandAdj and FullAdj models to NDHS2018 survey data.

Table 1: Parameter estimates and the corresponding 95% credible intervals obtained from fitting UnAdj, CovAdj, RandAdj and FullAdj models to NDHS-2018 secondary education completion data for 20-39 years old women. Numbers in the parantheses show the 95% credible interval lengths for the corresponding parameter estimates (except for the spatial range and the marginal variance).

	Models			
Parameter	UnAdj	CovAdj	RandAdj	FullAdj
ρ	68.60	111.62	77.81	114.26
σ^2	2.00	1.75	1.94	1.74
μ	-2.03 *(0.98)*	-1.97 *(1.14)*	-2.03 *(0.99)*	-1.98 *(1.11)*
β_{dist}	0.96 *(1.35)*	0.62 *(1.39)*	0.93 *(1.34)*	0.64 *(1.39)*
β_{tTime}	-0.39 *(0.15)*	-0.41 *(0.19)*	-0.38 *(0.14)*	-0.41 *(0.18)*
β_{elev}	-0.20 *(0.77)*	-0.01 *(0.71)*	-0.10 *(0.74)*	-0.01 *(0.71)*
β_{pop}	0.19 *(0.11)*	0.40 *(0.16)*	0.20 *(0.11)*	0.40 *(0.16)*
β_{urb}	-0.22 *(0.49)*	-1.53 *(0.85)*	-0.24 *(0.49)*	-1.54 *(0.84)*

3 Supplementary results for the simulation study

This section shows the average bias and RMSE values of the model parameter estimates that are obtained from the simulation study. Tables 2, 3 and 4 contain the results for SignalLow, SignalMed and SignalHigh signal strength levels, based on the parameter estimates obtained by fitting the FullAdj model to NDHS2018 data set. Smoothed model doesn't account for jittering, but it uses data from the smoothed versions of covariate rasters. The predictive measures sections of the tables show the bias and RMSE that are calculated from the predictions with the corresponding model.

Table 2: Average bias and RMSE of model parameter estimates across 49 simulations, together with the average predictive measures (RMSE and CRPS) for UnAdj, Smoothed, RandAdj, CovAdj and FullAdj models. The results belong to the signal strength level "SignalLow". The values in the "Parameter" section of the table show the biases for the corresponding parameter estimates, together with the RMSE in the parantheses .

	Unadjusted		**Adjusted**		
Model	**UnAdj**	**Smoothed**	**RandAdj**	**CovAdj**	**FullAdj**
Parameter					
ρ	-7.19 *(13.17)*	-6.97 *(12.94)*	0.52 *(11.23)*	-0.64 *(11.76)*	6.61 *(13.53)*
σ^2	0.03 *(0.10)*	0.04 *(0.10)*	0.03 *(0.10)*	0.03 *(0.10)*	0.03 *(0.10)*
μ	0.16 *(0.30)*	0.11 *(0.27)*	0.17 *(0.30)*	0.18 *(0.30)*	0.18 *(0.30*
β_{dist}	-0.09 *(0.39)*	-0.03 *(0.48)*	-0.09 *(0.39)*	-0.10 *(0.38)*	-0.09 *(0.38)*
β_{tTime}	0.02 *(0.03)*	-0.13 *(0.15)*	0.02 *(0.03)*	0.03 *(0.04)*	0.02 *(0.03)*
β_{elev}	0.01 *(0.19)*	-0.01 *(0.21)*	0.01 *(0.19)*	0.01 *(0.15)*	0.01 *(0.15)*
β_{pop}	-0.08 *(0.08)*	0.12 *(0.14)*	-0.08 *(0.08)*	-0.06 *(0.07)*	-0.06 *(0.06)*
β_{urb}	0.41 *(0.42)*	-0.94 *(1.05)*	0.40 *(0.42)*	0.32 *(0.34)*	0.29 *(0.31)*
Predictive measures					
RMSE	0.94	0.94	0.94	0.94	0.94
CRPS	0.51	0.51	0.51	0.51	0.50

Table 3: Average bias and RMSE of model parameter estimates across 49 simulations, together with the average predictive measures (RMSE and CRPS) for UnAdj, Smoothed, RandAdj, CovAdj and FullAdj models. The results belong to the signal strength level "SignalMed". The values in the "Parameter" section of the table show the biases for the corresponding parameter estimates, together with the RMSE in the parantheses .

	Unadjusted		**Adjusted**		
Model	**UnAdj**	**Smoothed**	**RandAdj**	**CovAdj**	**FullAdj**
Parameter					
ρ	-21.96 *(23.90)*	-20.01 *(21.85)*	-14.38 *(17.39)*	3.85 *(12.21)*	9.46 *(15.09)*
σ^2	0.01 *(0.09)*	0.03 *(0.10)*	0.005 *(0.09)*	-0.01 *(0.11)*	-0.01 *(0.11)*
μ	0.37 *(0.45)*	0.26 *(0.36)*	0.38 *(0.45)*	0.38 *(0.45)*	0.38 *(0.45*
β_{dist}	-0.06 *(0.36)*	0.05 *(0.44)*	-0.08 *(0.35)*	-0.10 *(0.36)*	-0.10 *(0.36)*
β_{tTime}	0.06 *(0.08)*	-0.24 *(0.26)*	0.06 *(0.07)*	0.04 *(0.05)*	0.03 *(0.05)*
β_{elev}	0.01 *(0.21)*	-0.02 *(0.25)*	0.01 *(0.20)*	0.007 *(0.17)*	0.01 *(0.17)*
β_{pop}	-0.16 *(0.16)*	0.26 *(0.28)*	-0.15 *(0.16)*	-0.10 *(0.11)*	-0.10 *(0.10)*
β_{urb}	0.88 *(0.89)*	-1.71 *(1.84)*	0.87 *(0.88)*	0.48 *(0.51)*	0.44 *(0.47)*
Predictive measures					
RMSE	1.02	1.02	1.02	1.00	1.00
CRPS	0.55	0.57	0.55	0.54	0.54

Table 4: Average bias and RMSE of model parameter estimates across 49 simulations, together with the average predictive measures (RMSE and CRPS) for UnAdj, Smoothed, RandAdj, CovAdj and FullAdj models. The results belong to the signal strength level "SignalHigh". The values in the "Parameter" section of the table show the biases for the corresponding parameter estimates, together with the RMSE in the parantheses .

