

Spatial prediction of naturally occurring gamma radiation in Great Britain

P. Chernyavskiy^a, G.M. Kendall^b, R. Wakeford^c, M.P. Little^{d,*}

^aRadiation Epidemiology Branch, National Cancer Institute, DHHS, NIH, Division of Cancer Epidemiology and Genetics, Bethesda, Maryland 20892-9778, USA, email pavel.chernyavskiy@nih.gov

^bCancer Epidemiology Unit, University of Oxford, Richard Doll Building, Old Road Campus, Headington, Oxford, OX3 7LF, UK, email gerald.kendall@ceu.ox.ac.uk

^cCentre for Occupational and Environmental Health, Institute of Population Health, The University of Manchester, Ellen Wilkinson Building, Oxford Road, Manchester, M13 9PL, UK, email richard.wakeford@gmail.com

^dRadiation Epidemiology Branch, National Cancer Institute, DHHS, NIH, Division of Cancer Epidemiology and Genetics, Bethesda, Maryland 20892-9778, USA, email mark.little@nih.gov

***Corresponding author:** Mark P. Little, National Cancer Institute, Room 7E546, 9609 Medical Center Drive, Rockville, MD 20892-9778, USA, tel +1 240 276 7375, mobile +1 301 875 3413, fax +1 240 276 7840, email mark.little@nih.gov

Abstract

Gamma radiation from natural sources is an important component of background radiation, and correlates with childhood leukaemia risk in Great Britain. The geographic variation of indoor gamma radiation dose-rates in Great Britain is explored using various geo-statistical methods. A multi-resolution Gaussian-process model using radial basis functions with 2, 4, or 8 components, is fitted via maximum likelihood, and a non-spatial model is also used, fitted by ordinary least squares. Because of the dataset size ($N=10,199$), four other parametric spatial models are fitted by variogram-fitting. A randomly selected 70:30 split is used for fitting:validation. The models are evaluated based on their predictive performance as measured by Mean Absolute Error, Mean Squared Error, as well as Pearson correlation and rank-correlation between predicted and actual dose-rates. Each of the four parametric models (Matérn, Gaussian, Bessel, Spherical) fitted the empirical variogram well, and yielded similar predictions at >50 km separation, although with more substantial differences in predicted variograms at <50 km. The multi-resolution Gaussian-process model with 8 components had the best predictive accuracy among the models considered. The Spherical, Bessel, Matérn, Gaussian and ordinary least squares models had progressively worse predictive performance, the ordinary least squares model being particularly poor in this respect.

Keywords: natural gamma radiation; kriging; variogram; multi-resolution Gaussian process; childhood cancer; ordinary least squares

Highlights

- Background γ radiation correlates with childhood cancer risk in Great Britain (GB).
- A multiresolution Gaussian process (MRGP) + 5 other models are fitted to GB γ data.
- The 8-component MRGP model has the best predictive accuracy of the 6 models fitted.
- The Spherical, Bessel, Matérn, Gaussian, least squares model are progressively worse.

Introduction

A causal link between high-level exposure to ionising radiation and cancer risk has long been established (e.g., (United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), 2008)), but the degree of risk produced by low-level exposure remains equivocal (Doss et al., 2014; Little et al., 2009; Tubiana et al., 2009). There have been many studies investigating correlations between exposure to low-dose-rate natural background gamma radiation and childhood cancer (e.g. (Evrard et al., 2006; Richardson et al., 1995; United Kingdom Childhood Cancer Study Investigators, 2002)), most of them reporting a null result. However, the lack of statistical power in these studies (Little et al., 2010) suggests that reliable conclusions cannot be made. A recent and more adequately powered case-control study in Great Britain (GB: England, Scotland and Wales) by Kendall *et al.* (Kendall et al., 2013) provided evidence of an association between the cumulative dose from indoor background gamma radiation and the risk of childhood leukaemia. In contrast, this study did not report any significant correlations of other childhood cancers with exposure to background gamma-rays or of any childhood cancer type with radon exposure.

A previous paper explored various strategies for predicting indoor natural background gamma-ray dose-rates from a dataset of measurements (Kendall et al., 2016). A geo-statistical model was developed, which assumed an underlying geologically-determined spatial variation superimposed on a Gaussian stochastic process with Matérn correlation structure that models the observed tendency of dose-rates in neighbouring houses to correlate. An empirical approach was also developed using linear regression models to fit predictor variables based on a number of dose-rate interpolation measures, calculated from averages over geologically- or administratively-defined areas, or distance-weighted averages of measurements at nearest-neighbour points. The

mean square error of the linear-regression model was lower than that of the Gaussian-Matérn model (Kendall et al., 2016).

As a follow-up to this study, we shall explore the large-scale geographic variation of indoor gamma radiation dose-rates in GB. We shall evaluate several geo-statistical methods of modelling indoor gamma radiation dose-rates. Because the number of observations is very large ($N=10,199$), a straightforward maximum likelihood approach is not currently feasible. As an alternative, we employed a variogram-based approach to spatial prediction, and a Gaussian process model which approximated spatial correlation among gamma-ray dose-rates using radial basis functions. Models were estimated using a randomly selected subset of the complete dataset and then validated using a holdout sample. The models were evaluated based on their predictive performance as measured by Mean Absolute Error, Mean Squared Error, as well as Pearson correlation and rank-correlation between predicted and actual dose-rates. We used the model with the best predictive performance to interpolate gamma radiation dose-rates on a fine grid for the entire country.

The analysis may be useful in a number of ways. For example, interpolated dose-rates on a fine grid will improve the geographical mapping of gamma-ray exposures. Further, we judge that this analysis may also be of public health interest in that dose-rate predictions will enable estimation of cumulative gamma-ray dose, as well as the subsequent risk for childhood leukaemia and other cancers in GB through improved epidemiological analyses. Finally, our statistical modelling efforts will shed light on the best modelling approach to high-dimensional environmental exposure data.

Data

Measurement data used in this study consisted of 2,283 indoor gamma radiation dose-rates from the National Survey of UK homes (Wrixon et al., 1988), and 7,916 indoor gamma radiation dose-

rates from the UK Childhood Cancer Study (United Kingdom Childhood Cancer Study Investigators, 2002). Kendall *et al.* (Kendall et al., 2016) discuss several factors that influence gamma-ray dose-rates inside a home. Outdoor dose-rates depend on cosmic-ray dose-rates, and levels of radionuclides present in the vicinity of the measurement. Indoor dose-rates tend to be a little higher than those measured outdoors, and are determined in part by the shielding (thickness and density) of the exterior walls and roof, as well as by the presence of naturally occurring radionuclides in the walls, floors, and ceilings. Because of the role of radionuclides in the environment in determining the gamma-ray dose-rates in a building, in a previous paper (Kendall et al., 2016) we considered two geological classifications, LEX-ROCK Bedrock codes and 50k-BEDSUP combined simplified bedrock and superficial codes. LEX-ROCK are codes for the underlying geology as defined by the British Geological Survey (BGS) (Smith, 2013), and provide the name of each rock unit or deposit (via its LEX or Lexicon code) and composition (via its ROCK or lithology code). The 50k-BEDSUP combined simplified bedrock and superficial codes were developed by the BGS in the context of an investigation into the variation of indoor radon concentrations across the UK (Miles and Appleton, 2005). For detailed geostatistical modelling both these coding schemes were found to be intractable and a simplified set of bedrock classes, a 16-level classification, was employed (Kendall et al., 2016). For the present work we also explored a simplified set of surface geological classes, broadly analogous to the bedrock classes (a 23-level variable), and also a combined bedrock and surface classification (a 23-level variable) that were still further reduced. Further details are given in Appendix A. The surface geology should broadly correlate with outdoor gamma-ray levels and possibly with locally-used building materials, which will also substantially modify indoor dose-rate. Figure 1 presents a map of measured indoor and

outdoor gamma-ray dose-rates in GB, taken from Kendall *et al.* (Kendall et al., 2016) (see also Figure A1 for a map of GB).

Statistical Methods

Since prediction was the primary analytic goal of this study, we considered two broad classes of spatial interpolation methods, namely variogram-based kriging (e.g., (Cressie, 1993)), and multi-resolution Gaussian processes (Nychka et al., 2014). For the sake of comparison, we also used a non-spatial ordinary least squares (OLS) model. All models were fitted using an “estimation” sample consisting of a random selection of approximately 70% of total observations, and validated using a holdout sample consisting of the remaining approximately 30% of total observations. Models were evaluated based on Mean Absolute Error (MAE), Mean Squared Error (MSE), as well as Pearson and Spearman (rank-order) correlations between predicted dose-rates and holdout sample measured dose-rates.

Variogram-based kriging

The variogram is a fundamental geo-statistical tool used to describe spatial variability in point-referenced data. The variogram is expressed as a function of the spatial separation vector $h = s - s'$, and is defined as:

$$\begin{aligned}\gamma(h) &= \frac{1}{2} \text{var}[Y(s) - Y(s')] = \text{var}[Y(s)] - \text{cov}[Y(s), Y(s')] \\ &= \tau^2 1_{h \neq 0} + \sigma^2 [1 - \rho(h)] = C(0) - C(h)\end{aligned}\tag{1}$$

where $C(0) = \tau^2 + \sigma^2$ is the so-called sill of the variogram, $C(h)$ is the value of the covariance function evaluated at a specific distance, and $\rho(h)$ is the spatial correlation at lag $h = s - s'$. Further details on the derivation of expression (1) are given in Appendix B.

