Received: 28 February 2022 | Revised: 17 May 2022 | Accepted: 28 June 2022

DOI: 10.1002/sim.9527

Statistics in Medicine WILEY

RESEARCH ARTICLE

Estimating cumulative spatial risk over time with low-rank kriging multiple membership models

Joseph Boyle[1] | **Mary H. Ward**[2] | **Stella Koutros**[2] | **Margaret R. Karagas**[3] | **Molly Schwenn**[4] | **Debra Silverman**[2] | **David C. Wheeler**[1]

[1] Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, USA

[2] Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA

[3] Department of Epidemiology, Dartmouth Geisel School of Medicine, Hanover, New Hampshire, USA

[4] Formerly of the Maine Department of Health and Human Services, Maine Cancer Registry, Augusta, Maine, USA

Correspondence
David C. Wheeler, Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.
Email: dcwheeler@vcu.edu

Funding information
National Institutes of Health; National Cancer Institute, Grant/Award Number: U01CA259376

Many health outcomes result from accumulated exposures to one or more environmental factors. Accordingly, spatial risk studies have begun to consider multiple residential locations of participants, acknowledging that participants move and thus are exposed to environmental factors in several places. However, novel methods are needed to estimate cumulative spatial risk for disease while accounting for other risk factors. To this end, we propose a Bayesian model (LRK-MMM) that embeds a multiple membership model (MMM) into a low-rank kriging (LRK) model in order to estimate cumulative spatial risk at the point level while allowing for multiple residential locations per subject. The LRK approach offers a more computationally efficient means to analyze spatial risk in case-control study data at the point level compared with a Bayesian generalized additive model, and as increased precision in spatial risk estimates by analyzing point locations instead of administrative areas. Through a simulation study, we demonstrate the efficacy of the model and its improvement upon an existing multiple membership model that uses area-level spatial random effects to estimate risk. The results show that our proposed method provides greater spatial sensitivity (improvements ranging from 0.12 to 0.54) and power (improvements ranging from 0.02 to 0.94) to detect regions of elevated risk for disease across a range of exposure scenarios. Finally, we apply our model to case-control data from the New England bladder cancer study to estimate cumulative spatial risk while adjusting for many covariates.

KEYWORDS

Bayesian, bladder cancer, residential history, spatial cluster

1 | INTRODUCTION

An emerging perspective in public health acknowledges that many health outcomes result from accumulated experiences over space and time. The concept of exposure is indeed quite broad and could apply to air quality, chemical mixtures, food access, education, occupational setting, and more. This viewpoint on health outcomes has been termed the "exposome",[1] a name chosen to mimic the genome due to its extensive health effects. Three domains of the exposome have been identified:

internal factors, which are unique to the individual; specific external factors, such as lifestyle choices and occupational exposures; and general external factors, such as socio-economic status and educational attainment.[2] Studies designed with the exposome in mind can thus compare the effects of different dimensions of exposure, including a comparison of neighborhood-level external factors and individual-level internal factors, on certain outcomes.[3-5] Measuring relevant components of the exposome at etiologically important timepoints can improve estimates of the association between exposures and outcome status.[6] One popular approach to assessing the associations of potential exposures—chemical, occupational, geographic, or otherwise—with health outcomes is spatial risk studies.

A goal of spatial risk studies is to determine if there are areas of significantly elevated risk for some disease, potentially after accounting for known or posited variables to be associated with the outcome. Identification of such areas can lead to further investigations to elucidate potential sources of this excess risk. However, many disease cluster studies do not uncover significant spatial signals.[7] Often in such studies, a single residential location has been used in the analysis, and two related factors challenge the utility of such a strategy. The first is disease latency. Many diseases, such as cancers, are characterized by a long latency period between exposure and diagnosis. For example, various studies have estimated long latency periods for lung cancer (19 to 25 years),[8] bladder cancer (20 to 40 years),[9] and breast cancer (15 and 20 years).[10,11] A second factor is that populations are mobile over time. In the United States, population mobility was between 10 and 15% annually from 1999 and 2012, and between 15 and 20% annually from 1965 to 1999.[12] Considering this, spatial cluster studies of cancers that rely on one residence will experience a misalignment between the true location of relevant exposures and the analysis location, typically at the time of diagnosis, for those subjects who have moved. In response to this, in recent years, disease cluster studies[13-19] and other epidemiological studies[20-22] have begun to explicitly include in modeling multiple residential locations of their subjects. The feasibility of doing so has increased, as public record databases such as LexisNexis have proven to be able to reconstruct residential histories over a geographically diverse set of study participants.[23]

One method that has used residential histories in case-control studies to detect spatial clustering over space and/or time has been Q-statistics.[13] These statistics use nearest-neighbor matrices and case/control status to quantify the degree of disease clustering among participants, assessing significance using Monte Carlo permutations. The statistics can adjust for risk factors using baseline predicted probabilities of case membership using techniques such as logistic regression[24] and can accommodate disease latency using exposure traces. However, the number of neighbors to consider as proximate in the analysis is unknown and can affect results. Further, the concept of spatial scale changes when going back in time in residential histories, as participants considered to be neighbors may have lived very far apart and potentially in different states, meaning their actual degree of spatial correlation was low.

More comprehensive inference can be performed with spatial regression models, such as generalized additive models (GAMs),[25] which can directly adjust for potential confounders and covariates that may be associated with the outcome and provide estimates of uncertainty in quantities of interest, such as odds ratios. GAMs commonly use thin plate regression splines[26,27] or locally-weighted scatterplot smoothing over spatial coordinates[18,28] to model the spatial variation in risk. Multiple residential locations have been incorporated into a GAM by using a subset of locations at intervals of time prior to study entry (eg, a time lag) and smoothing each set of locations at the time lag.[19] This approach allows for a relative comparison of model fit with and without including a given time lag and its smoother, allowing for inference on which time lags are more important than others for explaining disease risk. However, this approach has not accounted for all subject residential locations together.

Another approach with GAMs has been to include all residential locations for each subject as independent records in a model, replicating the case status and covariate values for each subject but using different residential locations.[17,18,29] While this approach includes each participant's complete residential history, it ignores the correlation between records for the same subject and could lead to spurious detections of areas of elevated spatial risk for disease due to highly mobile cases in a small area. The approach also weights all locations equally, regardless of a subject's duration. The correlation between residences for a given participant could be remedied with generalized additive mixed models, which have been used to estimate a catchment area for a cancer center by accounting for multiple cancer diagnoses for the same subject.[23,27]

The approach of nonparametric M-statistics combines the proportion of time individuals lived at each residential location with the likelihood that they were exposed in that location to test for disease clustering. Manjourides and Pagano incorporate an incubation distribution into the time-weighted distribution of residential locations for each subject to produce an incubation- and time-weighted quantity for each location, and compare the distribution of inter-point distances between weighted case locations to that of control locations to test for disease clustering.[15] Such a method is particularly effective if much is known about the latency period for the disease in question. If the incubation distribution used is relatively uninformative, however, the resulting weights are somewhat similar to the weights based on only residential duration. Further, this approach does not directly permit covariate adjustment, instead only allowing different incubation distributions that depend on covariates.

