

Extending multilevel spatial models to include spatially varying coefficients

Mark Janko^{a,b,*}, Varun Goel^{c,d}, Michael Emch^{c,d,e}

^a Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA

^b Duke Global Health Institute, Duke University, Durham, NC, USA

^c Department of Geography, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

^d Carolina Population Center, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

^e Department of Epidemiology, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

ARTICLE INFO

Keywords:

Health/medical geography
Spatially-varying coefficients
Multilevel models
Bayesian statistics
Disease ecology

ABSTRACT

Multilevel models have long been used by health geographers working on questions of space, place, and health. Similarly, health geographers have pursued interests in determining whether or not the effect of an exposure on a health outcome varies spatially. However, relatively little work has sought to use multilevel models to explore spatial variability in the effects of a contextual exposure on a health outcome. Methodologically, extending multilevel models to allow intercepts and slopes to vary spatially is straightforward. The purpose of this paper, therefore, is to show how multilevel spatial models can be extended to include spatially varying covariate effects. We provide an empirical example on the effect of agriculture on malaria risk in children under 5 years of age in the Democratic Republic of Congo.

1. Introduction

Geographers have long worked with both spatially referenced and multilevel data, and as the data environment continues to grow, geographers find themselves working with both simultaneously. For example, geographers frequently work with two-stage cluster surveys such as Demographic and Health Surveys, in which thousands of individuals are sampled from hundreds of spatially referenced clusters within a given country. As a result, health outcomes of interest are frequently collected at the individual level, while the spatially referenced primary sampling unit is collected at a higher level. Additionally, exposures of interest are often collected at this higher level, and health geographers have consistently been interested in exploring how the effects of exposures on health outcomes may vary spatially.

A wide range of methods has emerged in the last several decades to address each of these issues individually. For example, there is a vast literature on statistical methods for spatial modeling (Banerjee et al., 2014), as well as in multilevel modeling (Gelman and Hill, 2006). Both literatures present ways to allow covariate effects to vary. However, little work has addressed the need for models that are both multilevel and spatial within health geography, and less still on multilevel spatial models in which covariate effects vary. The purpose of this paper, therefore, is address this need by exploring the substantive questions

spatial multilevel models can help address, as well as to explore the structure of those models relative to spatial models and multilevel models.

The rest of this paper is organized as follows. Section 2 discusses recent trends in efforts to model spatial and multilevel data structures. Section 3 extends spatial multilevel models to include spatially-varying effects. In section 4, we turn to a real data example, where our focus is to understand the effect of two environmental risk factors on the probability of malaria infection among children under 5 years of age in the Democratic Republic of Congo (DRC). Section 5 concludes, and points to areas of future work.

2. Trends in recent statistical work on spatial and multilevel modeling

Public health researchers are paying increasing attention to the need to consider both space and contextual exposures in their work, and methodological developments have kept apace, with two basic model structures being widely employed. Earlier multilevel modeling work in this area was largely a spatial, with models fit to geographic data without geographic structure built into the model (Auchincloss et al., 2012; Owen et al., 2016). For example, a typical multilevel model has the following form:

* Corresponding author. Institute for Health Metrics and Evaluation, 2301 5th Ave #600, Seattle, WA, 98121, USA.

E-mail address: mjanko@uw.edu (M. Janko).

$$y_{ij} \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \theta_j, \sigma_y^2), \quad i = 1, \dots, n_j, \quad j = 1, \dots, J \quad (1)$$

$$\theta_j \sim N(0, \sigma_\theta^2), \quad j = 1, \dots, J$$

where y_{ij} represents a response variable for individual i in location j , \mathbf{x}_{ij}^T is a set of individual-level predictors and $\boldsymbol{\beta}$ a set of regression coefficients linking those covariates to the response, while θ_j is a coefficient or vector of coefficients that vary across locations with a variance σ_θ^2 . Finally, σ_y^2 is a residual variance term. In this case, the varying coefficients vary independently across locations, as no spatial structure is encoded in the model. Conversely, spatial methods have largely been employed primarily for distance calculations and spatial aggregations, with formal spatial modeling largely being done on aggregated data (Auchincloss et al., 2012). For example, a typical spatial model has the following form:

$$y_j \sim N(\mathbf{x}_j^T \boldsymbol{\beta} + \theta_j, \sigma_y^2), \quad j = 1, \dots, J \quad (2)$$

$$\theta_j \sim N(0, \sigma_\theta^2 \Sigma(\varphi)), \quad j = 1, \dots, J$$

where the index i disappears, with the response and covariates often representing averages (or counts or other summary) of individuals within each location j . Here, however, spatial variability of the θ_j term is introduced through a correlation function $\Sigma(\varphi)$ that depends on further hyperparameters φ . Examples include an exponential correlation for point-referenced data, or a CAR correlation structure for areal data.

The increase in software with the capacity to handle different correlation structures, however, has led to an increase in modeling that can address both concerns. Perhaps the most frequent (and often not explicitly mentioned) approach to incorporating both is to fit a multilevel model separate from a spatial model. For example, Author and colleagues fit a multilevel model to investigate the relationship between rurality and HIV infection prevalence in the Democratic Republic of Congo, but also provide descriptive maps of HIV prevalence using Bayesian kriging on aggregated data (Carrel et al., 2016). Work such as this is limited, however, in that the functional form of the multilevel model and the spatial model are different, with the correlation structure for the random effects and the covariate information being different across the two models. In the case of the former, missing spatial random effects precludes learning about a spatial process after accounting for individual and ecological covariates. In the case of the latter, individual covariate information is often missing in the spatial model, precluding our ability to learn about their effects, as well as the spatial process, since some of the spatial structure can likely be explained by conditioning on relevant covariate information.

