HHS Public Access

Author manuscript

Environ Sci Technol. Author manuscript; available in PMC 2024 September 28.

Published in final edited form as:

Environ Sci Technol. 2023 November 21; 57(46): 18001–18012. doi:10.1021/acs.est.2c08840.

Predicting Monthly Community-level Radon Concentrations with Spatial Random Forest in the Northeast and Midwest United States

Longxiang Li[1], **Rebeca Ariel Stern**[1], **Eric Garshick**[2,3,4], **Carolina L. Zilli Vieira**[1], **Brent Coull**[1,5], **Petros Koutrakis**[1]

[1]Department of Environmental Health, Harvard T.H Chan School of Public Health, 401 Park Drive, Boston, MA 02114, USA

[2]Pulmonary, Allergy, Sleep, and Critical Care Medicine Section, VA Boston Healthcare System, 1400 VFW Pkwy, West Roxbury, Boston, MA 02132, USA.

[3]Channing Division of Network Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA.

[4]Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA.

[5]Department of Biostatistics, Harvard T.H Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

Abstract

In 1987, United States Environmental Protection Agency recommended installing a mitigation system when indoor concentration of radon, a well-known carcinogenic radioactive gas, is at or above 148 Bq/m^3. In response, tens of millions of short-term radon measurements have been conducted in residential buildings over the past three decades either for disclosure or to initially evaluate the need for mitigation. These measurements, however, are currently underutilized to assess population radon exposure in epidemiological studies. Based on two relatively small radon surveys, Lawrence Berkeley National Laboratory developed a state-of-the-art national radon model. However, this model only provides coarse and invariant radon estimations, which limits the ability of epidemiological studies to accurately investigate the health effects of radon particularly the effects of acute exposure. This study involved obtaining over 2.8 million historical short-term

Correspondence to Longxiang Li, Department of Environmental Health, Harvard T.H. Chan School of Public Health, Landmark Center 4th West, 401 Park Drive, Boston, Massachusetts, USA 02114 (lol087@mail.harvard.edu).

Author Contributions

Petros Koutrakis and Longxiang Li initiated the study. Longxiang Li synthesized data and performed the analysis. Longxiang Li developed the model. Longxiang Li, Carolina L. Zilli Vieira, and Petros Koutrakis collected the data. Longxiang Li, Brent A. Coull, and Petros Koutrakis interpreted the results and wrote the manuscript, with the help from Rebecca A. Stern, Eric Garshick, Carolina L. Zilli Vieira.

Declare of Competing Interest

The authors declare no conflict of interest.

Code availability

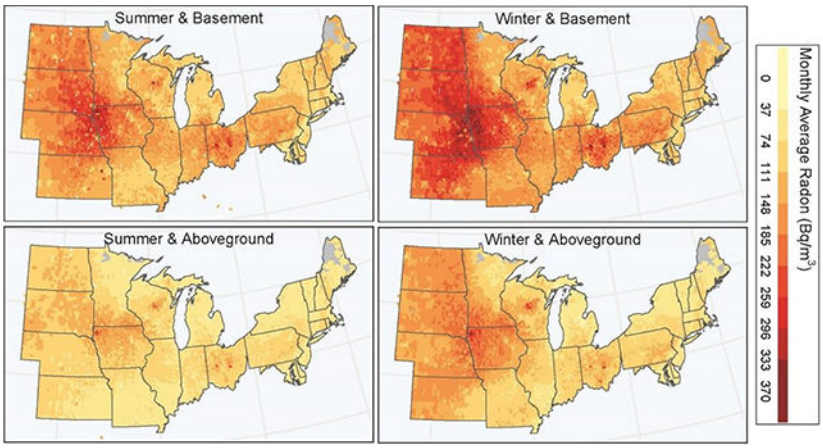
All model codes are available at the following link: https://github.com/longxiang1025/Radon_Mortality.

Supporting Information

Additional information regarding the sources of data, the spatiotemporal variance of predicting accuracy, the conceptual relationship between community-level observations and predictions, the workflow of model, the summary of observations, and the spatial heterogeneity in the importance of predictors, is provided in supplementary information free of charge.

radon measurements from independent laboratories. Using these measurements, an innovative spatial random forest (SRF) model was developed based on geological, architectural, socio-economical, and meteorological predictors. The model was used to estimate monthly community-level radon concentrations for ZIP Code Tabulation Areas (ZCTAs) in the Northeastern and Midwestern regions of the United States from 2001 to 2020. Via cross-validation, we found that our ZCTA-level predictions were highly correlated with observations. The prediction errors declined quickly as the number of radon measurements in a ZCTA increased. When ≥15 measurements existed, the mean absolute error was 24.6 Bq/m^3, or 26.5% of the observed concentrations (R^2=0.70). Our study demonstrates the potential of the large amount of historical short-term radon measurements that have been obtained to accurately estimate longitudinal ZCTA-level radon exposures at unprecedented levels of resolutions and accuracy.

Graphical Abstract:



Keywords

Radon; Exposure; Geographical; Map; Machine Learning

Introduction

Radon is the second leading cause of lung cancer after smoking in the United States, and it contributes to over 220,000 annual lung cancer mortalities worldwide [1]. Radon gas is naturally generated in Earth's crust and can move upwards into living spaces, accumulating to dangerous concentrations when ventilation is insufficient [2]. In 1987, the United States Environmental Protection Agency (EPA) recommended installing a radon mitigation system when the concentration in living space is at or above 148 Bq/m^3 (4 pCi/L) [3]. As a response, tens of millions of radon measurements have been collected in the U.S. either for disclosure or to initially evaluate the need for mitigation [4,5]. This extensive set of radon measurements has shown significant heterogeneity across regions and time [6–9], highlighting the importance of a detailed, spatially and temporally resolved model to estimate population radon exposure.

The Lawrence Berkeley National Laboratory developed the state-of-the-art national radon model for the U.S. (referred to as LBNL model) in the 1990s [10,11], using approximately 55,000 short-term measurements from the U.S. EPA/State Residential Radon Survey (SRRS) and 5,700 long-term measurements from the National Residential Radon Survey (NRRS) as data sources [12,13]. The LBNL model employed a Bayesian mixed-effects method to predict county-level, temporally invariant average radon concentrations based on geological provinces and building characteristics. Radon concentrations predicted by the LBNL model have been utilized in previous epidemiological studies to investigate the health effects of radon, including lung cancer [14], breast cancer [15], and chronic obstructive pulmonary disease [16].