A method to test for disease clusters[14] extended Kulldorff's spatial scan statistic[30] to account for population mobility. The approach uses a logistic function of weighted disease risks inside and outside all subregions and calculates a likelihood ratio test statistic for all subregions. Significance of the subregion with the greatest likelihood ratio test score is evaluated using Monte Carlo randomization. A disadvantage of this is that it requires specification of all subregions over which to test, and for each test, individuals' disease risk is simplified to the proportion of time they lived inside and outside of the subregion. Such a reduction to a dichotomous residential history, living inside or outside a subregion, loses a great deal of spatial information and is vulnerable to edge effects. Additionally, it does not consider covariates that could explain spatial risk.

Recent work has begun to account for multiple residential locations by combining a multiple membership model (MMM) into a Bayesian hierarchical regression model of disease data collected on the area level, in this case, to model mesothelioma risk for individuals using spatial random effects that operate on the municipality level in Belgium.[16] MMMs are suitable for use when response data come from units in the lowest level of some hierarchy, such as individuals living in counties within a state,[31] or when responses are generated from observations who spend different proportions of time in different units at the same level in the hierarchy.[32] The approach uses an MMM to weight multiple municipalities by the proportion of the time participants spent there in the model component that described the elevation in risk as a function of spatial location. Specifically, the spatial random effect operates on the level of the administrative areas (municipalities) in which the population lived and is given a conditional autoregressive (CAR) prior distribution. Through different choices of functions for the spatial risk, the approach allows for spatially unstructured, structured, or unstructured and structured risk. While this approach (CAR-MMM) accounts for the proportion of time spent in different administrative areas, the precision of the spatial risk estimation is limited by the administrative boundaries.

Each of the preceding methods has accounted in some way for the fact that individuals live in multiple locations, and thus derive their risk of disease from multiple locations. This is a step forward in considering cumulative spatial risk, particularly compared with the tradition in spatial risk analysis of using only one location, the location at time of diagnosis. However, none of the existing methods estimate cumulative spatial risk precisely, at the point level, and simultaneously adjust for covariates associated with disease status. In this article, we propose a flexible modeling strategy that uses point-level data for greater spatial precision and avoids the arbitrariness of political boundaries, allows for simultaneous adjustment for covariates, and weights each residential location by the proportion of time lived there. In contrast to previous methods in the literature, such a model uses all of participants' residential histories at the point level that were in the study area, as well as all of their relevant covariate information, allowing maximum usage of the data. Our approach embeds a MMM into a low-rank kriging (LRK) model, denoted as LRK-MMM, which aims to retain the inferential benefits of GAMs while reducing the computational burden of model fitting[33] by simplifying the representation of the spatial process into a lower dimension. We use LRK in place of a full Bayesian GAM for computational efficiency because obtaining full posterior inference through Markov chain Monte Carlo (MCMC) is computationally infeasible for large case-control studies. Additionally, the nature of case-control studies, in which the response variable is constant over time, represents a distinct scenario from other methods of dimension reduction in the literature for high-dimensional areal data, which act upon multivariate responses that vary over several discrete time points.[34] We perform a series of simulation studies to demonstrate the efficacy of the proposed model and its increased ability to detect areas of elevated spatial risk for disease relative to the CAR-MMM. We subsequently apply the LRK-MMM to model the risk of bladder cancer in the New England Bladder Cancer Study.

2 | METHODS

2.1 | Model specification

Following the supposition that individuals may be exposed to a variety of risks across geographic and temporal dimensions,[3,35] we embedded an MMM into an LRK model to estimate an individual's cumulative spatial risk for disease from their residential histories over an etiologically relevant time period. We specify a Bayesian LRK-MMM of the probability of being a case using a Bernoulli distribution for each subject. Specifically, for subject i, the case membership is distributed as $Y_i \sim \text{Bernoulli}\,(p_i)$, where we modeled the log-odds of the probability p_i as $log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{b=1}^{B}\beta_b x_{ib} +$

$\sum_{j\in A_{(i)}} w_{ij} \sum_{m=1}^{n_K} \psi_m C\left[\left|s_{ij} - k_m\right| / \rho\right]$. We note that, as in a traditional case-control study, a participant is classified as a case or control and then past exposure is assessed. Therefore, the outcome variable is treated as fixed over time in the study. Here, $\left\{\kappa_1, \ldots, \kappa_{n_\kappa}\right\}$ are the n_κ knot locations that geographically represent the distribution of case and control locations and are chosen by some knot selection algorithm (specified below). We do not include the set of covariates with regression coefficients β_b in the simulation study. The set of J residential locations for subject i is denoted by $A_{(i)} = (s_{i1}, \ldots, s_{iJ})$. The proportions of time that subject i lived in locations $A_{(i)}$ are given in w_{ij}, where $\sum_{j=1}^{J} w_{ij} = 1$. The term ψ_m is a spatially structured random effect, and the function $C[\cdot]$ is a member of the Matern family of covariance functions given by $C[f] = (1 + |f|)e^{-|f|}$ that is obtained by fixing parameters of the Matern family of m to 1 and ν to $\frac{3}{2}$. The Matern family of has been a popular choice of covariance function used in geostatistical models.[36,37] Thus, while the model incorporates all of each subject's residential locations to estimate their spatial risk, it encodes the information from the residential locations in terms of their model-specified covariance from the set of n_κ knot locations.

For priors in the LRK-MMM model, the regression intercept has a vague Normal prior $\beta_0 \sim N\left(0, \tau = 10^{-3}\right)$, where τ denotes the precision, or reciprocal of the variance. The random effects ψ_m receive a multivariate normal prior $\psi \sim MVN\left(0, \tau_R \Omega^{-1}\right)$, with precision matrix given by $\Omega = \left[C\left[\|\kappa_m - \kappa_{m'}\| / \rho\right]\right]$, for $1 \leq m, m' \leq n_\kappa$, with $C[\cdot]$ defined as above, and $\tau_R = \frac{1}{\sigma_R^2}$ and $\sigma_R \sim Uniform(1, 10)$. The spatial correlation parameter, ρ, receives a uniform prior, which for our study data described below is on $(0, 30)$ kilometers.

2.2 | Knot selection

One of the considerations when using the LRK-MMM is the location and the number of knots. A common method of knot selection in LRK models is the space-filling algorithm,[38] which minimizes a geometric space-filling criterion over the study region and has been implemented widely.[39-43] However, the design of case-control studies, which consist of realizations of a marked point process instead of continuous responses observed at fixed sampling sites, benefits from other methods of knot selection. Recent research has shown that space-filling is greatly outperformed by other methods of knot selection in case-control studies in terms of greater spatial sensitivity and power to detect regions of elevated risk in LRK models.[44] One such method is the Teitz and Bart heuristic.[45] It was developed to address the location-allocation problem in operations research, which seeks to minimize the distance between facilities and the clients they serve. The heuristic begins with an initial configuration of knot locations and utilizes an objective function that sums the total distance from demand points (here, case locations) to facilities (knot locations). It moves knot locations to candidate locations in an iterative fashion if doing so decreases the value of the objective function and continues until no further improvement in the objective function is possible. The heuristic has been used in operations research problems and has demonstrated improvements in spatial sensitivity and power to detect regions of elevated risk in case-control studies.[44,46] We use the Teitz and Bart heuristic to select knot locations in the simulation study and data analysis described below.