	Unadjusted		**Adjusted**		
	UnAdj	**Smoothed**	**RandAdj**	**CovAdj**	**FullAdj**
Parameter					
ρ	-58.59 *(59.19)*	-53.48 *(54.08)*	-52.17 *(52.77)*	9.41 *(16.79)*	11.74 *(18.44)*
σ^2	0.26 *(0.28)*	0.30 *(0.32)*	0.19 *(0.22)*	-0.02 *(0.14)*	-0.03 *(0.14)*
μ	0.68 *(0.74)*	0.36 *(0.46)*	0.67 *(0.73)*	0.51 *(0.58)*	0.52 *(0.58)*
β_{dist}	-0.13 *(0.39)*	0.02 *(0.45)*	-0.15 *(0.39)*	-0.14 *(0.37)*	-0.15 *(0.37)*
β_{tTime}	0.16 *(0.17)*	-0.49 *(0.50)*	0.15 *(0.16)*	0.04 *(0.06)*	0.04 *(0.05)*
β_{elev}	0.001 *(0.27)*	0.002 *(0.31)*	0.0009 *(0.26)*	-0.05 *(0.21)*	-0.05 *(0.21)*
β_{pop}	-0.37 *(0.37)*	0.43 *(0.44)*	-0.36 *(0.37)*	-0.16 *(0.16)*	0.15 *(0.16)*
β_{urb}	1.88 *(1.90)*	-3.22 *(3.33)*	1.85 *(1.87)*	0.53 *(0.57)*	0.50 *(0.54)*
Predictive measures					
RMSE	1.35	1.30	1.39	1.16	1.16
CRPS	0.75	0.75	0.90	0.65	0.65

4 Comparison of predictive measures against the smoothed covariates approach

This section shows the comparison of predictive measures that are obtained by fitting UnAdj, FullAdj and Smoothed models to the simulated data sets. The new data is simulated according to three different signal strength levels (SignalLow, SignalMed and SignalHigh, respectively), based on the parameter estimates from the FullAdj model on NDHS2018 data set. The blue numbers on top of each boxplot show the total number of simulations (out of 50) that ran without any numerical instabilities for the corresponding model.

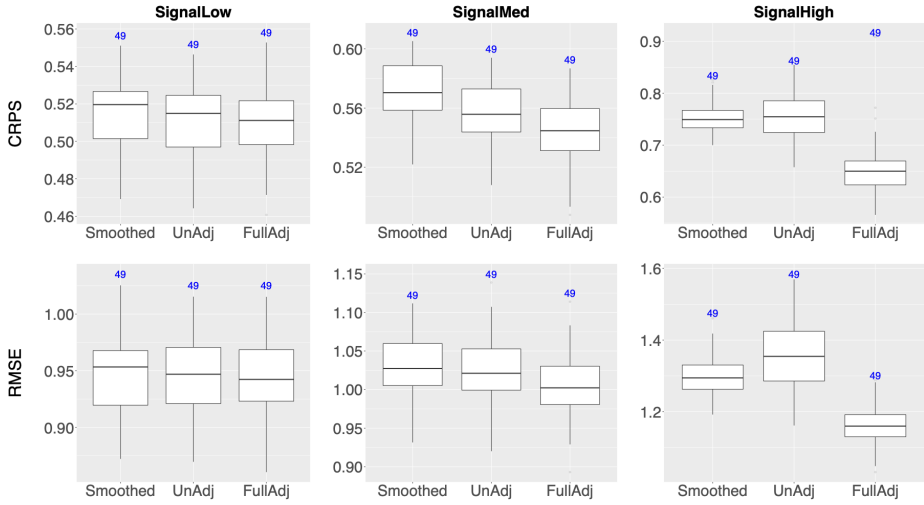


Figure 1: Box plots of CRPS and RMSE values that are obtained from the predictions with Smoothed, UnAdj and FullAdj models. SignalLow, SignalMed and SignalHigh indicate three different signal strength levels, respectively. Numbers written in blue on top of each box plot show the number of simulations (out of 50 for each level of signal strength) that ran without any numerical instabilities for the corresponding model.

Paper III

GeoAdjust: Adjusting for positional uncertainty in geostatistical analysis of DHS data

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A.

Submitted to R Journal on 21.03.2023

GeoAdjust: Adjusting for Positional Uncertainty in Geostatistial Analysis of DHS Data

Umut Altay John Paige Andrea Riebler
Geir-Arne Fuglstad

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

The R-package `GeoAdjust` implements fast empirical Bayesian geostatistical inference for household survey data from the Demographic and Health Surveys Program (DHS) using Template Model Builder (TMB). DHS household survey data is an important source of data for tracking demographic and health indicators, but positional uncertainty has been intentionally introduced in the GPS coordinates to preserve privacy. `GeoAdjust` accounts for such positional uncertainty in geostatistical models containing both spatial random effects and raster- and distance-based covariates. The R package supports Gaussian, binomial and Poisson likelihoods with identity link, logit link, and log link functions respectively. The user defines the desired model structure by setting a small number of function arguments, and can easily experiment with different hyperparameters for the priors. `GeoAdjust` is the first software package that is specifically designed to address positional uncertainty in the GPS coordinates of point referenced household survey data. The package provides inference for model parameters and can predict values at unobserved locations.

1 Introduction

In each demographic health survey collected by the DHS program, positional uncertainty is intentionally introduced in the GPS coordinates of the household cluster centers as a privacy protection measure(Burgert et al., 2013).

The random displacement procedure, or *jittering* scheme, is publicly known (Burgert et al., 2013). The jittering can be an issue because traditional geostatistical analyses assume locations are known exactly, and we

have recently shown that ignoring the positional error in DHS data may lead to attenuated estimates of the covariate effect sizes and reduced predictive performance (Altay et al., 2022a,b).

While common practice is to ignore jittering, some approaches have been put forward to account for it. With respect to the error induced in spatial covariates, Warren et al. (2016) proposed regression calibration for distance-based covariates, and Perez-Heydrich et al. (2013, 2016) proposed using a 5 km moving window (or buffer zone) for raster-based covariates. However, these approaches do not address the attenuation arising in the covariate effect sizes when replacing the true covariate with a proxy. With respect to the error induced in the spatial effect, Fanshawe and Diggle (2011) proposed a Bayesian approach in the limited setting of no covariates and Gaussian observation model. Wilson and Wakefield (2021) proposed a more complex approach using INLA-within-MCMC (Rue et al., 2009; Gómez-Rubio and Rue, 2018), which could handle the error induced in both the spatial random effect and in spatial covariates, but computation time is too extensive for routine use of the approach. None of the mentioned papers provide an R package for easy application of the methods.

With the package `GeoAdjust` we address the need for fast, flexible and user-friendly software to estimate geostatistical models for DHS data subject to positional uncertainty. `GeoAdjust` addresses the positional uncertainty by adjusting for jittering both in the spatial random effect and spatial covariates, and achieves fast inference by combining the computational feasibility of the stochastic partial differential equations (SPDE) approach (Lindgren et al., 2011) with the autodifferentiation feature of the Template Model Builder (TMB) R-package (Kristensen et al., 2016a). The R-package GeoAdjust is on CRAN (R Core Team, 2022) and can be installed by install.packages("GeoAdjust") command.