However, the LBNL model has several limitations. First, the county-level prediction is too coarse to reflect the small-scale variations driven by geological and architectural factors [17–19]. Second, the temporally invariant radon predictions are unable to represent seasonal and long-term patterns [20,21]. Third, the radon measurements used by the LBNL model were primarily conducted in unmitigated buildings over thirty years ago, and therefore do not account for recent trends due to the expanding installation of mitigation systems [22] and improved energy efficiency in new buildings [23]. These limitations likely introduce exposure misclassifications to the subsequent epidemiological studies, complicating the interpretations of the any observed exposure-response relationships.

Our study addresses the limitations of the LBNL model by estimating monthly radon concentrations at the ZIP Code Tabulation Area (ZCTA) level in Northeast and Midwest U.S. Compared to county-level prediction, ZCTA is a much smaller spatial unit that better captures the small-scale variations driven by geological and architectural factors. All obtained radon measurements were conducted with short-term detectors that lasted 2 to 7 days because of the time sensitive nature of property transactions and initial evaluation [24]. The short durations, though less useful in estimating long-term exposures, are particularly suitable to model the short-term fluctuations in radon concentrations. Radon disclosure during property transactions is required in 17 out of 21 states and highly recommended in the remaining 4 states in the study region [25]. This de facto nonvoluntary and comprehensive participation guaranteed a large and representative sample of residential buildings, avoiding potential volunteer bias [26]. Based on part of the data, Li et al (2021) developed an ensemble learning model to estimate monthly ZCTA-level radon concentrations for ZCTAs in Greater Boston, a densely populated fraction of this study region [9]. Multiple studies used other classic statistical learning methods to model the spatial distribution of radon concentrations in different parts of the world [6,27–33]. In this study, we developed an innovative geographical machine learning method to model the complex relationships between ZCTA-level radon concentrations and various predictors. This approach was particularly useful for studying large and heterogeneous areas, where the relationships between radon and predictors may vary significantly across space. Our goal is to enhance radon prediction models, thus facilitating prospective heath studies regarding the health effects of radon.

Methods and Materials

Radon Measurements

Our study region consists of nine Northeastern and twelve Midwestern states of the U.S. The study region encompasses an area of 2.6 million square kilometers (32.1% of the contiguous U.S.) and a population of 127 million (38.2% of the contiguous U.S.). According to the Köppen-Geiger climate classification system [34], most of the study region is located in the humid continental climate zone, which is characterized by cold winters and semi-humid summers. Our study region covers areas of the highest geological radon potential in the U.S, such as southwestern Iowa and southwestern Pennsylvania. The study period is from 2001 to 2020.

We obtained 2,867,120 short-term radon measurements (Figure 1A and 1B) from Spruce Environmental Technologies, Inc (Haverhill, MA). These measurements were conducted using three types of passive radon detectors approved both by the National Radon Safety Board (NRSB) and NRPP (National Radon Proficiency Program): the Air Chek foil bag charcoal kit (NRSB #10333, NRPP # AC-8200), AccuStar charcoal canister (AccuStar PicoCan-400; NRSB #10320; NRPP # AC-1159), and liquid scintillation vial (AccuStar CLS-2; NRSB# 12193; NRPP # LS-8088). All three types of detectors use activated charcoal, contained in three different kinds of containers, to passively absorb the radon in the surrounding environment during a 2-to-7-day period. The integral radon concentration during the period can be subsequently estimated either with gamma spectrometry (Air Chek and PicoCan-400) or scintillation counter (CLS-2) shortly after the completion of measurement in the three laboratories operated by Spruce Environmental Technologies, Inc. The relative errors of all three short-term passive detectors were below 0.2 in NRSB- and NRPP-accredited chambers [35,36]. Li et al. (2023) reported a strong correlation (nearly 0.8) between a short-term measurement and a collocated follow-up 90-day radon measurement in a U.S-based study, suggesting that short-term measurements can serve as a good proxy for the average radon concentration in the following months [37].

Each radon measurement record in our dataset included the observed concentration in Bq/m^3, an encrypted street address, the ZCTA of the residence, the starting and ending date and time of the measurement, testing floor, and the type of radon detector used. To protect the privacy of consumers, actual street addresses were encrypted by Spruce Environmental Technologies, Inc using a checksum function, which converted them into a semi-random string while keeping the measurements in the same building identifiable. We did not investigate floor-dependent gradients across aboveground floors because all measurements conducted on non-basement floors were labeled as aboveground. We were also unable to differentiate between aboveground concentrations in single-family and multi-family buildings due to a lack of building type information from the data provider. The study is part of a larger Harvard T.H. Chan School of Public Health IRB-approved project to reassess national residential radon exposure based on over ten million short-term measurements and one million long-term measurements (IRB21–0056).

Data Processing

Three types of ZCTA-level radon concentrations: actual, observed, and predicted concentrations, were analyzed in this study. Actual ZCTA-level radon concentration (C_a) equals the geometric mean of all unit-specific radon concentrations within a ZCTA in a month. Owing to the logistic difficulties to measure all units within a ZCTA in a month, C_a is de facto unobservable. Observed ZCTA-level radon concentration (C_o), which is the geometric mean of the N sampled individual unit-specific radon measurements in the ZCTA and month, is used to approximate C_a. Predicted ZCTA-level radon concentration (C_p) is the predicted concentration for a ZCTA and month of our model. The objective of our model is to minimize the differences between C_p and C_a. But due to the unobservability of C_a, we had to evaluate C_p against C_o while accounting for the similarities between C_o and C_a. The conceptual relationships between C_a, C_o, and C_p are detailed in Supplementary Figure S1.