Variogram models are generally parameterised in terms of the error variance term τ^2 (also called the nugget variance), the spatial variance σ^2 (also called the partial sill), and the correlation

function $\rho(h|\phi)$, which depends on some parameters ϕ . Parametric variogram models have the form:

$$\gamma(h|\phi) = \tau^2 1_{h \neq 0} + \sigma^2 (1 - \rho(h|\phi)) \quad (2)$$

Parametric variogram models that we considered in this study included:

- 1) the Matérn model (Matérn, 1960);;
- 2) the Gaussian model;
- 3) the Bessel model; and
- 4) the Spherical model

Further details of these four parametric models are given in Appendix B. Variogram models of these parametric forms were fitted to the empirical variogram constructed from pairs of observations $Y(s)$ and $Y(s')$, and denoted by $\hat{\gamma}(h)$. We used the robust Cressie empirical variogram estimator (Cressie, 1993) due to its reduced bias and resistance to outlying observations, compared to the standard empirical variogram. Detailed specifications of the four variogram models, and the Cressie empirical variogram estimator can be found in Appendix B and elsewhere (Banerjee et al., 2015). Estimates for τ^2 , σ^2 and ϕ are computed using weighted least squares, with weights $\frac{|N(h)|}{h^2}$. [$N(h)$ is the set of pairs of points (s, s') with spatial differences $h = \|s - s'\|$.]

Using parameter estimates from the variogram model in combination with estimates for spatial covariates (i.e., $\hat{\beta}$) we can compute the predicted dose-rate value at some unobserved location, s_0 , using the universal kriging equation (e.g., Banerjee *et al.* (Banerjee et al., 2015), Diggle and Ribeiro (Diggle and Ribeiro, 2007) p.136):

$$\hat{Y}(s_0) = X(s_0)^T \hat{\beta} + \hat{\gamma}^T \hat{\Sigma}^{-1} (Y - X(s)^T \hat{\beta}) \quad (3)$$

where $\hat{\gamma}^T = (\hat{\sigma}^2 \rho(\|s_0 - s_1\| \hat{\phi}), \dots, \hat{\sigma}^2 \rho(\|s_0 - s_N\| \hat{\phi}))$, and $\hat{\Sigma}$ is the estimated variance-covariance matrix of Y . The estimated variance-covariance matrix is defined as $\hat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi}) + \hat{\tau}^2 I_N$, where each element of the matrix $H(\cdot)$ is of the form: $H(\hat{\phi})_{ij} = \rho(\|s_i - s_j\| \hat{\phi})$. All computation was carried out using the `gstat` package (Pebesma and Graeler, 2015) in R (R Project version 3.2.2, 2015).

Multi-resolution Gaussian Process

Although it is a theoretically attractive alternative to variogram-based prediction, the implementation of maximum-likelihood kriging equations is known to be intractable for datasets with sample sizes on the order of $n \approx 10,000$ (Cressie and Johannesson, 2008). Spatial prediction based on a multi-resolution Gaussian-process model (Nychka et al., 2014) that approximates the true likelihood of the process provides one method that addresses this computational challenge. In this section, we give an overview of this approach in the context of our study; a more detailed description is found in the paper of Nychka *et al.* (Nychka et al., 2014).

We shall assume that the observed dose-rate at location s is given by:

$$Y(s) = X(s)^T \beta + g(s) + \varepsilon \quad (4)$$

where $X(s)^T \beta$ is the mean response, $g(s)$ is a spatial Gaussian process, and ε is a vector of independent zero-mean measurement errors. The unknown spatial process $g(s)$ is approximated by a sum of L independent processes:

$$g(s) = \sum_{l=1}^L g_l(s) \quad (5)$$

where $\text{var}[g_l(s)] = \sigma^2 \alpha_l$, and where σ^2 is a scaling term for the variance of the spatial process

$g(s)$, and α_l is a weighting term, selected so that $\sum_{l=1}^L \alpha_l = 1$. Each independent component $g_l(s)$

is expressed in turn as a sum of (non-stochastic) radial basis functions $\phi_j^l(s)$ weighted by a vector

of random coefficients, c_j^l , that is:

$$g_l(s) = \sum_{j=1}^{m(l)} c_j^l \phi_j^l(s) \quad (6)$$

where the coefficients c_j^l are multivariate normal, with mean zero and covariance

$\text{cov}[c_j^l, c_m^l] = \sigma^2 (P^l)_{jm}$, and cross covariance 0, i.e., $\text{cov}[c_j^l, c_m^p] = 0, l \neq p$. We assumed that the

$\phi_j^l(s)$ were the standard symmetric Wendland radial basis functions (Wendland, 1995).

Putting expressions (4)-(6) together, we therefore modeled $Y(s)$ approximately as:

$$Y(s) = X^T(s) \beta + \sum_{l=1}^L \sum_{j=1}^{m(l)} c_j^l \phi_j^l(s) + \varepsilon \quad (7)$$

where spatial covariance in $Y(s)$ is now expressed in relation to the covariance of the radial basis

coefficients c_j^l . As we show in Appendix B, the profile log-likelihood of $Y(s)$ (after

maximization with respect to σ) becomes:

$$l(Y(s)|\beta, \lambda) = C - N \ln \left[\left(Y(s) - X(s)^T \beta \right)^T \Sigma_\lambda^{-1} \left(Y(s) - X(s)^T \beta \right) \right] - \frac{1}{2} \ln |\Sigma_\lambda| \quad (8)$$

for a constant $C = -\frac{N}{2} [\ln(2\pi) - 2\ln(N) + 1]$. The approximate log-likelihood is maximised

numerically with respect to β and λ to obtain their respective maximum-likelihood estimates,

$\hat{\beta}$ and $\hat{\lambda}$.

One derives the predicted dose-rate at an unobserved location, s_0 , by considering $(Y(s_0), (Y(s_i))_{i=1}^N)$, which is a multivariate Gaussian process with mean $(X(s_0)^T \beta, (X(s_i)^T \beta)_{i=1}^N)$

and variance $\begin{bmatrix} \sigma^2 + \tau^2 & \sigma^2 r^T \\ \sigma^2 r & \sigma^2 V \end{bmatrix}$ where the vector r is given by

$r = (r_n)_{n=1}^N = \sum_{l=1}^L \sum_{j,p=1}^{m(l)} \phi_j^l(s_n) \phi_p^l(s_0) (P^l)_{jp}$. Then it is easily shown (e.g., p.136 in (Diggle and Ribeiro,

2007)) that the conditional distribution of $Y(s_0)$ given $(Y(s_n))_{n=1}^N$ is also multivariate Gaussian with mean:

$$\hat{Y}(s_0) = X(s_0)^T \hat{\beta} + r^T V^{-1} \left[Y(s_n) - X(s_n)^T \hat{\beta} \right]_{n=1}^N \quad (9)$$

All computation was carried out using the `LatticeKrig` package (Nychka, 2014) in R (R Project version 3.2.2, 2015).

Non-spatial Ordinary Least Squares

For the sake of completeness, in addition to models that explicitly accounted for the presence of spatial correlation among observed dose-rates, we considered the standard linear model that treats observations as independently normally distributed, with common variance. The observed dose-rate at location s is assumed to be:

$$Y(s) = X(s)^T \beta + \varepsilon \quad (10)$$

where $X^T(s)\beta$ is the mean response, and ε is the independent identically distributed error term with mean 0 and variance τ^2 . Parameter estimates, $\hat{\beta}$, are obtained by minimising the residual sum of squares. The predicted dose-rate at an unobserved location, s_0 , is:

$$\hat{Y}(s_0) = X(s_0)^T \hat{\beta} \quad (11)$$

All models included as explanatory variables the easting and northing coordinates and the outdoor gamma-ray dose-rate. The outdoor gamma-ray dose-rates are estimates for 10 km grid squares and were subject to interpolation and double smoothing (Kendall et al., 2016). All variogram models and the non-spatial OLS model included as additional explanatory variables either the 16-level bedrock geology code, the 23-level surface geology code, or the 23-level variable representing a combination of surface and bedrock geology; this could not be included in the multi-resolution Gaussian-process model fits because of singularities in the inverse of the design matrix. Results are presented in Table 1. Indoor gamma-ray dose-rates will depend on the details of construction of the dwelling in question. This information is not available for each of the buildings for which we have measurement data. However, we have obtained socioeconomic and demographic data from the 1981 UK census which allows us to assign parameters to each measurement location on the basis of the census ward within which it lies. There were 10,444 census wards in GB at the time of the 1981 census, with a mean population (all ages) of about 5,000. Within each census ward are a number of enumeration districts (EDs). The parameters that are available for analysis are:

- a) The Carstairs index of socioeconomic deprivation (Carstairs and Morris, 1991), where the more deprived the area the higher (more positive) is the score.
- b) The population density (i.e., number of inhabitants per square kilometer)
- c) Urban/rural status, a categorical variable running from 1 (most urban) to 6 (least urban) (Office of Population Censuses and Surveys (OPCS), 1984). Each ED within the census ward is categorised as urban or rural on the basis of the extent of urban development indicated on Ordnance Survey maps. The overall urban/rural score for the ward is 1 if all

EDs are urban, 2 if the proportion is above 75%, and so on until 6 where all EDs are rural.

Table 2 presents the results of analyses using these census data. As above, the urban-rural status variable could not be used in the multi-resolution Gaussian-process model because of singularities in the design matrix. The parameter estimates for the 16-level and 23-level geology composition variables are not shown in Tables 1 and 2, for clarity.