Efforts to extend the basic modeling setting above have been done through comparing the two approaches—typically by way of information criteria such as AIC—in an effort to more formally understand contextual and spatial forces underlying health phenomena. Chaix and colleagues, for example, conducted two studies in which multilevel models were compared to spatial models to learn about contextual and residual spatial effects on healthcare utilization and mental health (Chaix et al., 2005a, 2005b). Finally, other work is now emerging that identifies contextual effects on health while also learning about an underlying spatial process. An example of this can be found in the work of Hajat and colleagues, who investigate associations between air pollution and individual and neighborhood level socioeconomic status (SES) (Hajat et al., 2013). In that study, the authors write down a multilevel model to identify the contextual effect of SES on exposure to pollution, and incorporate spatial correlation in the random effects. Such modeling is straightforward, with the correlation structure for the random effects in equation (2) above being substituted in for that contained in equation (1). Perhaps the most extensive example of such modeling, however, can be seen in the work of Arcaya and colleagues, who investigate variability in county-level life expectancy across the United States. In that work,

they fit a series of multilevel spatial models in which counties are nested within states, but also extend the model to consider nesting within “spatial patches” (sets of neighboring counties, regardless of state membership) (Arcaya et al., 2012). After model fitting, they then turn to look at clustering of these different random effects, and hypothesize reasons why they observe the patterns they do. For example, the apparent clustering in random effects could be due to missing state-level policy variables (such as income requirements for government healthcare programs), with neighboring states often having similar policies (Arcaya et al., 2012).

While this brief review by no means covers the vast amount of recent work in spatial and multilevel modeling, it does highlight a trend showing that model sophistication is increasing to correspond with the complexity of the data we have at hand, as well as the complexity of the spatial and contextual questions we wish to ask, and on a wide variety of substantive areas (e.g. mental health, healthcare utilization, air pollution, life expectancy). Nevertheless, there are limitations and opportunities to overcome them. In the next section, we argue that an exposure’s effect on a health outcome may vary across space, and that spatially modeling such variability represents a critical next step both in the neighborhoods and health and spatial epidemiology subfields, but should be a central goal of inference within health geography as well.

3. Multilevel modeling with spatially varying coefficients

As noted, while model complexity (and software development) has grown such that a modeler can now consider contextual effects and residual spatial structure simultaneously, gaps remain. For example, the typical implementation of a multilevel spatial model only models the intercept (or set of intercepts) as spatially varying. We argue here that a fuller model should consider models in which the effects of covariates are modeled as spatially varying as well, and that fitting such models represents a core contribution geographers can make within public health broadly. Moreover, we can justify this modeling framework both from the perspective of health geography, as well as from a more classical epidemiological perspective.

From the perspective of health geography, we may imagine that the effect of an exposure on an outcome varies spatially as a result of the multiple relevant factors for disease described in the work of Jacques May (May 1950, 1959), and in particular the practical challenge of gathering data for all of them. For example, in the case of malaria or other vector-borne disease, data on the vector is generally not collected in large surveys, and environmental exposures such as temperature and precipitation, important to many vector populations, are used in models of disease risk instead. However, the vector population is itself heterogeneous, with different species of *Anopheles* mosquitoes responsible for the transmission of malaria. These different species, moreover, have their own habitat preferences and behaviors, both of which are spatially structured, and may respond to the same environmental conditions in different ways. Fitting a multilevel model in which the effects of ecological covariates are assumed to be invariant over space will not account for this unobserved spatial heterogeneity, which will instead remain hidden on the map, and possibly result in misleading epidemiological conclusions.

The theoretically-oriented example posed above relates to core epidemiological principles as well, particularly causal inference. In the example, mediation and its role in causal inference can clearly be seen when we consider the two directed acyclic graphs (DAGs) represented in Fig. 1.

In the first DAG, the malaria outcome is a direct result of temperature and precipitation, and essentially represents the typical implementation of models of malaria risk. However, the effects of these environmental variables are best viewed through their role on the vector population, whereby changes in precipitation and temperature may increase or decrease various characteristics of that population, including abundance and biting rates (Afrane et al., 2008; Kirby and Lindsay, 2009; Lindblade

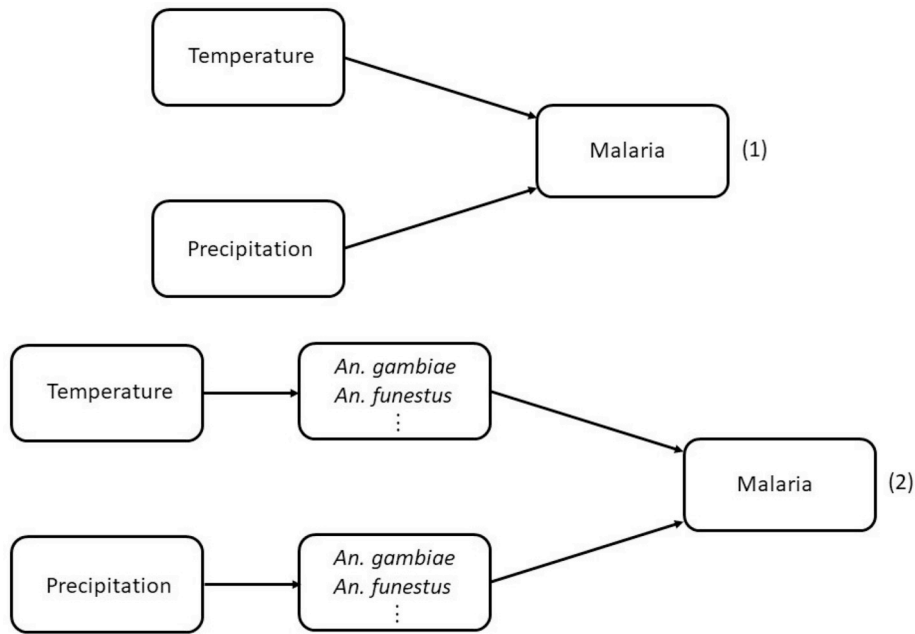


Fig. 1. Two basic DAGs for models of malaria transmission.

et al., 2000; Lyimo et al., 1992; Munga et al., 2006; Paaijmans et al., 2011; Patz and Olson, 2006; Stresman, 2010). This is represented generically in the second DAG, where different vector populations can respond to these environmental conditions in different ways, and, by extension, influence malaria transmission in different ways (Stresman, 2010).

To make this example clearer, consider the following generic example. Suppose we have two areas, each with 100 individuals (*i.e.* $n_1 = n_2 = 100$), and $y_1 = y_2 = 30$ individuals experiencing a given disease or other health outcome. Both overall prevalence and individual area prevalence of disease is 30%. Now suppose that we implement some form of intervention in each area, and that unobserved features (confounders, mediators, etc.) relevant to the overall disease process cause the effect of the intervention to vary in the two areas. For example, suppose in area 1 the intervention reduces occurrence of the health outcome by 50%, but in area 2, it increases the occurrence of the health outcome by 50%. Following the intervention, we will observe $y_1 = 15$ and $y_2 = 45$ individuals with the health outcome. However, overall prevalence is still 30%, and models assuming a constant treatment effect would yield an estimate of 0. This estimate would apply everywhere, but be true nowhere—an everywhere-is-nowhere effect (Duncan et al., 1996, 1998; Jones, 1993). If we allow the treatment effect to vary across these two different places, we can then uncover this variability, and, in the spirit of Arcaya and colleagues, begin to posit why this may be the case.

If we consider the investigation of spatial variability in contextual effects on health outcomes as an inferential goal of health geography, an immediate question arises as to how to obtain those inferences. Generally, there are two widely used methods for doing so: geographically weighted regression (GWR) and spatially-varying coefficient regression (Brunsdon et al., 1996; Gelfand et al., 2003). These two methods, however, are not equivalent. For example, the former applies a kernel smoother to the observed data to obtain local estimates of regression coefficients, whereas the latter obtains local estimates using a spatially-correlated prior distribution for the regression coefficients in a larger multilevel-modeling framework. Additionally, research in recent years has shown that GWR possesses a number of undesirable statistical properties, including collinearity (Wheeler and Tiefelsdorf, 2005), lack of a probabilistic structure for inference (Wheeler and Waller, 2009), and its failure to recover the data generating process in a variety of

simulation studies (Finley, 2011). Further, there is no clear way to model areal data using GWR, since the kernel smoothers used rely on distance-based measures to weight the observed data. To get around this issue, researchers using GWR typically compute the centroids of areal units before proceeding (Shoff et al., 2012).

Conversely, the spatially-varying coefficient regression modeling framework does not suffer from these limitations, and enjoys added benefits as well, such as the ability to incorporate additional prior information about the data-generating process into a model, thereby directly accommodating the need identified by Owen and colleagues for a more theoretical specification of a spatial process (Owen et al., 2016). Additionally, while initially developed and applied to point-referenced data, modeling spatially-varying relationships across areal units is also straightforward, whereby the researcher simply substitutes a CAR prior for the Gaussian process prior used in the point-referenced case, allowing for the data to be modeled more consistently with its actual structure. Further, as noted, spatially-varying coefficient regression is a special case of multilevel modeling more generally, meaning model comparisons are more straightforward, and, by extension, the ability to learn about the scientific questions under investigation. Thus, while GWR has been used extensively in studies modeling spatial variability in relationships between exposures and health outcomes (see, e.g. (Yang and Matthews, 2012)), the weaknesses of GWR as an inferential tool, coupled with the strengths of spatially-varying coefficient regression, suggests that future efforts to understand spatial variability will be more fruitful if done within the much richer multilevel modeling framework introduced earlier, and which we now turn to in greater detail.

As noted, spatially varying coefficient regression, which first emerged in 2003 with a seminal paper by Gelfand and colleagues (Gelfand et al., 2003), is based on extending the well-known spatial model introduced above, and described in greater detail here:

$$g(y_j) = \mathbf{x}_j^T \boldsymbol{\beta} + \theta_j + \varepsilon_j$$

Where y_j is a response (binary, count, continuous, etc.) at location j ($j = 1, \dots, J$), and is connected to the right hand side of the equation via a suitable link function $g(\cdot)$, \mathbf{x}_j^T is a $1 \times p$ vector of covariates at location j , $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients linking the covariates to the response, θ_j is a spatial random intercept, and ε_j white-noise error. We first extend the model by collecting observations on multiple individuals

within a single location, such that:

$$g(y_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \theta_j + \varepsilon_{ij}$$

Where everything is as before, but the model now represents a model for individual i ($i = 1, \dots, n_j$) within a location j . Note that the random effect still applies only to the location, meaning that it will be common to all individuals within that location. When we stack individuals across all locations, we have:

$$g(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Where \mathbf{Y} is an $n \times 1$ vector of responses (modeled through link function $g(\cdot)$), \mathbf{X} is an $n \times p$ matrix of covariates, $\boldsymbol{\beta}$ is as before, \mathbf{Z} is an $n \times J$ block-diagonal random effects design matrix that maps the random effect at each location to all individuals within that location. The vector $\boldsymbol{\theta}$ is thus $J \times 1$ and consists of random effects, with $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of white-noise errors. Under this specification, there is only a single random effect, which typically corresponds to the intercept. To extend this to the case to multiple random effects (i.e. intercepts and slopes), we have:

$$g(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathit{vec}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

Where the only changes are in the random effects design matrix \mathbf{Z} , which is now $n \times kJ$ ($k = 1, \dots, p$), with the number of columns equal to the number of varying covariates times the number of locations, and $\mathit{vec}(\boldsymbol{\theta})$, which takes the k -dimensional vectors of random effects for each location and stacks them into a $kJ \times 1$ vector. These models are typically implemented in a Bayesian setting, and thus we complete the model specification by assigning prior distributions to all unknown parameters. We do so for the case of a linear model with random effects assumed to be realizations from a zero-centered Gaussian process, which is typically used for point-referenced data, although other forms are readily available (e.g. a CAR prior):

$$\mathbf{Y} | \boldsymbol{\beta}, \mathit{vec}(\boldsymbol{\theta}), \tau^2 \sim N_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\theta}, \tau^2 \mathbf{I}_n)$$

$$\boldsymbol{\beta} | \mathbf{m}, \mathbf{S} \sim N_p(\mathbf{m}, \mathbf{S})$$

$$\mathit{vec}(\boldsymbol{\theta}) | \mathbf{H}, \boldsymbol{\varphi} \sim N_{kJ}(0, \boldsymbol{\Sigma}(\boldsymbol{\varphi}) \otimes \mathbf{H})$$

$$\mathbf{H} | \mathbf{S}_0, \nu \sim IW(\mathbf{S}_0, \nu)$$

$$\boldsymbol{\varphi} \sim U(a, b)$$

$$\tau^2 \sim IG(c, d)$$

where $N_r(\mathbf{m}, \mathbf{C})$ represents an r -dimensional normal distribution with mean vector \mathbf{m} and variance-covariance matrix \mathbf{C} , $IW(\mathbf{S}_0, \nu)$ is an inverse-wishart distribution with scale matrix \mathbf{S}_0 and degrees of freedom ν , $U(a, b)$ is a uniform distribution with bounds a and b , and $IG(c, d)$ is an inverse-gamma distribution with shape parameter c and scale parameter d . Importantly, the spatial structure in the above model is induced through the prior on $\mathit{vec}(\boldsymbol{\theta})$. In this setup, the prior mean at each spatial location is 0, such that the intercepts and slopes represent local spatial adjustments to the global intercept and slope contained in $\boldsymbol{\beta}$. In other words, $\boldsymbol{\theta}$ will adjust the overall effect $\boldsymbol{\beta}$ such that it is specific to each place, while further assuming that places are correlated in space.

The spatial association of these processes are modeled through $\boldsymbol{\Sigma}(\boldsymbol{\varphi}) \otimes \mathbf{H}$, where \otimes represents the kronecker product. The first term in this product represents a $J \times J$ matrix of pairwise distances between locations, with the rate of spatial association governed by the parameter $\boldsymbol{\varphi}$. With both spatially-varying intercepts and slopes, we will have multiple spatial variance components, which are on the diagonal entries of the $k \times k$ matrix \mathbf{H} , with the covariance between intercepts and slopes on the off-diagonals. Notably, we can encode any information we may have about these processes through their prior distributions. Moreover, one can modify the prior for the intercepts and slopes to be spatially-

uncorrelated and independently varying. Thus, an added benefit to working with multilevel spatial models is that they are simply an extension of the more traditional multilevel models. With this generic setup, we now turn to an applied example where we model the effects of temperature and precipitation on malaria risk. We also consider the underlying health geography theory motivating the model and link it to epidemiological principles as well.

4. Spatially modeling the effects of temperature and precipitation on malaria risk

Malaria is a vector-borne disease that exhibits considerable spatial heterogeneity, which is a frequently described, but poorly understood phenomenon (Bousema et al., 2010). This heterogeneity can be readily seen in Fig. 2, which shows malaria prevalence in children under 5 years of age across the Democratic Republic of Congo (DRC). It is thus naturally suited for study within a disease ecology framework, with risk factors that span behaviors such as bed net use, population characteristics such as age, and environmental factors such as altitude, temperature, and precipitation (Janko et al., 2018b; Messina et al., 2011). As noted previously, the unobserved mosquito population is sensitive to these environmental conditions. For example, *Anopheles arabiensis* and *Anopheles gambiae* have similar larval habitats, but different biting behaviors, with the former being relatively more zoophilic (Sinka et al., 2010). Given these similar habitats, environmental conditions will clearly affect both species. Precipitation will increase the number of available breeding pools, thereby leading to a possible decrease in competition between vectors, or to a more rapid expansion of one population such as *An. gambiae* over *An. funestus*. Further, while increased temperature is generally believed to increase development rates of the parasite, and thus favor transmission, this is not necessarily the case (Paaijmans et al., 2011). Given these complexities, and given that we do not observe the vector population, its composition, or its behaviors, linking a disease ecology framework with a spatially varying coefficient process framework to model the effects of temperature and precipitation on malaria risk may help us better understand transmission, identify areas where the effects of these exposures substantially increase risk, and develop hypotheses about why this may be so. This may lead to additional studies in those areas, with the ultimate aim to better guide malaria control.

The data for this example come from the 2013/14 Demographic and

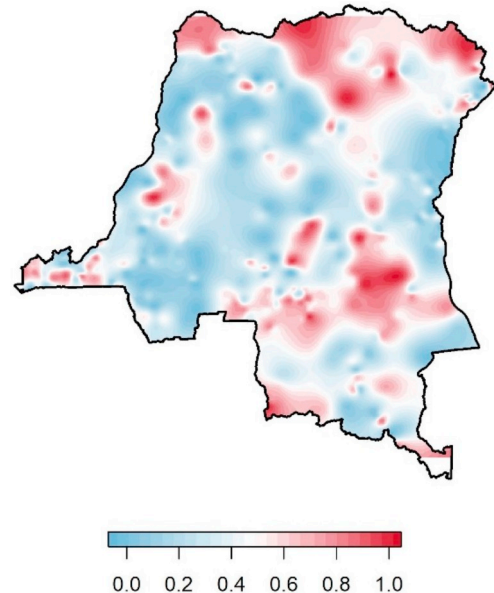


Fig. 2. Malaria prevalence in children under 5 years of age in the DRC.

Health Survey (DHS) conducted in the DRC. Briefly, DHSs are two-stage cluster surveys designed to provide representative estimates for a number of health conditions at national and regional levels, as well as across the urban/rural divide. In rural settings, each cluster represents a village, while in urban settings, a cluster represents either a city block or apartment building. Thus, individuals sampled within each cluster are nested, inducing the multilevel structure of the data, and any contextual exposures will be common to all individuals within a given location. In our example here, we work with 4612 children nested within 331 rural survey clusters.

The outcome of interest is the malaria status of children under 5 years of age, who were tested for malaria by rapid diagnostic test (RDT) as part of the survey. Temperature was measured as the average temperature (in Celsius) during the month of the survey using compiled data distributed by the National Oceanic and Atmospheric Administration (NOAA), while precipitation was measured as total rainfall the month prior to the survey using data from the Tropical Rainfall Monitoring Mission (TRMM). Each exposure was calculated around 10 km of each survey cluster, as this represents the maximum flight distance of a female, human-blood fed *An. gambiae* mosquito, and in turn represents the maximum spatial extent over which humans and mosquitos interact (Kaufmann and Briegel, 2004).

Our modeling strategy here follows closely from the formulation above, with one slight modification. Because our outcome is whether or not a child has malaria, a binary variable, we adopt a probit specification and introduce latent normal variables for the response variable Y . The remaining model specification is as before. We assign diffuse, mean-zero normal priors for the regression coefficients β . We model the spatial random effects as realizations from a zero-centered Gaussian process with separable covariance structure. We assign a low-precision inverse Wishart prior for the spatial variance-covariance matrix, and use an exponential covariance structure with a uniform prior for the spatial range, assuming that the range of spatial association is between 100 m and 225 km, corresponding to 10% of the maximum distance between survey clusters. We fit the model using Markov chain Monte Carlo (MCMC) and run the sampler for 120,000 iterations, discarding the first 20,000 as burn in and thinning the Markov chain to collect every tenth posterior sample. Inference for all parameters is thus based on 10,000

posterior draws.

Upon fitting the model, we predict the spatial intercepts and slopes across the DRC via Bayesian kriging. Doing so allows us to begin to understand how the effect of each exposure on the malaria outcome varies spatially from place to place. The way the model is parameterized (as a zero-centered Gaussian Process) further allows us to understand this variability in terms of local spatial adjustments to the overall average effect represented by β . For example, in Fig. 3, we can see that the effect of temperature on malaria risk is higher than the overall average along the southern border, as well as along a strip running from north to south. On either side of this strip, the effect is lower than the overall average. With regard to precipitation, the effect tends to decrease risk over the majority of the DRC, with a pocket of increased risk in south-central DRC. Finally, the intercept process, which captures the spatial structure of unmeasured confounding, suggests higher residual risk across large swaths of the north, east, and southeast regions of the country, as well as a pocket in the southwest (where the capital city of Kinshasa is located).

With these results in mind, we can begin to hypothesize why we have observed this variability, and suggest both where additional research may be needed, and what types of questions may need to be asked there. For example, with regard to the area of high precipitation effects in south-central DRC, we may posit that this may be due to the underlying land cover. This area of DRC has a large proportion of agriculture, and additional precipitation may be more inclined to develop small, transient, sunlit pools of standing water, the ideal breeding habitat for *An. gambiae* mosquitoes. From an epidemiological perspective, this may suggest possible effect measure modification, whereby the effect of precipitation on malaria risk is modified by the underlying land cover. Recent work has explored the links between agriculture, *An. gambiae* mosquitoes, and malaria risk in the DRC, and included a suite of demographic and behavioral variables, and found that increased exposure to agriculture was associated with increased malaria risk and increased indoor biting behavior among *An. gambiae* mosquitoes, but that the effect of agriculture did not appear to vary spatially, which in turn supports other modeling studies showing that the *An. gambiae* is the predominant vector in the region (Janko et al., 2018b; Sinka et al., 2010). In this way, observing a spatially-varying relationship can serve

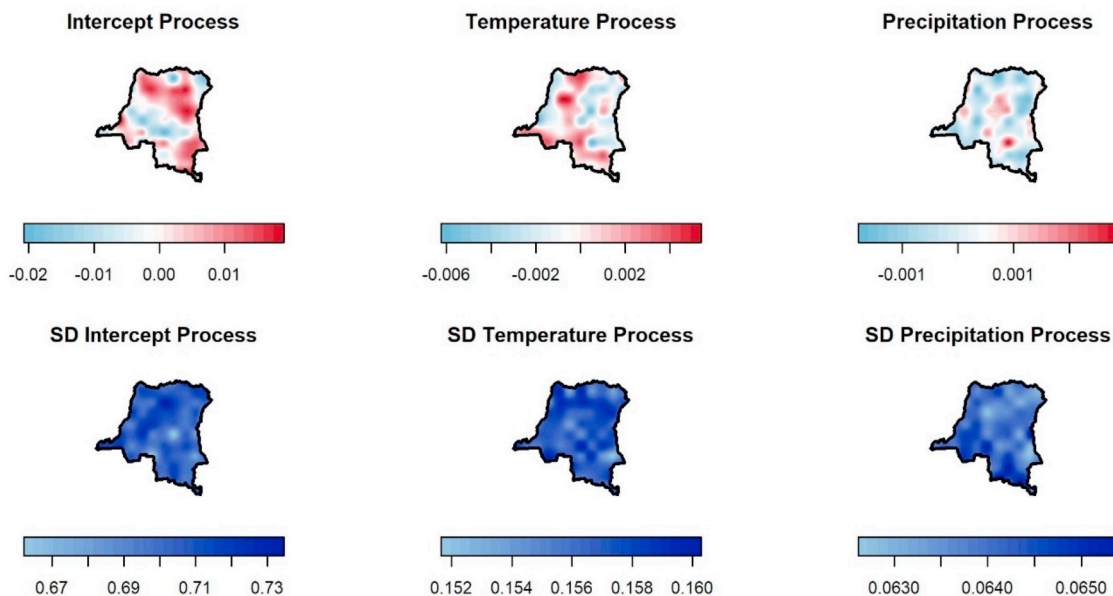


Fig. 3. Intercept and slope processes and their uncertainty across the DRC.

Results from spatially modeling the effects of temperature and precipitation on malaria risk in children under 5 in the DRC. The top row shows the local spatial adjustments to the overall intercept (Intercept Process), the effect of temperature on malaria risk (Temperature Process), and the effect of precipitation on malaria risk (Precipitation process). The bottom row shows the uncertainty (expressed as standard deviation) for all three processes.

to generate hypotheses that can lead to additional work to understand why that relationship may vary, which may be due to other relationships (e.g. agriculture), and how they may (or may not) vary spatially.

Importantly, we must take care to avoid over-interpreting these findings. First, there is considerable imprecision accompanying many of these estimates, which can be seen in the maps of the standard deviation of the intercept and slope processes in the bottom panels of Fig. 3, whereby the uncertainty in the spatial intercept and slope estimates is an order of magnitude or more greater than the estimates themselves. Notably, some of this uncertainty could be addressed by incorporating stronger prior information in both the mean and covariance structure about how much variability we expect to see in the effects, information presumably available from the literature or expert opinion. Second, we have not done any model comparison, a valuable tool in any effort to understand the underlying data-generating process. Indeed, we know that, like all models, our model is wrong, and that determining how useful our model is almost certainly requires comparing it to others. Our purpose here, however, is not to determine the 'best' model among a set of candidates, but to implement a multilevel model with spatially-varying coefficients in line with the broader goal of extending a multilevel modeling framework to address geographic questions. In our case, we began by looking at how two environmental processes may vary spatially, and then turned to hypothesize how another human-environment process (e.g. agriculture) might underlie what we observed. This in turn would serve further model building. For example, were we to move forward, we would certainly include human behavioral, socioeconomic, and community-level factors relevant to malaria transmission in the DRC, and there is a growing body of work to guide this effort (Janko et al., 2018a; Levitz et al., 2018; Messina et al., 2011; Mwandagaliirwa et al., 2017). Additionally, no modeling effort is complete without addressing model specification itself. Here, we chose to model the correlation between spatial intercepts and slopes via a separable covariance structure. This is far from the only option. A simpler model would allow the intercept and each slope to be independent of each other, with its own Gaussian process prior. This specification is likely unrealistic given that intercepts and slopes are almost always correlated, but it would be part of a wider modeling effort. Moving in the other direction, we could model the correlation between intercepts and slopes and their spatial variability through a linear model of coregionalization, which would introduce greater flexibility to learn about any underlying spatial effects. Again, however, our aim is not to reach any conclusions about environmental effects on malaria here, but to note that spatially-varying coefficient process models are a powerful and underutilized tool in the health geography arsenal, particularly regarding their ability to facilitate hypothesis generation about unobserved covariates and their potential effect on health outcomes.

5. Conclusion

Health geographers have long been interested in exploring spatial relationships, as well as in understanding contextual effects on health outcomes. The rapid evolution of the data and modeling environments in recent decades has led to the possibility to investigate both simultaneously. This situation has motivated us to track recent developments in multilevel and spatial modeling, and extend that effort to include spatially varying coefficient processes to identify how contextual effects may vary spatially. Additionally, while the exploration of spatial variability is a core interest of health geographers, we have further motivated the need to explore the spatial variability of contextual effects in epidemiological principles as well. For example, we note that unmeasured covariates can confound, modify, or mediate the effects of an exposure on a health outcome, and do so in ways that induce spatial variability in the exposure-outcome relationship.

Importantly, by extending multilevel models to the study of spatially-varying contextual effects, we do not suggest that this is the only inferential goal. Rather, this is hopefully the next step in a growing

body of health geographic work. There are a number of important avenues that need to be pursued further. For example, the spatial scale of a process is of considerable public health importance, and presents additional challenges when the spatial data consist of areal units. Kim and Subramanian (2016) discuss this in their work on variability in life expectancy across the United States, and demonstrate the importance of multiple geographic scales to health outcomes. Additionally, and more generally, there is an increasing focus on the inferential challenges posed by the variations in contextual effects on individual outcomes based on the geographic delineation of neighborhoods. This raises the ever-important question: what is a neighborhood? This spatial uncertainty, described by Kwan as the Uncertain Geographic Context Problem, may confound research studies and contribute to misleading findings if the measured contextual units deviate from the *true causally relevant* geographic context (Kwan, 2012). Recent work is addressing this problem from the perspective of measurement (Park and Kwan, 2017). That said, neighborhoods are not always explicitly or completely geographic, and individuals are not merely nested within a single neighborhood, as Owen and colleagues duly note (Owen et al., 2016). As a result, future measurement efforts will need to focus on constructing not just a singular neighborhood to which a study participant belongs, but the neighborhoods to which they belong. Additionally, future work will undoubtedly continue without the benefit of rigorously measured neighborhoods, a situation which should motivate further methodological efforts. In settings where measurements on neighborhood structure are unavailable, extending the methods presented here may show promise in addressing this uncertainty.

Declaration of competing interest

None.

Acknowledgements

The authors acknowledge support from the National Science Foundation (grant BCS-1339949 to ME). MJ received support from the Royster Society of Fellows at UNC-CH. MJ and VG were supported by the Population Research Infrastructure Program awarded to the Carolina Population Center (P2C HD050924) by the Eunice Kennedy Shriver National Institute of Child Health and Development.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.healthplace.2019.102235>.

References

- Afrane, Y.A., Little, T.J., Lawson, B.W., Githeko, A.K., Yan, G., 2008. Deforestation and vectorial capacity of *Anopheles gambiae* Giles mosquitoes in malaria transmission, Kenya. *Emerg. Infect. Dis.* 14, 1533–1538.
- Arcaya, M., Brewster, M., Zigler, C.M., Subramanian, S., 2012. Area variations in health: a spatial multilevel modeling approach. *Health Place* 18, 824–831.
- Auchincloss, A.H., Gebreab, S.Y., Mair, C., Roux, A.V.D., 2012. A review of spatial methods in epidemiology, 2000–2010. *Annu. Rev. Public Health* 33, 107.
- Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. *Hierarchical Modeling and Analysis for Spatial Data*. Crc Press.
- Bousema, T., Drakeley, C., Gesase, S., Hashim, R., Magesa, S., Masha, F., Otieno, S., Carneiro, I., Cox, J., Msuya, E., 2010. Identification of hot spots of malaria transmission for targeted malaria control. *JID (J. Infect. Dis.)* 201, 1764–1774.
- Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr. Anal.* 28, 281–298.
- Carrel, M., Janko, M., Mwandagaliirwa, M.K., Morgan, C., Fwamba, F., Muwonga, J., Tshefu, A.K., Meshnick, S., Emch, M., 2016. Changing spatial patterns and increasing rurality of HIV prevalence in the Democratic Republic of the Congo between 2007 and 2013. *Health Place* 39, 79–85.
- Chaix, B., Merlo, J., Chauvin, P., 2005. Comparison of a spatial approach with the multilevel approach for investigating place effects on health: the example of healthcare utilisation in France. *J. Epidemiol. Community Health* 59, 517–526.
- Chaix, B., Merlo, J., Subramanian, S., Lynch, J., Chauvin, P., 2005. Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies:

- the case of mental and behavioral disorders due to psychoactive substance use in Malmö, Sweden, 2001. *Am. J. Epidemiol.* 162, 171–182.
- Duncan, C., Jones, K., Moon, G., 1996. Health-related behaviour in context: a multilevel modelling approach. *Soc. Sci. Med.* 42, 817–830.
- Duncan, C., Jones, K., Moon, G., 1998. Context, composition and heterogeneity: using multilevel models in health research. *Soc. Sci. Med.* 46, 97–117.
- Finley, A.O., 2011. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol. Evol.* 2, 143–154.
- Gelfand, A.E., Kim, H.-J., Sirmans, C., Banerjee, S., 2003. Spatial modeling with spatially varying coefficient processes. *J. Am. Stat. Assoc.* 98, 387–396.
- Gelman, A., Hill, J., 2006. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge university press.
- Hajat, A., Diez-Roux, A.V., Adar, S.D., Auchincloss, A.H., Lovasi, G.S., O'Neill, M.S., Sheppard, L., Kaufman, J.D., 2013. Air pollution and individual and neighborhood socioeconomic status: evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). *Environ. Health Perspect.* 121, 1325.
- Janko, M.M., Churcher, T.S., Emch, M.E., Meshnick, S.R., 2018. Strengthening long-lasting insecticidal nets effectiveness monitoring using retrospective analysis of cross-sectional, population-based surveys across sub-Saharan Africa. *Sci. Rep.* 8, 17110.
- Janko, M.M., Irish, S.R., Reich, B.J., Peterson, M., Doctor, S.M., Mwandagaliwa, M.K., Likwela, J.L., Tshetu, A.K., Meshnick, S.R., Emch, M.E., 2018. The links between agriculture, Anopheles mosquitoes, and malaria risk in children younger than 5 years in the Democratic Republic of the Congo: a population-based, cross-sectional, spatial study. *Lancet Planet. Health* 2, e74–e82.
- Jones, K., 1993. Everywhere Is Nowhere: Multilevel Perspectives on the Importance of Place. *The University of Portsmouth Inaugural Lectures*.
- Kaufmann, C., Briegel, H., 2004. Flight performance of the malaria vectors *Anopheles gambiae* and *Anopheles atroparvus*. *J. Vector Ecol.* 29, 140–153.
- Kim, R., Subramanian, S., 2016. What's wrong with understanding variation using a single-geographic scale? A multilevel geographic assessment of life expectancy in the United States. *Procedia Environ. Sci.* 36, 4–11.
- Kirby, M.J., Lindsay, S.W., 2009. Effect of temperature and inter-specific competition on the development and survival of *Anopheles gambiae sensu stricto* and *An. arabiensis* larvae. *Acta Trop.* 109, 118–123.
- Kwan, M.-P., 2012. The uncertain geographic context problem. *Ann. Assoc. Am. Geogr.* 102, 958–968.
- Levitz, L., Janko, M., Mwandagaliwa, K., Thwai, K.L., Likwela, J.L., Tshetu, A.K., Emch, M., Meshnick, S.R., 2018. Effect of individual and community-level bed net usage on malaria prevalence among under-fives in the Democratic Republic of Congo. *Malar. J.* 17, 39.
- Lindblade, K.A., Walker, E.D., Onapa, A.W., Katungu, J., Wilson, M.L., 2000. Land use change alters malaria transmission parameters by modifying temperature in a highland area of Uganda. *Trop. Med. Int. Health* 5, 263–274.
- Lyimo, E., Takken, W., Koella, J., 1992. Effect of rearing temperature and larval density on larval survival, age at pupation and adult size of *Anopheles gambiae*. *Entomol. Exp. Appl.* 63, 265–271.
- May, J., 1950. Medical geography: its methods and objectives. *Geogr. Rev.* 40, 9–41.
- May, J., 1959. The ecology of human disease. *Ecol. Hum. Dis.*
- Messina, J.P., Taylor, S.M., Meshnick, S.R., Linke, A.M., Tshetu, A.K., Atua, B., Mwandagaliwa, K., Emch, M., 2011. Population, behavioural and environmental drivers of malaria prevalence in the Democratic Republic of Congo. *Malar. J.* 10, 161–161.
- Munga, S., Minakawa, N., Zhou, G., Mushinzimana, E., Barrack, O.-O.J., Githeko, A.K., Yan, G., 2006. Association between land cover and habitat productivity of malaria vectors in Western Kenyan highlands. *Am. J. Trop. Med. Hyg.* 74, 69–75.
- Mwandagaliwa, M.K., Levitz, L., Thwai, K.L., Parr, J.B., Goel, V., Janko, M., Tshetu, A., Emch, M., Meshnick, S.R., Carrel, M., 2017. Individual and household characteristics of persons with *Plasmodium falciparum* malaria in sites with varying endemicities in Kinshasa Province, Democratic Republic of the Congo. *Malar. J.* 16, 456.
- Owen, G., Harris, R., Jones, K., 2016. Under examination: multilevel models, geography and health research. *Prog. Hum. Geogr.* 40, 394–412.
- Paaijmans, K.P., Blanford, S., Chan, B.H., Thomas, M.B., 2011. Warmer temperatures reduce the vectorial capacity of malaria mosquitoes. *Biol. Lett.*, rsl20111075
- Park, Y.M., Kwan, M.-P., 2017. Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health Place* 43, 85–94.
- Patz, J.A., Olson, S.H., 2006. Malaria risk and temperature: influences from global climate change and local land use practices. *Proc. Natl. Acad. Sci.* 103, 5635–5636.
- Shoff, C., Yang, T.-C., Matthews, S.A., 2012. What has geography got to do with it? Using GWR to explore place-specific associations with prenatal care utilization. *Geojournal* 77, 331–341.
- Sinka, M.E., Bangs, M.J., Manguin, S., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W., 2010. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites Vectors* 3, 1.
- Stresman, G.H., 2010. Beyond temperature and precipitation: ecological risk factors that modify malaria transmission. *Acta Trop.* 116, 167–172.
- Wheeler, D., Tiefelsdorf, M., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J. Geogr. Syst.* 7, 161–187.
- Wheeler, D.C., Waller, L.A., 2009. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *J. Geogr. Syst.* 11, 1–22.
- Yang, T.-C., Matthews, S.A., 2012. Understanding the non-stationary associations between distrust of the health care system, health conditions, and self-rated health in the elderly: a geographically weighted regression approach. *Health Place* 18, 576–585.