We collaborated closely with Spruce Environmental Technologies, Inc. to design the workflow for filtering and processing the original dataset of 2,867,120 rows of measurements (as shown in Supplementary Figure S2). To ensure the validity of the measurements, we excluded 46,034 readings that did not adhere to the prescribed measurement duration instructions. These instructions specified a duration of 3–7 days for AirChek, 2–4 days for both AccuStar PicoCan and AccuStar CLS, and 2–7 days for Alpha Energy. We then averaged concurrent measurements conducted in the same building on the same floor during the same period, a practice recommended in regions such as New Jersey. Measurements conducted before 2000 (n=***) were excluded from the analysis due to a lack of matching radon predictors. We removed *** radon measurements that were conducted within three months after the previous measurements in the same buildings. These follow-up measurements were either conducted to confirm the necessity of mitigation as recommended by U.S EPA or to evaluate the effectiveness of the mitigation. As a result, these follow-up measurements were no longer conducted in a random and representative sample of buildings in a ZCTA and should be excluded to avoid bias. **** measurements over 3,700 Bq/m^3 (99.95th percentile) were excluded due to rarity of the measurements and the overshadowing impacts on the following model training. Measurements under the lower detection limit (LDL) of detectors (n=***) were imputed with random numbers uniformly distributed between 0.01 Bq/m^3 and the detector-specific LDLs (AirChek: 3.7 Bq/m^3; AccuStar PicoCan and CLS: 14.8 Bq/m^3; Alpha Energy: 19 Bq/m^3). We subsequently took the geometric mean of all measurements conducted within the same ZCTA during the same month, if more than two measurements were conducted. At the end of data processing, we obtained *** observed monthly ZCTA-level radon concentrations, which were subsequently log transformed and used in the model. It is important to note that we did not have access to building-specific mitigation records. As a result, we included all radon measurements, regardless of mitigation status, as long as they were taken from a representative sample of buildings in a ZCTA. This included radon concentrations from mitigated buildings, if they were measured during property transactions. The impact of an increasing proportion of mitigated residences was reflected in the ZCTA-level observations, although the presence of a specific mitigation system in a building could not be identified.

Radon Predictors

We compiled a database of 81 predictors to estimate monthly ZCTA-level radon concentrations. These radon predictors were categorized into five classes: detector-related, geological, architectural, socioeconomical, and meteorological factors. Most of these covariates have been used in our previous study to predict the monthly ZCTA-level basement radon concentrations in Greater Boston [9]. A few predictors, such as the spatially lagged radon concentrations, were excluded due to their over-smoothing effects in rural areas where radon measurements were sparse. The sources and resolutions of the 81 predictors are summarized in Supplementary Table S1.

Detector-related factors control for the difference among measurement due solely to the detector, independent of the other covariates. For each ZCTA-level observation, we calculated the proportions of measurements conducted in the basement, in aboveground spaces, and with three types of detectors. Incorporating the proportions of measuring floor enabled the model to detect floor-dependent gradients in concentrations, and subsequently predict radon concentrations in basement and aboveground floors respectively. Similarly, the proportions of three types of detectors allowed us to model the patterns in radon concentrations caused jointly by the systematic difference among detectors and regional market shares, therefore removing these artificial trends in the predicted ZCTA-level radon concentrations.

Geological factors govern the generation, emanation, and underground movement of radon, and therefore, determine the radon concentration in the soil gas [39]. These factors include: the ground surface concentration of Uranium-238, the parent element of radon, to account for the generation of radon [40]; radon potential, which accounts for the abundance of radon in soil gas [41]; the age and granularity of surficial materials, which influence the generation and emanation of radon in soil [42]; distance to the nearest geological fault, which account for the vertical movement of radon via fracture [7], magnetic and gravity anomality, which account for the underground structure [43]; and soil parameters, including available water capacity, percent of organic matter, saturated hydraulic conductivity, vertical permeability, bulk density, field capacity, porosity, erodibility, depth and the percent of soil components of different granularity (gravel, sand and clay) [42].

Architectural factors influence the infiltration of radon from soil to the indoor environment [44]. We also incorporated covariates related to other domestic radon sources such as construction materials, drinking water and natural gas as cooking or heating fuel. Four included factors are: building size, as a proxy of foundation depth, which governs the pressure difference between soil and domestic environment; building density, as a proxy of whether the dwelling has access to municipal water supply, which has lower radon level or relies on private well that generally has a higher radon concentration [45]; average building age, which accounts for the prevailing building structure, the construction materials applied and the condition of basement [46], and type of fuel, which accounts for the potential impact of heating or cooking with natural gas which has an elevated concentration of radon [47].

Socioeconomical factors likely influence energy efficiency and the prevalence of radon mitigation [19]. The covariates in our model are median property value, mean household

annual income, percent of residents living below poverty line, and percent of properties occupied by owner.

Meteorological factors influence the exhalation of radon gas from soil to the atmosphere, underground radon movement, and atmospheric radon concentrations. Included factors are: snow depth and accumulated precipitation, which depress radon exhalation [48]; barometric pressure, which is negatively related to radon exhalation [48]; soil temperature, which impacts radon concentration in soil gas via a temperature-sensitive gaseous/aqueous partitioning process [49]; soil moisture content, which can slow down the subterranean flow of soil gas and impact the emanation rate of radon [42], and gross beta radiation measurements at nearby RadNet monitors operated by U.S. EPA. Gross beta radiation is the sum of beta radiation from beta-emitting radionuclides that are bound to the ambient particulates [50]. Most of beta radiation is emitted by Pb-210, a decay product of radon gas [51]. Gross beta radiation can therefore be used as a proxy for atmospheric radon concentration, which has not been widely measured [52].

We included the calendar year to account for long-term trends and month of the year to account for seasonality that has not been fully captured by meteorological factors.

Data Analysis

We utilized a modified version of the random forest model called Spatial Random Forest (SRF) to predict the monthly average radon concentrations for each ZCTA in our study region. The original random forest method combines predictions from multiple decision tree models, each using a random subset of covariates and training dataset [53] This ensemble approach helps prevent the overfitting of individual decision trees, leading to a better overall accuracy. SRF builds upon the original model by incorporating two key characteristics of geographical phenomena: heterogeneity and dependence [54,55]. Geographical heterogeneity, also known as nonstationarity, occurs when the mechanisms affecting the phenomenon are not homogeneous across space and time [56]. This can lead to variation in the relationships between predictors and the outcome variable, as well as in the distribution of the outcome variable itself[56]. In the context of radon estimation, spatiotemporal heterogeneity leads to the varying performance of a universal prediction model in different regions during different periods. SRF addresses this issue by allowing the model to capture these nonstationarities and adapt to the local relationships between predictors and the outcome variable via place-based modelling. Geographic dependence was expressed by Tobler as “nearby things are more similar than distant things” [57]. The positive spatial correlation between radon measurements observed in neighboring communities is unable to be effectively modeled in the original random forest model that is more suitable for independent measurements [55]. SRF addresses this issue by weighing radon measurements based on proximity. SRF has been successfully used to estimate population density in Sub-Saharan Africa [54].