2 Geostatistical inference under jittering

We consider a country with spatial domain $\mathcal{D} \subset \mathbb{R}^2$, where C small groups of households, called *clusters*, are observed. For clusters $c = 1, \ldots, C$, we denote the true location by $\boldsymbol{s}_c^* \in \mathcal{D}$, and we denote the observed location, provided by DHS, by $\boldsymbol{s}_c \in \mathcal{D}$. Additionally, each cluster has a known classification as urban (U) or rural (R). The observed locations are linked to the true locations via a known jittering distribution $\pi_{\text{Urb}[c]}(\boldsymbol{s}_c|\boldsymbol{s}_c^*)$. The subscript $\text{Urb}[c] \in \{\text{U}, \text{R}\}$ is necessary since the DHS uses different jittering distributions in urban and rural clusters. Urban clusters are jittered up to 2 km, and rural clusters are jittered up to 5 km with probability 0.99 and jittered up to 10 km with probability 0.01 (Burgert et al., 2013). The angle and jittering distance are sampled from uniform distributions, but the boundaries of either the first or the second administrative level are respected.

We model responses $y_1, \ldots, y_C$ and observed locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_C$ jointly as

$$\begin{aligned} y_c \mid \eta_c, \boldsymbol{\phi} &\sim \pi(y_c \mid \eta_c, \boldsymbol{\phi}), \quad \boldsymbol{s}_c | \boldsymbol{s}_c^* \sim \pi_{\text{Urb}[c]}(\boldsymbol{s}_c | \boldsymbol{s}_c^*), \\ \eta_c &= \eta(\boldsymbol{s}_c^*), \end{aligned} \tag{2.1}$$

for $c = 1, \ldots, C$, where $\pi(y_c \mid \eta_c, \boldsymbol{\phi})$ is the likelihood of y_c with linear predictor η_c and likelihood parameter vector $\boldsymbol{\phi}$, and $\eta(\cdot)$ is a Gaussian random field describing spatial variation. The linear predictor is linked to the mean of the likelihood family through a link function. The package implements the identity link in the case of a Gaussian likelihood, the log-link for Poisson and the logit-link for the binomial likelihood.

The latent spatial variation is modelled as

$$\eta(\boldsymbol{s}^*) = \boldsymbol{x}(\boldsymbol{s}_c^*)^{\mathrm{T}} \boldsymbol{\beta} + u(\boldsymbol{s}_c^*), \quad \boldsymbol{s}^* \in \mathcal{D},$$

which combines p spatial covariates, $\boldsymbol{x}(\cdot)^{\mathrm{T}}$, with a Matérn Gaussian random field (GRF), $u(\cdot)$, with fixed smoothness $\nu = 1$. The coefficients of the covariates are assigned a Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, V\mathbf{I}_p)$, where V is a fixed variance. The parameters of the Matérn GRF, the spatial range ρ_{S} and marginal variance σ_{S}^2, are assigned penalized complexity (PC) priors with $\mathrm{P}(\rho_{\mathrm{S}} > \rho_0) = 0.50$ and $\mathrm{P}(\sigma_{\mathrm{S}} > 1) = 0.05$ (Fuglstad et al., 2019). We recommend choosing the median range ρ_0 as the 10% of the diameter of $\mathcal{D}$ to be able to capture the spatial variability at moderate distances.

To complete the specification of the model, we need to assign a prior for the true cluster locations. We choose a uniform prior $s_c^* \sim \mathcal{U}(\mathcal{D})$ so that all $\boldsymbol{s}_c^*$ compatible with $\boldsymbol{s}_c$ are equally likely *a priori*, $c = 1, \ldots, C$. More complicated choices taking population density or urban/rural status into account are possible, but such rasters would have to be estimated and could be biased and uncertain. `GeoAdjust` treats the unknown true locations as nuisance parameters and integrates them out,

$$\begin{aligned} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot)) &= \int_{\mathcal{D}} \pi(y_c, \boldsymbol{s}_c | \eta(\cdot), \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^* \\ &= \int_{\mathcal{D}} \pi(y_c | \eta(\boldsymbol{s}_c^*)) \pi_{\text{Urb}[c]}(\boldsymbol{s}_c | \boldsymbol{s}_c^*) \pi(\boldsymbol{s}_c^*) \, \mathrm{d}\boldsymbol{s}_c^*. \end{aligned} \tag{2.2}$$

We use the SPDE approach (Lindgren et al., 2011) to describe $u(\cdot)$ and use the speed and flexibility of autodifferentiation in TMB to perform inference quickly.

3 Package structure and functionality

`GeoAdjust` hides the complicated and technical steps in the algorithm from the user, to make the adjustment for jittering widely accessible. Figure

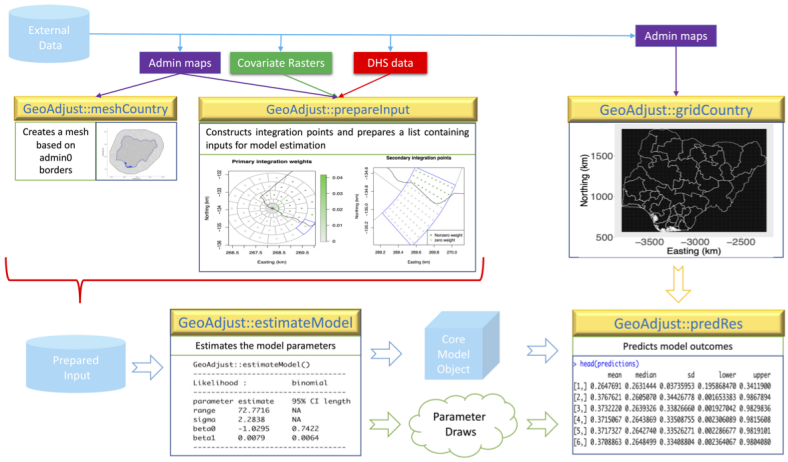


Figure 1: `GeoAdjust` R-package workflow

1 illustrates the structure of `GeoAdjust`, and how various data inputs are processed through the package workflow. The main functionality of the package is described in subsections.

3.1 Triangulation and mesh generation for the region under study

In `GeoAdjust`, the GRF $u(\cdot)$ is approximated using the so-called SPDE approach. This requires the construction of a constrained refined Delaunay triangulation (CRDT), in other words a mesh, over the country of interest. The approximated spatial field can then be projected from the mesh nodes to the cluster centers, by projector matrices (Lindgren et al., 2011). The function `meshCountry` creates a triangulation mesh based on the national borders. It has two key arguments: `max.edge` is a vector of two values, where its first and second elements represent the largest allowed triangle edge lengths for the inner and outer mesh, respectively, and `offset` stands for the extension distance outside the country borders.

3.2 Input data preparation

The integration in Equation (2.2) is done numerically and we need a set of integration points around each jittered survey cluster center. `GeoAdjust`

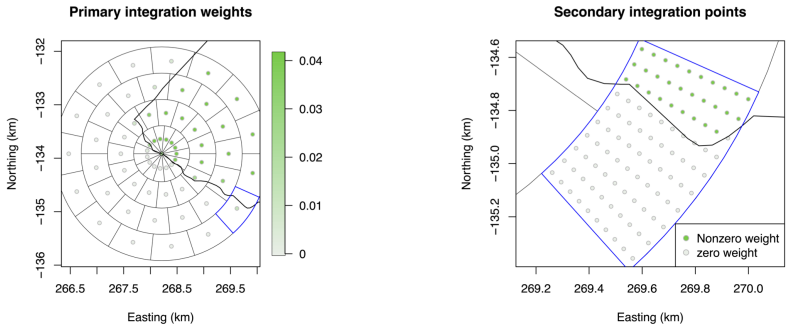


Figure 2: Illustration of primary (left) and secondary (right) integration weights for one cluster from Kenya 2014 DHS household survey.

specifies the cluster center itself as the first integration point and builds either 5 or 10 rings around it, depending on whether it is located in an urban or a rural stratum, respectively. Each ring contains a set of 15 angularly equidistant primary integration points.

