ENVIRONMETRICS
Environmetrics 2007; **18**: 481–497
Published online 9 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.816

Effect of housing factors and surficial uranium on the spatial prediction of residential radon in Iowa

Brian J. Smith[1*,†] and R. William Field[2]

[1]*Department of Biostatistics, University of Iowa, Iowa City, U.S.A.*
[2]*Departments of Occupational and Environmental Health & Epidemiology, University of Iowa, Iowa City, U.S.A.*

SUMMARY

Growing epidemiologic evidence suggests that residential radon is an important risk factor for lung cancer. Consequently, public health professionals have expressed interest in characterizing the spatial distribution of radon concentrations in order to identify geographic regions of high exposure. Ambient radon concentrations are a function of geologic features including soil radium content. Indoor radon concentrations can vary based on building characteristics that affect the entry of radon into the building and movement between rooms therein. We present a geostatistical hierarchical Bayesian model for radon that allows for spatial prediction based on geologic data and housing characteristics. Our model is applied to radon data from an epidemiologic study in Iowa that consist of 136 outdoor measurements and 2590 indoor measurements from 614 residential homes. Housing characteristics collected in the Iowa Study are included as predictors in the model. Geologic data in the form of county-average surficial uranium concentrations from the USGS National Uranium Resource Evaluation project are also considered. A 'change of support' approach is implemented to combine the radon measurements, collected at points in space, and the uranium concentrations, averaged over counties, so that point-source concentrations for the latter are available for the analysis. Estimates of the effect of select housing factors on radon are provided along with spatial maps of predicted radon concentrations in Iowa.

KEY WORDS: environmental exposure; hierarchical Bayesian model; Markov chain Monte Carlo simulation; radon gas; spatial statistics; uranium

1. INTRODUCTION

Radon-222 (radon) is a radioactive gas that originates from the radioactive decay of uranium-238 (uranium) and its subsequent decay into radium-226 (radium). Uranium and radium are found in varying concentrations in both rocks and soils. Two of the decay products of radon, polonium-218 and polonium-214, emit alpha particles which are potentially harmful to the bronchial epithelium. The extent to which radon is present in the environment depends on several factors. Ambient radon concentrations

*Correspondence to: Brian J. Smith, Department of Biostatistics, 200 Hawkins Drive, C22 GH, The University of Iowa, Iowa City, Iowa 52242-1009, U.S.A.
†E-mail: brian-j-smith@uiowa.edu

Received 5 August 2006
Accepted 10 August 2006

are a function of geologic features such as soil radium content and soil permeability. Indoor radon concentrations can vary based on building characteristics that affect the entry of radon into the building and movement between rooms therein.

Associations between radon concentrations and geologic factors, as well as housing factors, have been studied previously. Lévesque *et al*. (1997) examined statistical associations between radon measurements in Québec, Canada, and housing factors as well as geologic indicators of high radon potential. Similar associations have been studied using data from New Hampshire, Mid-Atlantic U.S. States, and Minnesota (Price, 1996; Price *et al*., 1996; Apte *et al*., 1999). These later studies also present methods for the prediction of residential radon. One pitfall noted by Price *et al*. (1996) is that the analytic approaches in the aforementioned studies do not account for spatial correlations among radon measurements. Other limitations include the use of covariates measured as regional averages and a focus on prediction at the aggregate (county or township) level. In particular, these studies provide mean estimates of radon that are constant within geographically defined region, whereas the underlying, true spatial distribution of radon is a continuous process that varies both within and across such regions. Preferred statistical methods for radon would allow predictions to vary continuously across individual geographic locations. Steck *et al*. (1999) used the statistical method of kriging to generate contour maps of the estimated mean radon concentrations in Iowa and Minnesota, although the effect of covariates on radon is not considered in their analysis. Moreover, prediction errors from traditional kriging approaches do not account for detector measurement error and uncertainties in estimating all model parameters.

Residential radon has been the subject of numerous studies due to its potential public health importance. Basement screening measurements collected by the United States Environmental Protection Agency and state survey projects provided early summaries of the distribution of residential radon concentrations in the U.S. (Ronca-Battista *et al*., 1988). Subsequent epidemiologic studies examined the lung cancer risk associated with individual subject-level exposures computed from multiple radon measurements taken in each subject's home (Field *et al*., 2000; Darby *et al*., 2005; Krewski *et al*., 2005).

In this paper, we develop a geostatistical model that can be applied to existing data sets in order to characterize the distribution of residential radon. The methodology is illustrated using 136 outdoor radon gas measurements collected in Iowa, radon gas measurements and housing factors from 614 Iowa homes, and county-average surficial uranium concentrations. We assume that radon measurements arise from a continuous spatial process, which allows for prediction at any given geographic location. The effect on radon of select housing factors and surficial uranium concentrations is considered. As an added level of complexity, our (point-source) radon and (regional) uranium data are measured at different geographic resolutions. The primary goals of our approach are to: (1) model the continuous spatial distribution of uranium using county-average data, (2) predict surficial uranium concentrations at the sites of outdoor and home radon measurements, and (3) model radon concentrations as a function of housing factors and predicted surficial uranium concentrations. The ability to combine point-source and regional data is an advantage of the proposed methodology. The model is fit within a Bayesian framework that allows for estimation of the joint posterior distribution of all model parameters, for an accurate account of different sources of variability, and for the inclusion of prior information.

Our paper is organized as follows. In Section 2, we describe our data sets and the Bayesian hierarchical model for residential radon. Spatial correlation, detector measurement error, and systematic mean differences in the radon and uranium measurements are considered in the model development. In addition, we show how the model can be used to predict uranium concentrations at unmeasured

DOI: 10.1002/env

geographic sites. Results of our analyses are contained in Section 3, and a final discussion of our research is presented in Section 4.

2. MATERIALS AND METHODS

In this section, we develop a geostatistical Bayesian model to characterize the distribution of residential radon. The model is motivated by radon gas measurements and housing factors collected in the Iowa Radon Lung Cancer Study (Field *et al*., 2000) and by uranium concentrations from the U.S. Geologic Survey (USGS) National Uranium Resource Evaluation (Duval *et al*., 1989).

2.1. Iowa Radon Lung Cancer Study

In 1993, a population-based, case-control study was initiated in Iowa to estimate the effect of residential radon on lung cancer risk. Although risk estimation was the primary goal of the Iowa Study, the detailed environmental data that were collected provide a unique opportunity to characterize the distribution of outdoor and home radon. Alpha-track detectors were used in the study to obtain year-long radon measurements. In order to predict the spatial distribution of radon in Iowa, we consider the following data in our analysis: (1) 136 radon measurements from outdoor sites across the state and (2) 2590 measurements from 614 disease-free study participant homes. The distribution of participant homes mirrors that of the general Iowa population since control subjects were a population-based sample, whereas the outdoor sites were chosen to be more uniformly distributed across the state, as shown in Figure 1. At least one radon measurement was taken on each floor of the home, resulting in an average of 4.2 measurements per home. Table 1 lists the total number of measurements taken on each floor of the home.