We fit local random forest sub-models to account for geographical heterogeneity using the place-based method. This approach was first introduced in geographical weighted regression and offers a flexible method to model the spatially varying relationship [58]. For each ZCTA, we fitted a local random forest sub-model based on a subset of training dataset that consists of all $\boldsymbol{C}_o$ within r km. Local random forest sub-models were built using a subset of the

81 radon predictors, which were selected based on their importance. The importance of each predictor was determined by measuring the relative increase in mean absolute error when an alternative random forest model was fitted using the predictor randomly permuted. Predictors that produced a significant increase in error were deemed more important than those that did not. Predictors whose permutation resulted in decreases in mean square error smaller than 0.01 log-transformed Bq/m^3 in each local random forest model were excluded. Consequently, the number and rank of important predictors varied depending on the local relationship between radon and the predictors. The permutation and selection process were carried out using the ranger package [59]. During the development of each tree, half of the selected important radon predictors were randomly chosen for each split. All other tuning parameters, including the number of trees (100), minimum size of node (1), and node splitting rule (minimize the residual), were kept constant across all local random forest models. Driven by the different training datasets, local random forest sub-models varied if local interactions among radon predictors changed.

The geographical dependence was modeled in each local random forest sub-model by weighting $\boldsymbol{C}_o$ according to their geographical distances to the point of prediction [60]. The weight of each $\boldsymbol{C}_o$ determined its likelihood of being selected in fitting individual decision trees; therefore, modifying the random sampling method in original random forest method. Nearby $\boldsymbol{C}_o$, relative to distant $\boldsymbol{C}_o$, had greater weights and were more frequently used to estimate $\boldsymbol{C}_p$ for the ZCTA upon which local sub-model is centered. The weight of one $\boldsymbol{C}_o$ $\left(j^{\text{th}}\right)$ in calculating $\boldsymbol{C}_p$ $\left(i^{\text{th}}\right)$ was calculated with the following formula:

$$w_j(i,j;b) = \frac{N_j}{\sigma_j} \times \frac{1}{\left(\sqrt{2\pi}b\right)} e^{-\frac{1}{2}\left(\frac{dist(i,j)}{b}\right)^2}$$

where b is the bandwidth parameter of the Gaussian kernel function that governs the decay in weights as the geographical distance between two points increases. The bandwidth parameter determines the rate at which the weight decreases. A larger b means a sharper decrease in weight across space, and greater relative leverages of nearby observations. N_j is the number of radon measurements based on which a $\boldsymbol{C}_o$ is calculated. σ_j is the standard deviation of the j-th ZCTA-level observation. We found that the performance of local random forest models relied on two parameters: r and b. We set r and b to 150 km and 75 km, respectively, after evaluating the performance of models with different combinations of parameters.

To evaluate the performance of the model, we employed the leave-one-out cross validation (LOOCV) method after converting the predicted log-transformed concentrations back to their original form. Each $\boldsymbol{C}_o$ was iteratively used as a testing set and predicted by a training set consisting of the remaining nearby observations. The local splitting of training and testing dataset guaranteed that each $\boldsymbol{C}_o$ was not used to predict itself; therefore, avoiding data leakage issues. We evaluated the performance by calculating three metrics: Mean Absolute Error (MAE), Mean Error (ME), and Mean Relative Error (MRE). The difference between $\boldsymbol{C}_p$ and $\boldsymbol{C}_o$ was separated into two sections: the difference between $\boldsymbol{C}_p$ and $\boldsymbol{C}_a$ and the

difference between C_a and C_o. The difference between C_a and C_o was a joint function of the sample size of C_o and the within-ZCTA variation in unit-specific radon concentrations. C_o based on a larger sample was more likely to represent C_a than a similar C_o based on a smaller sample. C_o can be used as a reliable estimation of C_a when C_o was based on a large sample. We, therefore, restricted the evaluation of C_p against C_o whose sample size was over a cutoff value (n). The n was determined by stratifying the pairwise difference between C_p and C_o according to N, investigating the trend of differences against N and locating a value above which the difference between C_p and C_o did not decrease. We furthermore fitted simple linear regression models to investigate the Pearson's correlation between C_p and C_o $\left(r^2\right)$. The R^2 of the linear models was used to evaluate the degree to which the variance of C_o can be explained by C_p. Local LOOCV-based metrics were also calculated to investigate the spatiotemporal patterns in the performance.

The prediction process followed a similar approach to the evaluation process, but with three key differences. Firstly, we assumed that the three types of detectors had equal proportions in order to remove spatial trends caused by regional differences in their market shares and systematic differences in performance. Secondly, we set the proportion of measurements in the basement to 100% and 0% to produce two separate prediction values for basement and aboveground floors. Finally, we used non-parametric bootstrapping to evaluate the uncertainty and the potential risk of overfitting for each C_p. This involved randomly resampling the nearby C_os with replacement, fitting local sub-models based on the resampled training sets, and subsequently generating predictions. We calculated 50 alternative predictions via resampling, then calculated the standard deviation of the predicted concentrations. When a local sub-model overfitted the local training dataset, the local predictions based on the randomly resampled local training dataset tended to vary because approximately one third of the training dataset were replaced. On the contrary, if a local sub-model characterized the underlying mechanism, the local predictions based on the randomly resampled local training dataset were insensitive to these changes. Local fitting (R^2) of each sub-model was also reported to further evaluate the risk of overfitting. A conjunction of high R^2 and great standard deviation of bootstrap-based predictions suggested higher risk of overfitting.

Results

The observed radon concentration in basements had a median value of 92.5 Bq/m^3 and an interquartile range (IQR) of 37.0–196.1 Bq/m^3, higher than the median concentration in aboveground floors (59.2 Bq/m^3, IQR: 25.9–129.5 Bq/m^3). The proportion of basement measurements that were above action level is 34.3%, higher than the corresponding proportion in aboveground floors (22.0%). The observed radon concentrations at the ZCTA level exhibit a log-normal distribution, as shown in Supplementary Figure S3. Observed radon concentrations in the basement (Figure 1A) and aboveground floors (Figure 1B) had similar spatial patterns. Midwestern states, especially those in the West North Central region, have higher observed radon concentrations than the rest of the study regions (Table 1). The region with the highest observed radon concentration was the northwestern part of Iowa, in agreement with the results of the Iowa Lung Cancer Study [61,62].