The first 5 rings are called the "inner rings" and the points located within them are weighted equally. The additional 5 rings are built for the rural cluster centers and are called the "outer rings". The primary integration points within them are also assigned equal weights, which are smaller than the ones assigned for the points within the inner rings. If an observed cluster location is closer to the nearest subnational border than the maximum jittering distance, the method deploys a set of secondary integration points, each with an associate primary integration point, and assigns zero weight to any that are across the border. The associated primary integration point weights are adjusted accordingly. Figure 2 shows the primary and secondary integration points and the corresponding integration weights, for a single cluster from the Kenya 2014 DHS household survey with observed location close to an administrative boundary. The supplementary materials section of (Altay et al., 2022a) provides a detailed mathematical explanation about the procedure.

The function `prepareInput` creates the set of integration points and the corresponding weights with respect to the urbanization strata, and constructs the urban and rural design matrices by extracting the covariate values at the coordinates of each integration point. The function returns a list containing the strata-wise design matrices and response vectors, together with the sparse matrix components of the SPDE model, and strata-wise projector matrices. Usage of the function will be shown based on the NDHS-2018 survey in Section 5.2.

3.3 Model parameter estimation

The input list returned by `prepareInput` function consists of the elements that will be processed by the autodifferentiation feature of TMB as implemented in the `estimateModel` function. The function `estimateModel` saves the users from writing a complex C++ code to run TMB, and allows estimating model parameters by setting a small number of arguments. The function is flexible and allows different prior choices for the model components, via its argument called `priors`.

The function `estimateModel` utilizes the C++ script of TMB, which integrates out the unknown true coordinates by computing the contribution of each integration point to the joint negative log-likelihood. Internally, once the TMB function `MakeADFun` constructs the core model object (Kristensen, 2022), `estimateModel` uses the `optim` function to optimize it. Afterwards, `estimateModel` extracts the estimated model parameters from the optimized core model object, and draws samples of size, that is controlled by the argument `n.sims` of `estimateModel`function, for each covariate effect. The function draws samples of size `n.sims` for the spatial random effect coefficients for each mesh node as well. The samples for the intercept and the covariate effect sizes are then used for constructing the 95% credible interval lengths as the measure of uncertainty corresponding to the estimated parameters.

The function `estimateModel` returns a list of four elements. The list contains a data frame of the estimated model parameters, together with the optimized core model object, a matrix containing the drawn samples of size `n.sims` for the covariate effect sizes and the random effect coefficients, and information about the type of the likelihood. The core model object and the drawn samples can then be passed to the function `predRes` to generate predictions at a set of prediction locations.

3.4 Prediction grid construction

Once the model parameters are estimated, the model can be used for predicting the model outcomes at a new set of locations. The function `gridCountry` in `GeoAdjust` helps with the construction of a set of prediction points. The function creates a raster of a preferred resolution within the bounding box of the national level shape file, extracts the coordinates of the cell centers and returns them together with the raster, as the elements of a list. A code example about the implementation of this function is given in Section 5.2.

3.5 Prediction

Obtaining predictions at a new set of locations the function `predRes` requires the optimized core model object, drawn samples of the parameters

and the random effect coefficients, triangulation mesh, a list of covariate rasters, coordinates of the prediction locations and an argument called "flag" as input arguments. The argument "flag" is used for passing the likelihood type into the function. The integers 0, 1 and 2 indicate the Gaussian, binomial and Poisson likelihoods, respectively, and the function deploys the corresponding link function as outlined before. The package allows the use of any number of covariates, as long as they are in a geospatial raster layer format. The covariates can be passed into the function within a single list. The function will extract the values from each one of them at the prediction locations and form a design matrix.

Internally, `predRes` combines the sampled covariate effect sizes and the random effect coefficients with the design matrix and forms one model per sample, `n.sims` models in total. Each model predicts outcomes across the set of prediction locations. Finally, the function calculates the mean, median, standard deviation, and the upper and lower bounds of 95% credible intervals of predictions for each prediction point. These results are returned in a matrix with a number of rows equal to the number of prediction points, and 5 columns.

4 DHS data acquisition

The data sets of DHS household surveys are semi-public, but access to them requires application for permission. A step by step guidance to the application procedure can be found in `https://dhsprogram.com/data/new-user-registration.cfm`. The application requires a brief project description explaining why the data set is needed and how it will be used within the provided project framework. Sharing the data sets with each one of the collaborating researchers requires permission as well. Once the permission is granted, the applicant receives a letter via email, which clearly states the content of the permission.

5 Preprocessing DHS data

Geospatial analysis of DHS household surveys usually require processing the individual level responses together with the cluster level information. Prior to using `GeoAdjust`, the DHS data needs to be preprocessed and certain variables need to be extracted. The package uses the clusterID, cluster center coordinates (both in degrees and in kilometers), administrative area names in which the clusters are located in, and the Gaussian, binomial or Poisson outcome variable aggregated from the individual responses into each cluster center. Two administrative border shape files are used in the analysis. One of them is the shape file that contains the national (admin0)

level borders of the country of interest. The second one contains the sub-national administrative level borders which are respected while jittering. Shape files of various administrative levels for different countries can be obtained from the website of "the Database of Global Administrative Areas (GADM)" (`https://gadm.org/data.html`). Once downloaded, the files can be read into R as "SpatialPolygonsDataFrame" objects, see Listing 1 for an example. Finally, if any raster- and distance-based covariates will be included in the model, they need to be read into R as separate raster layers. The R code for reading and further processing the administrative borders shape files and the covariate rasters will be shown in Section 5.2. Once the preprocessing is done, in other words, all the external data files are read in and the variables of interest are extracted and stored in a data frame, the functions `meshCountry` and `prepareInput` can be used (see Section 5.2).

5.1 Reading data

The individual responses and the cluster information are often contained in separate files in different formats. The survey responses are collected via questionnaires and the answers of the participants to each question are stored under the corresponding variable names within one large data file. Descriptions of the variables can be found from DHS recode manuals such as ICF (2018) and the response of interest can be aggregated into the cluster centers that is stored in the cluster level data file. The aggregation step must be adapted to the application. Listing 1 shows how to read the DHS data into R and set it in the working environment.

```
library(haven)
library(rgdal)
# Reading DHS data :
# path1 : the full path to the individual level data file (.DTA)
# path2 : the path to the folder where the cluster level file (.shp) is located

# individual level data (individual survey responses) :
individualData = read_dta(path1)

# cluster level data (clusterID, cluster center coordinates, strata, etc.)
corList = readOGR(dsn = path2,layer = "file name")

# extract cluster level information:
smallGeo = data.frame(clusterIdx = corList$DHSCLUST,
                      urban = corList$URBAN_RURA,
                      long = as.vector(corList@coords[,1]),
                      lat = as.vector(corList@coords[,2]),
                      admin1 = corList$ADM1NAME)

```

```
#  extract individual level information:
myData = data.frame(clusterIdx = individualData$v001,
                    variable1  = individualData$v1,
                    variable2  = individualData$v2)
```

Listing 1: Loading DHS data into R.

