Spatial Estimation of Radon Exposure for Epidemiologic Risk Assessment

Kyle Sorensen

Senior Honors Thesis
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

April 28, 2023

Approved:

Thesis Advisor

Reader #1

Reader #2

Contents

List of Figures

List of Tables

1 Introduction

Radon is a naturally occurring radioactive gas and an intermediate product of the decay of uranium. Exposure to radon is the second leading cause of lung cancer in the United States and is hypothesized to cause strokes [10] and other cardiovascular events such as clonal hematopoiesis of indeterminate potential [6], also known as CHIP. Additionally, radon levels appear to be rising across North America [5] and may be linked to climate change [13]. Thus, it is in our interest to model radon levels at a granular level to assess the epidemiological risks associated with radon exposure based on geographic location.

One attempt at fitting such a model occurred in the late 1980's and early 1990's, when the US Environmental Protection Agency created a map of three distinct radon zones. The zones were constructed using indoor radon measurement data as well as atmospheric, geologic and residential factors. This map is known as the EPA Map of Radon Zones but was also called the Geologic Radon Potential map by the US Geologic Survey [14], and sought to assign a metric to each county for the relative risk of harmful radon exposure. The primary weakness of this map is that it has fairly low spatial resolution, with one of three distinct levels assigned to each county. In 1996, researchers at Lawrence Berkeley labs used Bayesian hierarchical modeling on short term mean indoor radon measurement data to develop a more granular picture of radon levels [8]. But, the short-term data on indoor radon concentrations used in this study is highly variable and resulted in high predictive error estimates. In 1999, Gelman et al. used a survey of long-term mean indoor radon measurements to fit a similar hierarchical model with improved predictive accuracy [3]. However, this model could likely be improved by accounting for the other factors used in the EPA's radon zones map. Progress on statistical modeling of radon exposure using these data sets in the US has stagnated since the late 1990s, but a recent study by Li et al. used a comprehensive data set with over 500,000 observations and 68 covariates to develop a two-stage ensemble learning model with highly promising results [7]. However, we believe that there is still quite a bit of valuable information

that can extracted from the EPA's short-term survey and geologic radon potential data.

The goal of this analysis is to pursue the aforementioned areas for improvement and create a refined spatial model for the geographic distribution of indoor radon with some quantification of uncertainty accounting for geologic, atmospheric, and residential factors. Hence, we will provide more granular, improved estimates of radon exposure. To achieve this, we leverage kriging, latent process modeling and other spatial modeling techniques to analyze the State Residential Radon Survey and Geologic Radon Potential data.

2 Data

2.1 State residential radon survey (SRRS) data

The State Residential Radon Survey is a series of household-level short-term mean surveys of the lowest residential level of 63,291 homes in 42 US states and 6 native American territories conducted between 1986 and 1992 [2]. To collect the radon measurement data, a radon measurement device called a charcoal canister (shown in Figure 1) was placed in the home of each participant for two days, thus measuring the daily average indoor radon concentration. It is worth noting that the location of the measurement device within the home is a confounding variable here since some homes have basements, so the altitude of the sensor with respect to the home may vary.

This survey used a fairly complex sampling methodology. More specifically, researchers partitioned the United States into 22 strata across 10 regions to ensure that the survey provided coverage to areas that were expected to have differing levels of radon. More specifically, regions with higher expected radon measurements were biased in the sample. The effect of this sampling methodology is that the data is more densely packed in areas with high expected radon and more sparse in areas with low expected radon. The reason behind this seems to be that, in an epidemiological context, it may be better to overestimate radon in the interest

Figure 1: *Charcoal canister from Radon Testing Corporation of America, similar to those used in the State Residential Radon Survey*

of mitigating negative health effects as often as possible. We will take account of this when interpreting the final results of the analysis. The strata were constructed based on limited radon testing data available circa 1988 [2].

2.2 Geologic radon potential (GRP) and radon index (RI) data

The EPA's Geologic Radon Potential map (Figure 2) from 1993 was constructed using indoor radon measurements (pCi/L) from the State Residential Radon Survey, aerial radioactivity (ppm eU), geologic composition, soil permeability and architecture type and classified each county into one of three levels:

- "high" (estimated indoor radon concentration $>$ 4 picocuries per liter, or pCi/L) – zone 1
- "moderate/variable" (2–4 pCi/L) – zone 2
- "low" ($<$ 2 pCi/L) – zone 3

Notably, the GRP classification is a trichotimization of the radon index (RI) classification. With 15 distinct levels, the radon index provides a more granular but also somewhat less robust view of variability in geologic, atmospheric and residential features across the United States. This set of measures and classifications was selected since each of them has demon-

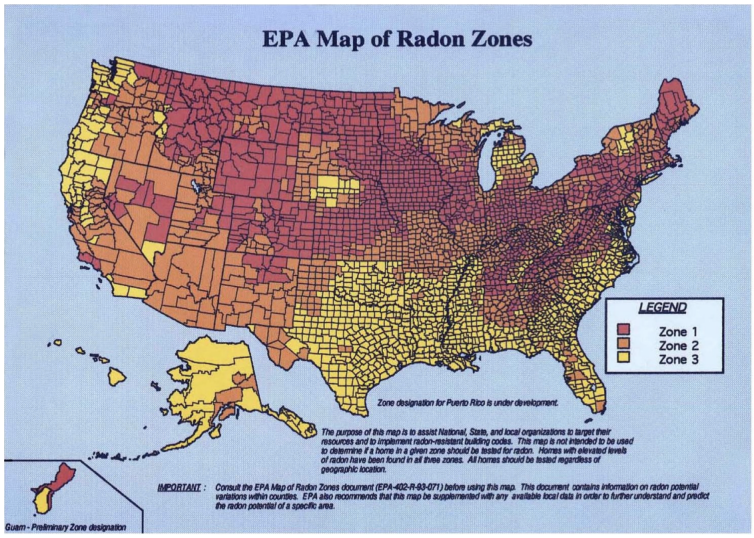


Figure 2: *Map of Geologic Radon Potential (GRP) zones from the EPA, 1993*

strated a strong correlation with radon measurements. The radon index and geologic radon potential values were computed for each county using the method outlined in Figure 3.

3 Methods

3.1 Transformation of the response

Before I begin describing the models used in this project, it is important to address the use of the transformed radon measurement values in the coming analyses. Across all of these methods, the models will be fit on the transformed radon values $Z = \log(\log(Y + 4))$ where Y is the vector of SRRS values in pCi/L and transformed back to the standard scale after prediction. We use this transformation to coerce the response to more closely follow a normal distribution according to a Kolmogorov-Smirnov test [15].

3.2 Kriging

Kriging, also known as Gaussian process modeling [18], is the most popular technique in spatial statistics and was created in 1979 by Noel Cressie [12]. The basic model is a Gaussian

TABLE 1. RADON INDEX MATRIX. "ppm eU" indicates parts per million of equivalent uranium, as indicated by NURE aerial radiometric data. See text discussion for details.

INCREASING RADON POTENTIAL →

FACTOR	POINT VALUE 1	2	3
INDOOR RADON (average)	< 2 pCi/L	2 - 4 pCi/L	> 4 pCi/L
AERIAL RADIOACTIVITY	< 1.5 ppm eU	1.5 - 2.5 ppm eU	> 2.5 ppm eU
GEOLOGY*	negative	variable	positive
SOIL PERMEABILITY	low	moderate	high
ARCHITECTURE TYPE	mostly slab	mixed	mostly basement

*GEOLOGIC FIELD EVIDENCE (GFE) POINTS: GFE points are assigned in addition to points for the "Geology" factor for specific, relevant geologic field studies. See text for details.

Geologic evidence supporting:	HIGH radon	+2 points
	MODERATE	+1 point
	LOW	-2 points
	No relevant geologic field studies	0 points

SCORING:

Radon potential category	Point range	Probable average screening indoor radon for area
LOW	3-8 points	< 2 pCi/L
MODERATE/VARIABLE	9-11 points	2 - 4 pCi/L
HIGH	12-17 points	> 4 pCi/L

POSSIBLE RANGE OF POINTS = 3 to 17

Figure 3: *Construction methodology for Radon Index (RI) values*

process whose mean is defined as a linear function of covariates, and whose covariance is a nonlinear function of the spatial coordinates; the latter feature gives the model its distinctive spatial structure [16]. As an aside, we will first discuss restricted maximum likelihood (REML) estimation since it is commonly used to obtain estimates for the parameters of this model.

3.2.1 Restricted maximum likelihood estimation

Restricted maximum likelihood or REML estimation is generally considered the best method for point estimation in Gaussian spatial models. While this is a computationally complex technique, the natural correspondence of maximum likelihood methods with Bayesian procedures allows for quantification of uncertainty in parameter estimates. Additionally, maximum likelihood methods are flexible in the sense that many models can be formulated and then compared using likelihood ratio tests or model selection criteria such as AIC or BIC. The purpose of using REML estimation as opposed to standard maximum likelihood estimation in this case is that it allows for approximately unbiased estimation of the kriging parameters. This description of REML estimation follows from Smith [16].