2.3 | Simulation study design

Data-generating process. To evaluate the ability of the proposed LRK-MMM to detect regions of elevated risk and accurately estimate cumulative spatial risk over time, we performed a simulation study that combined real residential histories with simulated disease outcomes. We obtained complete residential histories from participants in the New England bladder cancer study (NEBCS), a population-based case-control study that sought to identify environmental, occupational, and lifestyle risk factors responsible for the excess bladder cancer incidence in Maine, New Hampshire, and Vermont.[47,48]

We considered several simulation scenarios to investigate the performance of the proposed model under a variety of conditions. All scenarios shared a common design. First, we retained a random sample of 500 long-term residents from NEBCS and recorded their residential histories over the 20 years prior to study enrollment (1981-2000), a time period that may reflect the latency period for bladder or other cancers and some exposures.[9,49] In doing so, we assumed that all subjects entered the study in the same year (2000). Then, we activated a circular zone of elevated risk for disease of radius 75 km in a certain part of the study region for a fixed number of years. Participants in each simulated dataset living in the zone when it was active experienced a greater risk of disease, defined by an elevated odds ratio, relative to those who did not live in the zone when it was active. Using the odds ratios and locations of residences with respect to the zone when it was active, we then randomly generated case-control status from a Bernoulli distribution with baseline probability of being a case $P = 0.1$ for those who did not live in the zone of elevated risk when it was active, in order to reflect an adverse outcome that was neither common nor exceedingly rare. For each scenario, we simulated $D = 50$ datasets using the data-generating process, and fitted and evaluated models using the generated datasets.

The simulation scenarios differed with respect to the placement of the elevated risk area, population density, duration of exposure, and ratio of cases to controls. In Scenario 1, the zone was located in northern New Hampshire, an area of low to moderate population density, was active for the first 3 years (1981-1983) of residential histories and had a low case-control ratio of approximately 1:10 in order to create a scenario that was similar to case-control studies with the resources to include a large number of controls. In Scenario 2, the zone was located in southern New Hampshire, an area of higher population density, was active the first 9 years (1981-1989) of residential histories and had the same low case-control ratio. Figure 1 illustrates one simulated sample from this scenario. In Scenario 3, the zone was located in southern New Hampshire, was active the first year of residential histories, and had the same low case-control ratio. We used the same residential histories in each simulated dataset for Scenarios 2 and 3 to explore how detectability of the zone was affected by its duration of activity. Scenarios 4, 5, and 6 maintained the zone location and duration of Scenarios 1, 2, and 3, respectively, but increased the case-control ratio in each sample to approximately 1:3 by only retaining every third to fourth control. Additionally, in each scenario, odds ratios of disease of 1.5, 3.0, and 4.5 (subscenarios A, B, and C, respectively) were set for participants living in the active zone of elevated risk, creating 18 scenarios overall. A summary of the different scenarios is presented in Table 1.

2.4 | Model fitting

We fitted an LRK-MMM to each simulated dataset, choosing knot locations with the Teitz and Bart method. We used $n_\kappa = 60$ knots in all scenarios, as this was the approximate number of cases in the scenarios with the lowest proportion of cases.

We fitted models in a Bayesian framework using MCMC methods. For model estimation, we used just another Gibbs sampler (JAGS)[50] in R, version 3.6.1, using two chains that each had a burn-in period for 70 000 iterations and retained 10 000 iterations for sampling from the joint posterior distribution. We assessed convergence of model parameters using the Gelman-Rubin statistic, where a parameter was considered to have converged if its statistic was less than 1.2,[51] using

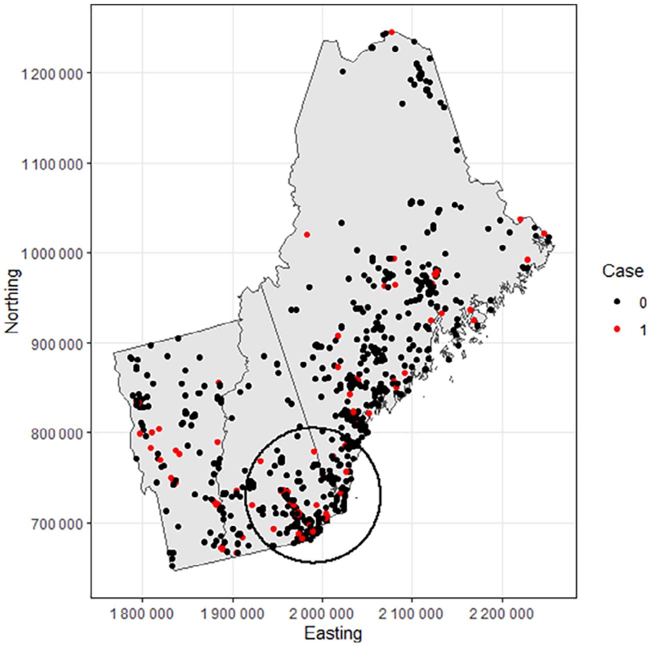

FIGURE 1 Simulated sample for Scenario 2, in which study participants living in a zone of elevated risk in southern New Hampshire when it was active experienced a higher odds ratio of being a case (odds ratio = 3.0) than those who did not live in the zone when it was active. Points are randomly jittered, and the zone of elevated risk is given by the black circle

TABLE 1 Summary of simulation scenarios, where OR = odds ratio and the case proportion is averaged over all 50 simulated datasets

Scenario	Zone location	Zone duration	OR in zone	Case proportion
1A	Northern	First 3 years (1981-1983)	1.5	0.111
1B			3.0	0.121
1C			4.5	0.128
2A	Southern	First 9 years (1981-1989)	1.5	0.116
2B			3.0	0.146
2C			4.5	0.168
3A	Southern	First 1 year (1981)	1.5	0.116
3B			3.0	0.145
3C			4.5	0.167
4A	Northern	First 3 years (1981 to 1983)	1.5	0.291
4B			3.0	0.315
4C			4.5	0.331
5A	Southern	First 9 years (1981-1989)	1.5	0.314
5B			3.0	0.374
5C			4.5	0.415
6A	Southern	First 1 year (1981)	1.5	0.314
6B			3.0	0.373
6C			4.5	0.413

the coda package in R.[52] Using the posterior samples of ψ, and the covariance function, we predicted the spatial odds of disease to a grid covering the extent of the study region, generating a posterior distribution of the spatial odds of disease at each grid cell. Each grid cell represented approximately a 6 km by 6 km square over the study region. We identified grid cells as being significantly elevated in risk using exceedance probabilities,[53] which are an estimate of how frequently the spatial odds at the ith location exceed the null value ($\theta_i = 1$). The estimate of this probability uses the posterior distribution of spatial odds at the ith location $\left(\theta_{i,m+1}, \ldots, \theta_{i,m+G}\right)$, where m represents the burn-in and G represents the number of posterior samples after the burn-in, and is calculated as $\widehat{q_{i,U}} = \frac{1}{G}\sum_{g=m+1}^{m+G} I\left(\theta_{i,g} > 1\right)$. Grid cells with an exceedance probability $\widehat{q_{i,U}} \geq 0.95$ were considered to be of significantly elevated risk.