5.2 Example for Nigeria

This section shows extracting and merging individual and cluster level data, based on Nigeria DHS 2018 (NDHS-2018) household survey. The example code in this section considers as outcome the completions of secondary education among 20-39 years old women in Nigeria. Population density is used as the only covariate and the corresponding raster file (Nga_ppp_v2c_2015.tif) can be downloaded from WorldPop (World Pop, 2022). The geography of Nigeria and the locations of the clusters are shown in the left-hand side panel of Figure 3.

This example will use the model

$$\begin{aligned} y_c|r_c, n_c &\sim \text{Binomial}(n_c, r_c), \quad \boldsymbol{s}_c|\boldsymbol{s}_c^* \sim \pi_{\text{Urb}[c]}(\boldsymbol{s}_c|\boldsymbol{s}_c^*), \\ r_c &= r(\boldsymbol{s}_c^*) = \text{logit}^{-1}(\eta(\boldsymbol{s}_c^*)), \end{aligned} \tag{5.1}$$

where y_c is the number of women who completed secondary education, n_c is the number of women interviewed, and r_c denotes the risk in cluster c, for $c = 1, \ldots, C$. The spatially varying risk $r(\cdot) = \text{logit}^{-1}(\eta(\cdot))$ is modelled through the linear predictor

$$\eta(\boldsymbol{s}^*) = \beta_0 + x(\boldsymbol{s}^*)\beta_1 + u(\boldsymbol{s}^*), \quad \boldsymbol{s}^* \in \mathcal{D},$$

where β_0 is the intercept, $x(\boldsymbol{s}^*)$ is the population density, β_1 is the effect of population density, and $u(\cdot)$ is the Matérn GRF with smoothness $\nu = 1$.

The shape file for Nigeria includes a large lake on its north-eastern corner. Lakes do not have any DHS household clusters within them, therefore it does not make sense to make any predictions at locations that are within the lake. Accordingly, we remove the polygon that corresponds to the lake from the admin2 level shape file. Listing 2 shows reading the administrative area shape files and DHS data and extracting the variables of interest based on the file and variable names of NDHS-2018.

```
# reading admin0 and admin2 shape files :
admin0 = readOGR(dsn = "dataFiles/gadm40_NGA_shp",
                           layer = "gadm40_NGA_0")

admin2 = readOGR(dsn = "dataFiles/gadm40_NGA_shp",
                 layer = "gadm40_NGA_2")

# remove the lake
admin2 = admin2[-160,] # Nigeria map has a large lake
```

```
                        # The lake corresponds to polygon
    160

# reading DHS data :
corList = readOGR(dsn = "dataFiles/DHS/NG_2018_DHS_
    02242022_98_147470/NGGE7BFL",
                                  layer = "NGGE7BFL")
educationData = read_dta("NGIR7BDT/NGIR7BFL.DTA")

# extract cluster level information:
smallGeo = data.frame(clusterIdx = corList$DHSCLUST,
                      urban = corList$URBAN_RURA,
                      long = as.vector(corList@coords[,1])
    ,
                      lat = as.vector(corList@coords[,2]),
                      admin1 = corList$ADM1NAME)

#  extract individual level information:
myData = data.frame(clusterIdx = educationData$v001,  #
    cluster ID
                    age = educationData$v012,
       # age
                    secondaryEducation = educationData$
    v106) # v106
                    #v106 : highest education level
                    # 0  : no education
                    # 1  : primary
                    # 2  : secondary
                    # >2 : higher

# reading the covariate raster:
library(raster)
r = raster::raster("Nga_ppp_v2c_2015.tif")
```

Listing 2: Data preprocessing: loading administrative shapefiles, DHS data, and covariate data into R for the Nigeria example.

Once the external data files are read into R, the data needs to be organized with respect to the content of the research. Accordingly, the individual survey answers contained in the data frame "`myData`" are first subsetted with respect to the age interval that we are interested in (20-39), and then merged with the cluster level information in the data frame "`smallGeo`". The merged data are then aggregated into the cluster centers. These steps can be followed through Listing 3.

```
# subset data with respect to the age interval of interest
    :
myData = subset(myData, age <= 39 & age >=20)

# number of 20-39 years old women who completed secondary
    education in each household
myData$ys = (myData$secondaryEducation>=2)+0
```

```
6
7  # merge the cluster level data with the subsetted
       individual level data,
8  # with respect to the cluster ID:
9  myData = merge(myData, smallGeo, by = "clusterIdx")
10
11 # add number of trials (for binomial response)
12 myData$Ntrials = 1
13
14 # aggregate the survey responses to the cluster centers
15 answers_x = aggregate(myData$ys,
16                       by = list(clusterID = myData[, 1]),
17                       FUN = sum)
18
19 answers_n= aggregate(myData$ys,
20                      by = list(clusterID = myData[, 1]),
21                      FUN = length)
22
23 # merge
24 answers_joint = merge(answers_x, answers_n, by="clusterID"
       )
25
26 # now we have the total number of women participants
       within the relevant age interval (ns),
27 # for each cluster. We also have the number of women among
        those who completed their secondary education (ys)
28 colnames(answers_joint) = c("clusterID", "ys", "ns")
```

Listing 3: Data preprocessing: subsetting and aggregating.

The main variables that are needed for the analysis are the ID numbers and coordinates of the cluster centers (both in degrees and in kilometers), their urbanicity stratum, and the aggregated response variable values. Accordingly, these are collected into a main data frame as in Listing 4.

```
1
2  # initial data frame
3  nigeria.data = data.frame(clusterID = corList@data[["
       DHSCLUST"]], long = as.vector(corList@coords[,1]), lat
       = as.vector(corList@coords[,2]))
4
5  # add ys and ns
6  nigeria.data = merge(nigeria.data, answers_joint, by="
       clusterID", all=T)
7
8  # add strata:
9  nigeria.data$urbanRuralDHS = corList@data[["URBAN_RURA"]]
10
11 # add coordinates in kilometers
12 nigeria.data$east = rep(NA, length(nigeria.data$long))
13 nigeria.data$north = rep(NA, length(nigeria.data$long))
14
```

```
nigeria.data[,c("east", "north")] = convertDegToKM(nigeria
    .data[,c("long", "lat")])
```

Listing 4: Data preprocessing: collecting data into a data frame.