All models were estimated after excluding the holdout sample from the dataset, using the remaining 7,037 observations. Models were subsequently used to predict the 3,162 dose-rates in the holdout sample, with the predictive accuracy of each model evaluated in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson correlation, and Spearman rank-correlation coefficients. Table 3 summarises these predictive accuracy metrics for each model; Table 4 gives the analogous results using the extended versions of the models with Carstairs score, population density and urban-rural status as extra explanatory variables.

Results

Figure 1 presents a map of indoor and outdoor gamma-ray dose-rates in GB. Indoor dose-rates tend to be higher in South-West and North-West England, as well as in the West Midlands region. On the other hand, indoor dose-rates tend to be lower in and around London, and in the East and South-East regions of England. It is difficult to discern meaningful patterns in the sparsely-populated and sparsely-measured areas in the North-East and Yorkshire and the Humber regions of England, and in rural Wales and Scotland.

Table 1 presents parameter estimates for the four parametric variogram models (Matérn, Gaussian, Bessel, Spherical), three multi-resolution Gaussian-process models ($L=2, 4, 8$), and the

non-spatial OLS model. The mean gamma-ray dose-rate for all models is expressed as a function of the intercept, Easting and Northing coordinates in kilometers, and the outdoor gamma-ray dose-rate.

Figure 2 depicts each of the four parametric variogram models taken from Table 1 fitted through the empirical sample variogram. All four models appeared to fit the empirical variogram well, and yielded similar long range (>200 km) predicted variograms, although there were more substantial differences among the predicted variograms at small spatial distances (<50 km). Specifically, the Gaussian model featured the most pronounced point of inflection around a distance of 50 km, whereas the Spherical model exhibited linear behavior and thus did not have a point of inflection at all. The estimated practical range (i.e., the distance at which dose-rates are virtually independent) varies between 200 km for the Gaussian model and 352 km for the Bessel model (Table 1). These estimates are little changed if the expanded set of explanatory variables (including Carstairs score, population density, urban-rural status) is employed (Table 2). Combined with a relatively large nugget effect (τ^2) and small partial sill (σ^2), the small ranges suggest that the spatial variation of indoor gamma-ray dose-rates was highly localised. The estimates for nugget (τ^2) for the multi-resolution Gaussian-process models were substantially lower compared to the six variogram models, although the partial sill (σ^2) was somewhat higher. These estimates suggest that the multi-resolution model produced a potentially over-smoothed predicted response surface. Again, these estimates are little changed if the expanded set of explanatory variables (including Carstairs score, population density, urban-rural status) is employed (Table 2). The regression coefficients for the outdoor gamma-ray dose-rate ($\beta_{Ext\ Gamma}$) varied considerably. Among models that accounted for spatial correlation, the estimated increase in the indoor dose-rate per unit of outdoor dose-rate was as low as 0.16 for the Matérn, Bessel, and

Spherical variogram models, and as large as 0.34 for the multi-resolution Gaussian-process model with $L=8$. The non-spatial model estimate was larger still at 0.50, but its validity was questionable due to the clear violations of OLS assumptions that resulted from spatially correlated data. When using the expanded set of explanatory variables (including Carstairs score, population density), the estimates for $\beta_{Ext\ Gamma}$ were little changed for the multi-resolution Gaussian-process model, although they were markedly larger for the variogram-based models (Table 2).

All the predictive accuracy metrics suggest that the multi-resolution Gaussian-process model with $L=8$ had the best predictive accuracy among the models considered (Tables 3 and 4). The multi-resolution Gaussian-process model fitted to the GB indoor gamma-ray dose-rate data is shown in Figure 3, and comparison with the leftmost panel of Figure 1 demonstrates that the model captures the essential features of the indoor gamma radiation dose-rate data. Unfortunately, models with a larger number of components ($L=12, 16$) proved computationally infeasible, requiring more than the allotted memory (192 GB) ($L=16$) or necessitating an excessively long CPU time (>2 months) for maximum-likelihood optimization ($L=12$). The Spherical variogram-based model had the second-best predictive performance (e.g., $MSE = 394.940$ vs 391.748) (Table 3). The predictive accuracy of all six variogram-based models, in particular the Spherical model, was further improved if the extra socio-economic variables (Carstairs score, population density, urban-rural status) were employed (e.g., $MSE = 394.940$ vs 391.921), although this was not the case for the Gaussian-process model (e.g. $MSE = 391.748$ vs 399.514) (Tables 3, 4). However, even when the expanded explanatory variables were used for the Spherical model, the Gaussian-process model without these variables had still marginally the best predictive performance of all models (e.g. $MSE = 391.921$ vs 391.748) (Tables 3, 4). There was little difference made by the choice of geological explanatory variables used (16-level bedrock geology, 23-level surface geology, 23-

level combination of surface and bedrock geology) in the predictive accuracy of the Spherical model (Tables 3, 4). The Bessel, Matérn, Gaussian and OLS models had progressively worse predictive performance, and the OLS model was particularly poor in this respect (Tables 3, 4). Model computation time was on the order of minutes for the Spherical model, compared to several hours for the multi-resolution Gaussian-process model. As expected, all models with a spatial component performed better than the non-spatial OLS model. Overall, the multi-resolution Gaussian process model with $L=8$ improved the predictive MAE and MSE over the OLS model by 4.4% and 7.0%, respectively (Table 3).

Discussion

We found that each of the four parametric models (Matérn, Gaussian, Bessel, Spherical) fitted the empirical variogram well, and yielded similar long-range (>200 km) predictions, although with more substantial differences in predicted variograms at smaller spatial distances (<50 km). The multi-resolution Gaussian-process model with $L=8$ had the best predictive accuracy among the models considered. The Spherical variogram-based model had the second-best predictive performance, and when the augmented set of explanatory socioeconomic variables was used (Carstairs score, population density, urban-rural status) the performance almost equalled that of the multi-resolution Gaussian-process model (Tables 3, 4). The Bessel, Matérn, Gaussian and OLS models had progressively worse predictive performance.

Even though variogram-based kriging is both familiar and able to handle high-dimensional data with relative ease, it is not without drawbacks. For example, estimation and inference was performed using the estimated sample variogram, not the original observations. Several arbitrary decisions must routinely be made during computation of the sample variogram, for example, selecting bin-widths for distance binning, and selecting the maximum distance for which to

compute the sample variogram. For these reasons, when it is feasible, most researchers prefer a likelihood-based approach to spatial parameter estimation and prediction (e.g., Diggle *et al.* (Diggle et al., 2003).

Warnery *et al* (Warnery et al., 2015) modelled natural background gamma dose rates in France using ordinary kriging (OK) and multi-collocated kriging (MCKK) (Rivoirard, 2001), in which dose rate was estimated using a weighted linear sum of nearby dose rates and certain ancillary variables, specifically geological uranium potential. Their dataset was much larger than ours, comprising 97,595 thermoluminescent dosimeter measurements in 17,404 locations. Warnery *et al* (Warnery et al., 2015) found that both OK and MCKK yielded very similar predictions, although MCKK gave slightly better predictive accuracy, evaluated using cross-validation.

The OLS model that was used here should be contrasted with the somewhat different OLS model fitted in the previous paper to a somewhat expanded set of variables (Kendall et al., 2016), which we shall refer to as the extended OLS (E-OLS) model. In contrast to the present OLS model, which uses only linear functions of the coordinates, the 23-level bedrock and surface geological predictor, the outdoor gamma-ray dose-rate and a few other variables (Table 2), this previous E-OLS model was derived by linear regression applied to predictor variables based on a number of dose-rate interpolation measures, calculated from averages over geologically- or administratively-defined areas, or distance-weighted averages of measurements at nearest-neighbour points. A very similar cross-validation methodology was used in that paper as here, indeed using the same estimation and test samples. The mean square error of the E-OLS model, 377.64, was substantially lower than that of the Gaussian-Matérn model, 410.92 (Kendall et al., 2016), and lower than any of the models fitted here – the best fitting multi-resolution Gaussian-process model with $L=8$ had

a mean square error of 391.748 (Table 3). Unfortunately, Gaussian-process models with a larger number of components ($L=12, 16$) were computationally infeasible. This suggests that the E-OLS model is still marginally the preferred model. Strictly, one would choose whichever model had the lowest MSE. Since:

$$MSE = \frac{1}{N} \sum_{i=1}^N E[(D_i - f_{D,i}(X))^2] = \frac{1}{N} \sum_{i=1}^N \text{var}[f_{D,i}(X)] + [E[f_{D,i}(X)] - D_i]^2 \quad (12)$$

the MSE is always bounded below by the variance of the estimator, which can be estimated to be ~366 (Kendall et al., 2016). Therefore the E-OLS estimator is close to being optimal, and the other estimators constructed here somewhat further from it. However, as discussed elsewhere (Kendall et al., 2016), it is likely on theoretical grounds that neither OLS model will estimate the variance correctly. That this is not merely of theoretical concern was confirmed by the evidence presented, suggesting that the particular E-OLS model used underestimated the estimated variance for the predicted dose-rates from those of the (admittedly quite poorly fitting) likelihood-based model by a factor of about 3 (Kendall et al., 2016). Geostatistical models can statistically account for the tendency of gamma radiation dose-rates in nearby houses to be similar to one another (i.e., to correlate), more than can be accounted for by the geological and other variables in the fixed-effect part of the model. As discussed elsewhere (Kendall et al., 2016), the gamma-ray dose-rate inside a house can be regarded as the outdoor gamma-ray dose-rate attenuated by the fabric of the house, but augmented by a contribution from radionuclides in the building materials. The second of these contributions is generally dominant (Kendall et al., 2016). To the extent that the housing stock in certain areas will tend to be built to similar design and using common materials (which would not necessarily be from that specific area), this would explain how the local correlation could arise. Set against that, the pattern of residuals derived from the E-OLS model that was developed in the previous analysis (Kendall et al., 2016) did not exhibit pronounced correlation, implying that the

rather larger number of parameters and variables that this model incorporated may be to some extent “explaining” the covariance structure in the raw data.