2.5 | Model evaluation

We evaluated model performance in several ways. The first metric was spatial sensitivity. Denoting the set of grid cells that are in the zone of elevated risk as S, the spatial sensitivity of a model for dataset d is given by $sen_d = \frac{1}{|S|} \sum_{s_i \in S} I\left(\widehat{q_{s_i}} > 0.95\right)$, where $\widehat{q_{s_i}}$ denotes the exceedance probability for grid cell s_i, and $I(\cdot)$ is an indicator function. The second metric is spatial specificity. Defining the set of grid cells that are not in the zone of elevated risk as NS, the specificity of a model for dataset d is given by $spec_d = \frac{1}{|NS|} \sum_{ns_i \in NS} 1 - I\left(\widehat{q_{s_i}} > 0.95\right)$. The spatial sensitivity and specificity were averaged over the D datasets.

Finally, spatial power is calculated according to a sensitivity threshold of zero. The LRK-MMM was considered to have identified the zone of elevated risk for dataset d if any of the grid cells defined to be of significantly elevated risk were identified as such. The spatial power was then calculated as $P = \frac{1}{D}\sum_{d=1}^{D} I\left(sen_d > 0\right)$.

2.6 | Comparison to CAR-MMM approach

For each scenario, we also fitted a CAR-MMM model to each dataset. This approach only considered the counties in which participants lived when estimating spatial risk and not the individual point locations. We chose to compare the performance of our model to the CAR-MMM because unlike such models as the standard CAR model, the CAR-MMM also assesses cumulative spatial risk using multiple residential locations. The outcome variable $Y_i \sim Bernoulli\,(p_i)$ was the same as above, and we modeled the log-odds of the probability of case membership as $log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j\in A_{(i)}} w_{ij} v_j$, where the spatial random effect v_j operated on the county level, and was given a proper CAR prior with zero mean and covariance matrix $\left[\sigma^2 (I - \phi C)^{-1} M\right]$. Here, σ^2 is a variance parameter and σ is assigned a $Uniform(0{,}100)$ prior distribution, I is the identity matrix, and C is a normalized neighborhood matrix of counties that uses binary neighborhood weighting and queen contiguity. Specifically, in C, the row corresponding to the ith county consists of entries $\frac{1}{n_i}$ for each neighbor of the county, where n_i is the number of neighbors of the ith county, and zeroes otherwise. The diagonal elements of C are zero, as a county is not considered to neighbor itself. M is a diagonal matrix of conditional variances, and φ is a spatial dependence parameter given a $Uniform\left(\lambda_{min}^{-1}, \lambda_{max}^{-1}\right)$ prior, where the $(\lambda_{min}, \lambda_{max})$ denote the smallest and largest eigenvalues of $M^{-1/2} C M^{1/2}$. The regression intercept was given the same prior distribution as above. We estimated model parameters in JAGS, using two chains that each had a burn-in period for 12 000 iterations and retained 8000 iterations for sampling from the joint posterior distribution. We used exceedance probabilities to assess the spatial risk for each county, considering county c to have elevated risk if $\widehat{q_c} \geq 0.95$. Specifically, we intersected the zone of elevated risk with the county boundaries and calculated sensitivity as the proportion of area in the intersection that belonged to counties estimated to have significant excess risk. We calculated specificity as the proportion of area in the study region and outside the intersection that belonged to counties estimated not to have significant excess risk. We calculated spatial power according to a sensitivity threshold of zero, considering the CAR-MMM to have identified the zone for a given dataset if any county in the intersection was estimated to have significant excess risk.

2.7 | Application to New England bladder cancer study

We applied the LRK-MMM to data from the NEBCS, a population-based case-control study in Maine, New Hampshire, and Vermont. In this analysis, we used the residential histories of long-term residents, who had been living in the study area for the 25-year period prior to study enrollment (500 cases and 602 controls), in order to focus on historic exposures within the study region. Rates of bladder cancer incidence and mortality have been elevated in the region, and the NEBCS sought to identify risk factors responsible for the elevated incidence of bladder cancer.[47,48] Cases were all newly diagnosed cases of bladder cancer among residents of the study region between 2001 and 2004, and controls were randomly selected from driver's license registration (under age 65 years)/CMS (age 65 to 79 years) records and frequency matched by state, sex, and approximate age at diagnosis. Complete residential histories were collected from study subjects based on in-person interviews with each subject using a standardized questionnaire. The current study estimated cumulative spatial risk for bladder cancer using the LRK-MMM. We included data from all residents who continuously resided in the study area (anywhere in Maine, New Hampshire, or Vermont) for the 25-year period before study enrollment and considered residential locations for subjects from 1970 to 1986, to reflect a potential latency period that is similar to an estimated latency period for an occupational exposure.[9,47] We adjusted our models for smoking status (former or current/occasional vs reference never), high-risk occupation, gender, age group (55-64 or 65-74 or 75+ vs reference <55), race, French-Canadian ancestry, ethnicity (Hispanic ethnicity vs reference no), educational attainment (high school degree or vocational or some college or college degree or postgraduate vs reference less than high school degree), cumulative estimated arsenic intake lagged 40 years,[54] cumulative total trihalomethanes (THM) intake from age 15+, average daily nitrate intake from public water supplies and private wells in 1970 and later, and drinking from an unconsolidated well in the study area before 1960. We fitted models in JAGS, using a burn-in period for 60 000 iterations and retaining 10 000 iterations for sampling from the joint posterior distribution. We assessed spatial risk over a 6 km × 6 km grid covering the study region, predicting spatial risk at each grid cell with the posterior estimates of the spatial random effects at the knot locations and the covariance function between the grid cell and the knot locations. In addition to identifying areas of elevated risk, we identified grid cells as being significantly lowered in risk, defining $\widehat{q_{i,L}} = \frac{1}{G}\sum_{g=m+1}^{m+G} I\left(\theta_{i,g} < 1\right)$ and considering grid cells with an exceedance probability above a threshold to be significantly lowered in risk. We determined significance of risk using 90% and 95% exceedance probabilities.

3 | RESULTS

The performance of the LRK-MMM and CAR-MMM models in the simulation study with respect to spatial sensitivity, specificity, and power are shown in Table 2.

TABLE 2 Simulation results comparing the performance of LRK-MMM with CAR-MMM with respect to spatial sensitivity, spatial specificity, and power. The scenario name gives the location of the zone and the number of years that it was active (eg, northern-3 denotes the zone in northern New Hampshire, active for 3 years). OR = odds ratio

	Case-control		LRK-MMM			CAR-MMM		
Scenario	**ratio**	**OR**	**Sensitivity**	**Specificity**	**Power**	**Sensitivity**	**Specificity**	**Power**
Northern-3	Lower	1.5	0.211	0.745	1.000	0.013	0.994	0.060
		3.0	0.328	0.736	1.000	0.045	0.993	0.240
		4.5	0.403	0.723	1.000	0.105	0.992	0.520
	Higher	1.5	0.168	0.878	0.900	0.005	0.993	0.040
		3.0	0.209	0.889	0.940	0.046	0.996	0.300
		4.5	0.336	0.864	1.000	0.212	0.988	0.740
Southern-9	Lower	1.5	0.569	0.756	1.000	0.045	0.997	0.220
		3.0	0.731	0.741	1.000	0.351	0.997	0.860
		4.5	0.809	0.737	1.000	0.640	0.991	0.980
	Higher	1.5	0.441	0.844	1.000	0.032	0.999	0.140
		3.0	0.587	0.820	0.960	0.180	0.998	0.540
		4.5	0.721	0.814	1.000	0.410	0.995	0.900
Southern-1	Lower	1.5	0.577	0.752	1.000	0.037	0.997	0.200
		3.0	0.728	0.738	1.000	0.335	0.998	0.880
		4.5	0.811	0.733	1.000	0.602	0.992	0.960
	Higher	1.5	0.432	0.841	1.000	0.022	0.999	0.080
		3.0	0.586	0.808	0.960	0.178	0.998	0.560
		4.5	0.691	0.817	1.000	0.405	0.998	0.940

3.1 | Sensitivity

The LRK-MMM exhibited greater spatial sensitivity than the CAR-MMM in every scenario, with increases in sensitivity ranging from 0.12 to 0.54 higher depending on the scenario. Generally, the LRK-MMM had the greatest improvement relative to the CAR-MMM when the spatial signal was the smallest (odds ratio = 1.5), suggesting that the greater precision gained by using residential locations in place of administrative areas allowed better identification of regions of elevated risk. Within each scenario, the sensitivity increased with the odds ratio for both the LRK-MMM and CAR-MMM models, reflecting the models' increased ability to detect regions of elevated risk with greater spatial signal. Additionally, sensitivity was greater for the scenarios with the zone centered in southern New Hampshire than those with the zone centered in northern New Hampshire, a consequence of the greater population density in the former region (approximately 100 to 500 residents per square mile) that supported more accurate estimation of excess risk than in the latter region (approximately 0 to 100 residents per square mile).[55] Finally, for both the LRK-MMM and CAR-MMM models, activating the zone centered in southern New Hampshire for 9 years generally led to increased sensitivity compared to activating it for only 1 year. Finally, for a given location and intensity of the zone of elevated risk, there were generally higher sensitivities for the low case-control ratio scenarios to the high case-control ratio ones, supporting the contention that a larger number of controls and thus larger sample size more accurately estimates baseline risk in the population.

3.2 | Specificity

Across all scenarios, the CAR-MMM showed greater spatial specificity than the LRK-MMM. The specificity values for the CAR-MMM were greater than 0.98 in all scenarios. This is partially explained by the model finding that fewer counties have significantly elevated risk in the first place, regardless of whether or not the county intersected the zone.

The specificities for the LRK-MMM ranged from 0.72 to 0.89 depending on the scenario, which is adequate to good. For the LRK-MMM, the specificity did not change much as the odds ratio increased in a given scenario. There were generally higher specificities for the high case-control ratio scenarios than for the low case-control ratio ones, as the smaller sample sizes found fewer areas of elevated risk and therefore fewer incorrectly-identified areas of elevated risk.

3.3 | Power

The LRK-MMM demonstrated significantly greater power than the CAR-MMM to detect any part of the region of elevated risk. Across all scenarios, the power of LRK-MMM ranged from 0.90 to 1.00, while the power of CAR-MMM ranged from 0.04 to 0.98. The LRK-MMM had increases in power ranging from 0.02 to 0.94 over the CAR-MMM. The only scenario that gave slightly lower power values for the LRK-MMM was when the zone was centered in northern New Hampshire, a region of lower population density, and was only active for 3 years, a relatively short amount of time. In almost every other scenario, the model identified the zone of elevated risk for every generated dataset every time. The CAR-MMM only exhibited spatial power values greater than 0.90 when the odds ratio in the zone was 4.5 and never had power above 0.25 when the odds ratio in the zone was 1.5.

3.4 | Application to NEBCS

Figure 2 illustrates the spatial risk pattern for bladder cancer over the study region, using the posterior median spatial odds ratios from the model. Most of the study region was characterized by odds ratios of approximately 1.0, indicating a lack of excess spatial risk for bladder cancer. However, there were many small areas of elevated risk with spatial odds ratios greater than 1.25 and ranging up to 3. These areas are found in western Vermont, northern and southern New Hampshire, and throughout Maine. A map of the areas that were significantly elevated or lowered in risk for bladder cancer based on 90% exceedance probabilities reveals one large spatial cluster of elevated risk in northern New Hampshire (Figure 3).

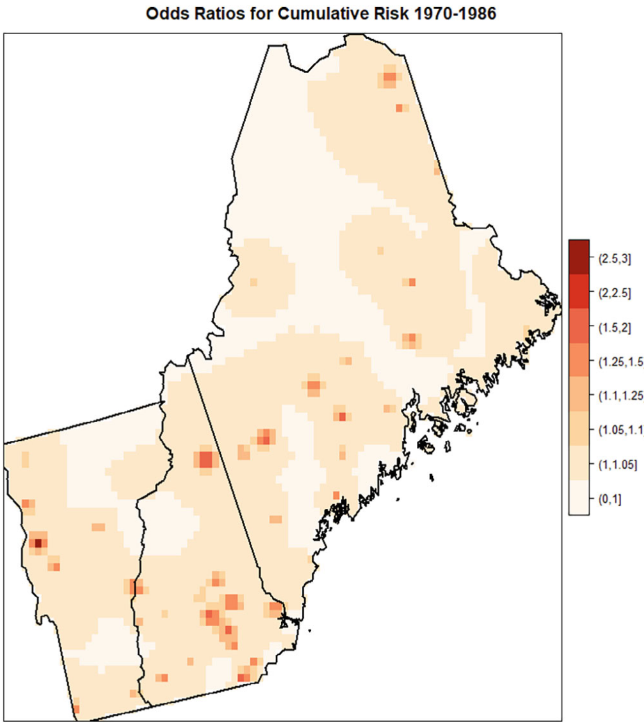

FIGURE 2 Surface plot showing estimated spatial odds ratios for bladder cancer in New England, over a 6 km by 6 km grid, for the time period 1980 to 1996

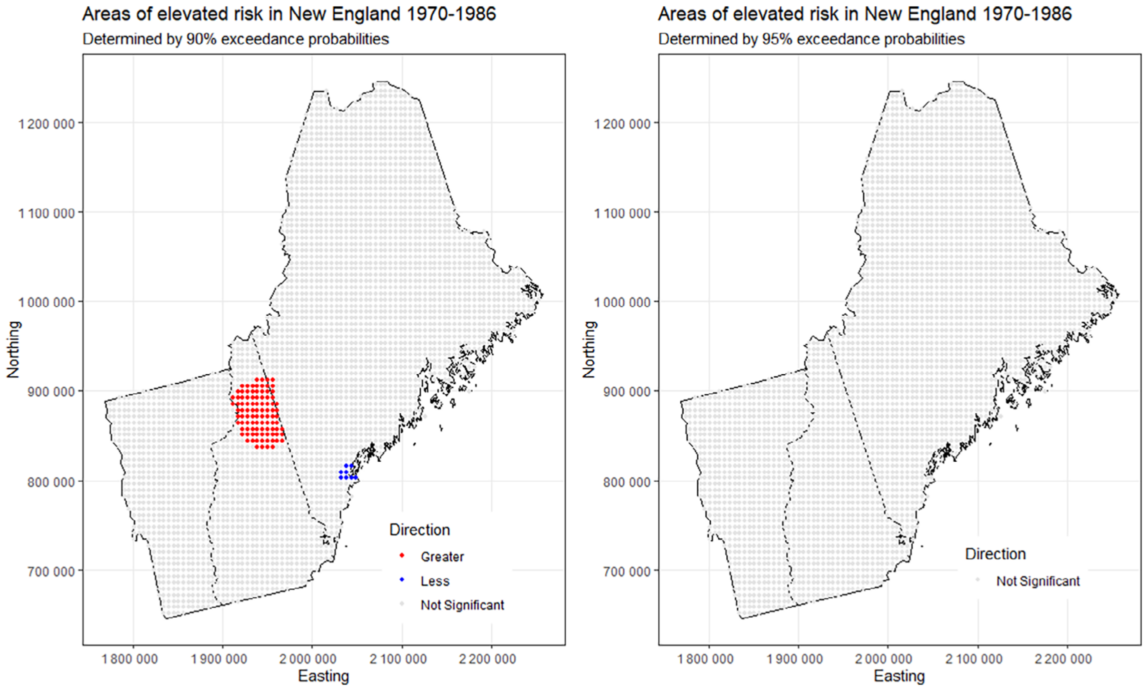
Areas of elevated risk in New England 1970-1986
Determined by 90% exceedance probabilities
Areas of elevated risk in New England 1970-1986
Determined by 95% exceedance probabilities
Northing
Easting
1 200 000
1 100 000
1 000 000
900 000
800 000
700 000
1 800 000
1 900 000
2 000 000
2 100 000
2 200 000
Direction
● Greater
● Less
● Not Significant

FIGURE 3 Regions of significant cumulative spatial risk for bladder cancer in New England in 1980 to 1996 based on the LRK-MMM and using 90% and 95% exceedance probabilities

The cluster was characterized by a central grid cell with an odds ratio between 2.0 and 2.5, diminishing to odds ratios between 1.1 and 1.5 spread out from the center and further to approximately 1.05 for the outermost areas of the cluster. There was a small region of significantly diminished risk for bladder cancer in southern coastal Maine, near Portland. Several small areas with moderately elevated odds ratios (Figure 2) ultimately did not have significantly elevated risk due to a low density of study participants nearby to precisely estimate the spatial risk, for example, in northeastern Maine and western Vermont. Using 95% exceedance probabilities, no areas of significantly elevated or lowered risk for bladder remain (Figure 3).

We calculated the annual empirical (observed) odds ratio based on the number of cases and controls living in the cluster to better characterize the significant area of elevated risk in New Hampshire when using 90% exceedance probabilities over the analysis range of 1970 to 1986 (Figure 4). The annual empirical odds ratio based on the observed number of cases and controls in the northern New Hampshire cluster was 1.33 for 13 of the 17 years and 1.60 for 2 years and was greater than or equal to one for all years in the analysis range. These results show a consistently higher number of cases to controls over the entire analysis range in the cluster (Figure 4), providing evidence that the excess risk persists over time for the area. However, this elevated risk area is based on a small number of cases and, combined with modest odds ratios, explains why the area is not significant when using a 95% exceedance probability threshold.

4 | DISCUSSION

This study proposed a novel method, the LRK-MMM, to estimate cumulative spatial risk for disease over time. Our method was developed in line with the philosophy of the exposome, which holds that individuals are exposed to a series of risks over their life course and that many outcomes, particularly those that relate to health, result from these accumulated exposures. This informed the two main assumptions of the model: that individuals could derive risk at any residential location, and that their cumulative risk derived from any location was proportional to the length of time they resided there. Thus, for each individual in the study, the LRK-MMM allowed their overall spatial random effect component in the model to consist of spatial random effects at each place they lived, which were weighted by the amount of time they lived in each place. The model used point locations in contrast to larger administrative units to enable greater spatial precision in estimating risk.

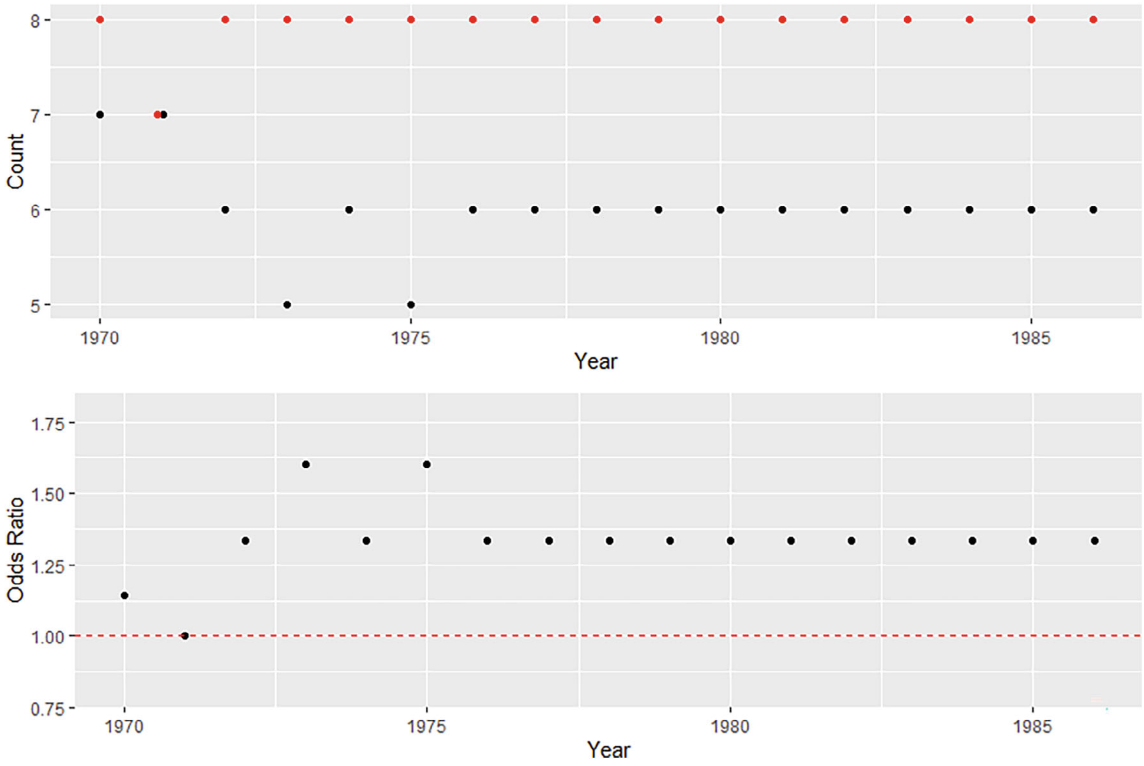

FIGURE 4 Number of cases and controls (top panel) and estimated spatial odds ratio (bottom panel) in the disease cluster in northern New Hampshire, from the LRK-MMM using residences from 1970 to 1986 and 90% exceedance probabilities. In the top panel, cases and controls are given by red and black dots, respectively. In the bottom panel, the red line indicates a null odds ratio of one

We demonstrated through a simulation study that the LRK-MMM enabled greater spatial sensitivity and power to detect regions of elevated risk for disease than a model that includes area-level spatial random effects to estimate risk (CAR-MMM). This improvement held regardless of the population density in the zone, the number of years the zone was active, and the strength of the spatial signal in the zone. In almost every scenario, the LRK-MMM detected some portion of the zone of elevated risk for every generated dataset in the scenario. Specifically, averaged over all scenarios, the power of the LRK-MMM was 0.99, and the power of the CAR-MMM was 0.51, much lower. The LRK-MMM performed best in terms of sensitivity and specificity when the zone of elevated risk was activated in areas of greater population density, when there were more controls per case in the sample, and when the zone was activated for a longer time period. Additionally, the LRK-MMM performed best relative to the CAR-MMM when the degree of elevated risk was smallest, suggesting that the increased precision offered by analyzing point locations led to an improved ability to detect areas of slightly elevated risk for disease. In the application of the LRK-MMM to the NEBCS, we found evidence of a small area of significantly elevated risk for bladder cancer in northern New Hampshire after adjusting for covariates in the model. One limitation in analyzing the NEBCS data is the potential for differential nonresponse bias between cases and controls that could artificially elevate risk in some areas. However, previous analysis of these data has concluded that because study participation was unlikely to have differed with respect to exposure between cases and controls, the potential for bias in the risk estimates was small.[56]

The CAR-MMM represents the most suitable model to compare to the LRK-MMM because both models seek to estimate cumulative spatial risk while controlling for covariates. The primary difference in the models is the scale of the spatial random effects. The CAR-MMM includes spatial random effects that operate on the municipality level and simplifies study participants' residential histories from the point level to the municipality level. Spatial dependence between proximate municipalities is acknowledged with a CAR prior distribution. Contrastingly, the LRK-MMM uses spatial random effects that operate on the point level and uses point locations for residential histories. We hypothesized that the increased spatial precision gained by using point locations would lead to improved performance over a model that used area-level residential data and random effects, and our simulation study supported this hypothesis. In addition, the application to the NEBCS demonstrated the ability to identify a small area of elevated risk based on a relatively small number of subjects. The primary criterion that represents the increase in performance of the LRK-MMM is spatial power. This model evaluation metric represents an important and often consequential outcome in a spatial analysis such as a case-control study. A model detecting some area in the study region with elevated risk for the disease outcome could spur further investigation into the unmeasured or unexplained factors that could have driven the excess risk. However, a model that does not correctly identify regions of elevated risk cannot set off the subsequent epidemiological investigations or environmental remediation. Increased spatial power is a notable advantage of using the LRK-MMM and has real-world consequences.

The additional strengths of the LRK-MMM are its computational efficiency and use of residential information. Regarding efficiency, the use of low-rank kriging makes the choice to use point-level locations feasible. Through simplifying the spatial process from the number of unique residential locations to a smaller number of knots, the spatial covariance matrix becomes more easily invertible, and these computational savings are repeated over many MCMC draws. Additionally, the LRK-MMM entails a more complete usage of residential histories than other approaches in the literature, which have used all participants residences without considering the duration lived in each, or which used residential locations only at certain and specific time lags.[17-19]

While our approach has several strengths, its limitations motivate future work. First, the approach estimates cumulative measure of disease risk over time. However, this model may not be optimal for use if the research goal is to identify particular years of elevated risk. For example, if a factory released toxic chemicals into the air for a two-year period in the middle of a twenty-year residential history, identifying the specific years associated with elevated risk could be helpful for clinical and public health purposes. Models developed in future work could address this other goal by learning from the data regarding the most important years in an entire residential history regarding disease risk and weight the years appropriately. A second limitation concerns the nature of covariate adjustment in the model. We controlled for potential exposures only using a single, fixed value, though more could be learned by modeling them as time-varying in nature. For example, annual exposure to certain chemicals in the past may allow more accurate inference than averaging individuals' exposure over an entire time period. The nature of such data, which is often similar in proximate years, will require an extension of the method that takes the between-individual and between-year correlation into account. Implementation of developments such as these may increase the accuracy of these models.

In summary, the LRK-MMM is a new model that estimates cumulative spatial risk for disease using precise point location residential histories and can simultaneously control for any number of relevant covariates. Fit in the Bayesian paradigm, our model provides full posterior inference on all model quantities of interest, allowing more informed decision-making processes. The LRK-MMM enables greater spatial sensitivity and power to detect regions of elevated risk than a similar MMM that uses residential histories and spatial random effects that operate on the area level. We envision that the LRK-MMM can be used in a wide range of spatial analyzes, particularly case-control studies, to address a litany of public health problems. The model is sufficiently general in that it requires only a disease outcome, the residential locations of study participants, and a set of potential confounders if applicable. Conclusions from the LRK-MMM center on the existence of regions of significant elevated risk for the outcome. Further, the outcome itself can be general and need not strictly be a binary health outcome such as incident cancer case vs noncancer control status. The LRK-MMM is suitable for public health practitioners to assess the associations between geographic locations and elevated likelihood of any health outcome while using many locations in the residential histories of study participants.

FUNDING INFORMATION

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award No. U01CA259376. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the author Dr. Debra Silverman. The data are not publicly available due to privacy or ethical restrictions.

ORCID

David C. Wheeler https://orcid.org/0000-0001-8121-5182

REFERENCES

1. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Prev Biomark*. 2005;14(8):1847-1850.
2. DeBord DG, Carreón T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. Use of the "exposome" in the practice of epidemiology: a primer on-omic technologies. *Am J Epidemiol*. 2016;184(4):302-314.
3. de Vuijst E, van Ham M, Kleinhans R. A life course approach to understanding neighbourhood effects. IZA Discussion Paper #10276; 2016.
4. Lian M, Madden PA, Lynskey MT, et al. Geographic variation in maternal smoking during pregnancy in the Missouri adolescent female twin study (MOAFTS). *PLoS One*. 2016;11(4):e0153930. doi:10.1371/journal.pone.0153930
5. Räisänen S, Kramer MR, Gissler M, Saari J, Hakulinen-Viitanen T, Heinonen S. Smoking during pregnancy was up to 70% more common in the most deprived municipalities — A multilevel analysis of all singleton births during 2005–2010 in Finland. *Prev Med (Baltim)*. 2014;67:6-11. doi:10.1016/j.ypmed.2014.06.026
6. Vrijheid M. The exposome: a new paradigm to study the impact of environment on health. *Thorax*. 2014;69(9):876-878.
7. Schulte PA, Ehrenberg RL, Singal M. Investigation of occupational cancer clusters: theory and practice. *Am J Public Health*. 1987;77(1):52-56.
8. Archer VE, Coons T, Saccomanno G, Hong D-Y. Latency and the lung cancer epidemic among United States uranium miners. *Health Phys*. 2004;87(5):480-489. https://journals.lww.com/health-physics/Fulltext/2004/11000/LATENCY_AND_THE_LUNG_CANCER_EPIDEMIC_AMONG_UNITED.4.aspx
9. Miyakawa M, Tachibana M, Miyakawa A, et al. Re-evaluation of the latent period of bladder cancer in dyestuff-plant workers in Japan. *Int J Urol*. 2001;8(8):423-430.
10. Aschengrau A, Paulu C, Ozonoff D. Tetrachloroethylene-contaminated drinking water and the risk of breast cancer. *Environ Health Perspect*. 1998;106(suppl 4):947-953.
11. Nadler DL, Zurbenko IG. Estimating cancer latency times using a Weibull model. *Adv Epidemiol*. 2014;2014:1-8.
12. Ihrke DK, Faber CS. Geographical mobility: 2005 to 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2021.
13. Jacquez GM, Kaufmann A, Meliker J, Goovaerts P, AvRuskin G, Nriagu J. Global, local and focused geographic clustering for case-control data with residential histories. *Environ Health*. 2005;4(1):4. doi:10.1186/1476-069X-4-4
14. Lan L, Malbasa V, Vucetic S. Spatial scan for disease mapping on a mobile population. Proceedings of the AAAI Conference on Artificial Intelligence; Vol 28, 2014; AAAI.
15. Manjourides J, Pagano M. Improving the power of chronic disease surveillance by incorporating residential history. *Stat Med*. 2011;30(18):2222-2233.
16. Petrof O, Neyens T, Nuyts V, Nackaerts K, Nemery B, Faes C. On the impact of residential history in the spatial analysis of diseases with a long latency period: a study of mesothelioma in Belgium. *Stat Med*. 2020;39(26):3840-3866.
17. Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: an application of generalized additive models to case-control data. *Environ Health*. 2005;4(1):11. doi:10.1186/1476-069X-4-11
18. Vieira V, Webster T, Weinberg J, Aschengrau A. Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. *Environ Health*. 2009;8(1):1-13.
19. Wheeler DC, Waller LA, Cozen W, Ward MH. Spatial–temporal analysis of non-Hodgkin lymphoma risk using multiple residential locations. *Spat Spatiotempl Epidemiol*. 2012;3(2):163-171.
20. Hurley S, Hertz A, Nelson DO, et al. Tracing a path to the past: exploring the use of commercial credit reporting data to construct residential histories for epidemiologic studies of environmental exposures. *Am J Epidemiol*. 2017;185(3):238-246.
21. Zahm SH, Hartge P, Hoover R. The National Bladder Cancer Study: employment in the chemical industry. *J Natl Cancer Inst*. 1987;79(2):217-222.
22. González CA, López-abente G, Errezola M, et al. Occupation and bladder cancer in Spain: a multi-Centre case-control study. *Int J Epidemiol*. 1989;18(3):569-577.
23. Wheeler DC, Wang A. Assessment of residential history generation using a public-record database. *Int J Environ Res Public Health*. 2015;12(9):11670-11682.
24. Meliker JR, Jacquez GM. Space–time clustering of case–control data with residential histories: insights into empirical induction periods, age-specific susceptibility, and calendar year-specific effects. *Stoch Environ Res Risk Assess*. 2007;21(5):625-634.
25. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Oxfordshire, England: Routledge; 2017.
26. Wood SN. Thin plate regression splines. *J R Stat Soc Ser B (Stat Methodol)*. 2003;65(1):95-114.
27. Wood SN. *Generalized Additive Models: an Introduction with R*. New York, NY: CRC Press; 2017.
28. Young RL, Weinberg J, Vieira V, Ozonoff A, Webster TF. A power comparison of generalized additive models and the spatial scan statistic in a case-control setting. *Int J Health Geogr*. 2010;9(1):1-12.
29. Vieira VM, Webster TF, Weinberg JM, Aschengrau A. Spatial-temporal analysis of breast cancer in upper Cape Cod, Massachusetts. *Int J Health Geogr*. 2008;7(1):1-12.
30. Kulldorff M. A spatial scan statistic. *Commun Stat - Theory Methods*. 1997;26(6):1481-1496. doi:10.1080/03610929708831995
31. Browne WJ, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. *Stat Modelling*. 2001;1(2):103-124.
32. Hill PW, Goldstein H. Multilevel modeling of educational data with cross-classification and missing identification for units. *J Educ Behav Stat*. 1998;23(2):117-128.
33. Nychka DW, Bailey BA, Ellner SP, Haaland PD, O'Connell MA. FUNFITS data analysis and statistical tools for estimating functions; 1996.

34. Bradley JR, Holan SH, Wikle CK. Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Ann Appl Stat*. 2015;9(4):1761-1791.
35. Prescott SL, Logan AC. Each meal matters in the exposome: biological and community considerations in fast-food-socioeconomic associations. *Econ Hum Biol*. 2017;27:328-335.
36. Shaddick G, Zidek JV. A case study in preferential sampling: long term monitoring of air pollution in the UK. *Spat Stat*. 2014;9:51-65.
37. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *J R Stat Soc Ser C Appl Stat*. 1998;47(3):299-350.
38. Johnson ME, Moore LM, Ylvisaker D. Minimax and maximin distance designs. *J Stat Plan Inference*. 1990;26(2):131-148.
39. Wang H, Ranalli MG. Low-rank smoothing splines on complicated domains. *Biometrics*. 2007;63(1):209-217.
40. Roy J, Stewart WF. Estimation of age-specific incidence rates from cross-sectional survey data. *Stat Med*. 2010;29(5):588-596.
41. Wheeler DC, Calder CA. Sociospatial epidemiology: residential history analysis. *Handbook of Spatial Epidemiology*. Sacramento, CA: Chapman and Hall/CRC; 2016:627-648.
42. Kim J, Lawson AB, McDermott S, Aelion CM. Bayesian spatial modeling of disease risk in relation to multivariate environmental risk fields. *Stat Med*. 2010;29(1):142-157.
43. Crainiceanu CM, Diggle PJ, Rowlingson B. Bivariate binomial spatial modeling of Loa loa prevalence in tropical Africa. *J Am Stat Assoc*. 2008;103(481):21-37.
44. Boyle J, Wheeler DC. Knot selection for low-rank kriging models of spatial risk in case-control studies. *Spat Spatiotemporal Epidemiol*. 2022;41:100483.
45. Teitz MB, Bart P. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Oper Res*. 1968;16(5):955-961.
46. Owen SH, Daskin MS. Strategic facility location: a review. *Eur J Oper Res*. 1998;111(3):423-447.
47. Baris D, Waddell R, Beane Freeman LE, et al. Elevated bladder cancer in northern New England: the role of drinking water and arsenic. *JNCI J Natl Cancer Inst*. 2016;108(9):1-9.
48. Baris D, Karagas MR, Verrill C, et al. A case–control study of smoking and bladder cancer risk: emergent patterns over time. *JNCI J Natl Cancer Inst*. 2009;101(22):1553-1561.
49. Mazeman E. Tumours of the upper urinary tract calyces, renal pelvis and ureter. *Eur Urol*. 1976;2:120-128.
50. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol 124. Vienna, Austria: Technische Universität Wien; 2003:1-10.
51. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-472.
52. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News*. 2006;6(1):7-11. https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf
53. Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease-mapping studies. *Environ Health Perspect*. 2004;112(9):1016-1025.
54. Nuckols JR, Freeman LEB, Lubin JH, et al. Estimating water supply arsenic levels in the New England bladder cancer study. *Environ Health Perspect*. 2011;119(9):1279-1285.
55. Census Bureau U.S. Population Density, 2010.; 2010.
56. Colt JS, Karagas MR, Schwenn M, et al. Occupation and bladder cancer in a population-based case-control study in northern New England. *Occup Environ Med*. 2011;68(4):239-249.

How to cite this article: Boyle J, Ward MH, Koutros S, et al. Estimating cumulative spatial risk over time with low-rank kriging multiple membership models. *Statistics in Medicine*. 2022;41(23):4593-4606. doi: 10.1002/sim.9527