To illustrate this process, begin with n elements of data and q covariates in the model and let $W = A^T Z$ be a vector of $n - q$ linearly independent contrasts. The $n - q$ columns of A are linearly independent and $A^T X = 0$, so we find that

$$W \sim N[0, A^T \Sigma A].$$

Thus, with $\Sigma = \alpha V(\theta)$, the joint density of W is given by

$$(2\pi)^{-(n-q)/2} \alpha^{-(n-q)/2} |A^T V(\theta) A|^{-1/2} e^{-\frac{1}{2\alpha} W^T (A^T V(\theta) A)^{-1} W}$$

and the negative log likelihood is given by

$$l_W(\alpha, \theta) = \frac{n-q}{2} \log(2\pi) + \frac{n-q}{2} \log(\alpha) + \frac{1}{2} |A^T V(\theta) A| + \frac{1}{2\alpha} W^T (A^T V(\theta) A)^{-1} W.$$

From here, we must minimize $l_W(\alpha, \theta)$ with respect to α. It can be shown that this is achieved in setting $\tilde{\alpha} = G^2(\theta)/(n-q)$. Hence, with some simplification, we have

$$l_W^*(\theta) = \frac{n-q}{2} \log(2\pi) + \frac{n-q}{2} \log(\frac{G^2(\theta)}{n-q}) - \frac{1}{2} \log|X^T X| +$$

$$\frac{1}{2} |X^T V(\theta)^{-1} X| + \frac{1}{2} \log|V(\theta)| + \frac{n-q}{2}$$

Now that we have simplified the restricted likelihood and given an overview of REML estimation, we can apply it to the specific case of spatial models constructed via a technique known as kriging.

3.2.2 Overview and derivation of the kriging estimators

The generalized problem we attempt to solve via kriging [12] is as follows: given observations of a vector field $z(s_1), ..., z(s_n)$, predict the value of $z(s_0)$ for some $s_0 \notin \{s_1, ..., s_n\}$.

To begin the derivation of the kriging estimators, let $Z = (z(s_1), ..., z(s_n))^T$ and $z_0 = z(s_0)$.

From this we need to know the joint covariance matrix of Z and z_0; let's suppose

$$\text{Cov}\left\{\begin{array}{c} Z \\ z_0 \end{array}\right\} = \left\{\begin{array}{cc} \Sigma & \tau \\ \tau^T & \sigma_0^2 \end{array}\right\},$$

where Σ is the covariance matrix of Z, σ_0^2 is the variance of z_0 and τ is the vector of cross-covariances between Z and z_0. In terms of the common scale parameter α and functions $V,\ \omega$, and ν_0 of a finite-dimensional parameter θ, we can write

$$\Sigma = \alpha V(\theta), \tau = \alpha\omega(\theta), \sigma_0^2 = \alpha\nu_0(\sigma)$$

The model in this case will be of the form $Z = X\beta + \eta$ for some matrix of covariates X and we also assume that $z_0 = x_0^T\beta + \eta_0$ with x_0 known and both η and η_0 representing random errors with mean 0. This is called the *universal kriging problem.* We consider predictors of the form

$$\hat{z}_0 = \lambda^T Z$$

subject to the constraint

$$\lambda^T X = x_0^T.$$

The reason for this constraint is so that the procedure will work without assuming β is known. The prediction error can be written as

$$z_0 - \hat{z}_0 = x_0^T\beta + \eta_0 - \lambda^T(X\beta + \eta)$$

$$= \eta_0 - \lambda^T\eta.$$

The mean squared prediction error then becomes

$$E\{(z_0 - \hat{z}_0)^2\} = \sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda.$$

Thus we must solve the constrained optimization problem of minimizing the mean squared prediction error subject to $\lambda^T X = x_0^T$. To do this, we will consider the Lagrangian

$$L = \sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda - 2(\lambda^T X - x_0^T)\gamma$$

where 2γ is a vector of Lagrange multipliers. According to Lagrange multiplier theory, the optimal λ is achieved at some stationary point of L. Differentiating with respect to λ yields

$$\gamma = (X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau)$$

with the final result being

$$\lambda = \Sigma^{-1}\tau + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau)$$

and the predictor

$$\hat{z}_0 = \lambda^T Z = (x_0 - X^T\Sigma^{-1}\tau)^T\hat{\beta} + \tau^T\Sigma^{-1}Z$$

where $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Z$ from the method of generalized least squares. As a final point of this section, the mean squared prediction error can be written as

$$E\{(z_0 - \hat{z}_0)^2\} = \sigma_0^2 - 2\lambda^T\tau + \lambda^T\Sigma\lambda.$$

$$= \sigma_0^2 - \tau^T\Sigma^{-1}\tau + (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau).$$

3.2.3 Model specification for SRRS analysis

For this analysis, I will only be using the EPA's State Residential Radon Survey (SRRS) data.

To take advantage of the SRRS data's zip-code level spatial resolution, we consider a model

of the form

$$Z_i = \sum_{k=1}^{q} x_{ik}\beta_k + \sum_{j=1}^{m} a_{ij}W_j + \epsilon_i$$

where

- Z_i is the transformed radon value at home $i, i \in \{1, ..., n\}$;
- $x_{ik}, k \in \{1, ..., q\}$ are the values of q covariates (needed for later analyses involving atmospheric and geologic factors as well as other surveys);
- $\beta_1, ..., \beta_q$ are the regression parameters, but note that currently we only have one β since there are no linear covariates used for this analysis;
- W_j is the value of an unobserved latent process at sampling point (zipcode) j;
- a_{ij} is 1 if home i is in zipcode j and 0 otherwise;
- ϵ_i is the random error.

The values $\{a_{ij}\}$ comprise an $m \times n$ incidence matrix A. We assume that $\{W_j\}$ form a Gaussian process with mean 0 and covariances

$$\text{Cov}\{W_j, W_{j'}\} = \sigma_W^2 c_{jj'}$$

where σ_W^2 is the variance of the latent process and $c_{jj'}, j, j' \in \{1, ..., m\}$ are spatial correlations. We use an exponential structure for the correlations such that $c_{jj'}(\rho) = \exp(-d_{jj'}/\rho)$, where $d_{jj'}$ is the Euclidean distance between sampling points j and j' and ρ is the *spatial range* parameter, which characterizes the extent of spatial autocorrelation in the data.

The final assumption of the model is that

$$\epsilon_i \sim N[0, \sigma_\epsilon^2].$$

Thus, we can represent the model in matrix notation as

$$\mathbf{Z} = X\beta + A\mathbf{W} + \mathbf{E},$$

$$\mathbf{W} \sim N_m[0, \sigma_W^2 C(\rho)],$$

$$\mathbf{E} \sim N_n[0, \sigma_\epsilon^2 I_n]$$

such that the covariance matrix of $\mathbf{Z}$ is

$$\Sigma = \sigma_W^2 AC(\rho)A^T + \sigma_\epsilon^2 I_n.$$

If we separate out a common scale parameter α such that $\Sigma = \alpha V(\theta), \alpha = \sigma_W^2, \theta_1 = \sigma_\epsilon^2/\sigma_W^2$ and $\theta_2 = \rho$, then we have

$$V(\theta) = AC(\theta_2)A^T + \theta_1 I_n.$$

From here, we use the Sherman-Morrison-Woodbury formulas [9] to evaluate $V(\theta)^{-1}$ and $\det(V(\theta))$ as needed for the kriging algorithm. In general, the Sherman-Morrison-Woodbury identities are

$$(A + BCD)^{-1} = A^{-1} - A^{-1}C(B^{-1} + DA^{-1}C)^{-1}DA^{-1}$$

and

$$\det(A + BCD) = \det(B^{-1} + DA^{-1}C)\det(B)\det(A)$$

assuming that the matrices A, B, C, and D have suitable dimensions for the associated matrix products. Notably, these are used since they allow for much faster computation of the determinant and inverse of the spatial covariance matrix [9].

To obtain the predicted values for the maps and model validation statistics, we will reverse the loglog transformation.

3.2.4 Integration of GRP data, model specification for SRRS+GRP integrated analysis

Before I describe this analysis in further detail, it is worth noting that it was eventually discarded and was an ad-hoc procedure used to handle the incorporation of the GRP values. However, it is important to describe the methodology because it influenced the development of the corresponding latent process model and allows us to compare the two approaches to integrating the GRP values.

For this analysis, I will be using the EPA's Geologic Radon Potential (GRP) data in conjunction with the SRRS data. Each coordinate observed in the SRRS data was assigned the GRP value for its respective county.

To properly specify this model, we note that the GRP data was formulated using the SRRS data and that their respective values agree about 45% of the time [9] at the level of individual observations and 72% of the time at the county level [1]. Hence, there is some very clear overlap between these measures so our methodology for combining the data sources will not treat the SRRS and GRP data as independent and will instead focus on the conditional distributions of SRRS on GRP and vice versa.

As in the original kriging analysis, we write $Z = \log(\log(Y + 4))$ where Y is the measured value from SRRS. Notably, GRP is a classification of the assumed background level of radon. We will call the true transformed background level of radon T. Thus, in this phase of analysis, we must allow for random variation between the background level and the SRRS observations. The model assumed here is

$$T \mid Z \sim \mathcal{N}[Z, \sigma^2]$$

$$GRP = \begin{cases} 1 & \text{if } T > t_1^* \\ 2 & \text{if } t_2^* < T < t_1^* \\ 3 & \text{if } T < t_2^* \end{cases}$$

where σ^2 is an additional variance parameter and t_1^* and t_2^* are the transformed values of the interval endpoints for the GRP ranges, or $\log(\log(8)) = 0.7321$ and $\log(\log(6)) = 0.5832$ respectively.

In this model, we require an estimate of the standard deviation σ, which we will obtain via maximum likelihood estimation. The probability mass function of GRP given a Z value is

$$P(Z|GRP) = \begin{cases} 1 - \Phi\left(\frac{t_1^* - Z}{\sigma}\right) & \text{if } GRP = 1 \\ \Phi\left(\frac{t_1^* - Z}{\sigma}\right) - \Phi\left(\frac{t_2^* - Z}{\sigma}\right) & \text{if } GRP = 2 \\ \Phi\left(\frac{t_2^* - Z}{\sigma}\right) & \text{if } GRP = 3 \end{cases}$$

where Φ is the standard normal cumulative distribution function. The likelihood function is given by taking the product

$$L(\sigma) = \prod_{i=1}^{3} P(GRP = i|Z, \sigma)$$

and the MLE is derived using a simple one-dimensional optimization of $\log L(\sigma)$. For Tennessee, this procedure yields an MLE of $\hat{\sigma} = 0.2997$.

Now, we consider the conditional expectation of T given Z and GRP. Given $T \sim \mathcal{N}[\mu, \sigma^2]$, we want $\mathrm{E}\{T \mid a < T < b\}$. For $\mu = 0, \sigma = 1$, we have

$$\mathrm{E}\{T \mid a < T < b\} = \frac{\frac{1}{\sqrt{2\pi}} \int_a^b t e^{-t^2/2} dt}{\Phi(b) - \Phi(a)} = \frac{\phi(a) - \phi(b)}{\Phi(b) - \Phi(a)}$$

where $\phi(t)$ is the standard normal density. We can extend this to general μ, σ with

$$\mathrm{E}\{T \mid a < T < b\} = \mu + \sigma \cdot \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

Applying this to each of the cases $a = -\infty, b = t_2^*; a = t_2^*, b = t_1^*; a = t_1^*, b = \infty$ for $GRP = 1, 2, \text{and } 3$ respectively with $\mu = Z$ and $\sigma = \hat{\sigma}$, we can obtain $Z^* = \mathrm{E}\{T \mid Z, GRP\}$ for each pair of (Z, GRP) in the data set and repeat the kriging algorithm using the newly constructed Z^* values.

To obtain the predicted values for the maps and model validation statistics, we will reverse the loglog transformation.

3.3 Latent process modeling

Also known as latent variable models or Gaussian process models, *latent process models* are a flexible Bayesian modeling technique that relate observed data to some set of unobservable parameters. Notably, in the SRRS+GRP kriging analysis, we condition the background indoor radon concentration on the observed values. However, it seems somewhat more rational to reverse this conditioning. The latent process modeling approach and associated computational tools will allow us to structure our model such that the observed radon values will be conditioned on the background indoor radon concentration. Conceptually, we are treating the observed radon values as noisy samples from the true distribution of radon.

3.3.1 Bayesian hierarchical models

Latent process models are often represented in a Bayesian hierarchical formulation, the general form of which can be described as follows: Let y_i be an observation and θ_i be a parameter that dictates the generating process for y_i. Assume also that each θ_i is iid

according to a hyperparameter ϕ. Thus, the model can be expressed as

$$y_i|\theta_i, \phi \sim P(y_i|\theta_i, \phi)$$

$$\theta_i|\phi \sim P(\theta_i|\phi)$$

$$\phi \sim P(\phi)$$

Thus according to Bayes' theorem, we note that $P(\theta_i, \phi|y) \propto P(y_i|\theta_i)P(\theta_i|\phi)P(\phi)$. The most common techniques used for this sort of sampling problem are called Markov Chain Monte Carlo (MCMC) methods.

3.3.2 Model specification for SRRS analysis

Let $Z \in \mathcal{R}^m$ be the vector of transformed latent process values for each zipcode. Additionally, let the Gaussian spatial process have the covariance matrix $K \in \mathcal{R}^{m \times m}$ defined such that

$$K(x|\alpha, \rho, \sigma)_{i,j} = \alpha^2 \exp\left(-\frac{1}{2\rho^2} d_{ij}^2\right) + \delta_{i,j}\sigma^2$$

where d_{ij} is the Euclidean distance between sampling locations i and j in terms of their coordinates, as defined in the kriging analysis. Further, α is the marginal or spatial standard deviation, σ is the random standard deviation, and ρ is the spatial range parameter and $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Note that this is known as the *Kronecker delta function* and is adds the fixed random variance to the diagonal of K to ensure that it is positive definite. Now, let $\tilde{Z} \in \mathcal{R}^{m \times d}$ be a matrix of transformed SRRS values where m is the number of zip codes and d is the maximum number of SRRS samples amongst the zip codes. Hence, each row is the vector of transformed SRRS observations at the corresponding zipcode. Note that we will ignore the empty cells of this matrix since the number of SRRS samples is not constant across zipcodes in the data set. Thus, we consider a model of the

form

$$\tilde{Z}_i | Z_i \overset{\text{iid}}{\sim} N(Z_i, \sigma^2)$$

where $\tilde{Z}_i$ is a vector of transformed radon values at zipcode i and $Z_i \in \mathcal{R}$ is the value of the latent process on the loglog scale at zipcode i with $Z = L_K \eta \in \mathcal{R}^m$ where η is a vector of iid standard normal random variables. Here, L_K is the Cholesky decomposition of the spatial covariance matrix K where d_{ij} is the Euclidean distance between sampling locations i and j in terms of their coordinates, as defined in the kriging analysis. The Cholesky decomposition is a decomposition of a positive definite matrix into the product of a lower triangular matrix and its transpose. The use of this decomposition is what allows us to simulate a spatial process with the desired covariance structure according to the kernel function $K(x|\alpha, \rho, \sigma)$. It is also worth noting that K corresponds to Σ in the kriging analysis. The difference here is that K uses the quadratic exponential covariance function and Σ uses the exponential covariance function. The reason for this distinction is related to our use of STAN, which I will describe in more detail in the coming sections. More specifically, STAN does not include the exponential covariance function as a precoded option. Additionally, the two functions seem to be near enough equivalent at this spatial scale for practical estimation and prediction problems.

Prior distributions for the associated model parameters are given by

$$\sigma \sim \Gamma(1, 1)$$

$$\eta_1, ..., \eta_m \overset{\text{iid}}{\sim} N(0, 1)$$

$$\alpha \sim \Gamma^{-1}(1, 1)$$

$$\rho \sim \Gamma^{-1}(101, 0.5)$$

One choice to note is that we place a strict prior on ρ to ensure convergence of the MCMC

on a numerical scale consistent with the range of interpolation. The other prior distributions are quite commonly used for this sort of analysis and are provided in various examples of similar models in the STAN manual [17], but could be subject to further tuning in the future. In general, we seek to select prior distributions that ensure a well-behaved MCMC.

To compute predicted values for new coordinate values, we will use the algorithm outlined in the appendix titled *Algorithm for predictions from latent process models.* At a high level, we use the kriging formula for predicted values in each iteration of the MCMC and take a simple average over the iterations for each new set of coordinates. Then we reverse the original loglog transformation to obtain the predicted values.

3.3.3 Model specification for SRRS+GRP integrated analysis

For this analysis, I will be using the EPA's Geologic Radon Potential (GRP) map in conjunction with the SRRS data. For use in this analysis, each coordinate observed in the SRRS data was assigned the GRP value for its respective county.

We start with the same model specifications as in the latent process model for the SRRS data. In addition, $G \in \{1, 2, 3\}^m$ where each value of G is the geologic radon potential of the associated zipcode.

Let $B = (b_1, b_2, b_3)$ represent the group mean observed transformed radon values associated with each level of GRP. Based on the EPA's indoor radon concentration ranges associated with each GRP zone, we expect that $b_3 \in [0, 2], b_2 \in [2, 4]$, and $b_1 \in [4, \infty]$. In practice, we will transform the endpoints of these intervals to the loglog scale.

We construct the prior distribution for b_i as

$$b_i \sim N[\log(\log(2 * (3 - i) + 5)), 0.01], i = 3, 2, 1$$

where 0.01 is a common fixed variance across the GRP levels. This choice was made to

constrain the latent process values to more closely follow the ranges associated with the GRP values. Note that these distributions are formulated such that the group means lie at the minimum of the ranges associated with each GRP level. We could choose the midpoint of these intervals, but then selecting the prior group mean for b_1 is not so straightforward.

To incorporate these group means into our estimation of the latent process, we define $\Phi \in \mathcal{R}^{m\times 3}$ where m is the number of zip codes. We define

$$\Phi_{i,j} = -1000|Z_i - b_j|$$

where Z_i is the transformed value of the latent process at zipcode i and b_j is the radon value associated with the j^{th} GRP group mean. The reason for scaling $\Phi_{i,j}$ by 1000 is to ensure that the latent process values more closely follow the ranges associated with the GRP values. For each zipcode let G_i follow a *categorical logit* distribution with the probability that an observation belongs to GRP level w, π_w computed using the *softmax* formulation of multi-class probabilities [17] such that

$$\pi_w = \frac{\exp(\Phi_{i,w})}{\sum_j \exp(\Phi_{i,j})}$$

.

To compute predicted values for new coordinate values, we will use the algorithm outlined in the appendix titled *Algorithm for predictions from latent process models.* At a high level, we use the kriging formula for predicted values in each iteration of the MCMC and take a simple average over the iterations for each new set of coordinates. Then we reverse the original loglog transformation to obtain the predicted values.

3.3.4 Overview of STAN

One concern with conditioning the transformed response on the true indoor radon concentration is the computational capacity required to conduct the downstream analysis [9]. More specifically, a naive coding of the latent process model in R where we iteratively update each individual data point would be far beyond our computational capacity. To remedy this, we will use a software platform for Bayesian inference called STAN, authored by Andrew Gelman and the STAN development group. At a high level, STAN utilizes a C++ back-end to run MCMC simulations to fit or simulate from a wide array of user-specified models [17] with the added benefit that it is much faster than any routine we would implement directly in R.

More specifically, STAN has the capacity to fit a latent process model as shown in section 10.3 of [17] with highly flexible model specifications including a wide variety of built-in prior and conditional distributions. Additionally, the covariance functions available in STAN can be readily used with our coordinate-oriented data to construct spatial covariance matrices, making STAN a strong tool for spatial analyses and highly suitable for our needs.

3.4 Alternative methods

3.4.1 Locally estimated scatterplot smoothing (LOESS)

LOESS or locally estimated scatterplot smoothing, is a non-parametric regression technique commonly used for interpolation [11]. It is a local method due to its decaying weight structure in which more distant data factors less into the estimated value than nearby data for a given station.

Notably, LOESS uses the *tri-cube weight function* specified by

$$w(x) = (1 - |d|^3)^3$$

where

$$|d| = \frac{||x - z||_2}{\max_{y \in D} ||y - z||_2} \in [0, 1]$$

is the scaled Euclidean distance between a station x and prediction location z. Observe that $\max_{y \in D} ||y - z||_2$ is the maximum distance between the prediction location z and the elements of D, so $|d|$ is at most 1. The result of LOESS in the context of longitude-latitude coordinates is a smoothed surface which is locally linear or quadratic with respect to the coordinates at each location in the data. At a high level, each point on the interpolated surface can be thought of as a weighted average of a local subset of the data.

In R, the implementation of LOESS requires specification of a span parameter, which represents the proportion of nearest data that is considered when interpolating at a given station and is thus used to construct the aforementioned local subset D. We choose a span of 0.3, since it roughly corresponds to the magnitude of the spatial range parameter estimates we have seen in prior analyses. For the sake of simplicity we will stick with a span of 0.3 since it results in smoothness that is consistent with our other analyses. However, in principle we could use a cross-validation approach to tune this parameter

3.4.2 Ensemble estimation

As an initial attempt at a multistage model, I took a simple mean of the fitted values from each model excluding the the kriging analysis with the SRRS and GRP values. The reason for this exclusion is that the model has an extremely strong positive bias and significantly worse accuracy than the other candidate models.

This approach could be improved by using a more complex modeling approach in the second stage, with the kriging, latent process and LOESS models being the *base learning model* of the two-stage approach as referenced by Li [7].

3.5 Model validation

When modeling based on spatial data, it is not so obvious what the most effective method for model validation is since spatial data typically demonstrates a tendency for nearby data to be more similar than distant data [4]. This is more or less an obvious property, but it is one that must be addressed when validating a spatial model since this dependence structure can violate the assumption of independence between training and testing data if not handled appropriately. Some methods for mitigating the effects of the spatial dependence structure on evaluations of model performance are spatial blocking and buffered leave-one-out cross-validation [4]. However, these methods are quite computationally intensive and could be used at a later stage of the project. Still though, we must address the potential issues regarding independence of the training and test sets.

To validate our set of spatial models while limiting the computational burdens of a full cross-validation, we will used an 80/20 split between the training and testing data. To mitigate concerns with the effect of spatial dependence in our training and testing sets, the split will be conducted on the zipcode level rather than the individual observation level. This will ensure that we are predicting to locations for which we do not have any information within the training set. Thus, our test error metrics should closely mirror those we could expect in practice when applying this model to unobserved stations. For repeatability purposes, we set a seed of 1 for the sample function in R. Moving forward, we will want to repeat this analysis using different seeds for the sample function or with a full cross-validation to ensure that we more robustly evaluate these models. The metrics we will use to evaluate these models are mean absolute error (MAE) which is given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

and mean bias which is given by

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

where n is the number of observations in the test set, $\hat{y}_i$ is the estimated indoor radon concentration at the zipcode associated with observation i, and y_i is the measured indoor radon concentration for observation i. Notably, a positive mean bias indicates that the model tends to overestimate the indoor radon concentration.

Conversely, the heat maps for each method will be constructed using all available data since our goal is construct maps that use the available data to the best effect possible for use in epidemiological risk assessment.

4 Results

For the primary case study in this dissertation, each model described above will be fit to a 3 by 3 coordinate region centered approximately on Nashville, TN specified by latitudes $34.5 - 37.5°$N and longitudes $85.5 - 88.5°$W. I selected this region due to the presence of several bands of varying GRP, suggesting greater variability in potential radon exposure across the state. This subset of the data contains 1247 observations across 256 zip codes, so it is relatively dense when compared with the entire United States [9]. Notably, the training data contains 1038 observations while the test data contains 209 observations. As mentioned in the description of the methods, we will fit all models on the transformed data $Z = \log(\log(Y + 4))$ where Y is the observed indoor radon concentration in the SRRS data. Note that this transformation was used to ensure that normality assumptions of our models were not too harshly violated. Note also that when we observe the results of these methods, we will reverse the transformation to formulate predicted values.

m	n	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\alpha}$	$\hat{\mu}$
256	1247	5.103	0.187	0.00602	0.530

Table 1: *Key parameter estimates from the SRRS kriging analysis*

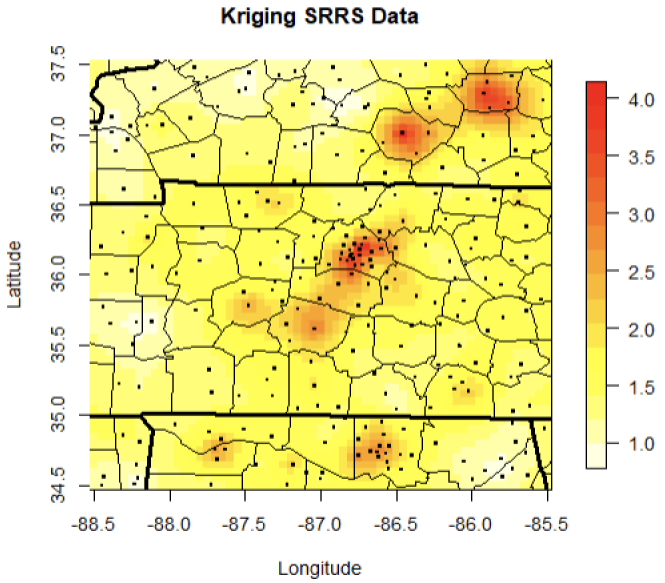


Figure 4: *Map of predicted radon exposure (pCi/L) over the 3 by 3 coordinate region in Tennessee*

4.1 Kriging

4.1.1 Kriging for SRRS data

Table 1 shows estimates for the kriging model parameters. These estimates seem reasonable and numerically stable, since the optimization function gives results consistent to 3 decimal places after multiple attempts. Hence, these parameter estimates should allow for relatively robust predictions.

The resulting heat-maps generated from kriging on the Tennessee SRRS data are shown in Figure 4.

The test MAE of this model is 2.112 pCi/L and the bias on the test set of this model is -0.397 pCi/L.

m	n	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\alpha}$	$\hat{\mu}$
256	1247	0.0208	1.459	0.1911	0.572

Table 2: *Key parameter estimates from the SRRS+GRP kriging analysis*

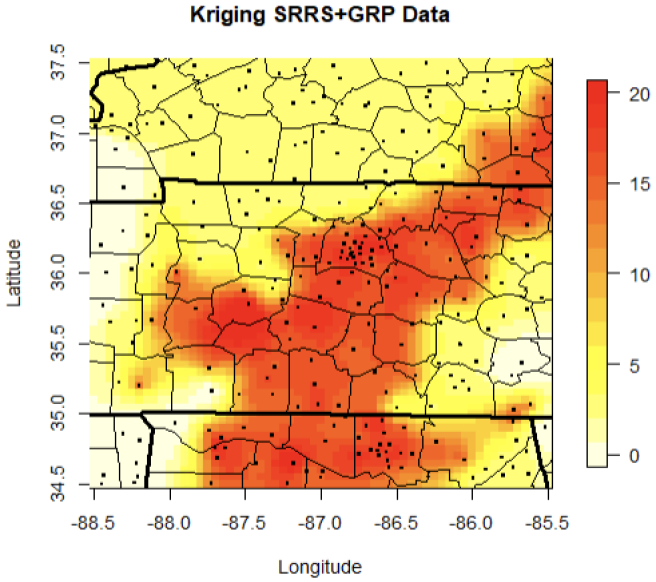


Figure 5: *Map of predicted radon exposure (pCi/L) over the 3 by 3 coordinate region in Tennessee*

4.1.2 Kriging for SRRS and GRP data

Table 2 shows estimates for the kriging model parameters. These estimates seem reasonable and numerically stable, since the optimization function gives results consistent to 3 decimal places after multiple attempts. Hence, these parameter estimates should allow for relatively robust predictions.

The resulting heat-maps generated from kriging on the Tennessee SRRS and GRP data are shown in Figure 5.

The test MAE of this model is 9.941 pCi/L and the bias on the test set of this model is +4.325 pCi/L.

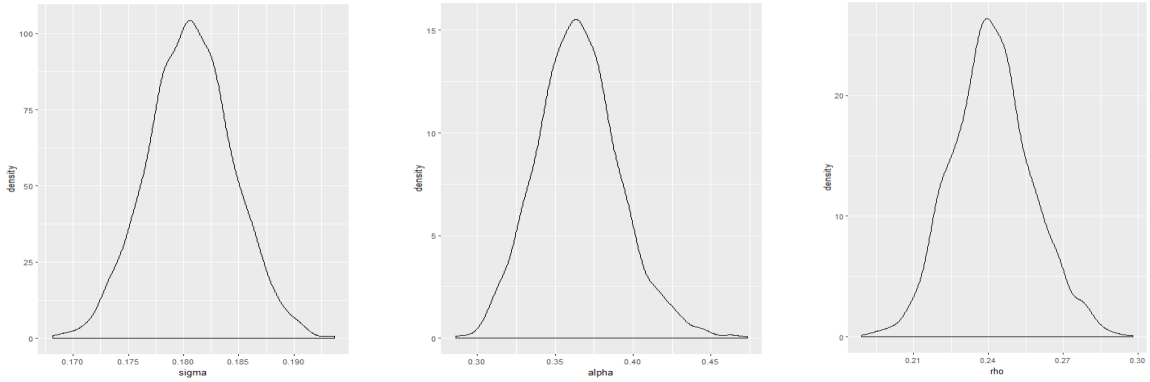
m	n	σ	α	ρ
256	1247	0.1806	0.3649	0.2415

Table 3: *Key parameter estimates from the SRRS latent process model*

4.2 Latent process modeling

4.2.1 LPM for SRRS data

The resulting posterior mean parameter estimates for σ, α and ρ can be seen in Table 3 with posterior distributions shown in Figure 6.



(a) Density plot for σ (b) Density plot for α (c) Density plot for ρ

Figure 6: *Posterior densities for (a) σ, (b) α and (c) ρ from the latent process model based on SRRS data*

The heat-map generated from the latent process modeling approach on the Tennessee SRRS data can be seen in Figure 7.

The test MAE of this model is 2.209 pCi/L and the bias on the test set of this model is -0.416 pCi/L.

4.2.2 LPM for SRRS and GRP data

The resulting posterior mean parameter estimates for σ, α and ρ can be seen in Table 4 with posterior distributions shown in Figure 8.

The heat-map generated from the latent process modeling approach on the Tennessee SRRS data can be seen in Figure 9.

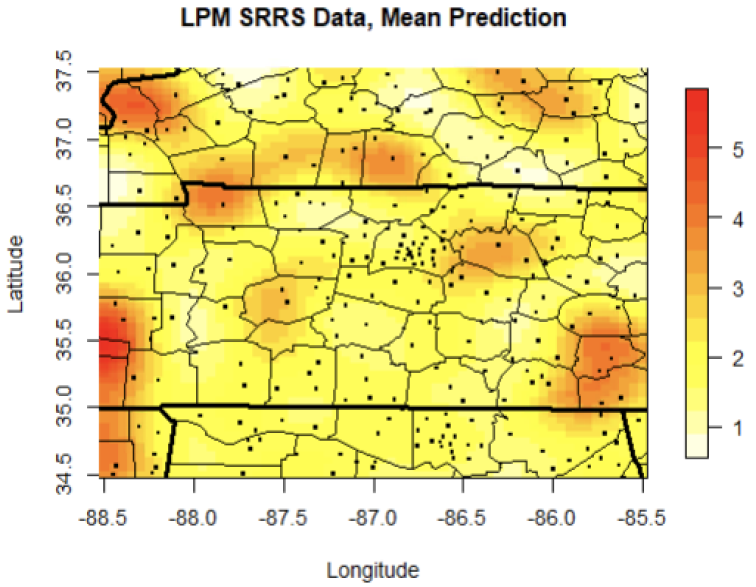
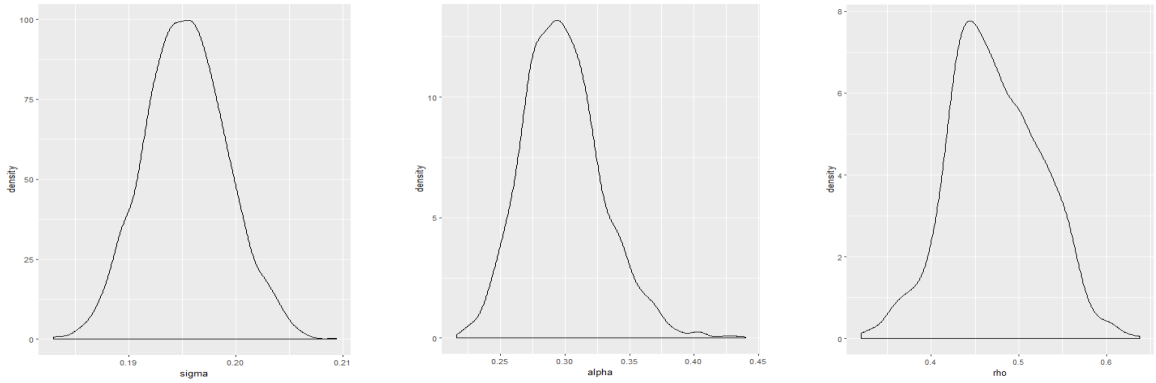


Figure 7: *Map of mean predicted radon exposure (pCi/L) over a 3 by 3 coordinate grid in central Tennessee from latent process model based on SRRS data*

m	n	σ	α	ρ
256	1247	0.1953	0.2984	0.4728

Table 4: *Key parameter estimates from the SRRS+GRP latent process model*



(a) Density plot for σ (b) Density plot for α (c) Density plot for ρ

Figure 8: *Posterior densities for (a) σ, (b) α and (c) ρ from the latent process model based on SRRS data*

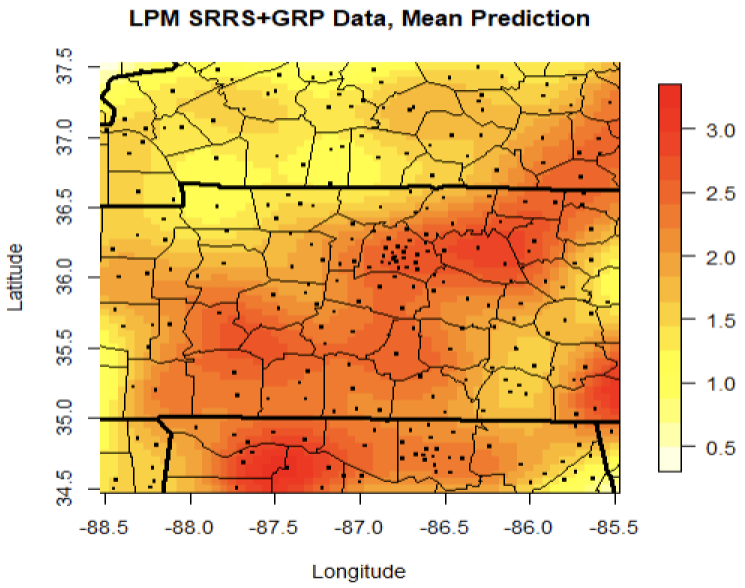


Figure 9: *Map of mean predicted radon exposure (pCi/L) over a 3 by 3 coordinate grid in central Tennessee from latent process model based on SRRS and GRP data*

The test MAE of this model is 2.071 pCi/L and the bias on the test set of this model is -0.462 pCi/L.

4.3 LOESS for SRRS data

The heatmap resulting from LOESS can be seen in Figure 10.

The test MAE of this model is 2.329 pCi/L and the bias on the test set of this model is -1.044 pCi/L.

4.4 Ensemble estimation

The heatmap resulting from the ensemble estimation approach can be seen in Figure 11.

The test MAE of this model is 2.059 pCi/L and the bias on the test set of this model is -0.579 pCi/L.

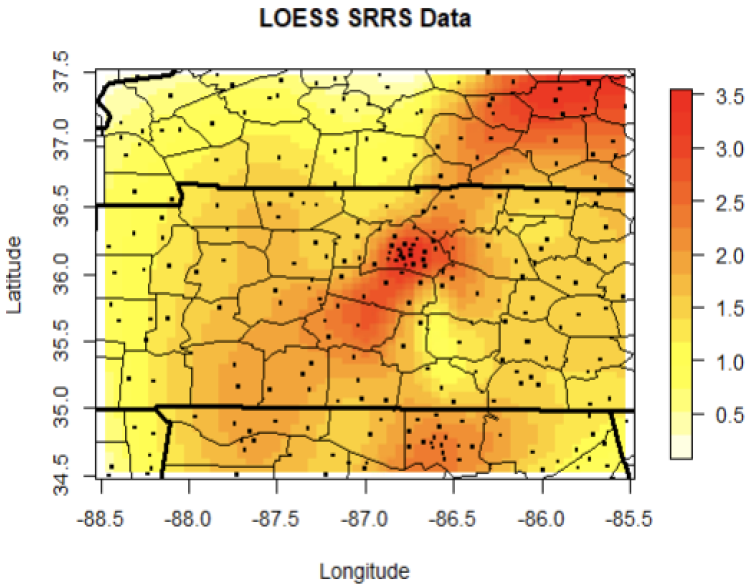


Figure 10: *Map of radon exposure (pCi/L) over a 3 by 3 coordinate grid in central Tennessee interpolated using LOESS regression*

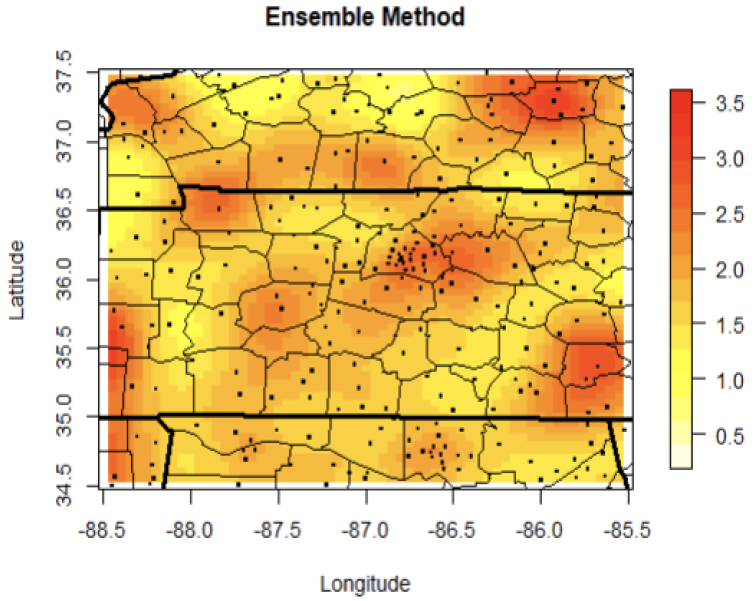


Figure 11: *Map of radon exposure (pCi/L) over a 3 by 3 coordinate grid in central Tennessee interpolated using ensemble estimation*

Model	Test MAE (pCi/L)	Test Set Bias (pCi/L)
Kriging (SRRS)	2.112	-0.397
Kriging (SRRS+GRP)	9.941	+4.325
LPM (SRRS)	2.209	-0.416
LPM (SRRS+GRP)	2.071	-0.462
LOESS	2.329	-1.044
Ensemble estimation	2.059	-0.579

Table 5: *Model validation results: MAE (pCi/L) and bias (pCi/L) calculated on test set (n=209)*

4.5 Summary of model validation results

For comparison purposes, I have displayed all of the model validation results in Table 5.

5 Discussion

There are several points worth discussing based on this regional case study. I will begin by commenting on the model validation results and move into a discussion of the maps generated from each analysis along with the key parameter estimates associated with each model. Then I will discuss some practical considerations regarding the strengths and weaknesses of this project as a whole.

The best model in terms of absolute accuracy on the test set is the ensemble estimation approach (2.059 pCi/L). Notably, all of the models aside from the kriging analysis using the SRRS and GRP data have comparable accuracy on the test set (< 2.4 pCi/L).

The best model in terms of testing set bias is the kriging model using only the SRRS data (-0.397 pCi/L). Notably, the test set bias of both latent process models and the kriging model using only the SRRS data are comparable. However, the LOESS model and kriging model with SRRS and GRP data have significantly stronger bias. All of the models besides the kriging model using the SRRS and GRP data have a negative bias on the test set. This tendency to underestimate radon levels is expected due to the presence of some extreme positive outliers in the SRRS data. Even though a positive bias is somewhat more desirable

than a negative bias in this context, the accuracy of this model (9.941 pCi/L) is significantly worse relative to the others in this case study.

Another observation to make from the model validation results is that the MAE of the latent process model marginally decreased upon the integration of the GRP values, while the MAE of the kriging analysis drastically increased. Thus, it seems that the latent process modeling approach more rigorously integrates the GRP values than the modified kriging analysis.

One potential drawback of the ensemble estimation method is that while is is the most accurate model, it may lack some interpretability. More specifically, we can not capture estimates of the spatial range parameter or within-zipcode variance as we can in the kriging or latent process modeling approaches. On a related note, the LOESS regression approach is fairly interpretable but is not as explicitly spatial as the kriging or latent process modeling approaches. However, interpretability is likely much less important than accuracy in this context.

Notably, the latent process modeling and kriging approaches seem to result in fairly different radon maps. The source of this discrepancy is unclear since the latent process model using only the SRRS values is conceptually a similar model to the analogous kriging model. However, the estimates of the key model parameters are similar in scale for the two models. Additionally, the effect of including the GRP values was similar for the two approaches. More specifically, inclusion of the GRP values resulted in a larger spatial range parameter estimate for both approaches (0.2415 to 0.4728 in the latent process model and 0.187 to 1.459 in the kriging model). This is reasonable since the GRP classification is discrete and is thus subject to less spatial variability than the continuous SRRS values.

Another note to make regarding the parameter estimates for these models is that they seem to be quite numerically stable. After fitting each of these models several times on the Tennessee data, we observed very little fluctuation in the parameter estimates across the iterations.

Based on the results of this analysis, the ensemble estimation approach seems to be the best

among these six choices for the estimation of radon levels since it has the lowest test MAE and reasonably small negative bias on the testing set. This model is also highly flexible since its individual components allow for the integration of more data sets. Additionally, the multi-stage approach allows for further flexibility in model selection and has the potential to increase predictive accuracy with further development.

There are few notable limitations of these modeling approaches as they have been implemented. First, there is an unaccounted for effect from the sampling design of the SRRS data. Since the original survey over-sampled areas of high population density and high suspected indoor radon concentration, we expect that estimates of radon concentration in these areas will be greater than those in other geographic regions. Thus, the accuracy of the models may fluctuate across the range of interpolation. Additionally, the sampling weights were not accounted for in these models because our selected region was not contained within the borders of a single state. If we were to account for these weights, we would likely see that the effect of large positive outliers in the data is reduced. Second, the observations are given as zipcode centroid coordinates rather than residential rooftop coordinates. This results in a jittering of the predicted radon surface and lessens the accuracy of estimates in terms of the exact location of coordinate regions with high indoor radon concentration. This effect worsens in rural areas since homes sampled within a rural zipcode may be quite far from the zipcode centroid. With access to residential rooftop coordinates, we could greatly increase the performance and specificity of these models.

Furthermore, there is a set of issues related to the computational constraints of this project. Notably, this data is stored on the Cardiovascular Epidemiology Research System, or CERES. Hence, the computation speed and storage limits are significantly more imposing than what we might see on another system, such as UNC's Longleaf computing cluster. Because of these computational limits, we were unable to fit all of the models on the entire state of Tennessee and elected to use a smaller 3 by 3 coordinate region for the case study. Notably,

this means we were unable to properly account for the differences in sampling weights by state. This likely affected our estimated indoor radon concentrations and may explain the discrepancy between the kriging and latent process model maps as they appear. With access to stronger computational tools outside of CERES, we could more easily fit these models on entire states or groups of states and account for the sampling weights accordingly. After making these improvements, we would expect that our maps will be more consistent with what is known about the distribution of radon in Tennessee and also more consistent with one another.

6 Future Work

Some of the original paths for future analysis outlined by Smith [9] remain. These include

- Integration of additional data sets: Now that we have established a base for the latent process and have a proof of concept with respect to incorporation of additional data sets, it would not be too difficult to bring in other data sources. Specifying the conditional distributions for the latent process is non-trivial but certainly doable.

- Temporal component: More recent radon measurement data exist, such as the US EPA's National Residential Radon Survey. Accurate forecasting of radon levels is obviously desirable, so it is worth investigating how we might integrate some temporal component into our models in the future.

- Bias in the sampling design, sampling weights: One piece of this analysis that we have not yet corrected for is the effect of the sampling design on our model fitting and predicted values. It is important to recall that areas with high expected radon exposure were biased in the SRRS.

- Other modeling approaches: Nearest neighboring measure and inverse distance weighted mean models could also be used to predict indoor radon concentrations. However, the

issue with these approaches is that they may be more drastically affected by us only having access to zipcode centroid coordinates rather than residential rooftop coordinates.

- More robust cross-validation of existing models: Due to the computational limitations of CERES, we did not implement a full cross-validation and opted for a training-validation set approach for model validation. Access to more powerful computational systems would grant us the capability to validate our models more rigorously.

An additional route for future work is to further develop the ensemble estimation approach. Li et al. use several second-stage models to improve their base learning model resulting in a weighted cross-validation RMSE of approximately 0.6 pCi/L. Notably, Li et al. also use a much larger data set than ours with over 500,000 observations [7]. However, this figure is still quite promising and suggests that a two-stage model could drastically improve our results on the SRRS data.

A Appendix

A.1 Training/test split in R

```
# get coordinate matrix by selecting unique coordinate pairs
# (these correspond to zipcodes in the data set)
coordinates_by_zipcode <- data %>%
  select(LAT, LON) %>%
  unique()

# define training and testing split
# (80/20 on zipcode level)
set.seed(1)
train <- sample(1:nrow(coordinates_by_zipcode),
    0.8*nrow(coordinates_by_zipcode))
```

```
test <- (-train)
train_zips <- coordinates_by_zipcode[train,]
test_zips <- coordinates_by_zipcode[test,]

# reset indices
rownames(train_zips) <- NULL
rownames(test_zips) <- NULL
# use right join to filter for observations from training zipcodes
train_data <- data %>%
  right_join(train_zips)
# use right join to filter for observations from testing zipcodes
test_data <- data %>%
  right_join(test_zips)
```

A.2 Kriging analysis for SRRS data in R

```
# Set up for likelihood calculation
n <- nrow(train_data)
X <- matrix(rep(1,n), ncol=1)
q <- ncol(X)
z <- log(log(4+train_data$ACTIVITY))
data_zip <- unique(train_data[,2:3])
index <- rep(0,nrow(train_data))
m <- nrow(data_zip)
for(i in 1:nrow(train_data)) {
  index[i] <- which(train_data[i,2] == data_zip[,1] & train_data[i,3] ==
   data_zip[,2])
}
num <- rep(0, nrow(data_zip))
for(j in 1:nrow(data_zip)) {
  num[j] <- sum(index == j)
}
C0 <- matrix(0, nrow=nrow(data_zip), ncol=nrow(data_zip))
```

```
V <- matrix(0, ncol=n, nrow=n)
par <- rep(0,2)
# Set initial values for parameter estimates for use in optim function
par[1] <- 0.5
par[2] <- 0.5

# Likelihood calculation helper function
helper <- function(par){
  diag(C0) <- 1
  for(j1 in 2:nrow(data_zip)){
    for(j2 in 1:(j1-1)){
      # What should I use for alpha here??
      distance <- sqrt((data_zip[j1,1] - data_zip[j2,1])^2 + (data_zip[j1
    ,2] - data_zip[j2,2])^2)
      C0[j1,j2] <- exp(-distance / par[2])
      C0[j2,j1] <- C0[j1,j2]
    }
  }
  B <- solve(par[1]*solve(C0)+diag(num))
  VI <- diag(rep(1,nrow(train_data)))
  for(i1 in 1:nrow(train_data)) {
    for(i2 in 1:nrow(train_data)) {
      VI[i1,i2] <- VI[i1,i2] - B[index[i1],index[i2]]
    }
  }
  VI <- VI / par[1]
  logdetV <- as.numeric(0-determinant(B)$modulus+
                          determinant(C0)$modulus+
                          (nrow(data)-nrow(data_zip))*log(par[1]))
  HVI <- solve (t(X) %*% VI %*% X)
  betahat <- as.numeric(HVI %*% t(X) %*% VI %*% z)
  H <- VI-VI %*% X %*% HVI %*% t(X) %*% VI
  G2 <- as.numeric(z %*% H %*% z)
```

```
  return(list(VI=VI,HVI=HVI,logdetV=logdetV,G2=G2,betahat=betahat))
}

# Compute likelihood for current values of parameter estimates
likelihood <- function(par) {
  if(par[1]<0 | par[1]>50 | par[2]<0.1 | par[2]>50) {
    lh <- 1e10
    return(lh)
  }
  lh_helper <- helper(par)
  lh <- as.numeric(0.5*(n-q)*log(lh_helper$G2) +
                     0.5*determinant(t(X) %*% lh_helper$VI %*% X)$modulus
   +
                     0.5*lh_helper$logdetV)
  return(lh)
}

# Restricted maximum likelihood estimation for kriging parameters
T1 <- Sys.time()
out <- optim(par, likelihood, method='BFGS', hessian=T,
             control=list(maxit=10000, ndeps=rep(1e-6, length(par))))
par <- out$par
T2 <- Sys.time()
print(T2-T1)
# Time difference of
lh_helper <- helper(par)
print(c(m,n,par,lh_helper$G2/(n-q),lh_helper$betahat))
likelihood(out$par)

# Define target coordinate sets for maps
y_pred <- latitude_s + 0.05*0:((latitude_n-latitude_s)/0.05)
x_pred <- longitude_w + 0.05*0:((longitude_e-longitude_w)/0.05)

```

```
pred <- function(y,par,testcoord,lh_helper){
  alf <- lh_helper$G2/(n-q)
  theta1 <- par[1]
  n <- nrow(y)
  tau <- rep(0,n)
  for(i in 1:n) {
    distance <- sqrt((y[i,2]-testcoord[1])^2+(y[i,3]-testcoord[2])^2)
    tau[i] <- alf * exp(-distance / par[2])
  }
  x0 <- 1
  X <- matrix(1,nrow=n,ncol=1)
  yhat <- as.numeric((x0-t(X) %*% as.matrix(lh_helper$VI) %*% as.numeric(
   tau)/alf) %*%
                       lh_helper$betahat + t(as.numeric(tau)) %*%
                       as.matrix(lh_helper$VI) %*% z/alf)
  return(list(yhat=yhat))
}
# Compute predictions for map
krig_srrs <- matrix(nrow=length(x_pred),ncol=length(y_pred))
T1 <- Sys.time()
for(i in 1:length(x_pred)){
  for(j in 1:length(y_pred)){
    testcoord <- c(x_pred[i],y_pred[j])
    krig_srrs[i,j] <- pred(train_data,out$par,testcoord,lh_helper)$yhat
  }
}
krig_srrs <- exp(exp(krig_srrs))-4
```

A.3 Integration of GRP data for kriging analysis in R

```
# Construct GRP-transformed response
sigma <- 1
grp <- train_data$GRP
```

```
z <- log(log(train_data$ACTIVITY+4))
aa <- z
bb <- z
# Applying log transformation to the endpoints
zz1 <- log(log(5.999))
zz2 <- log(log(8.001))
aa[grp==1] <- zz2
aa[grp==2] <- zz1
aa[grp==3] <- -1000
bb[grp==1] <- 1000
bb[grp==2] <- zz2
bb[grp==3] <- zz1

pnormdiff <- function(a,b){
  pp <- pnorm(b) - pnorm(a)
  pp[a>0 & b>0] <- pnorm(b[a>0 & b>0]) - pnorm(a[a>0 & b>0])
  return(pp)
}

transform <- function(a,b){
  pp <- pnormdiff(a,b)
  g0 <- (exp(-a^2/2) - exp(-b^2/2))/(sqrt(2*pi)*pnormdiff(a,b))
  u1 <- pnormdiff(a,b)<1e-20 & a>0 & b>0
  u2 <- pnormdiff(a,b)<1e-20 & a<0 & b<0
  g0[u1] <- pmin(a[u1],b[u1])
  g0[u2] <- pmax(a[u2],b[u2])
  return(g0)
}

sigma_helper <- function(sigma){
  return(-sum(log(pnormdiff((aa-z)/sigma,(bb-z)/sigma))))
}

```

```
# Find optimal value of sigma for use in transformation
sigma_helper_optim <- optimize(sigma_helper,lower=0.001,upper=10)

# preserve old z as z0 and construct transformed z
# for use in kriging analysis
z0 <- z
sigma <- sigma_helper_optim$minimum
z <- z0 + sigma*transform((aa-z0)/sigma,(bb-z0)/sigma)
```

A.4 Latent process model for SRRS data in STAN

```
data {
  int<lower=1> num_srrs;
  int<lower=1> num_zipcodes;
  int<lower=1> srrs_counts[num_zipcodes];
  vector[2] coordinates[num_zipcodes];
  vector[num_srrs] log_log_srrs;
}
parameters {
  real<lower=0,upper=50> rho;
  real<lower=0> alpha;
  real<lower=0> sigma;
  vector[num_zipcodes] eta;
}
model {
  int position = 1;
  real sq_sigma = square(sigma);
  vector[num_zipcodes] z;
  {
    matrix[num_zipcodes, num_zipcodes] L_K;

    matrix[num_zipcodes, num_zipcodes] K = cov_exp_quad(coordinates, alpha
   , rho);
```

```
    for (n in 1:num_zipcodes) {
      K[n, n] = K[n, n] + sq_sigma;
    }
    L_K = cholesky_decompose(K);
    z = L_K * eta;
  }

  rho ~ inv_gamma(101, 0.5);
  alpha ~ inv_gamma(1, 1);
  sigma ~ gamma(1, 1);
  eta ~ normal(0, 1);

  for (i in 1:num_zipcodes) {
    segment(log_log_srrs, position, srrs_counts[i]) ~ normal(z[i], sigma);
    position = position + srrs_counts[i];
  }
}
generated quantities {
  vector[num_zipcodes] lp_vals;
  real sq_sigma = square(sigma);
  vector[1] Beta_hat;
  matrix[num_zipcodes,num_zipcodes] K_inv;

  {
    matrix[num_zipcodes, num_zipcodes] K_lp = cov_exp_quad(coordinates,
   alpha, rho);
    matrix[num_zipcodes, num_zipcodes] L_K_lp;
    for (n in 1:num_zipcodes) {
        K_lp[n, n] = K_lp[n, n] + sq_sigma;
    }
    L_K_lp = cholesky_decompose(K_lp);
    lp_vals = L_K_lp * eta;
    K_inv = inverse(K_lp);
```

```
  }

  {
    matrix[num_zipcodes,1] X;
    X = rep_matrix(1.0, num_zipcodes, 1);
    Beta_hat = inverse(X'*K_inv*X)*X'*K_inv*lp_vals;
  }
}
```

A.5 Latent process model for SRRS+GRP data in STAN

```
data {
  int<lower=1> num_srrs;
  int<lower=1> num_zipcodes;
  int<lower=1> srrs_counts[num_zipcodes];
  vector[2] coordinates[num_zipcodes];
  vector[num_srrs] log_log_srrs;
  int<lower=1> grp[num_zipcodes];
}
parameters {
  real<lower=0,upper=50> rho;
  real<lower=0> alpha;
  real<lower=0> sigma;
  vector[num_zipcodes] eta;
  vector[3] b;
}
model {
  int position = 1;
  real sq_sigma = square(sigma);
  vector[num_zipcodes] z;
  {
    matrix[num_zipcodes, num_zipcodes] L_K;

```

```
23    matrix[num_zipcodes, num_zipcodes] K = cov_exp_quad(coordinates, alpha
      , rho);
24    for (n in 1:num_zipcodes) {
25      K[n, n] = K[n, n] + sq_sigma;
26    }
27    L_K = cholesky_decompose(K);
28    z = L_K * eta;
29  }
30
31  rho ~ inv_gamma(101, 0.5);
32  alpha ~ inv_gamma(1, 1);
33  sigma ~ gamma(1, 1);
34  eta ~ normal(0, 1);
35
36  for (i in 1:3) {
37    b[4-i] ~ normal(log(log(2*(i-1)+5)), 0.01);
38  }
39
40  {
41    matrix[num_zipcodes, 3] phi;
42    for (j in 1:num_zipcodes) {
43      for (k in 1:3) {
44        phi[j,k] = -1000 * fabs(z[j] - b[k]);
45      }
46    grp[j] ~ categorical_logit(phi[j]');
47    }
48  }
49
50  for (i in 1:num_zipcodes) {
51    segment(log_log_srrs, position, srrs_counts[i]) ~ normal(z[i], sigma);
52    position = position + srrs_counts[i];
53  }
54}
```

```
generated quantities {
  vector[num_zipcodes] lp_vals;
  vector[1] Beta_hat;
  real sq_sigma = square(sigma);
  matrix[num_zipcodes,num_zipcodes] K_inv;

  {
    matrix[num_zipcodes, num_zipcodes] K_lp = cov_exp_quad(coordinates,
   alpha, rho);
    matrix[num_zipcodes, num_zipcodes] L_K_lp;
    for (n in 1:num_zipcodes) {
        K_lp[n, n] = K_lp[n, n] + sq_sigma;
    }
    L_K_lp = cholesky_decompose(K_lp);
    lp_vals = L_K_lp * eta;
    K_inv = inverse(K_lp);
  }

  {
    matrix[num_zipcodes,1] X;
    X = rep_matrix(1.0, num_zipcodes, 1);
    Beta_hat = inverse(X'*K_inv*X)*X'*K_inv*lp_vals;
  }
}
```

A.6 Algorithm for predictions from latent process models

```
fit_stan <- stan(file = "", data = stan_list, iter = 2000, warmup = 1000,
   chains = 2, cores = 13)
T2 <- Sys.time()

# Formulation of predicted values for latent process
T1 <- Sys.time()
```

```
rownames(coordinates_by_zipcode) <- seq(from=1,to=nrow(
    coordinates_by_zipcode))
lpm <- matrix(0,nrow=length(x_pred),ncol=length(y_pred))
# Pull samples of parameters estimates and generated quantities from the
    MCMC
post_lp_stan <- extract(fit_stan, pars=c('lp_vals'))
post_bh_stan <- extract(fit_stan, pars=c('Beta_hat'))
post_Kinv_stan <- extract(fit_stan, pars=c('K_inv'))
post_alpha_stan <- extract(fit_stan, pars=c('alpha'))
post_rho_stan <- extract(fit_stan, pars=c('rho'))
X <- matrix(rep(1,nrow(train_zips)), ncol=1)
test_data$lpm_srrsgrp_pred <- 0
for (i in 1:2000) {
  beta_hat <- post_bh_stan$Beta_hat[i,]
  alpha <- post_alpha_stan$alpha[i]
  Sigma_inv <- post_Kinv_stan$K_inv[i,,]
  Z <- post_lp_stan$lp_vals[i,]
  rho <- post_rho_stan$rho[i]
  for(j in 1:length(x_pred)){
    for(k in 1:length(y_pred)){
      pred_coordinate <- c(x_pred[j], y_pred[k])
      tau <- matrix(rep(0, nrow(train_zips)), ncol=1)
      for(l in 1:nrow(train_zips)) {
        distance <- sqrt((pred_coordinate[1] - train_zips[l,2])^2 +
                           (pred_coordinate[2] - train_zips[l,1])^2)
        tau[l,] <-  alpha ^ 2 * exp(-(distance^2) / (2*rho^2))
      }
      x_0 <- 1
      current_pred <- t(x_0 - t(X) %*% Sigma_inv %*% tau) %*% beta_hat +
        t(tau) %*% Sigma_inv %*% Z
      current_pred <- as.numeric(current_pred)
      lpm[j,k] <- lpm[j,k] + current_pred
    }
```

```
  }
  for (m in 1:nrow(test_data)) {
      pred_coordinate <- as.numeric(test_data[m, 2:3])
      tau <- matrix(rep(0, nrow(train_zips)), ncol=1)
      for(b in 1:nrow(train_zips)) {
        distance <- sqrt((pred_coordinate[1] - train_zips[b,2])^2 +
                         (pred_coordinate[2] - train_zips[b,1])^2)
        tau[b,] <-  alpha ^ 2 * exp(-(distance^2) / (2*rho^2))
      }
      x_0 <- 1
      current_pred <- t(x_0 - t(X) %*% Sigma_inv %*% tau) %*% beta_hat +
        t(tau) %*% Sigma_inv %*% Z
      current_pred <- as.numeric(current_pred)
      test_data[m,11] <- test_data[m,11] + current_pred
   }
}
T2 <- Sys.time()
print(T2-T1)
lpm <- lpm / 2000
lpm <- exp(exp(lpm))-4
```

References

[1] United States Environmental Protection Agency. *Air and Radiation (6604J)*. data retrieved from National Radon Database, volume 1-5 State/EPA Residential Radon Survey. 1993.

[2] United States Environmental Protection Agency. "National Residential Radon Survey Summary Report". In: *Air and Radiation* 6604J.EPA-402-R-92-0011 (1993), pp. 1–36.

[3] A. Gelman et al. "Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation". In: *Statistical Science* 14.3 (1999), pp. 305–337. DOI: 10.1111/ecog.02881.

[4] D.R. Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: *Ecography* 40.1 (2017), pp. 913–929. DOI: 10.1111/ecog.02881.

[5] F.K.T. Stanley et al. "Radon exposure is rising steadily within the modern North American residential environment, and is increasingly uniform across seasons". In: *Scientific Reports* 9.18472 (2019). DOI: 10.1111/ecog.02881.

[6] K.M. Manson et al. "Radon is associated with clonal hematopoiesis of indeterminate potential in the Women's Health Initiative". In: *Circulation* 141.S1 (2020).

[7] L. Li et al. "Predicting monthly community-Level domestic radon concentrations in the greater Boston area with an ensemble learning model". In: *Environmental Science and Technology* 55 (2021), pp. 7157–7166.

[8] P.N. Price et al. "Bayesian prediction of mean indoor radon concentrations for Minnesota counties". In: *Health Phys* 71.6 (1996), pp. 922–936.

[9] R.L. Smith et al. "Spatial models for radon data analysis". In: *NIH Grant Proposal* (2021), pp. 1–14.

[10] S.H. Kim et al. "The prevalence of stroke according to indoor radon concentration in South Koreans: Nationwide cross section study". In: *Medicine (Baltimore)* 99.4 (2020). DOI: 10.1097/MD.0000000000018859.

[11] W.S. Cleveland. "Robust locally weighted regression and smoothing scatterplots". In: *Journal of the American Statistical Association* 74.368 (1979), pp. 829–836.

[12] N. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Wiley Press, 1993. ISBN: 9780471002550.

[13] W. Field. "Climate change and indoor air quality". In: *Contractor Report* (2010), pp. 1–15.

[14] L. Gundersen and R. Schumann. "Mapping the radon potential of the United States: Examples from the Appalachians". In: *Environment International* 22.1 (1996), pp. 829–837.

[15] F.J. Massey. "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American Statistical Association* 46.253 (1951), pp. 68–78.

[16] R.L. Smith. *Environmental Statistics*. Lecture Notes, University of North Carolina at Chapel Hill. http://rls.sites.oasis.unc.edu/postscript/rs/envnotes.pdf, 2001.

[17] Stan Development Team. *Stan User's Guide*. English. Version 2.31. Stan Development Team. 2022. 468 pp. URL: `https://mc-stan.org/users/documentation/`.

[18] H. Zhang and Y. Wang. "Kriging and cross-validation for massive spatial data". In: *Environmetrics* 21.1 (2010), pp. 290–304. DOI: `10.1002/env.1023`.