The DHS jittering scheme is implemented by respecting various levels of administrative borders in different countries. The function `prepareInput` creates the integration points and considers their proximity to the respected level of administrative borders to decide if a secondary set of points should also be deployed. In NDHS-2018, jittering is done by respecting the second administrative level borders in Nigeria. It is important to be sure that the admin2 level areas that each cluster center is located within can be identified, in other words, each cluster center matches with one of the areas. This is the information that will lead the function `prepareInput` to evaluate the proximity of each individual integration point to the borders of the corresponding particular administrative area. Accordingly, the cluster centers that do not match with any admin2 areas need to be dropped as shown in Listing 5.

```
# jittering is done by respecting admin2 borders in
    Nigeria.
# see if there are  cluster centers that doesn't match
    with any of the admin2 areas:

# first, add polygon IDs (some shape files may have it
    already) :
admin2@data[["OBJECTID"]] =1:774 # normally 775, we
    removed one (the lake)
                                 # this number might be
    different in other   countries

# the cluster coordinates:
latLon = cbind(nigeria.data[,"long"], nigeria.data[,"lat"
    ])
colnames(latLon) = c("long", "lat")

# make a SpatialPoints object
latLon = SpatialPoints(latLon, proj4string=CRS("+proj=
    longlat +datum=WGS84 +no_defs"), bbox = NULL)

# see if the points (cluster centers) are within the
    polygons (admin2 areas) :
check1 <- over(latLon, admin2, returnList = FALSE)

# drop the rows which don't match with none of the admin2
    areas.
# we will need them to match while creating the
    integration points later on.

# see which ones don't return a match :
```

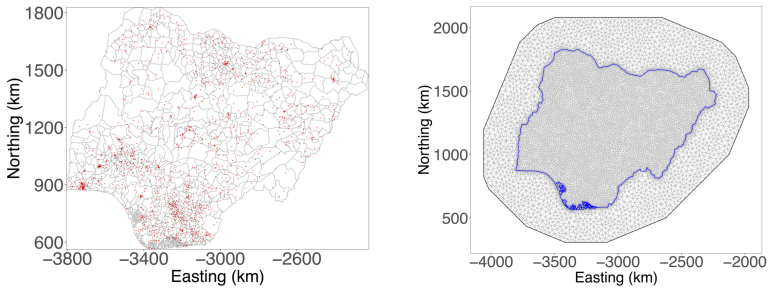


Figure 3: Nigeria subnational level map (left) and the triangulation mesh (right). The red points represent the jittered cluster centers.

```
# which(is.na(check1$NAME_2))  # see the rows that don't
    match:
# [1]   48  122  205  848  857 1116 1122 1287 1328

# drop the corresponding rows from the main data set :
nigeria.data = nigeria.data[-c( 48, 122, 205, 848,  857,
    1116, 1122, 1287, 1328),]
```

Listing 5: Data preprocessing: final adjustments.

Besides creating the integration points, the `prepareInput` function constructs the SPDE components and includes them in its returning list. The function implements this based on the triangulation mesh. This last step before running `prepareInput` is illustrated in Listing 6.

```
# transform admin0 into kilometers
proj = "+units=km +proj=utm +zone=37 +ellps=clrk80
                           +towgs84=-160,-6,-302,0,0,0,0 +
    no_defs"

admin0_trnsfrmd = spTransform(admin0,proj)

library(GeoAdjust)
# construct the mesh
mesh.s = meshCountry(admin0 = admin0_trnsfrmd,
                     max.edge = c(25, 50), offset=-.08)
```

Listing 6: Data preparation: constructing a triangulation mesh with the `meshCountry()` function.

Figure 3 shows the subnational (admin2 level) borders within Nigeria, together with the triangulation mesh that is constructed based on the national borders.

The function `prepareInput` saves the package user from various detailed long coding tasks and reduces the whole process into setting just couple of

arguments. The covariates in the model need to be passed into the function as raster layers within a list, via the argument "**covariateData**". The function needs the response variable values as well, so that it can construct the response vectors for the corresponding urban and rural integration points, separately. This example on NDHS-2018 has a binomial response. Accordingly, we need to pass both the aggregated binomial trials (ns) and the corresponding aggregated binomial successes (ys) for each cluster center. Here, ns represents the number of 20-39 years old women per cluster, and ys is the number of women who reported that their secondary education is completed, amongst them. If the response was Gaussian or Poisson, then the list passing to the "**response**" argument would only contain ys either as the Gaussian responses or the Poisson counts, respectively. Similarly, the likelihood type should be set via the argument "**likelihood**", so that the function processes the other arguments accordingly. Here, the values 0, 1 and 2 indicates that either Gaussian, binomial or Poisson likelihood is used in the model, respectively. The argument "**jScale**" sets the scaling of the default DHS maximum jittering distances. The function **prepareInput** multiplies the default distances by the value set to the argument, and evaluates the approximity of the primary integration points to their corresponding administrative area borders based on the scaled distances. The value 1 indicates that the default distances are in use. Different values can be set to this argument in order to experiment with them. Listing 7 shows how the arguments can be set and how the function can be used.

```
# read the covariate raster
library(raster)
r = raster::raster("Nga_ppp_v2c_2015.tif")

# the response variable
response = list(ys = nigeria.data$ys, # number of binomial successes
                ns = nigeria.data$ns)  # number of binomial trials

# cluster center coordinates in kilometers
locObs =cbind(nigeria.data[["east"]], nigeria.data[["north"]])

likelihood = 1 # binomial likelihood
               # (set 0, 1 or 2 for Gaussian, binomial or Poisson)
jScale = 1     # the maximum DHS jittering distances
               # can be scaled using this argument
               # 1 corresponds to the default DHS jittering


adminMap = admin2  # jittering is done by respecting admin2
```

```
                    # borders in Nigeria. This may be different
                    # for other countries. In Kenya, admin1 borders
                    # are respected instead.

inputData = prepareInput(response=response, locObs=locObs,
                         likelihood = likelihood,
                         jScale=jScale,
                         urban = nigeria.data$urbanRuralDHS,
                         mesh.s = mesh.s, adminMap=adminMap, nSubAPerPoint=10, nSubRPerPoint = 10,
                         covariateData=list(r=r))
```

Listing 7: Data preparation: preparing a list of input objects with the `prepareInput()` function.