We observed similar year-to-year declining trends in the basement and aboveground radon concentrations (Supplementary Figure S4). However, the year-to-year declines in radon concentrations were not monotonic. Radon concentrations decreased from 2001 to 2004, increased slightly during 2005–2010, and resumed decreasing until the end of study period. The seasonal patterns of monthly average radon concentrations on both floors were similar (Supplementary Figure S4). Highest monthly concentrations of each year were always observed in the winter, and lowest monthly concentrations were always observed in the summer. The average concentrations measured by AirChek were greater than those by the other two AccuStar detectors (Table 1), probably because of the different market share across regions. As shown in Supplementary Figure S5, AccuStar detectors were more commonly used in Northeastern states, especially New England, where the average radon concentrations were lower than Midwestern states.

Our ZCTA-level radon predictions (C_p) were highly correlated with the observations (C_o), especially when C_o were based on large samples of radon measurements. The predicted ZCTA-level basement and aboveground radon concentrations both follow log-normal distributions (Supplementary Figure S3). The overall MAE of C_p against C_o is 32.6 Bq/m^3, with an r^2 of 0.51 (Supplementary Table S2). By stratificating MAE with the sample size of C_o (N), we found that the difference between C_o and C_p decreased sharply when N increased from 5 to 15, then continued decreasing with smaller and inconsistent gradients when N was ≥ 15 (Figure 2A). Meanwhile, the correlation between C_o and C_p increased as N increased from 5 to 15 (Figure 2B–D). We observed a similar size-dependent pattern in the original (non-transformative) scale (Supplementary Figure S6). These patterns were likely driven by the diminishing difference between C_o and C_a as N increased from 5 to 15 as the difference between C_a and C_p was theoretically independent of N. When $N \geq 15$, the difference between C_o and C_a contributed a relatively smaller fraction of the difference between C_o and C_p, allowing us to approach the "true" prediction error (difference between C_p and C_a) with better accuracy. We therefore focused on comparing C_p with the corresponding C_o that were based on ≥15 radon measurements. The MAE of C_p against C_o based on ≥15 radon measurements was 24.6 Bq/m^3, much lower than that for C_o based on 5 radon measurements (32.6 Bq/m^3). The ME and MRE values for C_o based exclusively on ≥15 radon measurements were −5.6 Bq/m^3 and 26.5%, respectively, both smaller than the metrics for C_o based on smaller samples (Supplementary Table S2).

The results of the cross-validation analysis were further stratified based on subregion, season, long-term period, and population density to detect varying levels of accuracy (Supplementary Table S2). Our model had the smallest prediction error in the New England area, with a MAE of 17.1 Bq/m^3 and an MRE of 23.5%. The Mid-Atlantic region had the largest MAE of 31.5 Bq/m^3 and MRE of 35.4%. Summer had the lowest seasonal MAE of 16.9 Bq/m^3, while winter and autumn had the greatest seasonal MAE of 26.8 Bq/m^3. However, the seasonal r^2 between C_o and C_p in summer (0.43) was lower than the correlations in autumn (0.69) and winter (0.74), suggesting that the smaller MAEs in summer were likely due to low seasonal radon concentrations. No consistent long-term trends in the performance of our prediction model were observed. We used a threshold

value of 25,000 people per square mile to categorize ZCTAs as urban or non-urban [63]. Non-urban ZCTAs generally had greater MAEs and r^2 values than urban ZCTAs because radon measurements in non-urban buildings are more likely to be taken in the basement, where radon concentrations are generally higher than in aboveground spaces. This means that radon measurements in non-urban areas were less likely to be lower than the LDL and thus, less likely to be imputed with a random number between 0 and LDL. The trends in the accuracy of ZCTA-level radon estimations are similar in both urban and non-urban areas with regards to sample size, as shown in the Supplementary Figure S6.

Figure 1C and 1D showed the spatial distribution of predicted floor-specific ZCTA-level radon concentrations and the affiliated relative standard deviation across our study region. The spatial patterns in the predicted concentrations were largely in agreement with those in the original observation (Figure 1A and 1B). The relative standard deviations (Figure 1E and 1F) were greater in border areas, such as in the northwestern corner, where sub-models tended to rely on smaller local samples. The predicted concentrations in the basement (Figure 1C) were generally greater than the predictions for aboveground floors (Figure 1D), in agreement with the original observations (Table 1). The relative standard deviations in aboveground predictions (Figure 1F) were commonly greater than those in basement predictions (Figure 1E), likely because of the lower aboveground radon concentrations and similar prediction uncertainties. Figure 3 showed the seasonal distributions of predicted floor-specific ZCTA-level radon concentrations across our study region. The patterns in Figure 3 agreed with the observed seasonal variation shown in Supplementary Figure S4. Predicted radon concentrations in the summer (Figure 3A) were generally lower than those in the winter (Figure 3B). The predicted radon concentrations for basements were greater than the predicted radon concentrations in aboveground floors (Supplementary Figure S8)

By aggregating the importance of radon predictors in all sub-models, we found that geological radon potential was the most important covariate to predict the spatiotemporal distribution of radon. Excluding geological radon potential alone inflated the RMSE by 5.4% on average. The other most important radon predictors are: calendar year (4.3%), atmospheric temperature (4.1%), soil temperature (3.9%), gravity anomalies (3.6%), percent of measurements in the basement (3%), distance to the closest geological fault (2.8%), elevation (2.7%), barometric pressure (2.5%), population density (2.3%), and percent of units fueled by natural gas (2.0%). Figure 4 shows the heterogeneity of the importance of all radon predictors across local random forest sub-models. As shown in Figure 4A, the importance of geological radon potential was greater in northern Illinois and Ohio than other parts of the study region. The importance of the distance to the closest active geological fault was greater in border regions between Iowa and Indiana and the border regions between New York and Pennsylvania (Figure 4F) than other areas. The heterogeneity suggests that some features of the geological fault, such as depth and age, which were not characterized by the simple distance, also likely drive the spatial distribution of radon. Meanwhile, the importance of gravity anomaly was greater in southern Pennsylvania and New Jersey (Figure 4E), suggesting that the underground variation in density influenced the local distribution of radon. The importance of calendar year was greater in eastern Wisconsin (Figure 4B) than other areas, suggesting a greater year-to-year variation that count not be modelled by other meteorological factors. The variance in the importance of meteorological factors, such as

atmospheric temperature (Figure 4C), soil temperature (Figure 4D), and barometric pressure (Figure 4H), were generally smaller than the geological factors. Figure 4I showed the gradients in the importance of the percents of units fueled by natural gas. The importance in northern and western Pennsylvania was greater, suggesting that natural gas in these areas was more likely to influence indoor radon concentrations than other areas.