Extensive information about housing construction was collected via field personnel in the Iowa Study. Certain housing factors can affect the accumulation of radon within the home as well as the movement between floors. For instance, older homes may have lower concentrations because they are draftier, or homes with forced air furnaces may have higher concentrations in upper floors because of the increased movement of air between floors. Thus, our model for radon will allow for systematic differences between homes and floors. The housing factors that we selected *a priori* for inclusion in the analysis are year of home construction, number of above-ground floors, type of heating system, drinking water source, presence of a basement and whether it is finished, basement area in square-feet, and presence of a sump pump in the basement. These housing factors were identified as potentially important covariates in preliminary regression analyses. Water source has been considered previously as a predictor of indoor radon concentrations in several studies (Field and Kross, 1998; Apte *et al*., 1999). In our analysis, we allowed for mean differences between private wells and community water supplies. Community water supplies often consist of both ground (well) and surface water sources. Surface water sources generally contain only minimal, if any, waterborne radon, which reduces the overall radon concentrations in community water supplies as compared to ground water supplies. On the other hand, radon concentrations in ground water sources are fairly high. In a survey of waterborne radon concentrations in Iowa private wells, Field and Kross (1998) observed a geometric mean (standard deviation) concentration of 444 pCi/L (81 pCi/L). The well exhibiting the highest waterborne radon concentration slightly exceeded 2000 pCi/L. Nazaroff *et al*. (1987) estimated that 1000 pCi/L of radon in water contributes 0.1 pCi/L to indoor air, so the contribution of waterborne radon to the overall residential radon concentration would be expected to be low in Iowa. A summary of the select housing factors is provided in Table 1. Only the incremental effect of basement area above 750 ft^2 is considered

DOI: 10.1002/env

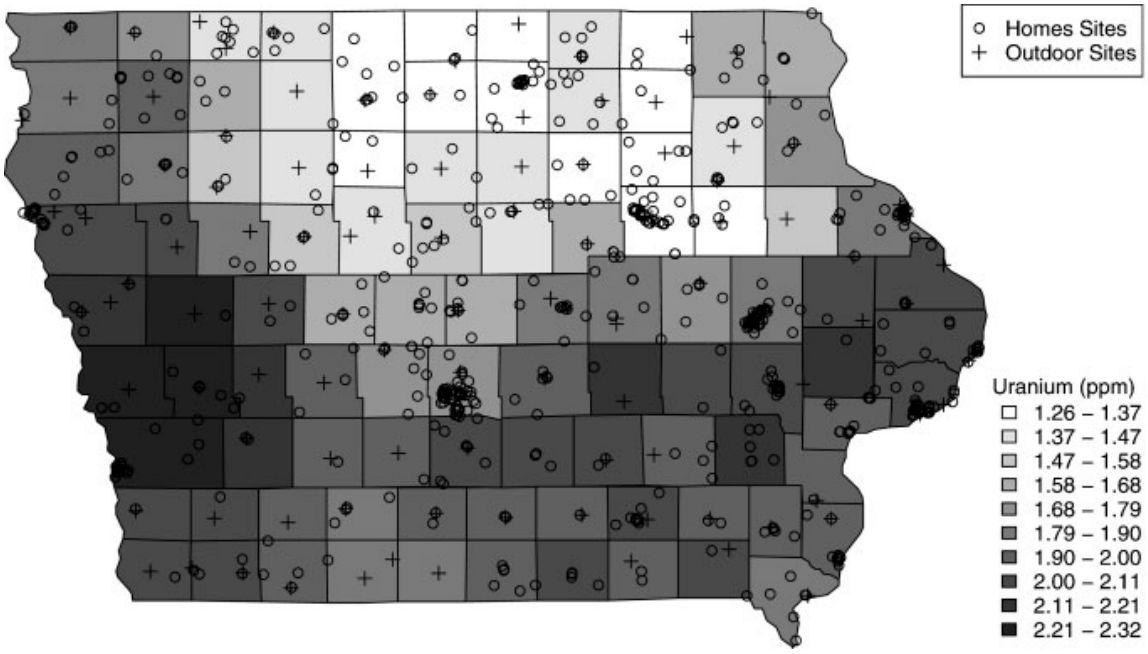

Figure 1. County mean uranium concentrations (ppm) from the NURE data and the geographic locations at which radon was measured in the Iowa Radon Study

in the analysis. Earlier analyses did not show a difference in radon levels between basement sizes below that level.

2.2. *National uranium resource evaluation*

The National Uranium Resource Evaluation (NURE) project was sponsored by the USGS in the 1980s to produce a nationwide map of surficial uranium concentrations (Duval *et al*., 1989). Maps were constructed from radiation spectra measured along airplane flight lines spaced 6–12 miles across the U.S. Since radon originates from uranium, we would like to include the NURE data as a predictor of radon in our model. Several authors have used NURE data to predict radon concentrations. Price *et al*. (1996) studied the association between mean county radon and NURE uranium concentrations in Minnesota. Apte *et al*. (1999) subsequently looked at the ability to predict individual indoor radon concentrations from NURE data averaged over New Hampshire townships. Like the Minnesota and New Hampshire studies, we propose the use of aggregate NURE data. In particular, our NURE data set contains mean uranium concentrations for each of the 99 counties in Iowa. The means are displayed on the map in Figure 1 to illustrate the different spatial resolutions at which the (regional) uranium and (point-source) radon data are available.

2.3. *Radon Model*

We would like to model the measured radon concentrations as a function of housing factors and uranium concentrations. Let $\{s_i : i = 1, \ldots, N\}$ denote the set of geographic locations, or *sites*, at which outdoor and home radon measurements were taken. Ideally, uranium concentrations at the same set of geographic

DOI: 10.1002/env

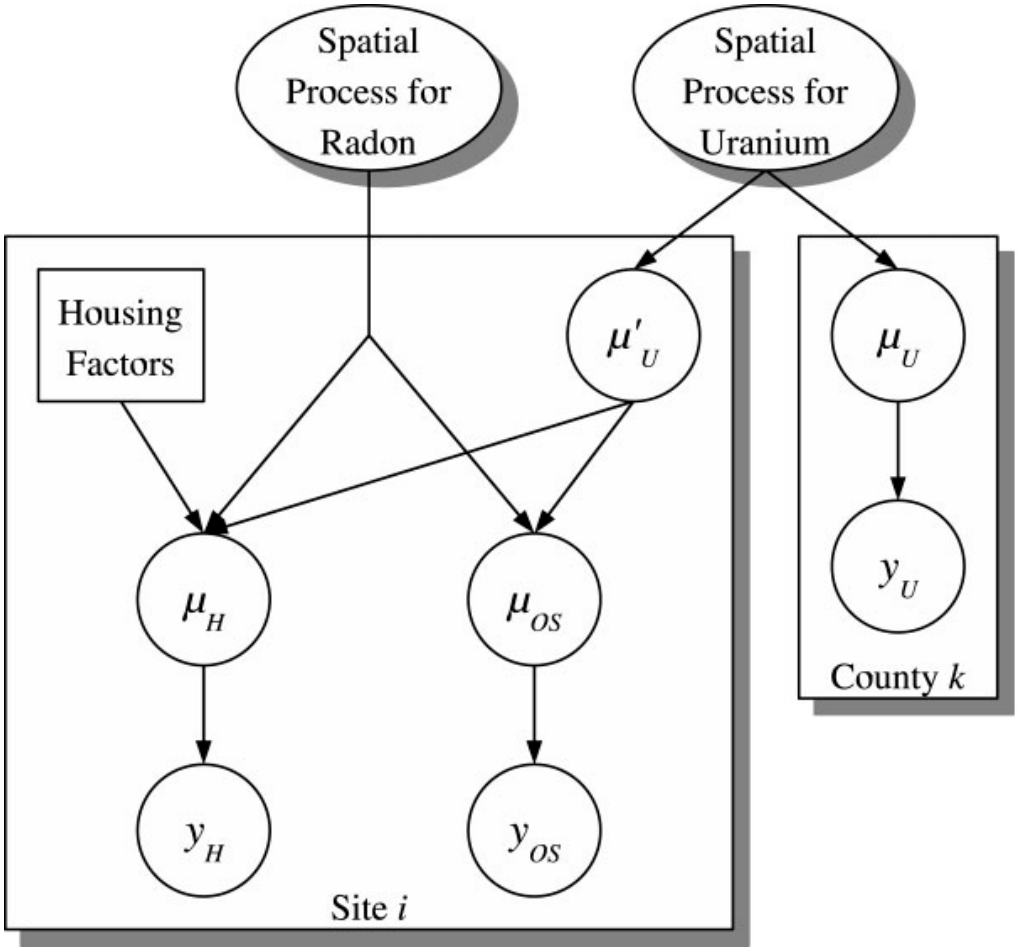

Figure 2. Directed graph diagram of the hierarchical geostatistical model for indoor home radon (H), outdoor radon (OS), and uranium (U) in Iowa