The content of the input list created by function `prepareInput` is shown in Listing 8. The list contains the response vectors (both ns and ys in binomial case), design matrices and the projector matrices all created separately for the urban and rural strata. The function `prepareInput` breaks the vectors ns and ys into the urban and rural vectors `num_iUrban`, `num_iRural`, `y_iUrban` and `y_iRural` with respect to the urban and rural integration points. The other elements are the coordinates of the urban and rural integration points, corresponding urban and rural integration weights, SPDE components and the likelihood and normalization flags.

```
The final input data list contains the following elements:
 inputData <- list(num_iUrban,  # Total numb. of urban obs.
              num_iRural,  # Total numb. of rural obs.
              num_s, # num. of vertices in SPDE mesh
              y_iUrban, # urban obs in the cluster
              y_iRural, # rural obs in the cluster
              n_iUrban,  # urban exposures in the cluster
              n_iRural,  # rural exposures in the cluster
              n_integrationPointsUrban,#num.of urb.int.pts.
              n_integrationPointsRural,#num.of rur.int.pts.
              wUrban = wUrban, # urban weights
              wRural = wRural, # rural weights
              X_betaUrban = desMatrixJittUrban, # urb. des. mat.
              X_betaRural = desMatrixJittRural, # rur. des. mat.
              M0, #=spde[['param.inla']][['M0']],
              M1, #=spde[['param.inla']][['M1']],
              M2, #=spde[['param.inla']][['M2']],
              AprojUrban,  # Projection matrix (urban)
```

```
                AprojRural,  # Projection matrix (rural)
                options = c(1, ## if 1, use normalization
                            1), ## if 1, run adreport
                flag1 = 1, # normalization flag.
                flag2 = flag2, #(0/1/2 for Gaussian/
  Binomial/Poisson)
 )
 )
```

Listing 8: Data preparation: the input list.

6 Model estimation and gridded spatial prediction

6.1 Estimation

The function `estimateModel` utilizes the `MakeADFun` function of TMB to construct a list we will refer to as the core model object, containing the objective functions with derivatives (Kristensen et al., 2016b), (Kristensen, 2022). Then `estimateModel` uses `optim` to optimize the core model object and estimate the model parameters, without requiring the user to write any C++ code. The main argument of the function is a list called "`data`", referring to the input list that has just been created above. The argument `nNodes` refers to the number of nodes that the triangulation mesh has. The remaining two other arguments are called `options` and `priors`.

The argument `options` specifies in which of the two model components, namely, the random effect and covariates, jittering should be accounted for. Jittering adjustment can be turned on and off either in the random effect or in covariates or both, by setting the values of "`random`" and "`covariates`" to 1 or 0, respectively.

The argument `priors` allows the user to specify the parameters of the Gaussian prior for covariate effect sizes, and of the penalized complexity (PC) priors for the spatial range and marginal variance. These values can be passed into the function as a list of six elements, namely, "`beta`", "`range`", "`Uspatial`", "`alphaSpatial`", "`UNugget`", and "`alphaNug`". The element beta needs to be a vector of length two. The first and the second elements of the vector beta are the mean and the standard deviation of the Gaussian priors that are assigned for the intercept and the covariate effect sizes. "`range`" is the *a priori* median range, and "`USpatial`" is the upper "`alphaSpatial`" percentile of the marginal standard deviation, and "`UNugget`" and "`alphaNug`" are the hyperparameters for the PC-prior on the nugget variance. The hyperparameters "`UNugget`" and "`alphaNug`" pass into the function as 1 and 0.05, by default, but they are only used in the calculations when the likelihood is Gaussian. The package user is free to

fix them to other values as well. Listing 9 shows how the `estimateModel` function can be used. The argument "`n.sims`" controls the number of samples that will be drawn for each model parameter and each random effect coefficient.

```
# number of nodes in the mesh:
nNodes = mesh.s[['n']]

# estimating the parameters
est = estimateModel(data = inputData,
      nNodes = nNodes,
      options = list(random = 1, covariates = 1), #
    account for jittering in random and covariate effects
      priors = list(beta = c(0,1),
        range = 114,
        USpatial = 1, alphaSpatial = 0.05, UNugget = 1,
    alphaNug = 0.05), n.sims = 1000)
```

Listing 9: Estimating model parameters: using the `estimateModel()` function.

`estimateModel` returns a list of four elements (see Listing 10). Two of them, namely, `obj` and `draws` will be passed into the `predRes` function for predictions on a new prediction grid. The element `obj` is the optimized core model object. The element `draws` contains `n.sims` draws for the effects of covariates and the random effect. The element `likelihood` indicates the likelihood type that is used in the model construction. Finally, the last element `res` contains the estimated model parameters and the lengths of 95% credible intervals, which are constructed using the sampled values in `draws`. The credible interval lengths are calculated within the `estimateModel` function as the difference between the 97.5% and 2.5% quantiles of the drawn samples for the corresponding parameter estimate. The result object `res` does not contain `CI_Length` values for the range and the marginal variance, as the inference is empirical Bayesian where these parameters are estimated to fixed values. The model estimates can be printed in a tidy way as in Listing 10

```
# the output of estimateModel() function:
names(est)
[1] "res"        "obj"        "draws"      "likelihood"

print(est)

GeoAdjust::estimateModel()
-----------------------------------
Likelihood :          binomial
-----------------------------------
parameter estimate    95% CI length
range     72.7716     NA
sigma     2.2838      NA
```

```
beta0      -1.0295      0.7488
beta1      0.0079       0.0064
-----------------------------------
```

Listing 10: Estimating model parameters: output.

6.2 Prediction grid

`GeoAdjust` provides the `gridCountry` function to create a grid of prediction points with respect to the national borders of the country of interest. The function has two arguments. The first argument `admin0` is the SpatialPolygonsDataFrame object containing the national borders. The second argument `res` indicates the resolution in kilometers. Internally, the function first creates a raster within the bounding box of the `admin0` object and with respect to the chosen resolution. Afterwards, it extracts the coordinates of the cell centroids and constructs a data frame containing the cell centroid coordinates both in kilometers and degrees. Finally, the function returns the coordinates and the prediction raster within a list. Having the prediction raster is necessary to use the function `plotPred`, which internally utilizes `geom_raster` from `ggplot2`, which is useful for plotting the predictions and the uncertainty across the country. Listing 11 shows how `gridCountry` function can be used.

```
# raster and the prediction coordinates:
predComponents = gridCountry(admin0 = admin0, res = 5)

names(predComponents)
[1] "loc.pred" "predRast"

# the prediction locations
loc.pred = predComponents[["loc.pred"]]

head(loc.pred)
       east    north     long      lat
1 -3803.287 1825.665 1.838939 13.27084
2 -3798.287 1825.665 1.875670 13.27712
3 -3793.287 1825.665 1.912420 13.28340
4 -3788.287 1825.665 1.949189 13.28967
5 -3783.287 1825.665 1.985977 13.29595
6 -3778.287 1825.665 2.022784 13.30222

> dim(loc.pred)
[1] 80201     4

predRast = predComponents[["predRast"]]
print(predRast)
class       : RasterLayer
dimensions : 253, 317, 80201  (nrow, ncol, ncell)
resolution : 5, 5  (x, y)
```

```
extent      : -3805.787, -2220.787, 563.1654, 1828.165  (
    xmin, xmax, ymin, ymax)
crs         : +proj=utm +zone=37 +ellps=clrk80 +units=km +
    no_defs
```

Listing 11: Prediction: the `gridCountry()` function.

6.3 Prediction

The `predRes` function uses the core model object that is created within `estimateModel` to predict the model outcomes at a set of prediction locations. The function `predRes` uses two elements from the output list of `estimateModel`, namely the core model object (`est[["obj"]]`) and the matrix containing the sampled covariate effect sizes together with the sampled random effect coefficients for each mesh node (`est[["draws"]]`).