The average local fitting (R^2) of local sub-models was 0.32 with an interquartile range of 0.24 to 0.27. The local fitting was generally lower than the overall fitting across the study region, which additionally accounted for large-scale agreement. Local fitting also showed strong spatial heterogeneity (Supplementary Figure S9). Areas with greater local R^2 were southern Pennsylvania, western New York, northern Ohio, eastern Wisconsin, and middle Nebraska. None of them overlapped with the areas with greater relative standard deviation (Figure 1E and 1F), suggesting a low risk of overfitting.

Discussion

We developed a geographical machine learning model to predict the ZCTA-level monthly radon concentrations for Northeastern and Midwestern U.S. based on nearly 3 million short-term radon measurements. Our model is able to predict ZCTA-level radon concentrations with a MAE of 24.5 Bq/m^3 and a MRE of 26.3%, suggesting that the predicted ZCTA-level concentrations correlate well with the observed ZCTA-level concentrations. Compared with LBNL model, our model has higher spatiotemporal resolutions, and is based on a more recent and larger sample, therefore, can be used to enhance the assessment of residential exposure to radon.

We compared our predictions with the results of LBNL, after averaging the monthly ZCTA-level predictions by county. Our predicted concentrations were moderately correlated to the LBNL predictions with an r^2 of 0.47. Additionally, our predicted concentrations were significantly higher than those of LBNL model with an average difference of 46.0 Bq/m^3 (95% CI: 44.2–47.8 Bq/m^3). This difference seemed contradictory to the observed pattern that radon concentrations decreased during the study period (Supplementary Figure S4). However, the pattern is likely caused by the differences in measuring protocols, measuring devices, and modeling methods. In specific, U.S. EPA did not implement the current short-term radon testing protocol, which requires close-building conditions, until the completion of two national radon surveys ahead of LBNL model [24,38]. Diffusion barrier is a regular design to lower the sensitivity to humidity and temperature conditions in all three types of detectors of our study. However, it was uncommonly used when the national radon surveys, on which LBNL model was based, were conducted in the 1980s [64]. Finally, the validity of Bayesian regression method used in the LBNL model depended on several strong assumptions, such as stationarity, linearity, and no interactions among radon predictors. The SRF method used in this study is more flexible in accommodating the spatially varying and nonlinear relationships, and the complex interactions among radon predictors.

We observed spatial heterogeneity in the relationship between measured radon concentrations and their predictors using sub-models (Figure 4). Geological radon potential, the most significant predictor on average, did not hold equal importance across the region,

illuminating the various mechanisms governing local variations in radon concentrations. For example, natural gas from cooktops sourced from nearby fields generally exhibited higher radon concentrations than gas transported via long-distance pipelines or liquified natural gas cargos, owing to radon's relatively short half-life (3 days). Consequently, natural gas from the Marcellus and Devonian shales in northern and southwestern Pennsylvania was more likely to influence distributions in adjacent areas than in more distant locations, despite equal reliance on natural gas, corroborating our observations (Figure 4I). [65] However, regional heterogeneity in predictor importance might also be attributed to the varying quality of radon predictors across the area. The diminished significance of coarsely gridded predictors, such as meteorological factors, could underestimate their actual importance since a sub-model only encompassed a portion of the study region.

Our study is the first of its kind to take advantage of the extensive previously underutilized short-term radon measurements to estimate temporally resolved community-level radon concentrations in a large area. Our model was built upon an existing data source, therefore was an economical way to enhance the assessment of residential exposure to radon. While individual short-term radon measurements may not be as accurate as a long-term measurement, their large sample size and the mandatory disclosure during property transactions make them more representative of the population than the LBNL model, which was based on only about 12 voluntary measurements per county, less than 1% of our sample on average. The improved data source allowed us to model the varying radon concentrations on finer spatial and temporal resolutions than those of LBNL model. Furthermore, the large samples enabled us to make floor-specific predictions, which can likely be used to lower the uncertainties in exposure assessment jointly with residential behavior records [66]. We modified the classic random forest method to account for the geographical heterogeneity and dependence. The place-based method enabled us to make local predictions based only on nearby observations, instead of by reference to all observations most of which were distant and likely to have different relationships among radon predictors. The place-based method can also lower the requirement for computation resources for local prediction and facilitated a fast implementation via parallel computing.

Limitations exist in our study. First, measurements by different types of detectors were used in our study. The accuracy and precision of different types of detectors varies under different circumstances, consequently complicating the ZCTA-level aggregation. To address this limitation, we used the proportions of three types of detectors as radon predictors to model the detector-specific difference. The predicted concentration was a synthetic concentration based on three detector-specific predictions under the assumption that three types of detectors have identical proportions and therefore are less likely to be biased than an unadjusted prediction. Second, building-level radon concentrations cannot be predicted with our model due to a lack of building-specific geological and architectural information in the training data set. We used ZCTA-level geological, architectural, and socioeconomical factors in the model, likely unable to fully represent the influences of these factors on individual buildings. As a result, our ZCTA-level predicted concentrations should not be generalized to individual buildings within the ZCTA due to varying geological and architectural factors. Third, aboveground floor-dependent gradients in multi-family buildings were not investigated. Multi-family buildings, such as high-rise residential buildings, are

getting increasingly popular in densely populated areas. But floor information (basement or aboveground) is only available for single-family buildings in our data. Previous studies have found that variation of radon among aboveground floors is generally smaller than the difference between basement and aboveground floors [46]. As a result, aggregating aboveground measurements regardless of building type has low risk to seriously bias the ZCTA-level observations. Further tuning of the machine learning model, including trying other learning algorithms, experimenting with different combinations of parameters in local random forest models, may lead to additional improvements in prediction accuracy. However, these potential refinements are beyond the scope of our study and are unlikely to alter our primary conclusions.