Purely likelihood-based geostatistical models have other advantages over the particular E-OLS models that were fitted by (Kendall et al., 2016) and over the six variogram-fitted models used here; in particular there is a more adequate likelihood-based framework for inference, so that model selection and estimates of prediction error can be more reliably performed. However, the poorer predictive performance of the likelihood-based models fitted here, in particular the multi-resolution Gaussian process model (which is admittedly only an approximation to a true likelihood-based model), suggests that they should probably not be the main method used for the estimation of doses for the purposes of epidemiological analysis. However, despite these disadvantages, the multi-resolution Gaussian-process model and the Spherical variogram-fitted model should be considered alongside the E-OLS model for epidemiological analyses using these data.

Acknowledgments

We are very grateful to for the detailed comments of the referee and to Dr Jill Simpson of the University of York and to the other UKCCS investigators for making available the indoor gamma-ray dose-rate measurements made for the United Kingdom Childhood Cancer Study and for advice on the interpretation of the data. We are also very grateful to Mr JD Appleton for advice on geological matters generally and in particular for suggesting geological classification schemes. We are grateful to Drs Phil Gilvin, Luke Hager and Rick Tanner at Public Health England (PHE) for advice on the dosimetry of the National Survey and the UKCCS, and to Dr Simon Bouffler and other colleagues at PHE for allowing the use of the National Survey data and for advice on its

interpretation. We are grateful to Tim Vincent for help with the census data. This work was supported by the Intramural Research Program of the National Institutes of Health, the National Cancer Institute, Division of Cancer Epidemiology and Genetics. The work of GMK was supported by Children with Cancer (UK) under grant number HTRVVK0.

Figure 1. Measured indoor and outdoor gamma-ray dose-rates (nGy/h) in Great Britain, using data taken from Kendall *et al.* (Kendall et al., 2016).

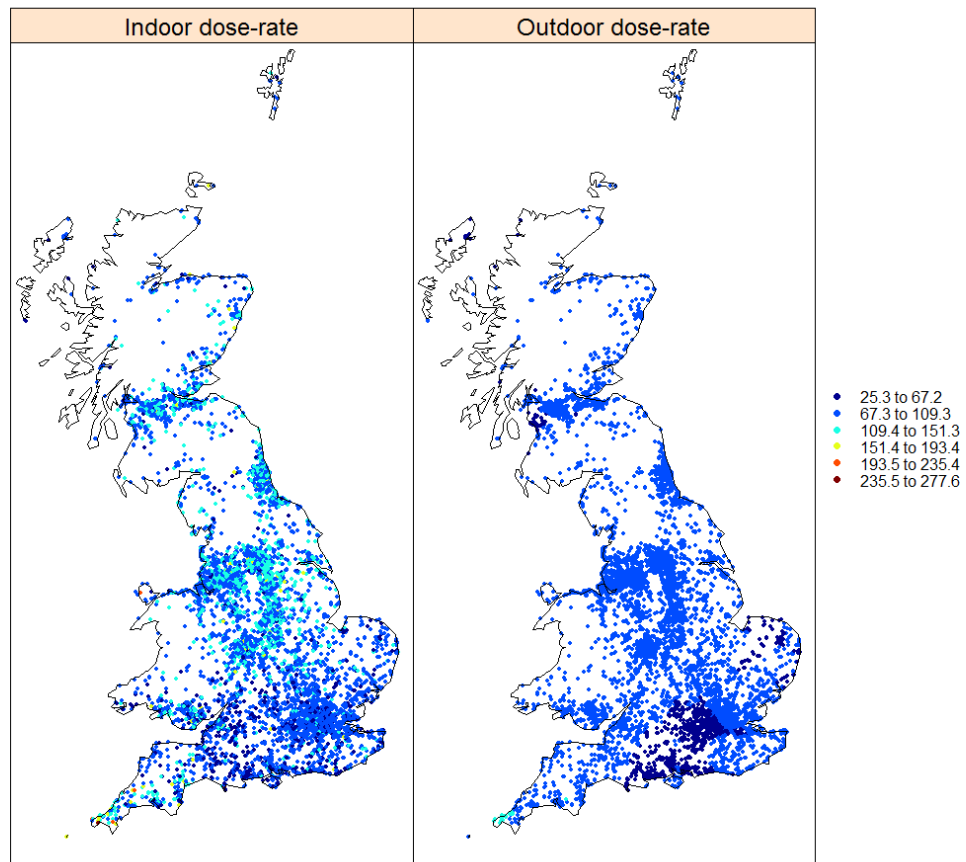


Figure 2. Variogram model fits for the Matérn, Gaussian, Bessel, and Spherical models.

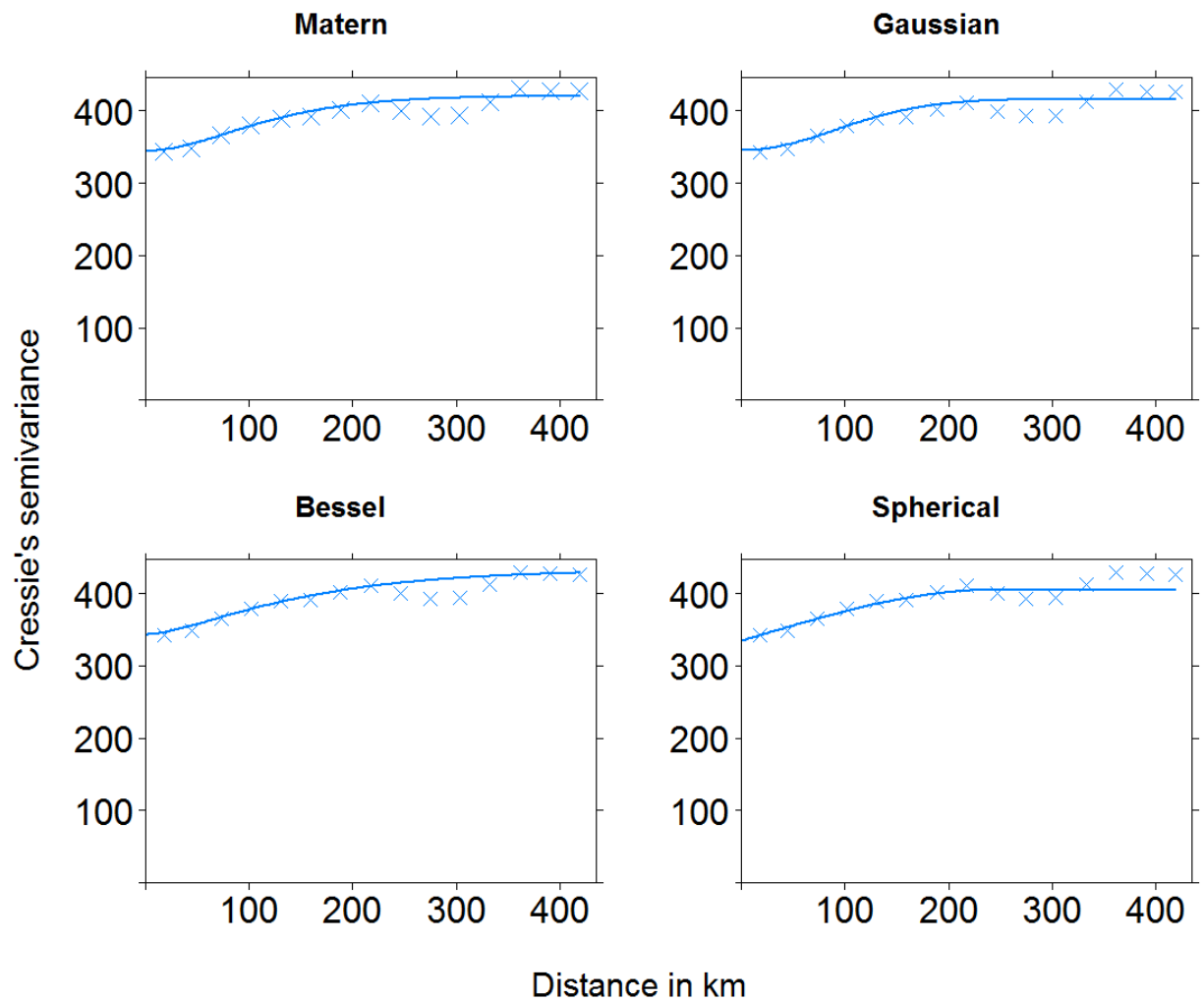


Figure 3. Fine-grid spatial prediction of indoor gamma-ray dose-rate (nGy/h) in Great Britain using the multi-resolution Gaussian-process model with $L=8$.