sites would be used as predictors of radon in the analysis. However, we are working with a NURE data set that provides uranium data in the form of county-average, rather than point-source, concentrations. In this and the following section, we develop a statistical model that appropriately relates the county-average NURE data to the point-source radon measurements from the Iowa Study.

Before formulating our model mathematically, we first provide a graphical summary by way of the abbreviated directed graph in Figure 2. The graph outlines the main components of our model and illustrates the interaction between data sources. The larger plate displays a home radon measurement y_{H} and outdoor radon measurement y_{OS} taken at geographic site i; whereas, the smaller plate displays the average uranium measurement y_{U} for county k. Our approach is to model the measurement-error-free home radon concentration μ_{H} at site i as a function of covariates specific to the corresponding home, an underlying continuous spatial process for radon, and a predicted uranium concentration μ'_{U} for the site. The outdoor concentration μ_{OS} at site i is modeled as a function of the same spatial process and predicted uranium concentration, but is not affected by housing factors. A predicted uranium concentration must be used to model radon because uranium measurements are not available for the sites at which radon

DOI: 10.1002/env

Table 1. Posterior mean (95% HPD) estimates of the relative change in residential radon concentrations for the predictors in the joint radon and uranium model

Predictors	N (Mean)	% (SD)	Posterior mean relative Change (95% HPD)	MC error
Log surficial uranium	(0.57)	(0.15)		—
Outside			1.03 (0.61, 1.49)*	0.0038
Home			1.04 (0.85, 1.24)*	0.0001
Location of measurements				
Outside	136	5.0	0.14 (0.10, 0.20)	0.0005
Basement	645	23.7	1.00	—
First floor	1366	50.1	0.53 (0.52, 0.55)	<0.0001
Second floor or higher	579	21.2	0.49 (0.48, 0.51)	<0.0001
Year of home construction				
1835–1924	194	31.6	0.84 (0.70, 0.98)	0.0016
1925–1949	104	16.9	1.00 (0.83, 1.17)	0.0021
1950–1976	316	51.5	1.00	—
Total above-ground floors				
1	316	51.5	1.00	—
2	271	44.1	0.85 (0.73, 0.96)	0.0014
3+	27	4.4	0.72 (0.52, 0.94)	0.0025
Type of heating system				
Forced air furnace	481	78.3	1.00	—
Gravity flow furnace	24	3.9	0.80 (0.58, 1.05)	0.0024
Other	109	17.8	0.87 (0.74, 1.00)	0.0014
Drinking water source				
Community supply	413	67.3	1.00	—
Private well	134	21.8	1.08 (0.92, 1.24)	0.0014
Other	67	10.9	0.93 (0.75, 1.12)	0.0018
Basement				
No	47	23.4	1.00	—
Yes	567	76.6	1.02 (0.80, 1.26)	0.0022
Finished Basement				
No	191	31.1	1.00	—
Yes	423	68.9	0.85 (0.73, 0.98)	0.0015
Basement area in ft^2	(728.4)	(399.2)	1.00 (0.98, 1.03)†	0.0002
Sump pump in basement				
No	483	78.7	1.00	—
Yes	131	21.3	1.21 (1.04, 1.39)	0.0020

* Estimates are for an increase of 1 unit in the predicted log-uranium concentrations.
† For an increase of 100 ft^2 in basement area above 750 ft^2.

measurements were taken. Indeed, uranium measurements are only available in the NURE data set as county averages. As the graph indicates, we will assume that the measurement-error-free uranium average μ_U for county k arises from an underlying continuous spatial process for uranium. The spatial process for uranium allows for prediction at any geographic site and, consequently, the implementation of our joint model for the Iowa radon and NURE uranium data.

We now turn to the mathematical formulation of our model. Although outdoor and home radon measurements are subject to the same detector measurement error in the Iowa Study, the distributions are not expected to be the same because of differences between the two types of environments and are thus modeled accordingly. The proposed model for the outdoor measurements

DOI: 10.1002/env

$\{y_{\mathrm{OS},i} : i = 1, \ldots, n_{\mathrm{OS}}\}$ is

$$\begin{aligned} \ln y_{\mathrm{OS},i} &= \mu_{\mathrm{OS}}(s_i) + \varepsilon_{\mathrm{OS},i} \\ &= \beta_{\mathrm{OS}} + \upsilon_{\mathrm{OS}} \ln \mu'_{\mathrm{U}}(s_i) + z_{\mathrm{R}}(s_i) + \varepsilon_{\mathrm{OS},i} \\ \varepsilon_{\mathrm{OS},i} &\overset{iid}{\sim} N\left(0, \sigma^2_{\mathrm{OS}}\right) \end{aligned} \tag{1}$$

where β_{OS} is an overall mean parameter; υ_{OS} is the mean effect of log-uranium concentrations; $\mu'_{\mathrm{U}}(s_i)$ is a predicted uranium concentration at site s_i; $z_{\mathrm{R}}(s_i)$ is a latent parameter that accounts for spatial correlation among radon concentrations; $\varepsilon_{\mathrm{OS},i}$ is an independent measurement error term; and σ^2_{OS} is the error variance. A detailed description of the distribution for the latent spatial parameters is given later in this section. The prediction of uranium concentrations at radon measurement sites is based on the county-averaged values from the NURE data set, as described in Section 2.4.

The distribution of radon concentrations indoors is more complex than the distribution outdoors. Radon primarily enters homes through the floor closest to the ground and dilutes as it rises up to higher floors, thus setting up a gradient of concentrations with the highest values occurring in basements. Furthermore, we would like to allow for the inclusion of housing factors and surficial uranium concentrations as predictors of indoor radon. Thus, we propose the following model for measurements at the $i = 1, \ldots, n_{\mathrm{H}}$ home sites:

$$\begin{aligned} \ln y_{\mathrm{H},ij} &= \mu_{\mathrm{H}}(s_i) + \varepsilon_{\mathrm{H},ij} \\ &= \boldsymbol{\beta}_{\mathrm{H}}^{\mathrm{T}} \boldsymbol{x}_{ij} + \upsilon_{\mathrm{H}} \ln \mu'_{\mathrm{U}}(s_i) + \gamma_i + z_{\mathrm{R}}(s_i) + \varepsilon_{\mathrm{H},ij} \\ \gamma_i &\overset{iid}{\sim} N\left(0, \sigma^2_{\mathrm{BH}}\right) \\ \varepsilon_{\mathrm{H},ij} &\overset{iid}{\sim} N\left(0, \sigma^2_{\mathrm{WH}}\right) \end{aligned} \tag{2}$$

where $y_{\mathrm{H},ij}$ is the jth measurement at the ith home, $j = 1, \ldots, n_{\mathrm{H},i}$; $\boldsymbol{\beta}_{\mathrm{H}}$ and $\boldsymbol{x}_{ij}$ are vectors of covariates and mean parameters, respectively; υ_{H} is the mean effect of log-uranium on home radon concentrations; $\mu'_{\mathrm{U}}(s_i)$ is a predicted uranium concentration; γ_i is an independent random effect for the home; $z_{\mathrm{R}}(s_i)$ is a latent spatial parameter; $\varepsilon_{\mathrm{H},ij}$ is an independent error term; and σ^2_{BH} and σ^2_{WH} are the between-home and within-home variances, respectively. In our analysis, $\boldsymbol{x}_{ij}$ will include housing factors and indicator variables for the floors on which measurements were taken. An estimate of the between-home variability not explained by covariates in the model is given by σ^2_{BH}. The variability within homes σ^2_{WH} includes variation due to detector measurement error as well as differences between rooms within a given floor. The relative magnitudes of the variance components can then be compared to estimate, for example, the amount of relative variability in radon measurements due to unexplained differences between homes.