In this example we use the cell center coordinates of the prediction raster which is just constructed by `gridCountry` function, but it is also possible for the package users to predict on a custom made grid or any other set of locations. Please note that, if the package user prefers to use a different location set, their coordinates need to be passed in kilometers as a matrix with the corresponding column names "east" and "north", respectively.

The function `prepareInput` used an argument called "`covariateData`". It was a list containing the raster layers of each covariate. The purpose of the argument there was to extract the covariate values at the integration points and to create urban and rural design matrices. Similarly, `predRes` function uses the same argument with the same name and content, but here the function creates a design matrix by extracting the covariate values at the prediction locations.

The argument `flag` takes one of 0, 1 or 2. The value of this argument indicates the type of the likelihood that the model includes. The values 0, 1 and 2 indicates the Gaussian, binomial and Poisson likelihoods, respectively. The function uses the value to decide which link function should be used. Listing 12 shows the usage of the function `prepareInput`.

```
predictions = predRes(obj = est[["obj"]] , predCoords =
    loc.pred,
                      draws = est [["draws"]] , nCov =
    nCov,
                      covariateData = covariateData,
                      mesh.s = mesh.s, flag = 1)

head(predictions)
          mean    median         sd       lower     upper
[1,] 0.2646259 0.2627431 0.03690252 0.196264273 0.3405225
[2,] 0.3742580 0.2584815 0.34268884 0.001753274 0.9855895
[3,] 0.3707435 0.2601383 0.33634250 0.002157952 0.9826507
[4,] 0.3687931 0.2615159 0.33290925 0.002405636 0.9801035
[5,] 0.3686910 0.2641275 0.33299652 0.002327227 0.9800920
```

```
[6,] 0.3677783 0.2623805 0.33203511 0.002436265 0.9803111

dim(predictions)
[1] 80201     5
```

Listing 12: Prediction: the `predRes()` function.

6.4 Plotting the predictions and uncertainty

This section shows how to plot the predicted values and the uncertainty accross the country map. We will use the predicted median values obtained from `predRes` function, as the point predictions. The plotted uncertainties will be the corresponding coefficient of variations calculated by $\frac{\sigma}{\mu} \times 100$, also obtained from `predRes` function. `GeoAdjust` uses the function `plotPred` to plot the predictions and the corresponding uncertainty accross the studied country. The first argument `pred` is the output of the function `predRes` which is obtained in Listing 12. The argument `predRaster` is the prediction raster that was constructed by `gridCountry` function in Listing 11. The arguments `admin0`, `admin1` and `admin2` stand for the `SpatialPolygonsDataFrame` objects representing the national, first level subnational and second level subnational administrative borders of the corresponding country. There might be a need to leave some of the admin2 level polygons uncolored as we did here for the polygon 160 (the lake). Then the number of the polygon that needs to be excluded can be pass into the function through the argument `rmPoly`. The argument doesn't remove the polygon from the map. The function still plots the polygon within the map, but it doesn't assign colors anywhere within that polygon. The arguments `rmPoly` and `admin2` can be set to `NULL` if there is no such need. The administrative borders that are overlaid on the map by this function are the admin1 level borders (see Figure 4). The argument `locObs` indicates the observed locations, or in other words, the DHS cluster centers. The function plots these as red dots on to the map. The function returns a list containing two `ggplot` objects, representing the plots for the predictions and uncertainty accross the country of interest. Listing 13 shows how to use the function `plotPred`.

```
admin1 = readOGR(dsn = "dataFiles/gadm40_NGA_shp",
                 layer = "gadm40_NGA_1")

plotPred(pred = predictions, predRaster = predRast, admin0 = admin0,
         admin1 = admin1, admin2 = admin2, rmPoly = 160, locObs = locObs)
```

Listing 13: Plotting: preparation.

Figure 4 shows the predicted risk and the corresponding uncertainty across Nigeria. Please note that since we assigned "NA" to the points that

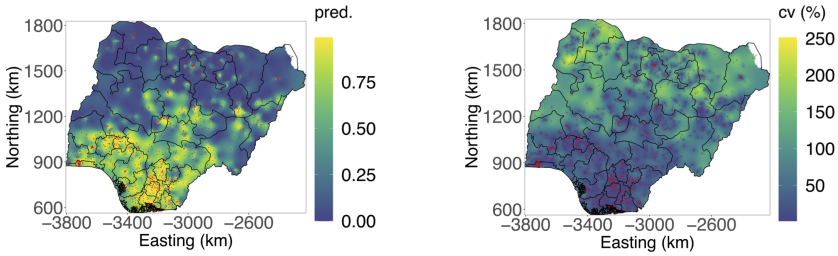


Figure 4: Predicted risk (left) and the CVs (right). The red points indicate the example survey cluster centers.

overlap with the lake, the north-east corner of the plots are not colored. This area is the area covered by the lake. This is specific to the geography of Nigeria and different features like this may need to be considered while plotting data on the maps of different countries.

7 Summary

`GeoAdjust` makes it easy to account for jittering, by isolating its user from complex code while still providing flexible control over the implementation. It is unique in a sense that it is the only package that specifically targets the positional uncertainty in the observed locations and also conveys a functionality emerging from a unique way of approaching to this problem. The package has a potential to be tested on and developed further for both areal and point referenced data from different areas involving the positional uncertainty and geomasking.

References

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022a). Fast geostatistical inference under positional uncertainty: Analysing DHS household survey data. *arXiv preprint arXiv:2202.11035*.

Altay, U., Paige, J., Riebler, A., and Fuglstad, G.-A. (2022b). Jittering impacts raster- and distance-based geostatistical analyses of dhs data. *arXiv preprint arXiv:2202.07442v1*.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). Geographic displacement procedure and georeferenced datarelease policy for the De-

mographic and Health Surveys. https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf. DHS Spatial Analysis Reports No. 7.

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

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114:445–452.

Gómez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051.

ICF (2018). Demographic and Health Surveys Standard Recode Manual for DHS7. *The Demographic and Health Surveys Program, Rockville, Maryland, U.S.A.: ICF.* https://dhsprogram.com/pubs/pdf/DHSG4/Recode7_DHS_10Sep2018_DHSG4.pdf/, last accessed on 2023-03-20.

Kristensen, K. (2022). The comprehensive tmb documentation.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016a). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70(5):1–21.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016b). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73:423–498.

Perez-Heydrich, C., Warren, J., Burgert, C., and Emch, M. (2013). Guidelines on the use of DHS GPS data. *ICF International, Calverton, Maryland.* https://dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf/, last accessed on 2023-03-20.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., and Emch, M. E. (2016). Influence of Demographic and Health Survey point displacements on raster-based analyses. *Spatial demography*, 4(2):135–153.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Sstatistical Methodology)*, 71(2):319–392.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016). Influence of demographic and health survey point displacements on distance-based analyses. *Spatial Demography*, 4(2):155–173.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100421.

World Pop (2022). Open spatial demographic data and research.

ISBN 978-82-326-7072-7 (printed ver.)
ISBN 978-82-326-7071-0 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology