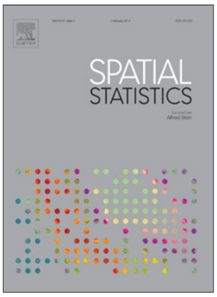

Combining school-catchment area models with geostatistical models for analysing school survey data from low-resource settings: Inferential benefits and limitations

Peter M. Macharia, Nicolas Ray, Caroline W. Gitonga, Robert W. Snow, Emanuele Giorgi

PII: S2211-6753(22)00048-3
DOI: https://doi.org/10.1016/j.spasta.2022.100679
Reference: SPASTA 100679

To appear in: *Spatial Statistics*

Received date : 28 March 2022
Revised date : 13 June 2022
Accepted date : 14 June 2022

Combining school-catchment area models with geostatistical models for analysing school survey data from low-resource settings: inferential benefits and limitations

Peter M. Macharia[a,b,*], Nicolas Ray[c,d], Caroline W. Gitonga[b], Robert W. Snow[b,e], Emanuele Giorgi[a]

[a] *Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, LA1 4YW, UK*
[b] *Population Health Unit, Kenya Medical Research Institute-Wellcome Trust Research Programme, PO, Box 43640, Nairobi, Kenya*
[c] *GeoHealth group, Institute of Global Health, University of Geneva, Geneva, Switzerland*
[d] *Institute for Environmental Sciences, University of Geneva, Geneva, Switzerland*
[e] *Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7LG, UK*

Abstract

School-based sampling has been used to inform targeted responses for malaria and neglected tropical diseases. Standard geostatistical methods for mapping disease prevalence use the school location to model spatial correlation, which is questionable since exposure to the disease is more likely to occur in the residential location. In this paper, we propose to overcome the limitations of standard geostatistical methods by introducing a modelling framework that accounts for the uncertainty in the location of the residence of the students. By using cost distance and cost allocation models to define spatial accessibility and in absence of any information on the travel mode of students to school, we consider three school catchment area models that assume walking only, walking and bicycling and, walking and motorized transport. We illustrate the use of this approach using two case studies of malaria in Kenya and compare it with the standard approach that uses the school locations to build geostatistical models. We argue that the proposed modelling frame-

[*] Corresponding author
Email address: pmacharia@kemri-wellcome.org.com (Peter M. Macharia)

work presents several inferential benefits, such as the ability to combine data from multiple surveys some of which may also record the residence location, and to deal with ecological bias when estimating the effects of malaria risk factors. However, our results show that invalid assumptions on the modes of travel to school can worsen the predictive performance of geostatistical models. Future research in this area should focus on collecting information on the modes of transportation to school which can then be used to better parametrize the catchment area models.

Keywords: catchment area models, disease mapping, school, school survey, missing locations, model-based geostatistics, prevalence

1. Background

In low resource settings, the prevalence of parasitic infections are important indicators to guide control. Surveys are traditionally undertaken among sampled residents in communities or from fixed locations that serve these communities, such as schools. School-based sampling has been used for decades in Sub Saharan Africa (SSA) to inform the targeted responses for helminth (Hodges et al., 2011; Tchuem Tchuenté et al., 2012; Soares Magalhães et al., 2011), schistosomiasis (Hodges et al., 2011; Tchuem Tchuenté et al., 2012; Soares Magalhães et al., 2011; Knowles et al., 2017; Fornace et al., 2020; Clements et al., 2006) and malaria (Gitonga et al., 2010; Brooker et al., 2009; Ashton et al., 2015; Mathanga et al., 2015) control. School-based surveys for parasitic diseases represent convenient and cost-effective sampling strategies to provide local disease information, where school attendance is high, and infections can be asymptomatic (Ashton et al., 2011; Brooker et al., 2009; Drake et al., 2011; Mathanga et al., 2015; Stevenson et al., 2013; Takem et al., 2013). Sample school surveys can be powered to provide estimates of malaria prevalence at geographical units (e.g., districts) used for decision making. For example, targeting certain districts with high malaria prevalence. More commonly, the lack of statistical power from minimally sampled schools has involved the applications of model-based geostatistical (MBG) methods (Diggle et al., 1998), using aggregated information at sampled school locations to provide information at unsampled locations (Soares Magalhães et al., 2011; Fornace et al., 2020; Clements et al., 2006; Ashton et al., 2011).

MBG methods for disease mapping have become an established set of modern and robust statistical tools (Diggle et al., 1998) that are used to inform disease control strategies (Macharia et al., 2018; Biggeri and Catelan, 2012), especially in low-resource settings where disease registries are non existent or incomplete (Alegana et al., 2020; Pop et al., 2019; Stefan et al., 2014). Typically, MBG for disease prevalence mapping aim to predict a disease risk surface using data consisting of a finite set of locations x_i, for $i = 1, \ldots, N$, where a number of n_i individuals are tested for a disease of interest and of which y_i test positive. Ideally, the x_i would correspond to the locations where individuals contracted the disease but, in practice, this is often difficult, if not impossible, to assess and access. In most geostatistical analyses of epidemiological data, the location of the school or village is used as the main location of exposure to the disease of interest.

However, in most school malaria surveys, due to resource constraints, precise spatial information on the residence of the children is rarely collected. Only the geographical location of the school is collected, hence the uncertainty in the household location of the children within a school catchment area. Consequently, when mapping malaria prevalence using school survey data, the school location is often used as the exposure location due to missing spatial information on the residential location. For example, in Ashton et al. (2015), Binomial geostatistical models are fitted to serological indicators collected from school surveys, while using the location of the school to define the spatial correlation between observations. The same approach has been used in other studies that have combined school-based data with community-based data (Macharia et al., 2018; Runge et al., 2020). In Stensgaard et al. (2011), the problem of the use of the school locations to estimate covariates effects on prevalence is alleviated by averaging the covariates within 1 km around each school location. However, this approach is questionable, since it implicitly assumes a circular school catchment area with a radius of 1km while still making use of the school location in the computation of the spatial correlation.

A common problem to these approaches is that, by allocating individuals to locations that are less representative of their actual exposure to malaria, it can potentially bias the spatial structure of the MBG model and thus invalidate the predictive inferences on prevalence. This is especially important for school going children since a bite from an infected *Anopheles* mosquito is more likely to occur during night, when children are usually at home (Maxwell et al., 1998).

The statistical problem addressed in this paper is related to the problem that arises in passive surveillance when dealing with spatially aggregated disease counts; see for example Cameron et al. (2021). In Sturrock et al. (2014), aggregated case data reported at health facility are modelled by defining the probability of observing a case at any given location as the product of being a case and the probability that individual would seek treatment. However, in this work the random effects that modulate the probability of being a case are modelled as a spatially discrete process that is tied to the specific definition of the hospital catchment areas. The fine scale predictions of prevalence in this case, are thus a product of spatially continuous covariates and spatially discrete random effects. In this paper, we overcome this limitation by adopting a spatially continuous random effect whose properties are independent of the catchment areas. However, one key difference between the problem addressed in this paper and previous work on aggregated counts at health facilities, is that the latter problem is more naturally addressed using spatial point patterns methods. In other words, if locations were fully observed, a Log-Gaussian Cox process (LGCP) would be a natural modelling option to model reported cases at health facilities, whilst in our case, the location is not the object of interest, but rather the prevalence associated with that. Diggle et al. (2013), using LGCPs, provide a principled solution for fine-scale mapping by modelling the spatially aggregated disease counts as the realization of an aggregated spatially continuous stochastic process. As in Cameron et al. (2021), the methods that have been developed in the context of spatially aggregated passive surveillance data are based on log-linear models do not provide a solution to the problem addressed in this paper. Nelli et al. (2020) develop methods to combine passive and active surveillance data, however the sampled locations of the latter are assumed to be observed.

To the best of our knowledge, no statistically rigorous solution has been proposed to handle the problem of missing residence locations from malaria school survey data in a geostatistical model. In this paper, we provide a first solution to this problem and compare the predictive inferences on prevalence resulting from this novel approach with standard statistical approaches that use the school location to model the spatial correlation.

2. Methods: combining school catchment area models with geostatistical models

In this study, the set of residence locations X, unlike in standard geostatistical analyses, must be treated as a random variable and a suitable distribution for this must first be defined. Once we have defined the statistical model for X, we can use this to impute likely values for the unobserved residence locations of children and incorporate these into a geostatistical model for disease prevalence. This process allows us to rigorously acknowledge the uncertainty arising from from the missing residence locations in the predictive inferences of disease prevalence. However, this begs the two following questions: 1) what is a suitable model for X? 2) How to combine the model for X with a geostatstical model for disease prevalence?

To answer these questions, our approach is to develop a marked Poisson process for the unobserved residence locations,X, with marks corresponding to each of the schools, and whose domain is restricted by a school catchment area (SCA) model informed by several factors that affect travel (road network, land use, protected areas, water bodies and travel speed). We then use the resulting model for X to generate samples of locations and feed these into a MBG model for disease prevalence. This approach presents several computational issues which we address using an stochastic partial differential equation (SPDE) approximation (Lindgren et al., 2011) for spatial Gaussian processes.

We first introduce the framework for accounting for missing residence locations in the context of disease prevalence mapping which entails creating school SCAs and generating samples of the most likely residential locations X. This is then followed by the estimation of the model parameters and spatial prediction of disease prevalence within a predefined geographical area of interest. The application of the proposed modelling framework is illustrated through two case studies of malaria mapping in Western Kenya using school survey data.

2.1. Accounting for missing residence locations in a geostatistical model for disease prevalence mapping

In this section, we formalize and provide a solution to the problem of how to propagate the uncertainty in geostatistical models for prevalence mapping, arising from the lack of residence locations in school survey data.

Let $[\cdot]$ be a shorthand notation for "the density function of the random variable $\cdot$". We then use $X = \{X_1, \ldots, X_n\}$ to denote the random variables representing the set of unobserved residence locations, and $Y = (Y_1, \ldots, Y_n)$ for the observed individual-level binary outcomes indicating a positive ($Y_i = 1$) or negative ($Y_i = 0$) test. Finally, let $S = \{S(x) : x \in \mathbb{R}^2\}$ be an isotropic and stationary Gaussian process with mean zero and covariance function

$$\mathrm{cov}\{S(x), S(x')\} = \sigma^2 \rho(u), u = \|x - x'\|.$$

Assuming that X and S are independent of each other, or in other words assuming a non-preferential sampling design, the joint distribution of X, S and Y can be written as

$$[X, S, Y] = [X][S][Y|S, X] = [X][S][Y|S(X)] \tag{1}$$

where $S(X) = \{S(x) : x \in X\}$ and $[Y|S(X)]$ is a set of mutually independent Bernoulli variables, i.e.

$$[Y|S(X)] = \prod_{i=1}^{n} [Y_i|S(X_i)].$$

When residence locations X are observed, then $[X]$ is irrelevant for drawing inferences on S, and thus can be ignored. In our case, instead, because of the missingess of X, the distribution $[X]$ must be integrated out from the likelihood function for the unknown vector of parameters θ, i.e.

$$L(\theta) = [Y; \theta] = \int_{A^{2n}} \int_{\mathbb{R}^n} [X][S][Y|S(X)] \, dS \, dX. \tag{2}$$

where $A \subset \mathbb{R}^2$ is our geographical region of interest.

To estimate θ from 2 and draw predictive inferences on S, we then first need to specify a model for the location process $[X]$.

2.2. Modelling $[X]$ *using school catchment area models*

To model the location process $[X]$, we propose to generate a school catchment area (SCA) based on travel time from the residence of children. The SCA is then used as the boundary of the area from which we shall draw samples of the residence locations using population density information. More specifically, an SCA is defined as the geographical area or zone around a school that draws majority of the students (Macharia et al., 2021b). One approach is to use so-called *gravity models* to express the decreasing likelihood of geographically accessing a school, as the distance or travel time to that school increase (Guagliardo, 2004) and they have been used to model catchment areas for healthcare planning (Macharia et al., 2017; Alegana et al., 2012; Guagliardo, 2004). Hence, *spatial accessibility* of a residence location x is defined as

$$a(x) = \sum_{j} \frac{c_j}{f(x, x_j)^{\gamma}}, \tag{3}$$

where: c_j is a constant that expresses the capacity of a school placed at location x_j, and can, for example, be quantified by the number of teachers in that school; $f(x, x_j)$ is the impedance (travel time) between locations x and x_j; γ is a gravity decay coefficient, also referred to as the travel friction coefficient.

Real SCAs vary in size according to the underlying population distribution, number of schools in the surrounding area, school capacity and other school *attractiveness* factors. School attendance and geocoded residential location X, when available, can be used to estimate the parameters in 3. However, in the absence of such data, as in the scenario considered in this paper, a feasible and useful alternative is to use cost distance to define the spatial accessibility function $a(x)$ and use optimization algorithms that allow to identify an optimal route of travel to school. Based on this definition of spatial accessibility, an SCA is then defined as the geographical area encompassing all locations closest, in terms of travel time, to that school than any other school. The travel time to school is dependent on the speed of the mode of travel to school which, in turn, is affected by several spatial layers, including the road network that students travel on, land cover layer to represent the travel impedance in spaces between the roads, and barriers to movement comprising water bodies, flooded areas, and protected areas. Barriers are considered impassable, except in the presence of a bridge where a road intersects a barrier such as a river. To model the speed of the mode of transport and identify the optimal route to school, we propose to use terrain-based least-cost path distance calculation (Ray and Ebener, 2008) and geospatial layers representing factors that affect travel to generate a travel time grid indicating the time taken by a student to travel from their residential location to the nearest school.

Each road class and land cover are assigned a mode of transport and travel speed, which ideally are informed by observational data of schools attendance behavior in a specific region. However, in low resource settings, such data are rarely available (Macharia et al., 2021b) and a common alternative is the review of literature in similar settings to assemble road speeds and modes of transport in one or more travel scenarios. We provide details on the parametrization of the speed functions for our study regions in Kenya, in Section 3.2.

After generating a SCA, we spatially link this with a population density raster (Stevens et al., 2015) which is then used to sample residential locations within the SCA. The population density rasters are constructed through dasymetric techniques that redistributes national census population counts from administrative units to high spatial resolution (e.g. 100 metres m x 100 metres) (Mennis, 2009; Stevens et al., 2015). Let $\mathcal{C}$ denote the area encompassed by the boundaries of a given SCA; we then generate samples $X_{(j)} = \{x_{1(j)}, \ldots, x_{n(j)}\}$, for $j = 1, \ldots, B$ for the unobserved residence locations, X, by using a fine regular grid covering C to approximate the probability of sampling a location $x \in \mathcal{C}$, given by

$$\frac{\lambda(x)}{\int_C \lambda(u)\, du},$$

where $\lambda(x)$ denotes the population density at x.

2.3. Approximating $[S]$ and the likelihood function: from parameter estimation to spatial prediction

The resulting Monte Carlo samples $X_{(j)}$, $j = 1, \ldots, B$, obtained as described in the previous section, are now used to approximate 2 as

$$L(\theta) \approx \int_{\mathbb{R}^n} \frac{1}{B} \sum_{j=1}^{B} [S][Y|S(X_{(j)})]\, dS. \tag{4}$$

To avoid the computation of B covariance matrices for each of the stimulated samples $X_{(j)}$, we approximate S using a piece-wise linear approximation for $S(x)$. More specifically, we partition the study region A into a set of non-intersecting triangles that share at most a common edge. For the generation of the triangles we follow the approach outline in Section 2.2.2 of Krainski et al. (2018).

We then approximate the spatial Gaussian process as

$$S(x) = \sum_{k=1}^{m} b_k(x) W_k \tag{5}$$

where $b_k(x)$ are basis functions and the W_k, for $k = 1, \ldots, m$ are Gaussian random variables. Based on the created mesh, we then define the basis functions $b_k(x)$ using barycentric coordinates, in which the location of a point is specified by reference to vertices of a triangle. It then follows that $b_k(x)$ takes a non-zero value whenever a point x falls inside a triangle identified by the vertex associated with $b_k(x)$ such that $\sum_{k=1}^{m} b_k(x) = 1$; for more details see Section 2.2.2 of Krainski et al. (2018). Following Lindgren et al. (2011), we assume that $W = (W_1, \dots, W_m)$ follows a zero-mean multivariate Gaussian distribution with precision matrix Q, which is chosen so as to approximate a Matérn spatial field with smoothness parameter $\kappa = 1$ and scale parameter ϕ. Hence, we write

$$Q = \left(\frac{\phi^{2\kappa}\Gamma(\kappa)}{4\pi\sigma^2\Gamma(\kappa+2)} \right)^2 (\phi^{-4}C + 2\phi^{-2}G_1 + G_2)$$

where C, G_1 and G_2 are sparse matrices whose entries are non-zero only for pairs of vertices that share the same triangles; for more details on how the entries of C, G_1 and G_2 are defined, we refer the reader to Lindgren et al. (2011).

Let $x_{i(j)}$ denote the i-th element of $X_{(j)}$ for $i = \dots, n$; following from 5, we then define $[Y|W, X_{(j)}]$ as a set of mutually independent Bernoulli variables with linear predictor

$$\log\left\{ \frac{p(x_{i(j)})}{1 - p(x_{i(j)})} \right\} = d^{\top}(x_{i(j)})\beta + \sum_{k=1}^{m} b_k(x_{i(j)})W_k \qquad (6)$$

where $d(x_{(j)})$ is a vector of spatial covariates, recorded at location $x_{(j)}$, with associated regression coefficients β.

Let us split the vector of unknown parameters θ, into covariance parameters $\psi = (\sigma^2, \phi)$ of the spatial process and regression coefficients β, and reparametrize 4 based on W to give

$$L(\theta) \approx \int_{\mathbb{R}^m} [W; \psi] \left(\frac{1}{B} \sum_{j=1}^{B} [Y|W, X_{(j)}; \beta] \right) dW. \qquad (7)$$

Since the above integral is intractable, we use Monte Carlo methods to approximate the likelihood function as follows. Let ψ_0 and β_0 be our initial guesses for the parameters ψ and β, respectively. To simplify the notation let $g(Y, W; \beta) = \sum_{j=1}^{B} [Y|W, X_{(j)}]/B$; we then rewrite 7 as

$$L(\theta) \approx \int_{\mathbb{R}^m} [W; \psi] g(Y, W; \beta)\, dW$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^m} [W;\psi] g(Y,W;\beta) \frac{[Y,W;\psi_0,\beta_0]}{[Y,W;\psi_0,\beta_0]}\, dW \\
&\propto \int_{\mathbb{R}^m} \frac{[W;\psi] g(Y,W;\beta)}{[W;\psi_0] g(Y,W;\beta_0)} [W|Y;\psi_0,\beta_0]\, dW \\
&= E_0\left[\frac{[W;\psi] g(Y,W;\beta)}{[W;\psi_0] g(Y,W;\beta_0)}\right] \qquad (8)
\end{aligned}
$$

where E_0 is the expectation taken with respect to the distribution of W, conditional on Y, with parameters ψ_0 and β_0 or, using the shorthand notation, $[W|Y;\beta_0,\psi_0]$. Hence, we approximate 7, by sampling from $[W|Y;\beta_0,\psi_0]$ using a Metropolis Hastings independence sampler with proposal distribution given by a multivariate Gaussian distribution with mean and covariance matrix corresponding to the mode and inverse of the negative Hessian of $[W;\psi_0]g(Y,W;\beta_0)$, respectively.

When the goal of the analysis is primarily focused on spatial prediction and not on drawing inferences on β, the computational burden can be alleviated by first estimating β using a simpler model that ignores spatial correlation, which is obtained by setting $S(x) = 0$ for all x to give $\log\{\tilde{p}(x_{(j)})/(1-\tilde{p}(x_{(j)})\} = d^\top(x_{(j)})\beta$. The likelihood function of this non-spatial model is given by

$$
\tilde{L}(\beta) = \frac{1}{B}\sum_{j=1}^{B}\prod_{i=1}^{n} \tilde{p}(x_{i(j)})[1-\tilde{p}(x_{i(j)})].
$$

After maximizing the above function with respect to β, we obtain $\tilde{\beta}$ which we now plug-in into the linear predictor 6. This then leads to a simplified likelihood function for the covariance parameters ψ, expressed by

$$
L_{\tilde{\beta}}(\psi) = E_0\left[\frac{[W;\psi] g(Y,W;\tilde{\beta})}{[W;\psi_0] g(Y,W;\tilde{\beta})}\right] = E_0\left[\frac{[W;\psi]}{[W;\psi_0]}\right]. \qquad (9)
$$

By maximizing the above function, we finally obtain $\tilde{\phi}$ as our point estimate of ψ. Since the primary objective of our case study in predicting malaria prevalence, we adopt this approach in our two applications.

Spatial prediction of prevalence is carried by plugging the point estimates $\tilde{\beta}$ and $\tilde{\phi}$ into 6.

3. Application 1: large scale mapping of malaria prevalence across eight counties of Western Kenya

3.1. Data

We analyse data from a national school-based survey of malaria prevalence conducted in 2009 in Kenya; full details of the survey are provided elsewhere (Gitonga et al., 2010). Here, we consider 84 sampled public day primary schools located in a high malaria transmission region of Western Kenya, covering eight counties and 62 sub-counties close to Lake Victoria. All eight counties had at least a school surveyed while 50 sub-counties (81%) had at least a surveyed school (Figure 1). At each school approximately 100 children aged 4-22 years were randomly sampled from classes 2-6 for a total of 9,103 children. The majority of the children (91%) were aged between 8 and 14 years while only 0.5% were aged at least 17 years. Each sampled child provided a finger-prick blood sample that was used to detect Histadine Rich Protein (HRP) as evidence of recent *Plasmodium falciparum* infection using a rapid diagnostic test (RDT) (Paracheck-Pf device). Slides from all RDT positive samples were examined using light microscopy and 10% of all RDT negatives. A child was deemed positive for malaria when parasites were detected on microscopy. The location of the school was recorded using a hand-held Global Positioning System (GPS) device (Gitonga et al., 2010).

In addition to the malaria school survey data, a set of spatial covariates were used to assist the spatial prediction of prevalence at unsampled locations. These are listed and described in Table 1. Before including these covariates in our model, we explored the association of each of the covariates listed in Table 1 by taking the *empirical logit transformation* (Stanton and Diggle, 2013) of the total number of cases recorded at each school and plotted this against each of the covariates, whose value on the x-axis was obtained by taking its average within the SCAs.

Table 1: Summary of the covariates in the analysis of Section 3.

Covariate	**Description**	**Source and resolution**
1. Annual mean Temperature	Temperature affects the survival and development of *P. falciparum* from larvae into viable adults. Temperatures greater than 34°C lead to almost 100% larval mortality, while temperatures less than 16°C, the larvae are unable to produce viable adults (Bayoh and Lindsay, 2004). Mortality of the anopheles mosquitoes increases at ambient temperatures approaching 40°C while temperatures between 25°C and 30°C are considered optimum for *Plasimodium falciparum* sporogony (Molineaux, 1988).	MODIS, 5.6 km grids (Busetto and Ranghetti, 2016)
2. Annual mean Precipitation	Rainfall combined with suitable ambient temperatures provides potential breeding environments for Anopheles (Noor et al., 2014; Dutta and Dutt, 1978).	CHIRPS, 5 km grids (Funk et al., 2015).
3. Urbanization	Malaria infection is usually lower in urban compared to rural areas due to reduced malaria vector density and biting rate (Kabaria et al., 2017). Urbanization was proxied by nighttime lights (NTL) (Savory et al., 2017). NTL are also associated with human activity, population distribution, reduced poverty rates and increased access to health facilities (Kabaria et al., 2017).	NTL 1 km square grids (Savory et al., 2017)
4. Enhanced vegetation index	Vegetation acts as a proxy for the presence of suitable mosquitoes breeding sites (Noor et al., 2014; Dutta and Dutt, 1978).	MODIS, 0.25 km grids (Busetto and Ranghetti, 2016)

3.2. Model specification for [*X*]

We first assembled a list of all public day primary schools in 2009 (Mulaku and Nyadimo, 2011) (Figure S1.1 in *Supplementary material 1*) and variables representing factors that are known to affect travel (Table 2 and Figures S1.2 to S1.4 in *Supplementary material 1*) in order to define the speed of a specific mode of travel at any location x of the study area (Figure 1). To compute travel time, we used AccessMod (version: 5.7.3-alpha) (Ray and Ebener, 2008), an open-source package that models geographical access using terrain-based least-cost path distance calculation (Ray and Ebener, 2008). We used the 'merge land cover' module of AccessMod to merge the land cover, road network, water bodies, national parks and reserves, and obtain a 'merged landcover' raster at 100-meter resolution. Travel speeds were then assigned to each land cover type and road class.

Table 2: Summary of factors that affect travel time to schools in Western Kenya.

Factor	Description	Type, Year and Resolution
1. Road Network	The road network was based on data from the Ministry of Transport that used the gold GPS technique to map coverage of roads in 2016. The layer was updated via OpenStreetMap and Google Map Maker and was cleaned by deleting duplicates and correcting digitization errors such as overshoots and undershoots at connection points or junctions and those that extended into water bodies. Further details in (Macharia et al., 2017, 2021a; Joseph et al., 2020).	Vector layer, 2016
2. Land use	The land use/cover information was obtained from 2016 Copernicus Sentinel-2 satellite containing five classes (bare areas,built-up areas, water bodies, cultivated and vegetation cover areas available at http://geoportal.rcmrd.org/	raster, 20 by 20 metres. 2016.
3. Digital elevation model (DEM)	The slope of the terrain was derived from a DEM based on Shuttle Radar Topographic Mission (SRTM) available at http://geoportal.rcmrd.org/	raster, 30 by 30 metres
4. Barriers to movement	Barriers to movement included major rivers, lakes, flooded areas and protected areas (Joseph et al., 2020).	Vector layers.

School going children can use different transport modes to reach their school. Since data on modes of travel were not available, we considered three different models corresponding to travel scenarios based on different assumptions for the means of transportation used by the children to reach their school.

- *Model W.* It assumes that all students walk to school with speeds ranging between 5 km/hr to 0 km/hr (Samimi and Ermagun, 2013; Macharia et al., 2017; Mehdizadeh et al., 2017; Alegana et al., 2021) as detailed in Table 3.
- *Model WB.* Travel to school is assumed to be a combination of walking and bicycling with a maximum speed of 10km/hr for bicycles.
- *Model WM.* Travel to school is assumed to be combination of walking and motorized transport (motorcycles, private and public vehicles). The maximum motorised speed was 50km/hr.

Table 3 shows how velocities vary in each transport scenario for different types of terrains.

Table 3: Speeds assigned to generate catchments for Models W (Walk only), WB (Walk and Bicycle) and WM (Walk and motorised). All speeds are in kilometer/hour and take a walking mode of transport unless where otherwise stated

Land and road type	**Model W**	**Model WB**	**Model WM**
Tree cover	3.5	3.5	3.5
Shrub Cover	4.5	4.5	4.5
Grassland	4	4	4
Cropland	3.5	3.5	3.5
Regularly Flooded	0	0	0
Sparse Vegetation	4.5	4.5	4.5
Bare Areas	5	10 (*Bicycling*)	5
Built Up Areas	5	10 (*Bicycling*)	5
Open Water	0	0	0
Primary road	5	5	50 (*Motorized*)
Secondary road	5	5	30 (*Motorized*)
County road	5	10 (*Bicycling*)	25 (*Motorized*)
Rural road	5	10 (*Bicycling*)	5

To generate SCAs for all schools in the region, we used the "accessibility" module of AccessMod to run an anisotropic travel time computation for each of three aforementioned travel models. The anisotropic option considers the slopes derived from the Digital Elevation Model (DEM) (Table 2) and travelling towards the school (direction of travel) to correct for walking 10 and bicycling speeds (Ray and Ebener, 2008; Tobler, 1993). Hence, we express the adjusted walking speed W_v as

$$w_v = w_f \times e^{\exp\{-3.5|k+0.05|\}} \tag{10}$$

where w_f is the speed on flat surface on the landcover considered and k is the slope derived from the elevation.

The adjustment for the walking mode of travel decreases walking velocities as the slope increases, while increasing the walking speed for a negative slope, according to the Tobler's hiking function (Tobler, 1993). Bicycling velocities are adjusted assuming that the increased speed on negative slopes does not exceed twice the speed on flat ground (Ray and Ebener, 2008). The motorized speeds were not adjusted for the slope as vehicles are powered by an engine. Figure 2 provides three example routes that help to illustrate how the chosen parametrizations for the different velocities affect travel time. Using a regular grid with spatial resolution of 20-metres, we estimate that a student riding a bicycle on Route 3 (blue-low class road), would travel at an average speed of 10km per hour taking a total of 14 minutes to reach their school. On Route 2, a student will first walk through the area without a road at 5km per hour (6 minutes), take a vehicle at the bus stop travelling at a speed of 50km/hr (2 minutes), taking a total of 8 minutes to reach their school. Finally, a student taking Route 1 (purple- high road class) would take 4 minutes using a motorized vehicle at 50 km per hour. The routes are least cost paths determined by the cost distance algorithm with speed adjusted as described above.

We used the 'cost allocation' option in AccessMod to compute the cost allocation grid delineating all SCAs. The cost allocation algorithm is similar to that used for a Voronoi diagram in Euclidean distance analysis, or to a location-allocation model in a network analysis (Ouma et al., 2021). The generated SCAs for the 84 sampled schools were then spatially linked to the 2009 population density maps (Stevens et al., 2015) and 10,000 samples for the residence residence locations, X, of the children were generated. The choice of least cost distance and cost allocation algorithms to model travel time and SCAs, respectively was appropriate given they account for transport factors (Table 2) and corresponding speeds (Table 3). This is unlike the use of straight line distances, Voronoi diagram and location-allocation model that do not account for these travel factors. See Ouma et al. (2021) and Macharia et al. (2021b) for further details.

Finally, three geostatistical models under the different travel scenarios described above were fitted; the mesh used to define the piece-wise linear approximation of the spatial Gaussian process is shown in Figure S1.12 of *Supplementary material 1.* In addition to these models, we also fit two geostatistical models that use the school location to model the spatial correlation in the data but make different use of the covariates: the SL model uses the value of the covariate at the location of the school; the SCLA model uses the averaged covariate within SCAs.

3.3. Simulation study

We carry out a simulation study to pursue two objectives:

1. to quantify the inferential benefit of accounting for the uncertainty in the location of residence;
2. to understand how the mis-specification of the mode of travel may affect the predictive inferences for prevalence.

To this end, we then proceed through the following iterative steps.

Step 1. Generate a data-set of binary outcomes indicating the malaria status of children, under the geostatistical model that assumes that children use “walking” only (Model W) as a mode of travel to school.

Step 2. Fit five statistical models to the simulated data-set from the previous step: model W (the true model), model WB, model WM, model SL and model SLCA.

Step 3. Generate prevalence predictions $\hat{p}(x)$ and exceedance probabilities $e(x)$ for a 30% prevalence threshold, over a regular grid covering the study area, for each of the five models.

Step 4. Using the exceedance probabilities $e(x)$ from the previous step, classify each pixel as being above 30% if $e(x) > l$ and below 30% if $e(x) < l$, where l is obtained by maximizing the specificity and sensitivity of the classification.

Step 5. Repeat Step 1 to Step 4, 1,000 times.

Let $p(x)$ denote the true prevalence. Using a regular grid $(x_1^*, \ldots, x_q^*)$ at 10-meter resolution, we then summarize the predictive performance using the following indices.

- Average bias: $\sum_{j=1}^{q}(\hat{p}(x_j^*) - p(x_j^*))/q$
- Root-mean-square prediction error: $[\sum_{j=1}^{q}(\hat{p}(x_j^*) - p(x_j^*))^2/q]^{1/2}$
- Sensitivity: the average proportion across all simulations of pixels that are correctly classified as exceeding 30% prevalence based on the classification in Step 4.
- Specificity: the average proportion across all simulations of pixels that are correctly classified as non-exceeding 30% prevalence based on the classification in Step 4.

3.4. Results

Geographic accessibility (travel time in minutes) to the nearest public day primary school in Western Kenya was highly heterogeneous in 2009. The combination of walking and motorized transport (*Model WM*) provided the fastest option of reaching the nearest school (ranging 0 to 128 minutes) (see Figure S1.7 in *Supplementary material 1*) while walking-only scenario (*Model W*) ranged between 0 and 233 minutes (Figure 3). Walking and bicycling combined (*Model WB*) ranged between 0 and 203 minutes (see Figure S1.6 in *Supplementary material 1*). The higher travel times was common in areas adjacent to the mountains and at county border regions. Overall, the majority of the school going children had good geographic access to their nearest primary schools in 2009. Across the eight counties, 68.4%, 74.1% and 78.3% of all the school going children (2.2 million) in the region in 2009 were within 30 minutes of the nearest primary school for models W, WB and WM, respectively.

A universal gold standard (threshold) of travel time/distance to the nearest primary school does exist due to differences in population distribution, context, geography, infrastructure, and resources between countries. In Kenya, The Ministry of Education aims to have a school within 2 km walking distance of every household which is about 24 minutes based on an average walking speed of 5 km/hr. The average travel time (distance) in the region was 28 minutes (2.33 km) while 65% of the school-going children were within 24 minutes. The western Kenya average is fairly comparable to Rwanda (1.7K), higher than Gutemala (1.1km), South Africa (1.1k) and Peru (1.4 km) but lower than Tanzania (5.9km) (Rodriguez-Segura and Kim, 2021). However, these estimates should be interpreted with caution given differences in methods, input data and context.

The travel time was used as the basis for generating SCAs for each of the three likely travel scenarios to a school. The catchment areas were generated for all the 2170 schools in Western Kenya in 2009 to account for competition of the neighbouring schools for the sampled schools. A subset comprising the 84 sampled schools was then retained for the subsequent analysis. The results of the catchment areas are shown in Figures S1.8 to S1.10 of *Supplementary material 1.* Each school catchment area covered all areas nearer to it (based on travel time) than any other school based on the cost allocation algorithm. This meant that all areas and school going children were covered by their nearest school. The size and shape of the catchment areas were variable within and between models. However, neither of the models provided a systematically smaller or bigger sized catchment area relative to the other catchments from alternative travel scenarios. Figure 4A shows school catchment areas generated using the three travel scenarios (*Models W, WB and WM*) for a Nasianda Primary School where 108 students were surveyed. Figure 4B shows generated residential locations, X, for one iteration for each child sampled from Nasianda Primary school, based on population density as the intensity of a inhomogenous Poisson process.

From the explanatory analysis, enhanced vegetation index was excluded from the analysis, since this was found to be highly correlated with both the precipitation and the temperature. The three remaining covariates showed an approximately linear relationship with the empirical logit (see Figure S1.11 in *Supplementary material 1*). Consequently, temperature, precipitation and night time lights were used as spatial predictors for all the five models considered (*Models SL, SLCA, W, WB and WM*) for school going children in Western Kenya.

All Five models considered provide a similar spatial pattern and identify the Western region as an area of high malaria prevalence in 2009. The north-Western region had the highest predicted prevalence (over 50%), while in the north-east the values range between 30% and 50%. In the Southern region, we find the lowest values of prevalence, ranging approximately between 10% and 30%. Only few small areas showed a predicted prevalence below 10%.

Overall across the models, the highest predicted values of prevalence is 77.5% according to model W (X sampled within SCAs based on walking scenario) followed by *Models WM and WB* with slightly lower values of 75.6% and 75.1%, respectively. The SLCA (school's location - SCA averaged covariates) model had a maximum predicted prevalence of 72.6% while model SL (school's location) had 69.15% (Figure 5).

Based on each of the five models, we also estimated the population of school-going children living in areas having a prevalence of at least 30% in 2009. The SL and SLCA models would have the smallest proportion of about 42.1 % and 42.6% respectively. On the other hand, the W, WB and WM models yields similar estimates of 44.3%, 44.6% and 43.5% respectively.

These results suggest that difference between the five models considered can be found in localized areas of the study region. A close inspection of Figure 5 confirms this, where we used purple ovals to highlight those areas. For example, in Busia county, corresponding to the leftmost ring of Figure 5), the SL and SLCA models classified areas in the ring as over 50% while the other models predicted the prevalence to be between 30% and 50%

In addition to the maps of the predicted pravalence, we also compared the exceedance probabilities (EPs) that malaria prevalence lies above 30%, as shown in Figure 6. The differences between the five models are more stark in these maps, where we highlighted areas that are at least 90% likely to exceed 30% in red. The extents where malaria prevalence was greater than 30% with over 90% probability was dominant in the north west region. The maps clearly show that the SL and SLCA models identifies a large contiguous areas of EPs larger than 90%, whilst the other three models yield a more heterogeneous pattern in the Northern part of the study region. Unlike the SL and SLCA models, in the Southern part, a relatively large area is shown to have an EP larger than 90% based on W, WB and WM. Approximately 23.1 % and 22.6% of the school going children were within the areas where prevalence was classified as over 30 % with a probability of at least 90% based on *SL and SLCA models* respectively. On the other hand, *models W, WB, WM* has approximately 19.2 %, 16.7 % and 14.3 %, respectively school going children with the same margins.

The results of the simulation study are presented in Table 3.3. Except for WM model which provides the worst performance, the differences between the other models in terms of the four metrics used, are rather small. These small differences can be explained by the fact that the catchment areas are relatively small to the scale parameter ϕ, hence school locations can be used to approximate the residual spatial correlation structure reasonably well in this instance.

Table 4: Summaries of the simulation study of the application presented in Section 3; for more details see Section 3.3.

Model	Bias	RMSE	Sensitivity	Specificity
W	0.002	0.176	0.741	0.757
WB	-0.005	0.190	0.716	0.729
WM	-0.008	0.257	0.569	0.581
SL	-0.001	0.169	0.764	0.756
SLCA	-0.001	0.169	0.765	0.756

4. Application 2: Small scale mapping of malaria prevalence in Western Highlands of Kenya

4.1. Data

In this section, we analyse data from a school-based survey of malaria prevalence conducted in 2010 in a small area within the region encompassed by the first application as shown in Figure 1. The full details of the survey are provided elsewhere (Stevenson et al., 2013). The data consists of 46 sampled primary schools, randomly selected from a census of all public primary schools in the area. At each school, 11 boys and 11 girls per class from classes 2 to 6 were selected randomly for a total of 4,852 children. Those aged 8-14 and over 17 years were 90% and 1.2% of the total children, respectively. Each sampled child provided a finger-prick blood sample that was used to detect HRP as evidence of recent *Plasmodium falciparum* infection using RDT (Paracheck, Orchid Biomedical Systems, India). The compound of each child sampled at school was located and mapped using a personal digital assistant (PDA) with GPS receiver (Stevenson et al., 2013) and as illustrated in Figure 1. Important distinction between the first and the second application, is that the latter mapped the compound/households of the school children, the gold standard, while the former did not.

The set of spatial covariates in the first application listed in Table 1 were used. However, because of the low variation of these across the study site of this second application, we carry out the analysis without using any covariates. Factors that affect travel to school are shown in Figure S2.1 (*Supplementary material 2*).

4.2. Model specification for [X]

We used the same approaches to estimate travel time and model SCAs as defined in the first application in terms of factors that affect travel (Table 2) and corresponding travel speeds (Table 3) for the three apriori defined travel scenarios (*W, WB and WM*). Likewise, 10,000 sample residence locations, X, of the children were generated constrained by the SCAs from the 46 sampled schools.

Five geostatistical models were then fitted. The first three were based on sampled residence locations, X using the mesh shown in Figure S2.6 of *Supplementary material* 2. The other two models were based on school location (*model SL*) and household location, the *gold standard* model denoted as *model HL.*

In this application, the HL model, which was not available in the first application, represents our gold-standard reference which we use to discriminate which model delivers the best predictive performance. Hence, we used the following indicators to quantify which model yields a predictive surface for prevalence that follows more closely the HL model.

- Average bias: $\sum_{j=1}^{q}(\hat{p}(x_j^*) - \hat{p}_{HL}(x_j^*))/q$
- Root-mean-square prediction error: $[\sum_{j=1}^{q}(\hat{p}(x_j^*) - \hat{p}_{HL}(x_j^*))^2/q]^{1/2}$

In the above expressions, $\hat{p}_{HL}(x_j^*)$ denotes the predicted prevalence a grid location x_j^* from the HL model.

4.3. Results

Travel time (in minutes) to the nearest primary school and the corresponding SCAs are shown in Figure S2.2 of the *Supplementary material 2.* Travel time to the nearest school was high and heterogeneous across the three models. Almost all the children were within half hour of the nearest primary school in the three travel scenarios. Specifically, 97.6%, 98.3% and 97.8% of all school going children in the area (347,013) were within half-hour threshold for models *W, WB and WM*, respectively. Those outside the threshold were adjacent to a river and on the edge of the study area.

All the five geostatistical models provide a similar spatial pattern of malaria prevalence in 2010 (Figure 7). Overall, the malaria prevalence in the area is highly variable. The north-western region had the highest predicted prevalence of over 30% with a contiguous area of over 50%. The rest of the areas had prevalence of between 10% and 30% and a few areas in the southern and eastern border with less than 10% (Figure 7). The gold standard model (*model SL*), had a maximum prevalence of 83.2%, the highest recorded among the five models in the area. This prevalence was closer to the maximum values recorded by the models accounting for uncertainty (75.3%, 77.4%, 76.0% for *models W, WB and WM*, respectively) relative to model (*SL*) estimate of 70.8%.

When we spatially overlaid population distribution maps with the predicated prevalence for each model, there were differences in the proportion of school-going children living in high malarious areas (at least 30% prevalence) in 2010. A third (33.1%) of the school-going children lived in areas with a prevalence of at least 30% based on model HL. The WB model had the closest proportion (27%) to model HL while, models SL W, and WM yields the lowest proportion of children within similar margins, 24.1%, 24.4% and 23.2%, respectively.

Similarly, we also compared EPs that malaria prevalence lies above 30%. In Figure 8, we highlight areas (in red) that are at least 90% likely to exceed 30% which were predominant in the north-west region. All the models identify a large contiguous area meeting the criterion in north-west region. However, the contiguous area has smaller geographical extents for model HL relative to the other 4 models. In addition to the large area, the models HL, W, WB and WM identify other smaller patches meeting the criteria which are not identified by the school location model. Approximately 10.3 % of the school going children resided in areas where prevalence was over 30 % with a probability of at least 90% based on models HL and SL. The models accounting for location uncertainty, W, WB and WB had slightly higher proportion, 16.7 %, 16.8 %, and 15.8 % of the school going children, within the same margins, respectively.

Table 5 shows the average root-mean-square and bias metrics (see Section 4.2) which are used to compare how well each of the four considered models can yield a predictive risk surface for prevalence that more closely follows most the one generated by the geostatistical model based on the actual residence locations of the students. The results indicate that, albeit marginally, the model based on the school locations generates a predictive prevalence surface that better approximate the predictions generated by the HL model than the other three models that use school catchment areas. This can be explained in this case by the fact that the accuracy with which estimated catchment areas can approximate the actual catchment areas varies greatly across the study site. Figure S2.3 to S2.5 in *Supplementary material 2* shows an overlay of the modelled SCAs and the actual household location. Overall, the models fairly approximated the actual SCAs represented by household locations. Model W had the highest number of household within modelled SCAs (74.4%) while model WM had the least (68.8%) as shown in Table 1 of *Supplementary material 2*. Per modelled SCA, the number of households within modelled SCA was highly variable ranging from only 22.8 % to 100% of the households. The performance was especially poor where there were several schools in close proximity.

Table 5: Average root-mean-square-error (RMSE) and bias as specified in Section 4.2.

Model	Average RMSE	Average Bias
SL	0.071	0.005
W	0.082	0.007
WB	0.085	0.007
WM	0.082	0.007

5. Discussion

In this paper, we have introduced a geospatial framework for the geostatistical analysis of school malaria surveys data with incomplete spatial information on the residential addresses where disease exposure occurs. The solution that we have introduced in the paper can be summarized in three main steps. The first step requires the formulation of a suitable statistical model for the unobserved residence locations of the children attending school. This was achieved by generating school catchment areas based on factors that affect travel to primary schools and sampling possible residential locations within the catchments. In the second step, we used the proposed statistical model for the unobserved residence locations to generate samples of residence locations to average the likelihood function and carry out parameter estimation. The third step consisted of carrying out prediction for disease prevalence by plugging-in the parameter estimates obtained in the second step and generating predictive samples of prevalence at selected prediction locations. This framework provides a statistically rigorous approach to propagate the uncertainty arising from the missingness of residence locations, which in standard analyses of school malaria data is reduced to the location of the school (Macharia et al., 2018; Runge et al., 2020; Stensgaard et al., 2011).

One of the main benefits of the proposed modelling framework is it provides a solution to several other statistical problems that have been covered in this paper: combining data from multiple surveys, some of which may have missing information on the residence of sampled individuals; reducing the bias in the estimation of regression relationships between prevalence and disease risk factors, induced by the aggregation of spatial information to a single location. However, as suggested by the results of the applications, the extent to which those issue can be successfully tackled is largely dependent on how well the school-catchment-area models allow for reliable inferences on the residence locations.

In our first case study, whilst areal-level summaries, such as the total population falling in areas with an exceedance probability over 90% for a 30% prevalence thresholds, were moderately similar (ranging between 23% and 14%) across the models considered, we found substantial localised differences in both the predicted prevalence (Figure 5) and the exceedance probabilities (Figure 6) between models that accounted for location-uncertainty (*models W, WB and WM*) and those that did not, based on school locations (*models SL and SLCA*. However small these areas may be, this aspect is especially important to consider when targeting and prioritizing specific areas with suitable malaria control activities. Among the models that accounted for location-uncertainty, the differences were less strong, suggesting that the assumptions made by different geographical access models may not strongly affect the inferences on disease prevalence.

In the second application, we assessed how similar are the inference between geostatistical models that incorporate school-catchment area models and geostaistical models that use the actual residence locations. The results suggested that inaccurate school catchment areas can have a material impact on geostatistical inferences and a simpler model that only uses the school locations can deliver better predictive performances. Future research should thus focus on improving the catchment area models which provide a way of developing more realistic geostatistical models than those simply use school locations. For example, in the first application, if more accurate information were available, such as the *ward* (the smallest administrative unit in Kenya) or enumeration area in which the students resided, the boundaries of these administrative areas could be used as an alternative to the modelled school catchment areas. However, the improvement accrued by exploiting this information will vary according to the size of those administrative areas relative to the school catchment area.

In both applications, among the three models considered for the school catchment area, the one assuming “walking only” (i.e. Model W) as a mode of travel to school is the one more strongly supported by previous studies. It is in fact more plausible that majority of the children walk to reach their school, since this mobility pattern has been observed in relation to other service points (e.g., healthcare providers) in Western Kenya (Salon and Gulyani, 2010; Dixit et al., 2016), as a result of low levels of ownership of motor vehicles Macharia et al. (2021a). However, more recently, it has also been observed that the number of motorcycles, locally known also known as “boda boda”, have been increasing especially in Western Kenya (Macharia et al., 2021a), with a smaller proportion of children more likely to use the public transport or private vehicles. For these reasons, more accurate information on the mode of travel of children could help to significantly improve the methods presented in this paper, which can be applied to other infectious diseases that are monitored using school data, as in the case of helminth and schistosomiasis infections (Hodges et al., 2011; Tchuem Tchuenté et al., 2012; Soares Magalhães et al., 2011; Gitonga et al., 2010; Brooker et al., 2009; Ashton et al., 2015; Mathanga et al., 2015).

The framework developed is not only applicable to school survey data but also to other data-scenarios where there is missing information on the location where most of the exposure to the disease is likely to occur. For example, in some households surveys and routinely collected data, due to confidentiality reasons, the residence location cannot be made available for analysis. An example of this is given by the Demographic and Health Survey (DHS). DHS are nationally representative household surveys that have been conducted in more than 85 countries since 1984 to collect demographic and health data (Corsi et al., 2012). However, to reduce disclosure risk in DHS, a cluster is assigned the coordinates of the center of the sampled enumeration area and further randomly displaced. The framework developed can be adapted to partially account for this geomasking. This is vital because in low resource settings, such household surveys are the only source of development indicators. An important aspect of the proposed framework is that it can also be used to combine data from multiple sources with varying accuracy for the information of the residence of the sampled individuals.

As shown in this study, catchment areas models are an integral component of disease mapping and are essential in order to yield reliable inferences and summaries of uncertainty for the health outcome under investigation (Macharia et al., 2021b). While the three types of catchment areas we generated did not yield substantially different results, further research is required to improve these models. Here, each of the models assumed non overlapping catchments but, in reality, students from the same location can attend different schools and thus create overlapping catchments. To accommodate this, the attractiveness factor should not only be a function of distance or travel time, as in our case, but should account for school capacity, classroom size, number of teachers, perceived teaching quality, and previous examinations results. In our analysis, every student was assumed to attend their nearest school. This could be further improved if school attendance data were available, which would allow us to empirically identify a threshold for the distance or travel time below which students do in fact go to the nearest school. Such an approach has been implemented in the construction of health facility catchment areas by using health seeking behaviour information from DHS (Alegana et al., 2012).

In addition to the limitations outlined in the previous section, there are other limitations in the illustrated applications. In our analysis, we excluded boarding schools, which may lead to the underestimation of the travel time and thus overestimate the size of SCAs. However, there were only nine schools that were were purely boarding out of 2170 in Western Kenya in 2009. The access models were parametrized using speeds and modes of transport from other studies (Samimi and Ermagun, 2013; Macharia et al., 2017; Mehdizadeh et al., 2017; Alegana et al., 2021) since we did not have data for Western Kenya. Future studies should consider collecting data on mode of transport, speeds and the utilized school in this region for improved modelling of accessing metrics and SCAs. The chosen travel route to school is a complex process influenced by socio-economic factors and facilitators of movement such as roads and obstacles. We assumed that students do not bypass their nearest school, but a small proportion likely bypass their nearest school due to poverty, affordability, or parents' past education experience among other factors.

Acknowledgements

The authors wish to thank members of the KEMRI – Eastern and Southern Africa Centre of International Parasite Control (ESACIPAC), Drs Charles Mwandawiro, Jimmy Kihara and Sammy Njenga for their assistance during the 2009 school surveys and Gillian H. Stresman for assistance with data used in the second application.

Conflict of interest

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Study protocol for application one (Gitonga et al., 2010) received ethical approval from the Kenya Medical Research Institute and National Ethics Review Committee (numbers 1407 and 1596). Additional approval was provided by the Permanent Secretary's office of the Ministry of Education (MoE) and the Division of Malaria Control, Ministry of Public Health and Sanitation. All national, provincial and district-level health and education authorities were briefed about the survey purpose and selected schools. Official letters of support were prepared by Provincial MoE officers. Study protocol for application two (Stevenson et al., 2013) was approved by the ethical committees of the London School of Hygiene & Tropical Medicine and the Kenya Medical Research Institute and was part of a larger government-lead, national school survey programme (Gitonga et al., 2010).

Availability of data and materials

Data that support the findings of this study are available at Population Health Dataverse Gitonga et al. (2022), generated gridded surfaces are located at https://figshare.com/s/2150ec75f50dda325d59 while the codes used in this analysis are available at https://github.com/giorgilancs/mbgmissinglocations.

Authors' contributions

EG and PMM: Conceptualization, data curation, formal analysis, funding acquisition,investigation, methodology, software, validation, visualization, writing original draft, review and editing. NR: software, methodology, writing - review and editing. CWG: Data curation, funding acquisition, writing - review and editing. RWS: Conceptualization, data curation, funding acquisition, investigation, writing - review and editing

Funding

This work was supported by Newton International Fellowship (number NIF/R1/201418) of the Royal Society awarded to PMM and EG. EG was also supported by a Springboard grant (number SBF004/1009) of the Academy of Medical Sciences. RWS is supported as a Wellcome Trust Principal Fellow (number 212176). CWG was supported by a Commonwealth PhD scholarship from the UK's Department for International Development. Field work for the 2009 school survey was provided by the Division of Malaria Control, Ministry of Public Health and Sanitation through a grant from the Department for International Development (DfID) through the WHO Kenya Country Office and the Wellcome Trust UK as part of fellowship support to AM Noor (number 081829) and RW Snow (number 079080). PMM, CWG and RWS are grateful to the support of the Wellcome Trust to the Kenya Major Overseas Programme (number 203077).

References

Alegana, V., Pezzulo, C., Tatem, A., Omar, B., Christensen, A., 2021. Mapping out-of-school adolescents and youths in low-and middle-income countries. Humanities and Social Sciences Communications 8, 1–10.

Alegana, V.A., Okiro, E.A., Snow, R.W., 2020. Routine data for malaria morbidity estimation in africa: challenges and prospects. BMC medicine 18, 121.

Alegana, V.A., Wright, J., Petrina, U., Noor, A.M., Snow, R.W., Atkinson, P.M., 2012. Spatial modelling of healthcare utilisation for treatment of fever in Namibia. International Journal of Health Geographics 11, 6.

Ashton, R.A., Kefyalew, T., Rand, A., Sime, H., Assefa, A., Mekasha, A., Edosa, W., Tesfaye, G., Cano, J., Teka, H., Reithinger, R., Pullan, R.L., Drakeley, C.J., Brooker, S.J., 2015. Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys. American Journal of Tropical Medicine and Hygiene 93, 168–177. doi:10.4269/ajtmh.14-0620.

Ashton, R.A., Kefyalew, T., Tesfaye, G., Pullan, R.L., Yadeta, D., Reithinger, R., Kolaczinski, J.H., Brooker, S., 2011. School-based surveys of malaria in Oromia Regional State, Ethiopia: A rapid survey method for malaria in low transmission settings. Malaria Journal 10, 1–13. doi:10.1186/1475-2875-10-25.

Bayoh, M.N., Lindsay, S.W., 2004. Temperature-related duration of aquatic stages of the Afrotropical malaria vector mosquito Anopheles gambiae in the laboratory. Medical and Veterinary Entomology 18, 174–179. doi:10.1111/j.0269-283X.2004.00495.x.

Biggeri, A., Catelan, D., 2012. Disease Mapping. 1979. doi:10.1002/9781119940012.ch9.

Brooker, S., Kolaczinski, J.H., Gitonga, C.W., Noor, A.M., Snow, R.W., 2009. The use of schools for malaria surveillance and programme evaluation in Africa. Malaria Journal 8, 1–9. doi:10.1186/1475-2875-8-231.

Busetto, L., Ranghetti, L., 2016. MODIStsp: An R package for automatic preprocessing of MODIS Land Products time series. Computers and Geosciences 97, 40–48. doi:10.1016/j.cageo.2016.08.020.

Cameron, E., Young, A.J., Twohig, K.A., Pothin, E., Bhavnani, D., Dismer, A., Merilien, J.B., Hamre, K., Meyer, P., Le Menach, A., Cohen, J.M., Marseille, S., Lemoine, J.F., Telfort, M.A., Chang, M.A., Won, K., Knipes, A., Rogier, E., Amratia, P., Weiss, D.J., Gething, P.W., Battle, K.E., 2021. Mapping the endemicity and seasonality of clinical malaria for intervention targeting in haiti using routine case data. eLife 10, e62122. URL: https://doi.org/10.7554/eLife.62122, doi:10.7554/eLife.62122.

Clements, A.C., Lwambo, N.J., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi, S., Webster, J.P., Fenwick, A., Brooker, S., 2006. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in tanzania. Tropical medicine & international health 11, 490–503.

Corsi, D.J., Neuman, M., Finlay, J.E., Subramanian, S., 2012. Demographic and health surveys: a profile. International journal of epidemiology 41, 1602–1613.

Diggle, P.J., Moraga, P., Rowlingson, B., Taylor, B.M., 2013. Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. Statistical Science 28, 542 – 563. URL: https://doi.org/10.1214/13-STS441, doi:10.1214/13-STS441.

Diggle, P.J., Tawn, J., Moyeed, R., 1998. Model-based Geostatistics. Applied StatisticsStatistics 47, 299–350. doi:10.1007/978-0-387-98135-2, arXiv:arXiv:1011.1669v3.

Dixit, A., Lee, M.C., Goettsch, B., Afrane, Y., Githeko, A.K., Yan, G., 2016. Discovering the cost of care: consumer, provider, and retailer surveys shed light on the determinants of malaria health-seeking behaviours. Malaria journal 15, 179. doi:10.1186/s12936-016-1232-7.

Drake, T.L., Okello, G., Njagi, K., Halliday, K.E., Jukes, M.C., Mangham, L., Brooker, S., 2011. Cost analysis of school-based intermittent screening and treatment of malaria in Kenya. Malaria Journal 10, 1–11. doi:10.1186/1475-2875-10-273.

Dutta, H.M., Dutt, A.K., 1978. Malarial ecology: a global perspective. Social Science and Medicine 12, 69–84.

Fornace, K.M., Fronterrè, C., Fleming, F.M., Simpson, H., Zoure, H., Rebollo, M., Mwinzi, P., Vounatsou, P., Pullan, R.L., 2020. Evaluating survey designs for targeting preventive chemotherapy against schistosoma haematobium and schistosoma mansoni across sub-saharan africa: a geostatistical analysis and modelling study. Parasites & vectors 13, 1–13.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. Scientific Data 2, 1–21. doi:10.1038/sdata.2015.66.

Gitonga, C., Noor, A., Brooker, S., Snow, R., 2022. Nationwide school health surveys 2009-2010 in Kenya. URL: https://doi.org/10.7910/DVN/UQLTO5, doi:10.7910/DVN/UQLTO5.

Gitonga, C.W., Karanja, P.N., Kihara, J., Mwanje, M., Juma, E., Snow, R.W., Noor, A.M., Brooker, S., 2010. Implementing school malaria surveys in Kenya: Towards a national surveillance system. Malaria Journal 9, 1–13. doi:10.1186/1475-2875-9-306.

Guagliardo, M.F., 2004. Spatial accessibility of primary care: concepts, methods and challenges. International journal of health geographics

3, 3. URL: http://www.ij-healthgeographics.com/content/3/1/3, doi:10.1186/1476-072X-3-3.

Hodges, M., Dada, N., Wamsley, A., Paye, J., Nyorkor, E., Sonnie, M., Barnish, G., Bockarie, M., Zhang, Y., 2011. Improved mapping strategy to better inform policy on the control of schistosomiasis and soil-transmitted helminthiasis in sierra leone. Parasites & vectors 4, 1–7.

Joseph, N.K., Macharia, P.M., Ouma, P.O., Mumo, J., Jalang'o, R., Wagacha, P.W., Achieng, V.O., Ndung'u, E., Okoth, P., Muñiz, M., Guigoz, Y., Panciera, R., Ray, N., Okiro, E.A., 2020. Spatial access inequities and childhood immunisation uptake in Kenya. BMC public health 20, 1407. doi:10.1186/s12889-020-09486-8.

Kabaria, C.W., Gilbert, M., Noor, A.M., Snow, R.W., Linard, C., 2017. The impact of urbanization and population density on childhood Plasmodium falciparum parasite prevalence rates in Africa. Malaria Journal 16, 1–10.

Knowles, S.C., Sturrock, H.J., Turner, H., Whitton, J.M., Gower, C.M., Jemu, S., Phillips, A.E., Meite, A., Thomas, B., Kollie, K., et al., 2017. Optimising cluster survey design for planning schistosomiasis preventive chemotherapy. PLoS neglected tropical diseases 11, e0005599.

Krainski, E., Gómez Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H., 2018. Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. doi:10.1201/9780429031892.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 423–498.

Macharia, P.M., Giorgi, E., Noor, A.M., Waqo, E., Kiptui, R., Okiro, E.A., Snow, R.W., 2018. Spatio-temporal analysis of Plasmodium falciparum prevalence to understand the past and chart the future of malaria control in Kenya. Malaria journal 17, 340. doi:10.1186/s12936-018-2489-9.

Macharia, P.M., Mumo, E., Okiro, E.A., 2021a. Modelling geographical accessibility to urban centres in Kenya in 2019. Plos One 16, e0251624. doi:10.1371/journal.pone.0251624.

Macharia, P.M., Odera, P.A., Snow, R.W., Noor, A.M., 2017. Spatial models for the rational allocation of routinely distributed bed nets to public health facilities in Western Kenya. Malaria Journal 16, 367. doi:10.1186/s12936-017-2009-3.

Macharia, P.M., Ray, N., Giorgi, E., Okiro, E.A., Snow, R.W., 2021b. Defining service catchment areas in low-resource settings. BMJ Global Health 6, e006381.

Mathanga, D.P., Halliday, K.E., Jawati, M., Verney, A., Bauleni, A., Sande, J., Ali, D., Jones, R., Witek-McManus, S., Roschnik, N., Brooker, S.J., 2015. The high burden of malaria in primary school children in Southern Malawi. American Journal of Tropical Medicine and Hygiene 93, 779–789. doi:10.4269/ajtmh.14-0618.

Maxwell, C.A., Wakibara, J., Tho, S., Curtis, C.F., 1998. Malaria-infective biting at different hours of the night. Medical and Veterinary Entomology 12, 325–327. doi:10.1046/j.1365-2915.1998.00108.x.

Mehdizadeh, M., Mamdoohi, A., Nordfjaern, T., 2017. Walking time to school, children's active school travel and their related factors. Journal of Transport & Health 6, 313–326.

Mennis, J., 2009. Dasymetric mapping for estimating population in small areas. Geography Compass 3, 727–745. doi:10.1111/j.1749-8198.2009.00220.x.

Molineaux, L., 1988. The epidemiology of human malaria as an explanation of its distribution including some implications for its control., in: Wernsdorfer, W.H.M.I. (Ed.), Malaria: principles and practice of malariology.. volume 2. ed.. Churchill Livingstone, Edinburgh, pp. 913–998.

Mulaku, G., Nyadimo, E., 2011. Gis in education planning: The kenyan school mapping project. Survey Review 43, 567–578.

Nelli, L., Ferguson, H.M., Matthiopoulos, J., 2020. Achieving explanatory depth and spatial breadth in infectious disease modelling: Integrating active and passive case surveillance. Statistical Methods in Medical Research 29, 1273–1287. URL: https://doi.org/10.1177/0962280219856380, doi:10.1177/0962280219856380, arXiv:https://doi.org/10.1177/0962280219856380. pMID: 31213191.

Noor, A.M., Kinyoki, D.K., Mundia, C.W., Kabaria, C.W., Mutua, J.W., Alegana, V.A., Fall, I.S., Snow, R.W., 2014. The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity. Lancet 383, 1739–1747. doi:10.1016/s0140-6736(13)62566-0.

Ouma, P., Macharia, P.M., Okiro, E., Alegana, V., 2021. Methods of measuring spatial accessibility to health care in uganda, in: Makanga, P.T. (Ed.), Practicing Health Geography: The African Context. Springer International Publishing, Cham, pp. 77–90.

Pop, B., Fetica, B., Blaga, M.L., Trifa, A.P., Achimas-Cadariu, P., Vlad, C.I., Achimas-Cadariu, A., 2019. The role of medical registries, potential applications and limitations. Medicine and Pharmacy Reports 92, 7.

Ray, N., Ebener, S., 2008. AccessMod 3.0: Computing geographic coverage and accessibility to health care services using anisotropic movemen of patients. International Journal of Health Geographics 7, 63. doi:10.1186/1476-072X-7-63.

Rodriguez-Segura, D., Kim, B.H., 2021. The last mile in school access: Mapping education deserts in developing countries. Development Engineering 6, 100064. doi:https://doi.org/10.1016/j.deveng.2021.100064.

Runge, M., Snow, R.W., Molteni, F., Thawer, S., Mohamed, A., Mandike, R., Giorgi, E., Macharia, P.M., Smith, T.A., Lengeler, C., Pothin, E., 2020. Simulating the council-specific impact of anti-malaria interventions: A tool to support malaria strategic planning in Tanzania. PLoS ONE 15. doi:10.1371/journal.pone.0228469.

Salon, D., Gulyani, S., 2010. Mobility, poverty, and gender: travel 'choices' of slum residents in Nairobi, Kenya. Transport Reviews 30, 641–657. doi:10.1080/01441640903298998.

Samimi, A., Ermagun, A., 2013. Students' tendency to walk to school: case study of tehran. Journal of urban planning and development 139, 144–152.

Savory, D.J., Andrade-Pacheco, R., Gething, P.W., Midekisa, A., Bennett, A., Sturrock, H.J.W., 2017. Intercalibration and Gaussian Process Modeling of Nighttime Lights Imagery for Measuring Urbanization Trends in Africa 2000–2013. Remote Sensing 9, 713. doi:10.3390/rs9070713.

Soares Magalhães, R.J., Biritwum, N.K., Gyapong, J.O., Brooker, S., Zhang, Y., Blair, L., Fenwick, A., Clements, A.C., 2011. Mapping helminth co-infection and co-intensity: geostatistical prediction in ghana. PLoS neglected tropical diseases 5, e1200.

Stanton, M.C., Diggle, P.J., 2013. Geostatistical analysis of binomial data: generalised linear or transformed Gaussian modelling? Environmetrics 24, 158–171. doi:10.1002/env.2205.

Stefan, D.C., Baadjes, B., Kruger, M., 2014. Incidence of childhood cancer in namibia: the need for registries in africa. The Pan African Medical Journal 17, 191.

Stensgaard, A.S., Vounatsou, P., Onapa, A.W., Simonsen, P.E., Pedersen, E.M., Rahbek, C., Kristensen, T.K., 2011. Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in Uganda: predictors of risk and geographical patterns of co-endemicity. Malaria Journal 10, 298. doi:10.1186/1475-2875-10-298.

Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., 2015. Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. PLoS ONE 10, e0107042. doi:10.1371/journal.pone.0107042.

Stevenson, J.C., Stresman, G.H., Gitonga, C.W., Gillig, J., Owaga, C., Marube, E., Odongo, W., Okoth, A., China, P., Oriango, R., Brooker, S.J., Bousema, T., Drakeley, C., Cox, J., 2013. Reliability of School Surveys in Estimating Geographic Variation in Malaria Transmission in the Western Kenyan Highlands. PLoS ONE 8, e77641. doi:10.1371/journal.pone.0077641.

Sturrock, H., Cohen, J., Keil, P., Tatem, A.J., Le Menach, A., Ntshalintshali, N.E., Hsiang, M.S., Gosling, R.D., 2014. Fine-scale malaria risk mapping from routine aggregated case data. Malaria Journal 13, 421. doi:10.1186/1475-2875-13-421.

Takem, E.N., Affara, M., Amambua-Ngwa, A., Okebe, J., Ceesay, S.J., Jawara, M., Oriero, E., Nwakanma, D., Pinder, M., Clifford, C., Taal, M., Sowe, M., Suso, P., Mendy, A., Mbaye, A., Drakeley, C., D'Alessandro, U., 2013. Detecting Foci of Malaria Transmission with School Surveys: A

Pilot Study in the Gambia. PLoS ONE 8, e67108. doi:10.1371/journal.pone.0067108.

Tchuem Tchuenté, L.A., Kamwa Ngassam, R.I., Sumo, L., Ngassam, P., Dongmo Noumedem, C., Nzu, D.D.L., Dankoni, E., Kenfack, C.M., Gipwe, N.F., Akame, J., et al., 2012. Mapping of schistosomiasis and soil-transmitted helminthiasis in the regions of centre, east and west cameroon. PLoS neglected tropical diseases 6, e1553.

Tobler, W., 1993. Three presentations on geographical analysis and modeling:Non-Isotropic Geographic Modeling; Speculations on the Geometry of Geography;Global Spatial Analysis. Technical Report 93-1. National Center for Geographic Information and Analysis, University of California. Santa Barbara. URL: http://www.geodyssey.com/papers/tobler93.html.

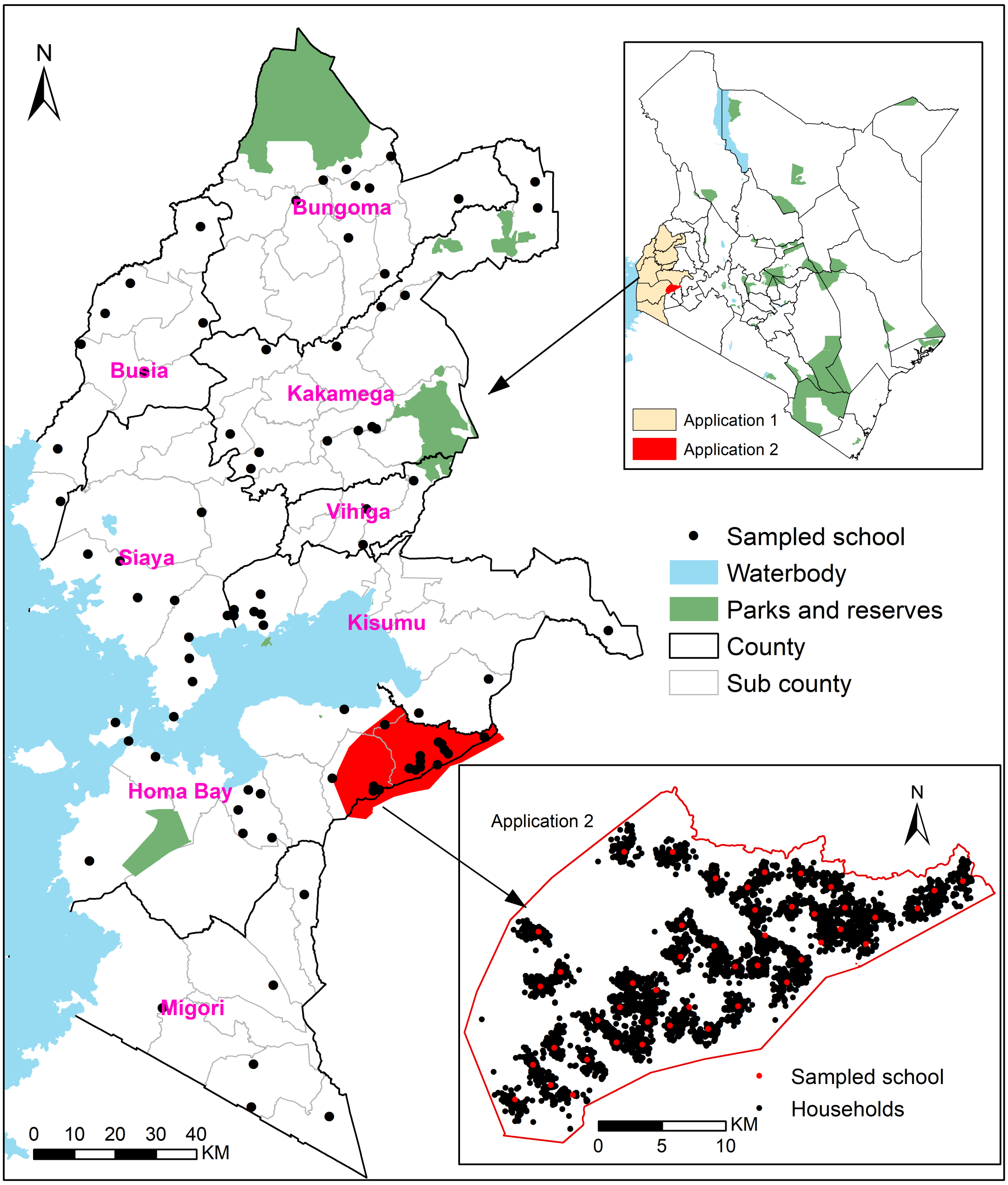

Figure 1: Study areas of the applications of Section 3 and Section 4, both located in Western Kenya. Solid black (counties) and light grey (sub-counties) lines represent administrative units . The map in the lower right corner shows the study area and sampled households from the application of Section 4.

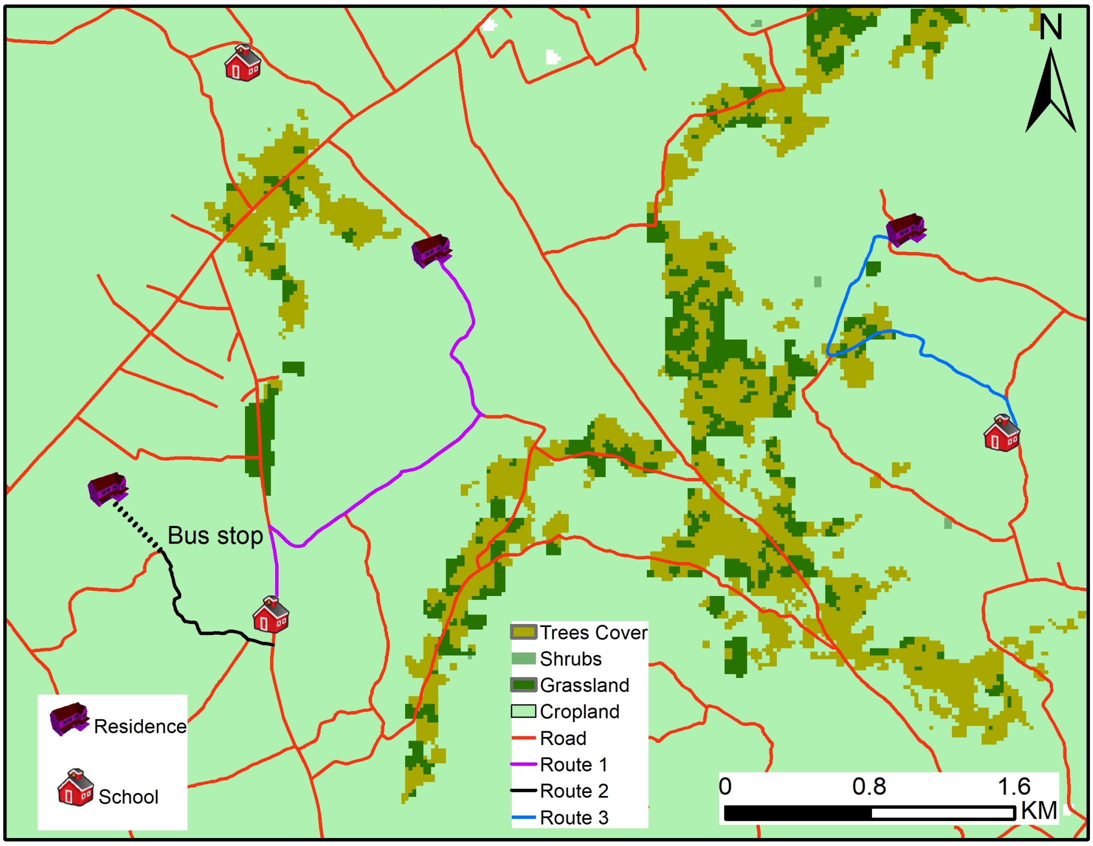

Figure 2: A map showing examples of optimal routes taken by students, under different mode of travel. For a detailed explanation, we refer the reader to the main text.

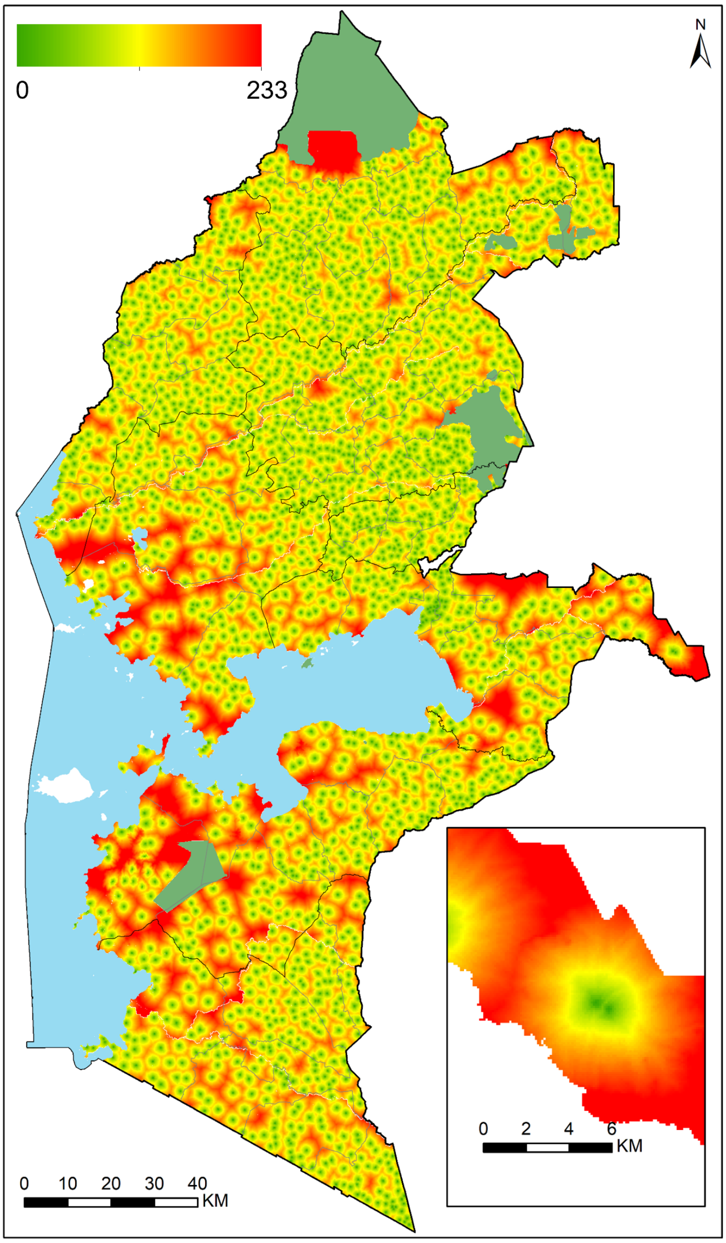

Figure 3: Travel time (in minutes) to the nearest public day primary school for all 2170 public primary schools in Western Kenya in 2009 ranging from 0 minutes (light green) to 233 minutes (red) *Model W*. Results of *Model WB* and *Model WM* are shown in *Supplementary material 1*

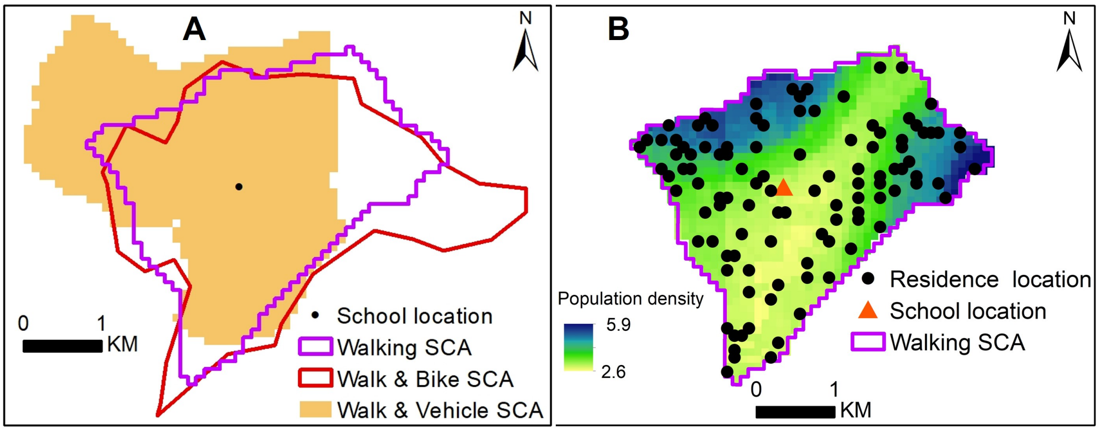

Figure 4: A: Sample of school catchment areas generated from each of transport scenarios; walking (*Model W*), walking and bicycling (*Model WB*) and walking and motorised (*Model WM*). The school is shown as a black dot; B sampled locations for a single iteration from one of the catchment areas (*Model WB*) overlaid on a population distribution map Stevens et al. (2015) (green to blue shades). In B, the school is shown as an orange triangle. All the school catchment areas are shown in *Supplementary material 1*

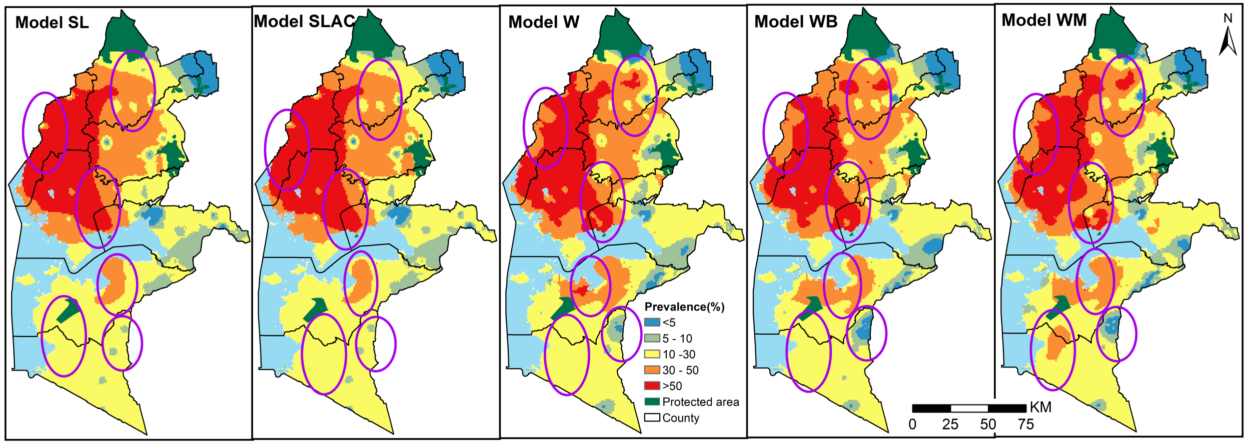

Figure 5: Annual predicted mean malaria prevalence at 1 × 1 km spatial resolution ranging from 0% (blue) to 77.5% (dark red) in Western Kenya in 2009 for 2 standard statistical models (*SL and SLCA*) and three models accounting for locations uncertainty (*W, WB, WM.* Ellipses show areas with differences. The protected areas are shown in green

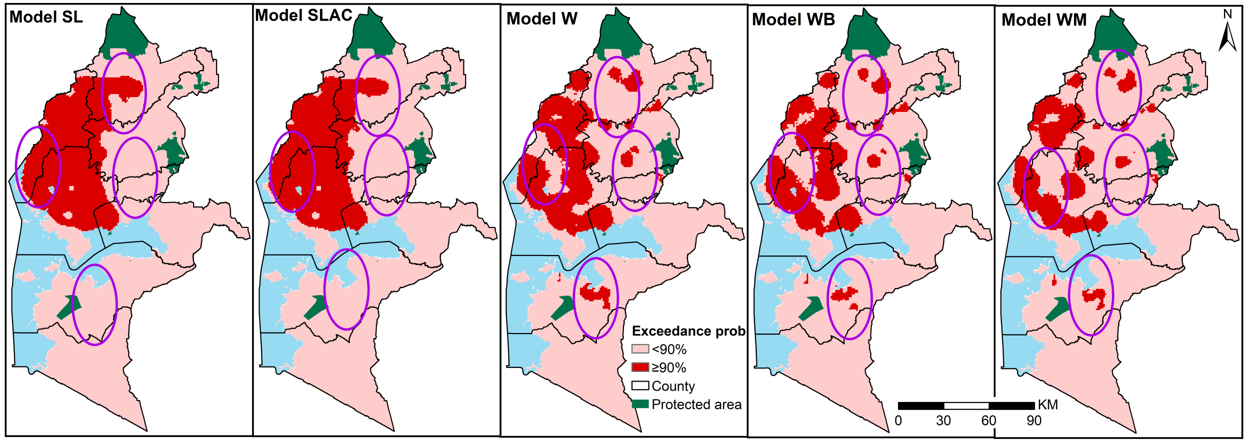

Figure 6: Maps of the predicted exceedance probability for a 30% prevalence threshold with a 90% probability on a 1 by 1 km regular grid in Western Kenya in 2009 for 2 standard statistical models (*SL and SLCA*) and three models accounting for locations uncertainty (*W, WB, WM.* Ellipses show areas with differences. The protected areas are shown in green

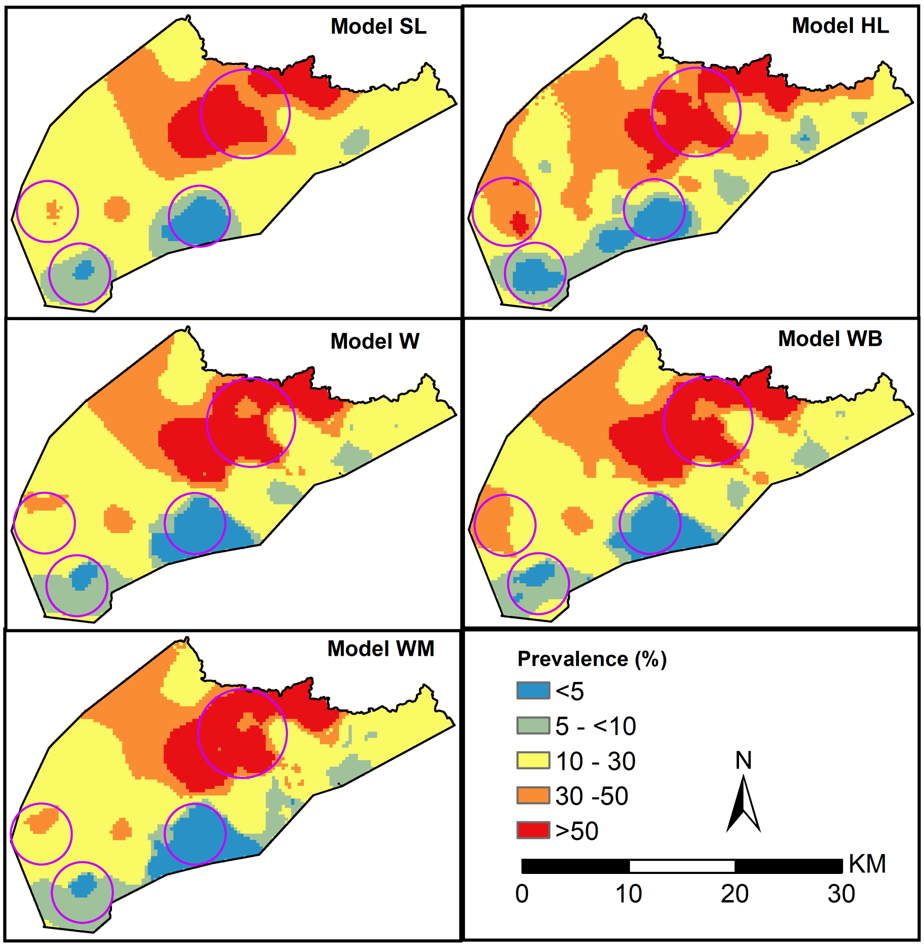

Figure 7: Predicted mean malaria prevalence at 0.3 × 0.3 km spatial resolution ranging from 0% (blue) to 83.2% (dark red) in 2010 for 5 geostatistical models based on school location (*SL*), gold standard model (*HL*) based on household location and three models accounting for locations uncertainty (*W, WB, WM*. Circles show areas with differences.

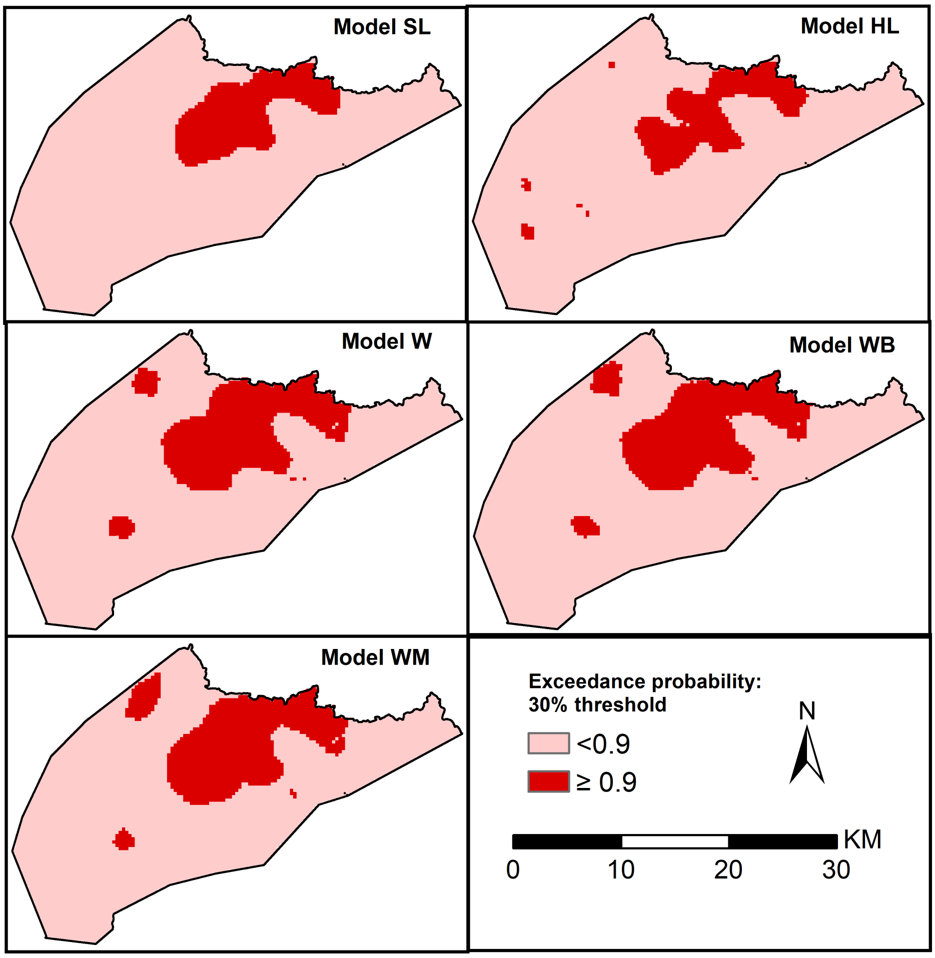

Figure 8: Maps of exceedance probability for a 30% prevalence threshold with a 90% probability on a 0.3 km regular grid for 5 geostatistical models based on school location (*SL*), gold standard model (*HL*) based on household location and three models accounting for locations uncertainty, models *W, WB and WM*

Appendix A. Supplementary material 1

Additional results for the first application.

Appendix B. Supplementary material 2

Additional results for the second application.

Credit Authors Statement

EG and PMM: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, writing original draft, review and editing. **NR**: software, methodology, writing - review and editing. **CWG**: Data curation, funding acquisition, writing - review and editing**.** **RWS**: Conceptualization, data curation, funding acquisition, investigation, writing - review and editing