Note that the specification of our outdoor (Eq. 1) and home (Eq. 2) radon models assume measurements follow a log-Normal distribution—a distribution commonly used in the analysis of residential radon (Nero *et al.*, 1990). We selected these model formulations based on preliminary frequentist analyses of the Iowa data. In particular, we examined the residuals from a linear regression of log-transformed home radon measurements on the housing factors described in Section 2.1. A

DOI: 10.1002/env

subsequent Kolmogorov–Smirnov test of the residuals provided no statistically significant evidence of non-Normality (p-value = 0.9487). Likewise, there was insufficient evidence to reject that the log-transformed outdoor measurements were Normally distributed (p-value = 0.9456). Given the standard practice of analyzing log-transformed radon measurements and our non-significant test results, we elected to assume log-Normality in our approach. In Section 3.3, the appropriateness of this assumption is further explored with diagnostic checks of the final hierarchical Bayesian model.

Radon originates from soil deposits of uranium which remain relatively stable over time. The determinants of surficial radon cannot be measured directly but are known to induce spatial correlation among observed concentrations. The latent parameters $\mathbf{z}_{\mathrm{R},s} = (z_{\mathrm{R}}(s_1), \ldots, z_{\mathrm{R}}(s_{\mathrm{N}}))^{\mathrm{T}}$ are included in our radon model to account for the correlation associated with the underlying latent spatial process. These parameters are assumed to have the following Gaussian distribution:

$$\mathbf{z}_{\mathrm{R},s} \mid \boldsymbol{\theta}_{\mathrm{R}} \sim N\left(\mathbf{0}, \Sigma_s(\boldsymbol{\theta}_{\mathrm{R}})\right) \tag{3}$$

where spatial patterns in the latent parameters are characterized by the variance matrix $\Sigma_s(\boldsymbol{\theta}_{\mathrm{R}})$. In particular, we define the spatial variance between radon measurements at sites s_i and $s_{i'}$ to be a continuous function $c_{\mathrm{R}}(s_i - s_{i'}; \boldsymbol{\theta}_{\mathrm{R}})$ of the geographic distance between sites and of the parameter vector $\boldsymbol{\theta}_{\mathrm{R}}$; that is $(\Sigma_s(\boldsymbol{\theta}_{\mathrm{R}}))_{ii'} = c_{\mathrm{R}}(s_i - s_{i'}; \boldsymbol{\theta}_{\mathrm{R}})$. Note that the $\mathbf{z}_{\mathrm{R},s}$ parameters represent the residual spatial correlation between log-radon measurements. In other words, they represent the spatial correlation not explained by covariates in the mean structure of the likelihood in Equation (2) for home radon.

For the spatial covariance function, we use an exponential structure of the form

$$c_{\mathrm{R}}\left(s_i - s_{i'}; \boldsymbol{\theta}_{\mathrm{R}}\right) = \tau_{\mathrm{R}}^2 \exp\left\{-\left\|s_i - s_{i'}\right\| / \rho_{\mathrm{R}}\right\}$$

where the geographic distance between sites $\|s_i - s_{i'}\|$ will be computed as the great circle distance in miles (Banerjee, 2003), the covariance function is parameterized by $\boldsymbol{\theta}_{\mathrm{R}} = (\tau_{\mathrm{R}}^2, \rho_{\mathrm{R}})$, and ρ_{R} is restricted to values greater than zero to ensure a valid correlation function. Under the specified structure, the correlation between sites decays exponentially as a function of distance. The rate of exponential decay is controlled by the ρ_{R} parameter. The τ_{R}^2 parameter provides a measure of the variability in the radon measurements that is attributable to residual spatial correlation. Many other correlation functions can be considered (Banerjee *et al.*, 2004). The exponential function is a common choice in geostatistical analyses. It was selected for this research based on our experience with the radon data suggesting no appreciable gain in prediction accuracy when using more flexible structures such as the two-parameter Matérn-class correlation function.

2.4. NURE Uranium Model

In our approach for the NURE county-averaged data, we start by specifying a theoretical model for a point-referenced uranium measurement taken at an arbitrary site s,

$$\begin{aligned} \ln y_{\mathrm{U}}(s) &= \mu_{\mathrm{U}}(s) + \varepsilon_{\mathrm{U}} \\ &= \beta_{\mathrm{U}} + z_{\mathrm{U}}(s) + \varepsilon_{\mathrm{U}} \\ \varepsilon_{\mathrm{U}} &\overset{iid}{\sim} N\left(0, \sigma_{\mathrm{U}}^2\right) \end{aligned} \tag{4}$$

DOI: 10.1002/env

where $y_{U(s)}$ is the uranium measurement, β_U is an overall mean, $z_U(s)$ is a latent parameter that accounts for spatial correlation among uranium measurements, ε_U is an independent measurement error term, and σ_U^2 is the error variance. However, we do not have access to such point-referenced uranium measurements and, instead, must work with county-average measurements. We treat the aggregate data as integrated geometric means over the theoretic point-references measurements define in Equation (4). Consequently, the county measurements $\{y_{U,k} : k = 1, \ldots, K\}$ have the form

$$\begin{aligned} \ln y_{U,k} &= \frac{1}{|B_k|}\int_{B_k} (\beta_U + z_U(s) + \varepsilon_U)\,ds \\ &= \beta_U + z_U(B_k) + \varepsilon_{U,k} \\ \varepsilon_{U,k} &\overset{ind}{\sim} N\left(0, \sigma_U^2/|B_k|\right) \end{aligned} \tag{5}$$

where $z_U(B_k)$ is a latent spatial parameter for county B_k, and $|B_k|$ denotes the total surface area in square-miles.

The 'change of support' method of Gelfand *et al.* (2001) can be applied to the average uranium concentrations for the counties, or geographic *blocks*, $\{B_k : k = 1, \ldots, K\}$ to predict uranium concentrations at the radon measurement sites $\{s_i : i = 1, \ldots, N\}$. A predicted log-uranium concentration at geographic site s_i is defined to be free of measurement error and of the form

$$\ln \mu'_U(s_i) = \beta_U + z'_U(s_i)$$

Thus, prediction at the desired sites involves the vector of latent parameters $\mathbf{z}'_{U,s} = (z'_U(s_1), \ldots, z'_U(s_N))^T$ which, in turn, is related to the vector $\mathbf{z}_{U,B} = (z_U(B_1), \ldots, z_U(B_K))^T$ of parameters for the observed county-average uranium concentrations. Analogous to our approach for radon, we assume a zero-mean, Gaussian process with continuous covariance function for the latent spatial parameters association with log-uranium concentrations. Based on this assumption, the joint distribution for the latent parameters at measured county blocks and prediction sites can be written as

$$\begin{pmatrix} \mathbf{z}_{U,B} \\ \mathbf{z}'_{U,s} \end{pmatrix} \Bigg| \boldsymbol{\theta}_U \sim N\left(\begin{pmatrix} \mathbf{0}_K \\ \mathbf{0}_N \end{pmatrix}, \begin{pmatrix} \Sigma_B(\boldsymbol{\theta}_U) & \Sigma_{B,s}(\boldsymbol{\theta}_U) \\ \Sigma_{B,s}^T(\boldsymbol{\theta}_U) & \Sigma_s(\boldsymbol{\theta}_U) \end{pmatrix}\right) \tag{6}$$

where the covariance matrices have elements given by

$$\begin{aligned} (\Sigma_B(\boldsymbol{\theta}_U))_{kk'} &= \frac{1}{|B_k|\,|B_{k'}|}\int_{B_k}\int_{B_{k'}} c_U\left(s - s'; \boldsymbol{\theta}_U\right) ds'\,ds \\ \left(\Sigma_{B,s}(\boldsymbol{\theta}_U)\right)_{ki} &= \frac{1}{|B_k|}\int_{B_k} c_U(s_i - s; \boldsymbol{\theta}_U)\,ds \\ (\Sigma_s(\boldsymbol{\theta}_U))_{ii'} &= c_U(s_i - s_{i'}; \boldsymbol{\theta}_U) \end{aligned}$$

and c_U models the covariance between uranium concentrations as a function of the geographic distance between sites and parameter vector $\boldsymbol{\theta}_U$. Specification of a continuous covariance function allows for prediction of uranium concentrations at any geographic site. The latent parameters $(\mathbf{z}_{U,B}, \mathbf{z}'_{U,s})$ represent systematic spatial deviations from the overall mean uranium parameter β_U defined in Equation (5). The

DOI: 10.1002/env

Σ matrices characterize the covariance between county blocks and the sites at which predictions are to be made. Uranium concentrations from the NURE data set are assumed to be integrated averages over the counties. The spatial covariance matrices for the county uranium data can be approximated via numerical integration techniques. In particular, integration over B_k is replaced with summation over an independent and uniform grid of fixed geographic sites $\{s_{kj} : j = 1, \ldots, L_k\}$ such that

$$\left(\hat{\Sigma}_{\mathrm{B}}\left(\boldsymbol{\theta}_{\mathrm{U}}\right)\right)_{kk'} = \frac{1}{L_k L_{k'}} \sum_{j=1}^{L_k} \sum_{j'=1}^{L_{k'}} c_{\mathrm{U}}\left(s_{kj} - s_{k'j'}; \boldsymbol{\theta}_{\mathrm{U}}\right)$$

$$\left(\hat{\Sigma}_{B,s}\left(\boldsymbol{\theta}_{\mathrm{U}}\right)\right)_{ki} = \frac{1}{L_k} \sum_{j=1}^{L_k} c_{\mathrm{U}}\left(s_{kj} - s_i; \boldsymbol{\theta}_{\mathrm{U}}\right)$$

Prediction of uranium concentrations at geographic sites is based on the distribution for $\ln \boldsymbol{\mu}'_{\mathrm{U},s} \mid \beta_{\mathrm{U}}, \mathbf{z}_{\mathrm{U},B}, \boldsymbol{\theta}_{\mathrm{U}}$ which is given by

$$N\left(\beta_{\mathrm{U}} \mathbf{I}_N + \hat{\Sigma}_{B,s}^T\left(\boldsymbol{\theta}_{\mathrm{U}}\right) \hat{\Sigma}_B^{-1}\left(\boldsymbol{\theta}_{\mathrm{U}}\right) \mathbf{z}_{\mathrm{U},B},\ \Sigma_s\left(\boldsymbol{\theta}_{\mathrm{U}}\right) - \hat{\Sigma}_{B,s}^T\left(\boldsymbol{\theta}_{\mathrm{U}}\right) \hat{\Sigma}_{\mathrm{B}}^{-1}\left(\boldsymbol{\theta}_{\mathrm{U}}\right) \hat{\Sigma}_{B,s}\left(\boldsymbol{\theta}_{\mathrm{U}}\right)\right).$$

Predicted log-uranium concentrations can then be included as additional parameters in the models for outdoor (Eq. 1) and home (Eq. 2) radon.

As was the case in the radon setting, an exponential structure is used to model the spatial covariance for the latent uranium parameters. Here, the form of the structure is given by

$$c_{\mathrm{U}}\left(s - s'; \boldsymbol{\theta}_{\mathrm{U}}\right) = \tau_{\mathrm{U}}^2 \ \exp\left\{-\left\|s - s'\right\| / \rho_{\mathrm{U}}\right\}$$

where the parameters $\boldsymbol{\theta}_{\mathrm{U}} = (\tau_{\mathrm{U}}^2, \rho_{\mathrm{U}})$ are allowed to differ from those in the covariance structure for the radon model, and $\rho_{\mathrm{U}} > 0$.

2.5. Bayesian methods

A fully Bayesian approach is taken to obtain the joint posterior distribution of all model parameters. The joint posterior is proportional to the product of the likelihood functions associated with the models in Equations (1), (2), and (5); the distributions for the latent spatial parameters in Equations (3) and (6); and a specified joint prior distribution for the model parameters. We employ independent priors for the analysis. Vague, conjugate $N(0, 0.001)$ priors are specified for each of the β mean parameters and $Gamma(0.001, 0.001)$ for the inverse variance parameters $1/\sigma^2$ and $1/\tau^2$, where the $Gamma(\alpha, \beta)$ distribution is parameterized with mean equal to α/β. A *Uniform* $(0, 250)$ prior is specified for each of the ρ_{R} and ρ_{U} spatial decay parameters. The upper bound of 250 was subjectively chosen, based on prior experience with radon data, to allow for a maximum spatial correlation of approximately 25% at the largest possible distance between geographic sites in Iowa.

Markov chain Monte Carlo (MCMC) methods were used to simulate draws from the joint posterior distribution. Three parallel chains with dispersed starting values were generated for the analysis. Fifty-thousand iterations were run for each chain, of which 2500 were discarded as a burn-in sequence and every 5th subsequent iteration retained. Thus, posterior estimates are based on 28 500 autocorrelated

DOI: 10.1002/env

samples from the posterior distribution. The diagnostics of Gelman and Rubin (1992) as well as graphical checks were performed in order to assess convergence of the MCMC draws. The method of Chen and Shao (1999) was used to compute 95% highest probability density intervals (HPD). The HPD intervals are Bayesian probability intervals. Unlike a frequentist 95% confidence interval, they have the interpretation that there is a 95% probability that the parameter is in the interval. As a measure of the precision of the posterior estimates from our MCMC sampler, we present Monte Carlo standard errors based on the spectral density at zero frequency (Heidelberger and Welch, 1981). Convergence diagnostics and posterior summaries were performed with the BOA software (Smith, 2005). The software programs used to fit the Bayesian model are available from the first author (BJS).

3. IOWA RADON ANALYSIS

A Bayesian approach was employed to fit the full model described in Section 2 to the Iowa radon and NURE uranium data. The results of this analysis are presented in the following sections.

3.1. Results

For the analysis, housing factors and indicator variables for floors were included as covariates in the mean structure of Equation (2) for home radon. A summary of the radon predictor variables is provided in Table 1. The number and percentage of measurements at each level of the categorical variables are given in the second and third columns. For the continuous basement area variable, the sample mean and standard deviation are reported. Posterior mean (95% HPD) estimates and associated Monte Carlo errors from the Bayesian analysis appear in the final two columns. The estimates are given in terms of the relative change in radon across the levels of each predictor. For example, according to the posterior mean, outdoor concentrations at a given site tend to be 0.14 times the basement concentrations.

The log-uranium effect on radon deserves special attention since it is based on the predicted uranium concentrations at measured radon sites. In general, our Bayesian analysis provides the full predictive distribution for point-source uranium concentrations across the state. Since this predictive distribution defines a continuous spatial process for uranium, maps of uranium concentrations can be constructed by predicting at a grid of sites over Iowa. The predicted geometric mean surficial uranium concentrations are mapped in Figure 3. The estimated posterior mean and standard deviation for the predictive distribution are given in columns two and three of Table 1. The relative change in radon associated with a one unit change in the predicted log-uranium concentrations is summarized in the fourth column. Since the posterior mean is close to unity (no effect) and the probability interval wide, the association between radon and uranium concentrations appears to be weak.

A summary of the variance parameters is presented in Table 2. Separate variance components are included in the radon model to account for detector measurement error, unexplained differences between homes, and spatial dependencies. The model parameter σ^2_{BH} provides a measure of the between-home variance not explained by covariates in the model. Within-home variance is captured by the σ^2_{WH} parameter and includes both random measurement error and room-to-room variability. Posterior mean estimates indicate that the variance between homes is approximately six times higher than that within homes. Posterior estimates are similar for the within-home variance and the outdoor measurement error variance σ^2_{OS}, which may indicate that the former is primarily a function of random detector

DOI: 10.1002/env

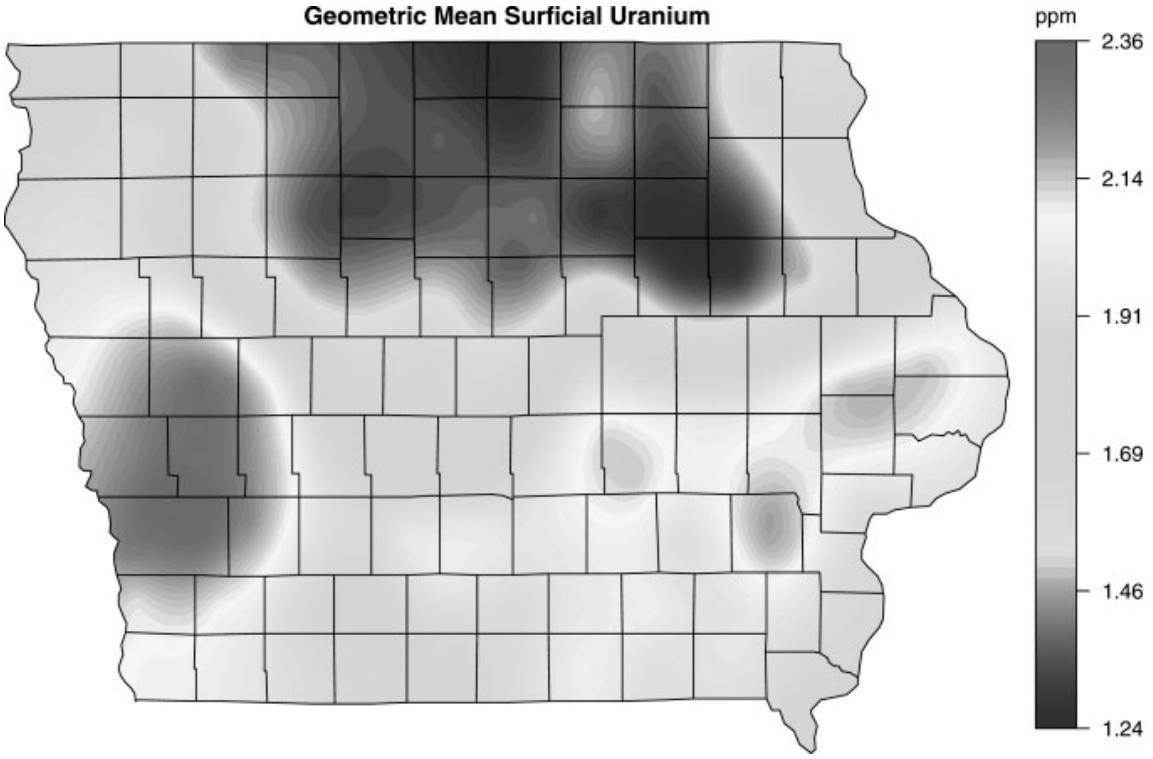

Figure 3. Posterior estimates of the mean surficial uranium concentrations in Iowa

measurement error. The τ_R^2 variance parameter provides a measure of the variability in the data due to the underlying spatial process. Its posterior mean falls in-between the means for the between- and within-home variances. Values of the measurement error parameter τ_U^2 are somewhat larger than those seen for the radon data. This makes intuitive sense because areal methods were used to measure uranium, whereas more direct ground-based detectors were used for radon. Posterior mean estimates of the spatial correlation parameters for the Iowa radon and NURE uranium data are $\rho_R = 94$ and $\rho_U = 226$ miles, respectively. The estimates suggest that radon measurements are more locally correlated than the uranium concentrations. However, the correlation parameter for uranium is apt to be overestimated because averaged data tend to exhibit less spatial variability or, equivalently, higher spatial correlation. The mean range of the spatial correlation (distance at which the correlation equals 5%) is 337.9 miles for radon and 458.3 miles for uranium.

Our model was used to construct the surface map of the geometric mean basement radon concentrations displayed in Figure 4(a). The distributions of outdoor, first floor, and second floor and above radon are also available but not shown here since they differ only by a multiplicative factor that is constant across space. An advantage of a Bayesian analysis is the ability to sample from the joint posterior distribution and to subsequently compute any distributional summary of interest. Radon concentrations above 4 pCi/L are of interest to public health professionals because the U.S. Environmental Protection Agency recommends mitigation steps be taken in homes with measurements in this range. Thus, there is practical interest in identifying areas above the EPA action level of 4 pCi/L. Figure 4(b) provides a map of the estimated probabilities that geometric mean basement concentrations exceed

DOI: 10.1002/env

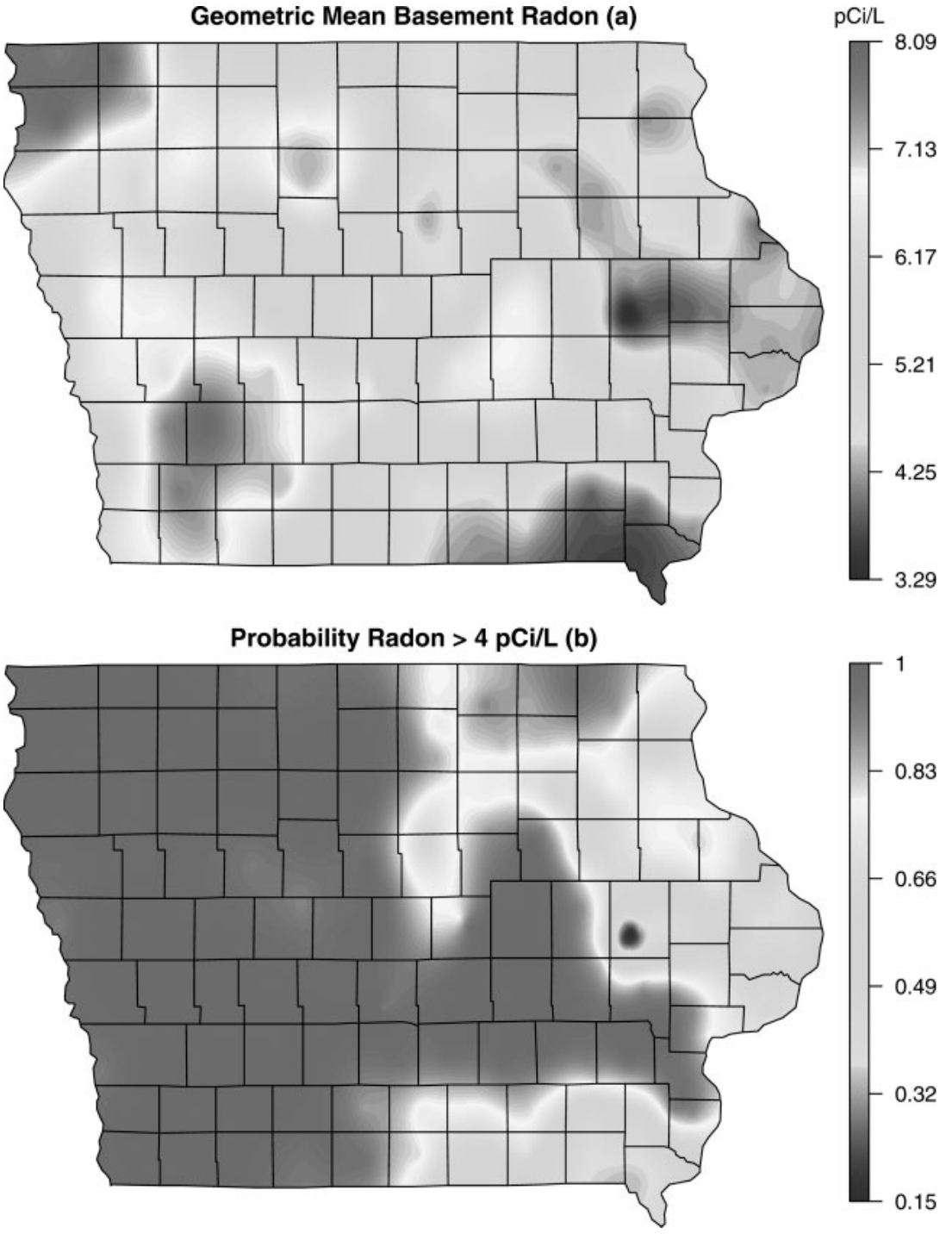

Figure 4. Posterior estimates of the (a) mean basement radon concentrations in Iowa and (b) probability that the mean concentrations exceed 4 pCi/L

DOI: 10.1002/env

Table 2. Posterior estimates for the variance parameters in joint radon and uranium model

Parameter	Mean	Standard deviation	95% HPD	MC error
σ^2_{OS}	0.073	0.015	(0.044, 0.102)	0.0005
σ^2_{BH}	0.463	0.032	(0.402, 0.526)	0.0007
σ^2_{WH}	0.073	0.002	(0.069, 0.077)	<0.0001
τ^2_R	0.126	0.076	(0.031, 0.280)	0.0052
ρ_R	94	66	(13, 228)	6.2
σ^2_U	0.284	0.174	(0.004, 0.601)	0.0039
τ^2_U	0.072	0.015	(0.045, 0.102)	<0.0001
ρ_U	226	20	(185, 250)	0.3

4 pCi/L.

3.2. Variability explained by the radon predictors

In order to study the role played by covariates in the model, we compared different analysis results obtained by varying the set of radon predictor variables in the model. Let HFPU denote the full model described previously in which radon is a function of housing factors and predicted log-uranium concentrations. Consider the following modifications to the predictor variables in the full model:

- exclude both the housing factors and predicted uranium (NFNU),
- exclude the predicted uranium (HFNU), and
- include the observed county-average uranium as a fixed covariate in place of the predicted uranium (HFCU).

The covariates in models NFNU, HFNU, and HFCU are all fixed. Hence, these three new models do not involve the prediction of uranium concentrations. In model HFCU, the county averages from the NURE data set are inserted directly into the mean structures for outdoor and home radon. Consequently, HFCU links the same uranium value to all radon measurements from a given county and ignores the measurement error inherent in the uranium measurements. This is how the NURE data might be treated in traditional analyses that do not reconcile the different spatial resolutions of the aggregate uranium and point-source radon data.

Table 3 provides the posterior mean between-home variance σ^2_{BH} and spatial variance τ^2_R for each of the four models. Recall that σ^2_{BH} represents systematic between-home differences that are not accounted for by covariates in the model. Likewise, τ^2_R measures the residual variability explained by the spatial correlation structure. In comparing models NFNU and HFNU, it can be seen that the addition of housing factor results in the largest decrease in the variance parameters. Thus, these predictors explain some of the between-home variability and spatial variability in the observed radon measurements. More interesting is the decrease in spatial variance in moving from model HFCU to HFPU. The results indicate that the use of predicted, point-referenced uranium concentrations, rather than the county-average concentrations, leads to a decrease in the unexplained spatial variability.

3.3. Model assessment

In order to gage the strengths and limitations of our model, various posterior predictive checks were performed. First, every 5th posterior predictive percentile from the 5th to the 95th was computed to assess

DOI: 10.1002/env

Table 3. Comparison of the posterior mean (95% HPD) between-home variance and spatial variance for radon models that include covariates for housing factors (HFNU), housing factors and observed county-average uranium concentrations (HFCU), housing factors and predicted uranium (HFPU), and neither (NFNU)

Model	σ^2_{BH}	τ^2_R
NFNU	0.490 (0.429, 0.558)	0.146 (0.031, 0.323)
HFNU	0.464 (0.403, 0.527)	0.123 (0.028, 0.282)
HFCU	0.463 (0.403, 0.526)	0.155 (0.032, 0.355)
HFPU	0.463 (0.402, 0.526)	0.126 (0.031, 0.280)

Table 4. Observed proportion of home radon measurements >4 pCi/L (posterior predictive 95% HPD intervals) by geographic regions in Iowa

	West	Central	East
North	0.46 (0.42, 0.48)	0.32 (0.28, 0.34)	0.29 (0.27, 0.34)
Central	0.45 (0.36, 0.46)	0.50 (0.47, 0.53)	0.28 (0.25, 0.30)
South	0.48 (0.42, 0.53)	0.26 (0.21, 0.29)	0.20 (0.17, 0.24)

the appropriateness of the assumed log-Normal distribution for radon measurements. All of the 95% HPD intervals for the posterior percentiles, except for those at the 5th and 95th percentiles, contained the observed data (not shown). Although this may suggest a lack-of-fit in the tails of the distribution, the extreme percentiles in the Iowa data are not of practical importance in the determination of risk because they represents either very low concentrations at which the differences are relatively small or high concentrations from basements where individuals spend very little time. Furthermore, lack-of-fit in the tails is more likely to be a problem in analyzing Iowa data where state average radon concentrations are the highest in the nation.

In addition to the analysis of percentiles, we examined posterior predictive p-values (Gelman *et al.*, 1996) based on the goodness-of-fit discrepancy measure

$$X^2(y;\theta) = \sum \frac{(y_i - E(y_i \mid \theta))^2}{\mathrm{Var}(y_i \mid \theta)}$$

where θ represents the model parameters. Posterior predictive p-values that deviate from a value of 0.5 indicate differences between radon measurements that were observed and those that would be predicted by the model. p-values of 0.51 and 0.55 were obtained for outdoor and home radon, respectively. We also examined the consistency of the observed data with tail probabilities from the model by focusing on the predictive probabilities that home radon measurements exceed the EPA action level of 4 pCi/L. Tail probabilities were predicted within each block of a 3×3 grid of equally sized geographic regions in Iowa. The associated 95% HPD intervals and observed proportion of measurements above 4 pCi/L are displayed in Table 4. In all cases, the data are consistent with the predicted probabilities from the model.

Finally, sensitivity to the specified priors was assessed. Particular attention was paid to the priors for the spatial correlation parameters. Posterior estimates for the parameters appearing in the radon

DOI: 10.1002/env

models (1) and (2) did not change noticeably when we reran the MCMC sampler with $Uniform(0, 150)$ or with $Uniform(0, 500)$ priors instead of the $Uniform(0, 250)$ described in Section 2.5. An additional run of the sampler was performed with vaguer prior specifications for the mean and variance parameters. This was done to verify that the original $N(0, 0.001)$ and $Gamma(0.001, 0.001)$ priors had minimal effect on the posterior estimates, as intended. The additional run produced indistinguishable results.

4. DISCUSSION

Epidemiologic research plays a major role in uncovering causal links between environmental exposures and human disease. Pew Environmetal Health Commission (2000) issued a plea for the creation of a coordinated nationwide health tracking network that would monitor and track potential links between environmental exposures and adverse health outcomes. A health tracking system would endeavor to link health and exposure data on an on-going basis; thus enabling public health practitioners to evaluate the spatial and temporal relations between environmental factors and adverse health outcomes. However, linkages between toxicants and diseases are impeded by the lack of geographic information on the distribution of both man-made and naturally occurring agents. When environmental data are available, they are generally collected at a county or regional level that often poorly reflect the spatial distribution of the toxicant. Human-made environments (e.g., building construction) add to the complexity of modeling exposure as does the spatio-temporal mobility of an individual within the natural or built environment.

The analytic methodology described in this paper utilizes the continuous nature of the spatial data without the limitation of relying on information at the county level or some other artificially defined geographic boundary. In our radon-related example, the unified Bayesian framework allows for the joint modeling of radon and uranium data measured at different spatial resolutions. This approach provides additional information over traditional kriging techniques by allowing the examination of the effects of multiple covariates. Moreover, the methods appropriately account for errors in prediction resulting from both uncertainties in estimating model parameters and due to known sources of variability, including measurement error, unmeasured differences between homes, and spatial dependencies.

As presented, spatial predictive modeling can also be used to assess the degree of random uncertainty in the estimation of individual residential radon concentrations and in turn exposure. It logically follows that improving the reliability of predictive models contributes to a reduction in uncertainty of an estimated exposure. The use of multiple data sources of inter-related information enhances investigators' ability to improve the predictive ability of a model, which is often needed when limited information is available for a parameter. For example, in relation to the example above, maps incorporating the effects of covariates could be used to map radon exposure in various parts of the world, which could have direct application such as contributing to the World Health Organization's effort to estimate the world burden of radon-related lung disease (WHO, 2005). Predictive models can also be used to prioritize areas that have higher levels of exposure so that educational and mitigation efforts can be targeted. In addition, a global pooling of residential radon studies is currently underway that could be enhanced by the use of predictive models incorporating various covariates unique to each study. In summary, this paper suggests that use of covariates at the aggregate level may serve as a useful tool to improve predictive models by accounting for errors such as uncertainties in estimating model parameters and random measurement error.

DOI: 10.1002/env

ACKNOWLEDGEMENTS

This publication was made possible in part by grant numbers R01 ES05653 and P30 ES05605 from the National Institute of Environmental Health Sciences, NIH and grant number R01 CA85942 from the National Cancer Institute, NIH.

REFERENCES

Apte MG, Price PN, Nero AV, Revzan KL. 1999. Predicting New Hampshire indoor radon concentrations from geologic information and other covariates. *Environmental Geology* **37**: 181–194.

Banerjee S. 2003. Essential geodesics for the spatial statistician. Technical Report 009, University of Minnesota Division of Biostatistics.

Banerjee S, Carlin BP, Gelfan AE. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC: New York.

Chen M-H, Shao Q-M. 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **8**(1): 69–92.

Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Deo H, Falk R, Forastiere F, Hakma M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, Mäkeläinen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Rosario AS, Tirmarche M, Tomāsek L, Whitley E, Wichmann H-E, Doll R. 2005. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *British Medical Journal* **330**(7485): 223.

Duval JS, Jones WJ, Riggle FR, Pitkin JA. 1989. Equivalent uranium map of the conterminous United States. Technical report, United States Geological Survey, Denver, CO.

Field RW, Kross BC. 1998. Iowa survey of waterborne 222Rn concentrations in private wells. *Health Physics* **74**: 249–252.

Field RW, Steck DJ, Smith BJ, Brus CP, Fisher EL, Neuberger JS, Platz CE, Robinson RA, Woolson RF, Lynch CF. 2000. Residential radon gas exposure and lung cancer: the Iowa radon lung cancer study. *American Journal of Epidemiology* **151**: 1091–1102.

Gelfand AE, Zhu L, Carlin BP. 2001. On the change of support problem for spatio-temporal data. *Biostatistics* **2**: 31–45.

Gelman A, Meng X-L, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**: 733–807.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**: 457–511.

Heidelberger P, Welch PD. 1981. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* **24**(4): 233–245.

Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Letourneau EG, Lynch CF, Lyon JI, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB. 2005. Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. *Epidemiology* **16**(2): 137–145.

Lévesque B, Gauvin D, McGregor RG, Martel R, Gingras S, Dontigny A, Walker W, Lajoie P, Létourneau E. 1997. Radon in residences: influences of geological and housing characteristics. *Health Physics* **72**: 907–914.

Nazaroff WW, Doyle SM, Nero AV, Sextro RG. 1987. Potable water as a source of airborne 222Rn in U.S. dwellings: a review and assessment. *Health Physics* **52**: 281–295.

Nero AV, Gadgil AJ, Nazaroff VOW, Revzan KL. 1990. Indoor radon and decay products: concentrations, causes and control strategies. Technical Report DOE/ER-0480P, Department of Energy, Washington, DC.

Pew Environmetal Health Commission. 2000. America's environmental health gap: why the country needs a nationwide health tracking network. Technical report, Pew Environmental Health Commission, Baltimore, MD.

Price PN. 1996. Predictions and maps of county mean indoor radon concentrations in the mid-atlantic states. *Health Physics* **72**: 893–906.

Price PN, Nero AV, Gelman A. 1996. Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Health Physics* **71**: 922–936.

Ronca-Battista M, Moon M, Bergsten J, White SB, Holt N, Alexander B. 1988. Radon-222 concentrations in the United States—results of sample surveys in five states. *Radiation Protection Dosimetry* **24**: 307–312.

Smith BJ. 2005. Bayesian Output Analysis Program (BOA), version 1.1.5. The University of Iowa. http://www.public-health.uiowa.edu/boa [Accessed 5 September 2006].

Steck DJ, Field RW, Lynch CF. 1999. Exposure to atmospheric radon. *Environmental Health Perspectives* **107**: 123–127.

WHO. 2005. WHO launches project to minimize risks of radon. http://www.who.int/mediacentre/news/notes/2005/np15/en/index.html.