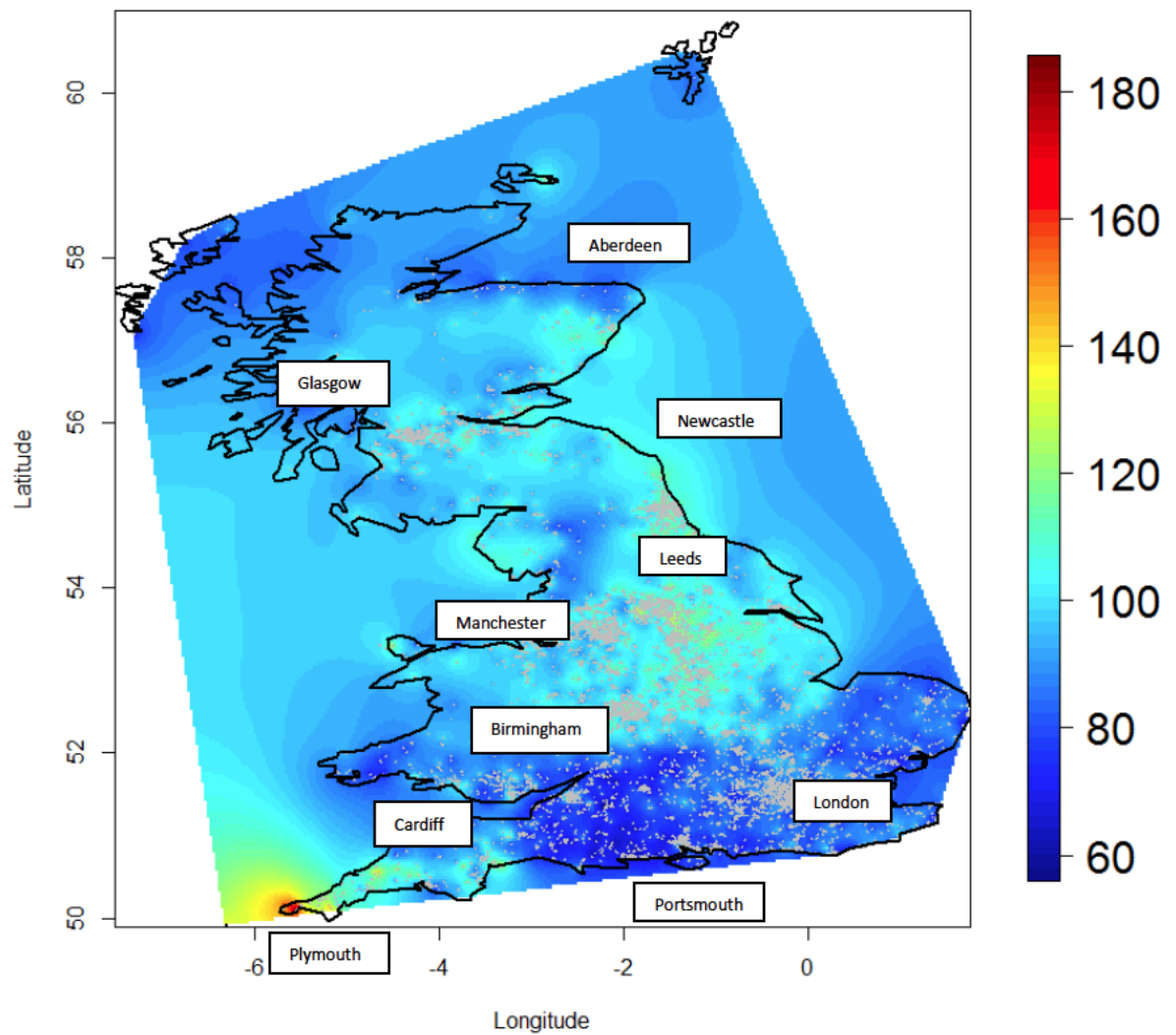


Table 1. Parameter estimates for parametric variogram-based models, multi-resolution Gaussian-process models, and non-spatial OLS model.

Parameter Estimates	Variogram-based models						Multi-resolution Gaussian-process models			Non-spatial model
	Matérn ^c	Gaussian ^c	Bessel ^c	Spherical			<i>L</i> =2	<i>L</i> =4	<i>L</i> =8	
				16-level bedrock geology ^a	23-level surface geology ^b	23-level combined bedrock-surface geology ^c				
Intercept	89.085	86.478	90.804	87.790	82.248	87.470	83.026	83.197	74.050	35.658
$\beta_{Easting}$	-0.001	0.003	-0.004	0.0009	-0.005	-0.002	-0.006	-0.003	-0.0002	0.036
$\beta_{Northing}$	-0.008	-0.008	-0.010	-0.009	-0.013	-0.007	-0.016	-0.015	-0.012	0.020
$\beta_{Ext\ gamma}$	0.166	0.184	0.162	0.125	0.208	0.165	0.191	0.256	0.339	0.504
Nugget (τ^2)	341.75	342.41	340.06	334.44	338.80	336.17	19.86	19.48	18.46	413.31
Partial sill (σ^2)	64.98	61.24	74.30	70.77	78.48	67.03	105.97	90.53	107.29	
Scale parameter (ϕ)	42.32	115.58	87.96	244.98	287.83	251.70				
Practical Range (km)	250.50	200.13	351.74	198.85	233.62	204.22				

^aestimates for the 16-category bedrock geology covariates are not reported in the table in the interest of clarity.

^bestimates for the 23-category surface geology covariates are not reported in the table in the interest of clarity.

^cestimates for the 23-category bedrock and surface geology covariates are not reported in the table in the interest of clarity.

Table 2. Parameter estimates for parametric variogram-based models, multi-resolution Gaussian-process models, and non-spatial OLS model. Models have been specified with additional covariates for the standardised 1981 Carstairs score, 1981 Population Density, and categorical Urban/Rural status.

Parameter Estimates	Variogram-based models						Multi-resolution Gaussian-process models			Non-spatial model
	Matérn ^c	Gaussian ^c	Bessel ^c	Spherical			L=2	L=4	L=8	OLS ^c
				16-level bedrock geology ^a	23-level surface geology ^b	23-level combined bedrock-surface geology ^c				
Intercept	0.003	0.004	0.002	0.001	0.003	0.003	91.498	99.225	82.409	36.809
$\beta_{Easting}$	0.071	0.071	0.069	0.066	0.067	0.065	-0.008	-0.007	-0.003	0.032
$\beta_{Northing}$	0.024	0.024	0.024	0.023	0.023	0.022	-0.011	-0.022	-0.016	0.023
$\beta_{Ext\ gamma}$	0.532	0.542	0.516	0.585	0.572	0.595	0.180	0.124	0.270	0.510
$\beta_{Carstairs1981}$	0.046	0.054	0.046	0.038	0.049	0.046	0.061	0.038	0.093	0.065
$\beta_{Popdensity1981}$	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0004	0.0004	0.0001
β_{Urban1}	ref.	ref.	ref.	ref.	ref.	ref.				ref.
β_{Urban2}	-1.551	-1.454	-1.541	-1.471	-1.468	-1.464				-2.928
β_{Urban3}	0.280	0.259	0.489	0.572	0.515	0.507				-2.300
β_{Urban4}	0.083	0.075	0.010	-0.055	-0.014	-0.100				-4.616
β_{Urban5}	0.016	0.012	0.011	0.010	0.006	0.006				-9.755
β_{Urban6}	0.224	0.208	0.123	0.060	0.092	0.096				-4.752
Nugget (τ^2)	339.77	340.41	338.28	333.38	336.31	334.57	19.90	19.59	18.46	409.25
Partial sill (σ^2)	61.49	58.15	70.84	66.00	69.47	62.89	136.27	90.26	107.29	
Scale parameter (ϕ)	43.52	119.54	91.79	255.80	273.17	256.79				
Practical Range (km)	257.74	206.80	366.98	207.60	221.61	208.35				

^aestimates for the 16-category bedrock geology covariates are not reported in the table in the interest of clarity.

^bestimates for the 23-category surface geology covariates are not reported in the table in the interest of clarity.

^cestimates for the 23-category bedrock and surface geology covariates are not reported in the table in the interest of clarity.

Table 3. Test sample predictive accuracy measures for parametric variogram-based models, multi-resolution Gaussian-process models, and the non-spatial OLS model.

Metric	Variogram-based models						Multi-resolution Gaussian-process models			Non-spatial model
	Matérn	Gaussian	Bessel	Spherical			$L=2$	$L=4$	$L=8$	OLS
				16-level bedrock geology	23-level surface geology	23-level combined bedrock- surface geology				
MAE	15.305	15.357	15.262	15.223	15.208	15.227	15.301	15.231	15.155	15.806
MSE	398.634	400.646	396.840	395.461	394.940	395.064	399.014	395.847	391.748	418.696
Pearson correlation	0.472	0.468	0.476	0.479	0.480	0.480	0.471	0.478	0.486	0.429
Spearman correlation	0.483	0.479	0.486	0.487	0.494	0.491	0.475	0.483	0.494	0.440

Table 4. Test sample predictive accuracy measures for parametric variogram-based models, multi-resolution Gaussian-process models, and the non-spatial OLS model. Models have been specified with additional covariates for the standardised 1981 Carstairs score, 1981 Population Density, and categorical Urban/Rural status.

Metric	Variogram-based models						Multi-resolution Gaussian-process models			Non-spatial model
	Matérn	Gaussian	Bessel	Spherical			<i>L</i> =2	<i>L</i> =4	<i>L</i> =8	OLS
				16-level bedrock geology	23-level surface geology	23-level combined bedrock- surface geology				
MAE	15.214	15.260	15.174	15.153	15.144	15.147	15.310	15.271	15.294	15.659
MSE	395.090	396.532	393.600	392.702	391.921	392.029	398.191	397.787	399.514	411.736
Pearson correlation	0.480	0.477	0.482	0.484	0.486	0.486	0.473	0.476	0.476	0.444
Spearman correlation	0.490	0.489	0.493	0.492	0.495	0.496	0.478	0.478	0.485	0.459

Appendix A. Supplementary tables on dose statistics for simplified bedrock and surface classes and reduced geologies of the measurement set

The geological classifications used by Kendall *et al* (Kendall et al., 2016) were based on simplified 1:50,000 geological mapping developed by the British Geological Survey for predicting radon concentrations (Appleton, 2005; Miles and Appleton, 2005). These classifications combined bedrock and surface components. In about 40% of cases the surface geology was judged the same as the bedrock while elsewhere a superficial geology overlays the bedrock. For purposes of gamma-ray modelling the bedrock classifications were further simplified to sixteen “Bedrock Classes”. These broadly correspond to geological periods. However, six rock types that correlate with patterns of gamma-ray dose-rate were also included.

For the present work two additional geological classifications were explored. Since very few gamma-rays penetrate more than a metre in rock it is plausible that surface geology will be more relevant than bedrock in determining indoor gamma-ray dose-rates. Twenty-three “Surface Classes” were therefore defined by combining similar types and removing superficial geologies, which were likely to be very thin. Sixteen of these surface classes were bedrock which was found at the surface.

It was also thought desirable to explore the predictive power of combinations of bedrock and surface geologies. However, simple combinations of the bedrock and surface classes above would result in too many classifications to be tractable and would likely result in overfitted models. In a further level of simplification, the bedrock classifications were simplified to five combinations of geological Periods plus granite, and the superficial geologies were reduced to three classifications resulting in twenty-three bedrock/surface combinations. Details of these geological classifications

are given in Appendix Tables A1-A5. Data given are number of measurements, mean dose-rate (nGy/h) with standard deviation, minimum and maximum.

Appendix Table A1. Simplified bedrock classes

Geological Coding	Abbreviation	Number		Dose-rate (nGy/h)			
				Mean	SD	Min	Max
Granite	GRANITE	51	0.5%	117.4	37.3	53.9	212.3
Precambrian	PRECAM	88	0.9%	95.2	20	44	136.3
Cambrian	CAMB	31	0.3%	90.8	25.7	54.4	199.3
Ordovician	ORDOV	28	0.3%	91.6	26.7	46.2	151.2
Silurian	SILUR	126	1.2%	93.2	21.8	41.5	171.7
Devonian	DEV	457	4.5%	96.1	24.9	35.7	277.6
Carboniferous	CARB	1196	11.7%	103.6	22.9	31.9	177.5
Carboniferous Coal Measures	CARBMEAS	1837	18.0%	104.2	20.9	31.6	202.9
Permian-Triassic	PERTRI	1319	12.9%	103.8	19.3	45	192.2
Permian Mercia Mudstone	PTMD	862	8.5%	101	20.4	37.6	193
Jurassic	JUR	816	8.0%	89.4	24.4	35.2	270.5
Jurassic Ironstone	JURFE	47	0.5%	96.1	20	49.3	131.6
Jurassic Oxford clay	JUOXCLAY	173	1.7%	84.8	20.5	30	133.2
Cretaceous	CRET	1249	12.2%	86.6	19.5	25.5	191.2
Cretaceous Clay	CRETCLAY	75	0.7%	84.6	20	43.9	156.7
Tertiary	TERT	1844	18.1%	83.3	18	25.3	203.9
Total		10199	100.0%	95.6	22.7	25.3	277.6

Appendix Table A2. Simplified surface classes

Geological Coding	Abbreviation	Number		Dose-rate (nGy/h)			
				Mean	SD	Min	Max
Granite	GRANITE	20	0.2%	131.2	48.1	53.9	212.3
Precambrian	PRECAM	12	0.1%	85.1	27.2	44	135.4
Cambrian	CAMB	13	0.1%	95.7	34.5	66.9	199.3
Ordovician	ORDOV	8	0.1%	75.6	19.7	57.9	114.8
Silurian	SILUR	50	0.5%	92	20.5	56.4	150.3
Devonian	DEV	193	1.9%	97.9	30.4	35.7	277.6
Carboniferous	CARB	602	5.9%	103.4	24	31.9	177.5
Carboniferous Coal Measures	CARBMEAS	450	4.4%	108.2	22.5	31.6	202.9
Permian-Triassic	PERTRI	380	3.7%	105.6	20.9	45	192.2
Permian Mercia Mudstone	PTMD	261	2.6%	99.6	22.6	37.6	161.7
Jurassic	JUR	495	4.9%	84.3	24.9	35.2	270.5
Jurassic Ironstone	JURFE	45	0.4%	96	20.4	49.3	131.6
Jurassic Oxford clay	JUOXCLAY	71	0.7%	79.8	22.2	30	132.6
Cretaceous	CRET	615	6.0%	83.1	19.4	25.5	191.2
Cretaceous Clay	CRETCLAY	58	0.6%	84.3	20.9	43.9	156.7
Tertiary	TERT	858	8.4%	83.6	18.4	25.8	203.9
Clay with flints	C-DMTN	73	0.7%	83.2	18.4	35.8	122
Clay-silt (mainly alluvium)	CLSI	1009	9.9%	96.9	22	32.4	167.5
Diamicton (mainly glacial till)	DMTN	2764	27.1%	101.1	20.1	41.5	193.4
Head - congeliturbate deposits (clay-silt)	H-CLSI	262	2.6%	86.5	18.6	33.3	144.9
Head - congeliturbate deposits (sand and gravel)	H-SAGR	13	0.1%	72.4	16.1	43.9	96.4
Sand and gravel	SAGR	1938	19.0%	93.8	20.9	25.3	186.7
Peat	PEAT	9	0.1%	86.4	7.3	78.8	99
Total		10199	100.0%	95.6	22.7	25.3	277.6

Appendix Table A3. Reduced bedrock classifications

Geological Coding	Abbreviation	Number		Dose-rate (nGy/h)			
				Mean	SD	Min	Max
Granite	Granite	51	0.5%	117.4	37.3	53.9	212.3
Devonian and earlier periods	To_Devonian	730	7.2%	95.1	24	35.7	277.6
Carboniferous	Carbonif	3033	29.7%	103.9	21.7	31.6	202.9
Permian to Triassic	Perm-Tri	2181	21.4%	102.7	19.8	37.6	193
Jurassic, Cretaceous and Tertiary	JurCretTert	4157	40.8%	85.6	20.1	25.3	270.5
Jurassic Ironstone	Jurassic Fe	47	0.5%	96.1	20	49.3	131.6
Total		10199	100.0%	95.6	22.7	25.3	277.6

Appendix Table A4. Reduced surface classifications

Geological Coding	Abbreviation	Number		Dose-rate (nGy/h)			
				Mean	SD	Min	Max
Granite	Granite	20	0.2%	131.2	48.1	53.9	212.3
Devonian and earlier periods	To_Devonian	283	2.8%	95.3	28.8	35.7	277.6
Carboniferous	Carbonif	1057	10.4%	105.4	23.5	31.6	202.9
Permian to Triassic	Perm-Tri	668	6.5%	102.9	21.7	37.6	192.2
Jurassic, Cretaceous and Tertiary	JurCretTert	2415	23.7%	83.6	20.2	25.5	270.5
Jurassic Ironstone	Jurassic Fe	45	0.4%	96	20.4	49.3	131.6
Clay-silt (mainly alluvium)	CLSI	1009	9.9%	96.9	22	32.4	167.5
Diamicton (mainly glacial till)	DMTN	2764	27.1%	101.1	20.1	41.5	193.4
Sand and Gravel	SAGR	1938	19.0%	93.8	20.9	25.3	186.7
Total		10199	100.0%	95.6	22.7	25.3	277.6

Appendix Table A5. Reduced bedrock and surface combinations

Geological Coding^a	Number		Dose-rate (nGy/h)			
			Mean	SD	Min	Max
Granite/Granite	20	0.2%	131.2	48.1	53.9	212.3
Granite/CLSI	2	0.0%	135	47.9	103.4	167.5
Granite/DMTN	19	0.2%	101.8	16.7	64.6	126.4
Granite/SAGR	10	0.1%	115.8	32.1	69.9	186.7
To_Devonian/To_Devonian	283	2.8%	95.3	28.8	35.7	277.6
To_Devonian/CLSI	70	0.7%	93.9	20	49.8	141.4
To_Devonian/DMTN	233	2.3%	94.5	19.4	41.5	153.5
To_Devonian/SAGR	144	1.4%	96.4	22.1	46.2	171.7
Carbonif/Carbonif	1057	10.4%	105.4	23.5	31.6	202.9
Carbonif/CLSI	287	2.8%	102.5	22.1	38.7	162.7
Carbonif/DMTN	1387	13.6%	103.4	20.6	43.1	193.4
Carbonif/SAGR	302	3.0%	102.7	19.8	52.5	169
Perm-Tri/Perm-Tri	668	6.5%	102.9	21.7	37.6	192.2
Perm-Tri/CLSI	254	2.5%	103.6	22.4	47.7	158.1
Perm-Tri/DMTN	754	7.4%	102.2	17.9	51.6	193
Perm-Tri/SAGR	505	5.0%	102.7	18.5	41.6	170.2
JurCretTert/JurCretTert	2415	23.7%	83.6	20.2	25.5	270.5
JurCretTert/CLSI	396	3.9%	88.8	18.9	32.4	133.3
JurCretTert/DMTN	370	3.6%	94.5	20.5	46.1	166.7
JurCretTert/SAGR	976	9.6%	85.8	18.9	25.3	170.2
Jurassic Fe/Jurassic Fe	45	0.4%	96.0	20.4	49.3	131.6
Jurassic Fe/DMTN	1	0.0%	96.4	0	96.4	96.4
Jurassic Fe/SAGR	1	0.0%	99.4	0	99.4	99.4
Total	10199	100.0%	95.6	22.7	25.3	277.6

^alisted using abbreviated codes (bedrock/surface) from Tables A3-A4.

Figure A1. Map of Great Britain



Appendix B. Detailed statistical methods.

Variogram-based kriging

The variogram is a fundamental geo-statistical tool used to describe spatial variability in point-referenced data. Formally, let D represent the spatial domain in \mathbb{R}^2 indexed by a two-dimensional coordinate system, and let $Y(s)$ denote the observed gamma-ray dose-rate collected at location $s \in D$. We assume that there is an underlying spatial process $S(s)$ which generates $Y(s)$, so that:

$$Y(s) = S(s) + Z(s) \quad (\text{B1})$$

where $Z(s) \sim N(0, \tau^2)$ are independent and identically distributed (for different values of s), and all independent of $S(s)$, and correspond to the measurement error of the observed process. The variance of this process, τ^2 , is also called the nugget variance.

We assume that $S(\cdot)$ is second-order stationary and isotropic, and thus for any two locations $s, s' \in D$:

$$E[Y(s)] = E[S(s)] = \mu(s) = X(s)^T \beta \quad (\text{B2})$$

$$\begin{aligned} \text{cov}[Y(s), Y(s')] &= \text{cov}[S(s) + Z(s), S(s') + Z(s')] \\ &= V(\|s - s'\|) + \tau^2 1_{s=s'} = \sigma^2 \rho(\|s - s'\|) + \tau^2 1_{s=s'} \\ &= C(\|s - s'\|) \end{aligned}$$

where $X(s)$ is the design matrix of spatial covariates, β is a vector of fixed effect coefficients, $\|s - s'\|$ is the Euclidean distance between locations s and s' , and $V(\cdot)$ is a valid covariance function for the underlying process $S(\cdot)$. σ^2 is the variance of the underlying process $S(\cdot)$, and

$\rho(\cdot)$ the spatial correlation of this process. The variogram is expressed as a function of the spatial separation vector $h = s - s'$, and is defined as:

$$\begin{aligned}\gamma(h) &= \frac{1}{2} \text{var}[Y(s) - Y(s')] = \text{var}[Y(s)] - \text{cov}[Y(s), Y(s')] \\ &= \tau^2 1_{h \neq 0} + \sigma^2 [1 - \rho(h)] = C(0) - C(h)\end{aligned}\tag{B3}$$

where $C(0) = \tau^2 + \sigma^2$ is the so-called sill of the variogram, $C(h)$ is the value of the covariance function evaluated at a specific distance.

Variogram models are generally parameterised in terms of the error variance term τ^2 (also called the nugget variance), the spatial variance σ^2 (also called the partial sill), and the correlation function $\rho(h|\phi)$, which depends on some covariance parameters ϕ . Parametric variogram models have the form:

$$\gamma(h|\phi) = \tau^2 1_{h \neq 0} + \sigma^2 (1 - \rho(h|\phi))\tag{B4}$$

where the correlation function $\rho(h|\phi)$ must have the properties that: $\rho(0|\phi) = 1$, $\rho(h|\phi) \rightarrow 0$ as $h \rightarrow \infty$. Covariance parameters in ϕ control how quickly $\rho(h|\phi)$ decays to 0 and its overall shape. Parametric variogram models that were considered in this study included the following spatial correlation functions:

1) the Matérn model (Matérn, 1960), in which:

$$\gamma(h|\phi, \kappa) = 2^{\kappa-1} (h/\phi)^\kappa K_\kappa(h/\phi)\tag{B5}$$

and where $K_\kappa(\cdot)$ is the modified Bessel function of the second kind of order κ ; we shall generally fix $\kappa = 2.5$;

2) the Gaussian model, in which:

$$\gamma(h|\phi) = \phi [1 - \exp(-3h^2 / \phi^2)]\tag{B6}$$

3) the Bessel model, in which:

$$\gamma(h | \phi, \kappa) = 2^{\kappa-1} (h / \phi)^{-\kappa} \Gamma(\kappa + 1) J_{\kappa}(h / \phi) \quad (\text{B7})$$

and where $J_{\kappa}(\cdot)$ is the modified Bessel function of first kind of order κ ; and

4) the Spherical model, in which:

$$\gamma(h | \phi) = \begin{cases} \phi [1.5(h / \phi) - 0.5(h / \phi)^3] & \text{if } h \leq \phi \\ \phi & \text{if } h > \phi \end{cases} \quad (\text{B8}).$$

Variogram models using these parametric forms are fitted to the empirical variogram, $\hat{\gamma}(h)$, constructed from pairs of observations $Y(s)$ and $Y(s')$. For this purpose we employ the robust Cressie empirical variogram estimator (Cressie and Hawkins, 1980; Cressie, 1993; Hawkins and Cressie, 1984):

$$\hat{\gamma}(h) = \frac{1}{2(0.457 + 0.494 / |N(h)|)} \left[\frac{1}{|N(h)|} \sum_{s, s' \in N(h)} |Y(s) - Y(s')|^{1/2} \right]^4 \quad (\text{B9})$$

where $N(h)$ is the set of pairs of points (s, s') with spatial differences $h = \|s - s'\|$. This estimator is employed because of its reduced bias and resistance to outlying observations, compared to the standard empirical variogram. Estimates for τ^2 , σ^2 and ϕ are derived via weighted least squares, with weights $\frac{|N(h)|}{h^2}$.

Using parameter estimates from the variogram model in combination with estimates for spatial covariates (i.e., $\hat{\beta}$) we can compute the predicted dose-rate value at some unobserved location, s_0 , using the universal kriging equation (e.g., Banerjee *et al.* (Banerjee et al., 2015), Diggle and Ribeiro (Diggle and Ribeiro, 2007) p.136):

$$\hat{Y}(s_0) = X(s_0)^T \hat{\beta} + \hat{\gamma}^T \hat{\Sigma}^{-1} (Y - X(s)^T \hat{\beta}) \quad (\text{B10})$$

where $\hat{\gamma}^T = (\hat{\sigma}^2 \rho(\|s_0 - s_1\| \hat{\phi}), \dots, \hat{\sigma}^2 \rho(\|s_0 - s_N\| \hat{\phi}))$, and $\hat{\Sigma}$ is the estimated variance-covariance matrix of Y . The estimated variance-covariance matrix is defined as $\hat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi}) + \hat{\tau}^2 I_N$, where each element of the matrix $H(\cdot)$ is of the form: $H(\hat{\phi})_{ij} = \rho(\|s_i - s_j\| \hat{\phi})$. All computation was carried out using the `gstat` package (Pebesma and Graeler, 2015) in R (R Project version 3.2.2, 2015).

Multi-resolution Gaussian Process

Although it is a theoretically attractive alternative to variogram-based prediction, the implementation of maximum-likelihood kriging equations is known to be intractable for datasets with sample sizes on the order of $n \approx 10,000$ (Cressie and Johannesson, 2008). Spatial prediction based on a multi-resolution Gaussian-process model (Nychka et al., 2014) that approximates the true likelihood of the process provides one method that circumvents this computational challenge. In this section, we give an overview of this approach in the context of our study; a more detailed description is found in the paper of Nychka *et al.* (Nychka et al., 2014).

We shall assume that the observed dose-rate at location s is given by:

$$Y(s) = X(s)^T \beta + g(s) + \varepsilon \quad (\text{B11})$$

where $X(s)^T \beta$ is the mean response, $g(s)$ is a spatial Gaussian process, and ε is a vector of independent zero-mean measurement errors. The unknown spatial process $g(s)$ is approximated by a sum of L independent processes:

$$g(s) = \sum_{l=1}^L g_l(s) \quad (\text{B12})$$

where $\text{var}[g_l(s)] = \sigma^2 \alpha_l$, and where σ^2 is a scaling term for the variance of the spatial process

$g(s)$, and α_l is a weighting term, selected so that $\sum_{l=1}^L \alpha_l = 1$. Increasing the number of independent processes (i.e., L) improves how accurately the sum of $g_l(s)$ approximates $g(s)$, but also increases the computational burden. Each independent component $g_l(s)$ is expressed in turn as a sum of (non-stochastic) radial basis functions $\phi_j^l(s)$ weighted by a vector of random coefficients, c_j^l , that is:

$$g_l(s) = \sum_{j=1}^{m(l)} c_j^l \phi_j^l(s) \quad (\text{B13})$$

where coefficients c_j^l are multivariate Normal, with mean zero and covariance $\text{cov}[c_j^l, c_m^l] = \sigma^2 (P^l)_{jm}$, and cross covariance 0, i.e., $\text{cov}[c_j^l, c_m^p] = 0, l \neq p$. As in the recent paper of Nychka *et al.* (Nychka et al., 2014), we shall assume that the $\phi_j^l(s)$ are the standard symmetric Wendland radial basis functions (Wendland, 1995).

Putting expressions (B11)-(B13) together, we therefore model $Y(s)$ approximately as:

$$Y(s) = X^T(s) \beta + \sum_{l=1}^L \sum_{j=1}^{m(l)} c_j^l \phi_j^l(s) + \varepsilon \quad (\text{B14})$$

where spatial covariance in $Y(s)$ is now derived from the covariance of the radial basis coefficients c_j^l . Thus, we can estimate covariances among spatial Gaussian process terms at any two locations $g(s_n)$ and $g(s_w)$ ($n \neq w = 1, \dots, N$) via:

$$\sigma^2 V_{nw} = \text{cov}[g(s_n), g(s_w)] = \text{cov}\left[\sum_{l=1}^L g_l(s_n), \sum_{k=1}^L g_k(s_w)\right] \quad (\text{B15})$$

$$\begin{aligned}
&= \sum_{l,k=1}^L \text{cov} \left[\sum_{j=1}^{m(l)} c_j^l \phi_j^l(s_n), \sum_{p=1}^{m(k)} c_p^k \phi_p^k(s_w) \right] \\
&= \sum_{l=1}^L \sum_{j,p=1}^{m(l)} \phi_j^l(s_n) \phi_p^l(s_w) \text{cov} [c_j^l, c_p^l] = \sigma^2 \sum_{l=1}^L \sum_{j,p=1}^{m(l)} \phi_j^l(s_n) \phi_p^l(s_w) (P^l)_{jp} \\
&= \sigma^2 \sum_{l=1}^L [\Phi^{lT} P^l \Phi^l]_{nw}
\end{aligned}$$

where the matrix $(\Phi^l_{pw}) = \phi_p^l(s_w)$ ($p, w = 1, m(l)$) indexes the radial functions by location.

Using the above notation we can write the distribution of the observed gamma-ray dose-rate as:

$$Y \sim MVN \left(X(s)^T \beta, \sigma^2 \sum_{l=1}^L \Phi^{lT} P^l \Phi^l + \tau^2 I \right) \quad (\text{B16})$$

where τ^2 is the variance of the uncorrelated error term ε . To derive the full likelihood it is helpful to re-parameterise the variance of $Y(s)$ in terms of the quantities:

$$\lambda = \tau^2 / \sigma^2 \quad (\text{B17})$$

$$\Sigma_\lambda = \sum_{l=1}^L \Phi^{lT} P^l \Phi^l + \lambda I$$

After doing so, the log-likelihood of $Y(s)$ becomes:

$$\begin{aligned}
l(Y(s)|\beta, \lambda, \sigma) &= -\frac{N}{2} \ln(2\pi) - N \ln(\sigma) - \frac{1}{2} \ln |\Sigma_\lambda| \\
&\quad - \frac{1}{2\sigma^2} (Y(s) - X(s)^T \beta)^T \Sigma_\lambda^{-1} (Y(s) - X(s)^T \beta)
\end{aligned} \quad (\text{B18})$$

so that the profile log-likelihood of $Y(s)$ (after maximisation with respect to σ) becomes:

$$l(Y(s)|\beta, \lambda) = C - N \ln \left[(Y(s) - X(s)^T \beta)^T \Sigma_\lambda^{-1} (Y(s) - X(s)^T \beta) \right] - \frac{1}{2} \ln |\Sigma_\lambda| \quad (\text{B19})$$

for a constant $C = -\frac{N}{2}[\ln(2\pi) - 2\ln(N) + 1]$. The approximate log-likelihood is maximised numerically with respect to β and λ to obtain their respective maximum-likelihood estimates, $\hat{\beta}$ and $\hat{\lambda}$.

We derive the predicted dose-rate at an unobserved location, s_0 , by considering $(Y(s_0), (Y(s_i))_{i=1}^N)$, which is multivariate Gaussian with mean $(X(s_0)^T \beta, (X(s_i)^T \beta)_{i=1}^N)$ and variance $\begin{bmatrix} \sigma^2 + \tau^2 & \sigma^2 r^T \\ \sigma^2 r & \sigma^2 V \end{bmatrix}$, and where the vector r is given by

$r = (r_n)_{n=1}^N = \sum_{l=1}^L \sum_{j,p=1}^{m(l)} \phi_j^l(s_n) \phi_p^l(s_0) (P^l)_{jp}$. Then it is easily shown (e.g., p.136 in (Diggle and Ribeiro, 2007)) that the conditional distribution of $Y(s_0)$ given $(Y(s_n))_{n=1}^N$ is also multivariate Gaussian

with mean:

$$\hat{Y}(s_0) = X(s_0)^T \hat{\beta} + r^T V^{-1} \left[Y(s_n) - X(s_n)^T \hat{\beta} \right]_{n=1}^N \quad (\text{B20})$$

All computation was carried out using the `LatticeKrig` package (Nychka, 2014) in R (R Project version 3.2.2, 2015).

References

- Appleton, J.D., 2005. Simplified geological classification for radon potential mapping in England and Wales (based on DiGMapGB-50 V3.12). British Geological Survey, Natural Environment Research Council, Keyworth, UK.
- Banerjee, S., Carlin, B.P., Gelfand, A.E., 2015. Hierarchical modeling and analysis for spatial data. Second Edition. Chapman and Hall/CRC, Boca Raton, FL.
- Carstairs, V., Morris, R., 1991. Deprivation and health in Scotland. Aberdeen University Press, Aberdeen.
- Cressie, N., Hawkins, D.M., 1980. Robust estimation of the variogram: I. Math. Geology 12, 115-125.
- Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. Royal Statist. Soc. Series B 70, 209-226.
- Cressie, N.A.C., 1993. Statistics for spatial data. John Wiley & Sons, Inc., New York.
- Diggle, P.J., Ribeiro, P.J., Jr, 2007. Model-based geostatistics, Springer Series in Statistics. Springer, 233 Spring Street, New York, NY 10013, pp. 1-228.
- Diggle, P.J., Ribeiro, P.J., Jr., Christensen, O.F., 2003. An introduction to model-based geostatistics, in: Møller, J. (Ed.), Spatial statistics and computational methods, pp. 43-86.
- Doss, M., Little, M.P., Orton, C.G., 2014. Point/Counterpoint: low-dose radiation is beneficial, not harmful. Med Phys 41, 070601.
- Evrard, A.S., Hemon, D., Billon, S., Laurier, D., Jougl, E., Tirmarche, M., Clavel, J., 2006. Childhood leukemia incidence and exposure to indoor radon, terrestrial and cosmic gamma radiation. Health Phys. 90, 569-579.
- Hawkins, D.M., Cressie, N., 1984. Robust kriging — A proposal. J. Int. Assoc. Math. Geology 16, 3-18.
- Kendall, G.M., Little, M.P., Wakeford, R., Bunch, K.J., Miles, J.C.H., Vincent, T.J., Meara, J.R., Murphy, M.F.G., 2013. A record-based case-control study of natural background radiation and the incidence of childhood leukaemia and other cancers in Great Britain during 1980-2006. Leukemia 27, 3-9.
- Kendall, G.M., Wakeford, R., Athanson, M., Vincent, T.J., Carter, E.J., McColl, N.P., Little, M.P., 2016. Levels of naturally occurring gamma radiation measured in British homes and their prediction in particular residences. Radiat Environ Biophys 55, 103-124.
- Little, M.P., Wakeford, R., Lubin, J.H., Kendall, G.M., 2010. The statistical power of epidemiological studies analyzing the relationship between exposure to ionizing radiation and cancer, with special reference to childhood leukemia and natural background radiation. Radiat. Res. 174, 387-402.
- Little, M.P., Wakeford, R., Tawn, E.J., Bouffler, S.D., Berrington de, G.A., 2009. Risks associated with low doses and low dose rates of ionizing radiation: why linearity may be (almost) the best we can do. Radiology 251, 6-12.
- Matérn, B., 1960. Spatial variation, Meddelanden fran Statens Skogsforsknings institut, Stockholm, pp. 1-144.
- Miles, J.C.H., Appleton, J.D., 2005. Mapping variation in radon potential both between and within geological units. Journal of radiological protection : official journal of the Society for Radiological Protection 25, 257-276.
- Nychka, D., 2014. R package 'LatticeKrig': multiresolution kriging based on Markov random fields. Version 3.4. Comprehensive R Archive Network (CRAN).
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S., 2014. A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Computat. Graph. Statist. 24, 579-599.
- Office of Population Censuses and Surveys (OPCS), 1984. Census 1981: Key Statistics for Urban Areas. . Her Majesty's Stationery Office (HMSO), London.
- Pebesma, E., Graeler, B., 2015. R package 'gstat'. Version 1.0-26. Comprehensive R Archive Network (CRAN).

R Project version 3.2.2, 2015. R version 3.2.2 <http://www.r-project.org/>. Comprehensive R Archive Network (CRAN).

Richardson, S., Monfort, C., Green, M., Draper, G., Muirhead, C., 1995. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Stat.Med.* 14, 2487-2501.

Rivoirard, J., 2001. Which models for collocated cokriging? *Mathematical Geology* 33, 117-131.

Smith, A., 2013. Digital Geological Map of Great Britain, information notes, 2013. British Geological Survey Open Report OR/13/007. British Geological Survey, Keyworth, Nottingham, UK.

Tubiana, M., Feinendegen, L.E., Yang, C., Kaminski, J.M., 2009. The linear no-threshold relationship is inconsistent with radiation biologic and experimental data. *Radiology* 251, 13-22.

United Kingdom Childhood Cancer Study Investigators, 2002. The United Kingdom Childhood Cancer Study of exposure to domestic sources of ionising radiation: 2: gamma radiation. *Br. J. Cancer* 86, 1727-1731.

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), 2008. UNSCEAR 2006 Report. Annex A. Epidemiological Studies of Radiation and Cancer. United Nations, New York, pp. 13-322.

Warnery, E., Ielsch, G., Lajaunie, C., Cale, E., Wackernagel, H., Debayle, C., Guillevic, J., 2015. Indoor terrestrial gamma dose rate mapping in France: a case study using two different geostatistical models. *Journal of environmental radioactivity* 139, 140-148.

Wendland, H., 1995. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Computat. Math.* 4, 389-396.

Wrixon, A.D., Green, B.M.R., Lomas, P.R., Miles, J.C.H., Cliff, K.D., Francis, E.A., Driscoll, C.M.H., James, A.C., O'Riordan, M.C., 1988. Natural radiation exposure in UK dwellings. National Radiological Protection Board, Chilton, Didcot, Oxon OX11 0RQ, UK.