The results of our study have illustrated the feasibility of applying extensive short-term radon measurements in assessing residential exposure to radon. Our model can be readily expanded to include more measurements from different providers without requiring much more computing resources, making it a candidate for centralizing all radon measurements from the past three decades. Currently, much less is known about the nonmalignant effects of radon, other than its well-known relationship with lung cancer. The accurate longitudinal exposure assessment, as presented here, can be used in future studies to investigate the nonmalignant effects of radon, thus improving our understanding of this omnipresent air pollutant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

Shawn Price, Director of Laboratory Operations at Spruce Environmental Technologies, Inc, provided us with the short-term radon measurements used in this study and reviewed the manuscript. This publication is made possible by U.S. EPA grant RD-835872, NIH grant R21ES029637, and NIH grant K99ES034459. Its contents are solely the responsibility of the grantee and do not necessarily represent the official view of the U.S. EPA or NIH. Further, U.S. EPA and NIH do not endorse the purchase of any commercial products or services mentioned in the publication.

Data availability

All data sources are summarized in Supplementary Information Table. S1. The final research data will be provided upon reasonably requested.

References

(1). Gaskin J; Coyle D; Whyte J; Krewksi D Global Estimate of Lung Cancer Mortality Attributable to Residential Radon. Environ Health Perspect 2018, 126 (5). 10.1289/EHP2503.

(2). National Council on Radiation Protection and Measurements (NCRP). Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2; The National Academies Press: Washington D.C, 2006.

(3). Indoor Radon Abatement Act (Title 15 United States Constitution § 2661–2671); Unites States House of Representatives: United States, 1988.

(4). George AC The History, Development and the Present Status of the Radon Measurement Programme in the United States of America. Radiat Prot Dosimetry 2015, 167 (1–3), 8–14. 10.1093/rpd/ncv213. [PubMed: 25911413]

(5). Gregory B; Jalbert PP National Radon Results: 1985 to 2003; Washington, DC, 2004.

(6). Haneberg WC; Wiggins A; Curl DC; Greb SF; Andrews WM; Rademacher K; Rayens MK; Hahn EJ A Geologically Based Indoor-Radon Potential Map of Kentucky. Geohealth 2020, 4 (11), e2020GH000263. 10.1029/2020GH000263.

(7). Dai D; Neal FB; Diem J; Deocampo DM; Stauber C; Dignam T Confluent Impact of Housing and Geology on Indoor Radon Concentrations in Atlanta, Georgia, United States. Science of the Total Environment 2019, 668, 500–511. 10.1016/j.scitotenv.2019.02.257. [PubMed: 30852225]

(8). Casey JA; Ogburn EL; Rasmussen SG; Irving JK; Pollak J; Locke PA; Schwartz BS Predictors of Indoor Radon Concentrations in Pennsylvania, 1989–2013. Environ Health Perspect 2015, 123 (11), 1130–1137. 10.1289/ehp.1409014. [PubMed: 25856050]

(9). Li L; Blomberg AJ; Stern RA; Kang C-M; Papatheodorou S; Wei Y; Liu M; Peralta AA; Vieira CLZ; Koutrakis P Predicting Monthly Community-Level Domestic Radon Concentrations in Greater Boston Area with an Ensemble Learning Model. Environ Sci Technol 2021, 55 (10), 7157–7166. [PubMed: 33939421]

(10). Price PN Predictions and Maps of County Mean Indoor Radon Concentrations in the Mid-Atlantic States. Health Phys 1997, 72 (6), 893–906. [PubMed: 9169931]

(11). Price P; Nero A Mapping of Mean Radon Concentrations, Using Survey Data and Covariates. In International radon symposium. Lawrence Berkeley National Laboratory; 1996.

(12). White SB; Alexander B. v ; Rodman NF Predicting the Annual Concentration of Indoor 222Rn from One or More Short-Term Measurements. Health Phys 1994, 66 (1), 55–62. [PubMed: 8253579]

(13). Marcinowski F; Lucas RM; Yeager WM National and Regional Distributions of Airborne Radon Concentrations in U.S. Homes. Health Phys 1994, 66 (6).

(14). Turner MC; Krewski D; Chen Y; Pope III CA; Gapstur S; Thun MJ Radon and Lung Cancer in the American Cancer Society Cohort. Cancer Epidemiology, Biomarkers & Prevention 2011, 20 (3), 438–448. 10.1158/1055-9965.EPI-10-1153.

(15). VoPham T; DuPré N; Tamimi RM; James P; Bertrand KA; Vieira V; Laden F; Hart JE Environmental Radon Exposure and Breast Cancer Risk in the Nurses' Health Study II. Environmental Health 2017, 16 (1), 97. 10.1186/s12940-017-0305-6. [PubMed: 28882148]

(16). Turner MC; Krewski D; Chen Y; Pope CA; Gapstur SM; Thun MJ Radon and COPD Mortality in the American Cancer Society Cohort. European Respiratory Journal 2012, 39 (5), 1113–1119. 10.1183/09031936.00058211. [PubMed: 22005921]

(17). Ball TK; Cameron DG; Colman TB; Roberts PD Behaviour of Radon in the Geological Environment: A Review. Quarterly Journal of Engineering Geology 1991, 24 (2), 169–182. 10.1144/GSL.QJEG.1991.024.02.01.

(18). Barros-Dios JM; Ruano-Ravina A; Gastelu-Iturri J; Figueiras A Factors Underlying Residential Radon Concentration: Results from Galicia, Spain. Environ Res 2007, 103 (2), 185–190. 10.1016/J.ENVRES.2006.04.008. [PubMed: 16729995]

(19). Kendall GM; Miles JCH; Rees D; Wakeford R; Bunch KJ; Vincent TJ; Little MP Variation with Socioeconomic Status of Indoor Radon Levels in Great Britain: The Less Affluent Have Less Radon. J Environ Radioact 2016, 164, 84–90. 10.1016/j.jenvrad.2016.07.001. [PubMed: 27442258]

(20). Sesana L; Caprioli E; Marcazzan GM Long Period Study of Outdoor Radon Concentration in Milan and Correlation between Its Temporal Variations and Dispersion Properties of Atmosphere. J Environ Radioact 2003, 65 (2), 147–160. 10.1016/S0265-931X(02)00093-0. [PubMed: 12527232]

(21). Miles JCH; Howarth CB; Hunter N Seasonal Variation of Radon Concentrations in UK Homes. Journal of Radiological Protection 2012, 32 (3), 275. 10.1088/0952-4746/32/3/275. [PubMed: 22809737]

(22). Vázquez BF; Adán MO; Quindós Poncela LS; Fernandez CS; Merino IF Experimental Study of Effectiveness of Four Radon Mitigation Solutions, Based on Underground Depressurization, Tested in Prototype Housing Built in a High Radon Area in Spain. J Environ Radioact 2011, 102 (4), 378–385. 10.1016/J.JENVRAD.2011.02.006. [PubMed: 21382656]

(23). Symonds P; Rees D; Daraktchieva Z; McColl N; Bradley J; Hamilton I; Davies M Home Energy Efficiency and Radon: An Observational Study. Indoor Air 2019, 29 (5), 854–864. 10.1111/INA.12575. [PubMed: 31127966]

(24). United States Environmental Protection Agency. Protocols for Radon and Radon Decay Product Measurements in Homes; 1993.

(25). Elizabeth Ann Glass Geltman. State Radon Laws https://lawatlas.org/datasets/state-radon-laws (accessed 2021-11-28).

(26). Tripepi G; Jager KJ; Dekker FW; Zoccali C Selection Bias and Information Bias in Clinical Research. Nephron Clin Pract 2010, 115 (2), c94–c99. 10.1159/000312871. [PubMed: 20407272]

(27). Timkova J; Fojtikova I; Pacherova P Bagged Neural Network Model for Prediction of the Mean Indoor Radon Concentration in the Municipalities in Czech Republic. J Environ Radioact 2017, 166, 398–402. 10.1016/j.jenvrad.2016.07.008. [PubMed: 27440462]

(28). Kropat G; Bochud F; Jaboyedoff M; Laedermann JP; Murith C; Palacios M; Baechler S Improved Predictive Mapping of Indoor Radon Concentrations Using Ensemble Regression Trees Based on Automatic Clustering of Geological Units. J Environ Radioact 2015, 147, 51–62. 10.1016/j.jenvrad.2015.05.006. [PubMed: 26042833]

(29). Ambrosino F; Sabbarese C; Roca V; Giudicepietro F; Chiodini G Analysis of 7-Years Radon Time Series at Campi Flegrei Area (Naples, Italy) Using Artificial Neural Network Method. Applied Radiation and Isotopes 2020, 163, 109239. 10.1016/J.APRADISO.2020.109239. [PubMed: 32561065]

(30). Sabbarese C; Ambrosino F; D'Onofrio A; Pugliese M; la Verde G; D'Avino V; Roca V The First Radon Potential Map of the Campania Region (Southern Italy). Applied Geochemistry 2021, 126, 104890. 10.1016/J.APGEOCHEM.2021.104890.

(31). Panahi M; Yariyan P; Rezaie F; Kim SW; Sharifi A; Alesheikh AA; Lee J; Lee J; Kim S; Yoo J; Lee S Spatial Modeling of Radon Potential Mapping Using Deep Learning Algorithms. 10.1080/10106049.2021.2022011 2022. 10.1080/10106049.2021.2022011.

(32). Rezaie F; Panahi M; Bateni SM; Kim S; Lee J; Lee J; Yoo J; Kim H; Won Kim S; Lee S Spatial Modeling of Geogenic Indoor Radon Distribution in Chungcheongnam-Do, South Korea Using Enhanced Machine Learning Algorithms. Environ Int 2023, 171, 107724. 10.1016/J.ENVINT.2022.107724. [PubMed: 36608375]

(33). Petermann E; Meyer H; Nussbaum M; Bossew P Mapping the Geogenic Radon Potential for Germany by Machine Learning. Science of The Total Environment 2021, 754, 142291. 10.1016/J.SCITOTENV.2020.142291. [PubMed: 33254926]

(34). Peel MC; Finlayson BL; McMahon TA Updated World Map of the Köppen-Geiger Climate Classification. Hydrol. Earth Syst. Sci. 2007, 11 (5), 1633–1644. 10.5194/hess-11-1633-2007.

(35). National Radon Proficiency Program. Manufacturer Device Performance Test (DPT) Requirements. https://nrpp.info/manufacturer-device-performance-test-dpt-requirements/ (accessed 2021-12-11).

(36). National Radon Safety Board. Device Evaluation Request. https://www.nrsb.org/wp-content/uploads/2018/11/DeviceEvaluationApp082418.pdf (accessed 2021-12-11).

(37). Li L; Coull BA; Koutrakis P A National Comparison between the Collocated Short- and Long-Term Radon Measurements in the United States. J Expo Sci Environ Epidemiol 2023. 10.1038/s41370-023-00521-5.

(38). United States. Environmental Protection Agency; Centers for Disease Control (US). A Citizen's Guide to Radon: The Guide to Protecting Yourself and Your Family from Radon; 2016.

(39). Hahn EJ; Gokun Y; Andrews WM; Overfield BL; Robertson H; Wiggins A; Rayens MK Radon Potential, Geologic Formations, and Lung Cancer Risk. Prev Med Rep 2015, 2, 342–346. 10.1016/J.PMEDR.2015.04.009. [PubMed: 26844090]

(40). Otton JK The Geology of Radon; United States Geological Survey: Washington, D.C, 1992.

(41). Gundersen LCS; Schumann RR Mapping the Radon Potential of the United States: Examples from the Appalachians. Environ Int 1996, 22, 829–837. 10.1016/S0160-4120(96)00190-0.

(42). Tanner AB Radon Migration in the Ground: A Supplementary Review. In Natural radiation environment III: proceedings of a symposium held at Houston, Texas, April 23–28, 1978, Volume 1 (DOE Symposium Series 51); U.S. Department of Energy: Washington, D.C., 1980; pp 5–56.

(43). Wysocka M; Kotyrba A Radon Mapping with the Support of Geophysical Methods. Journal of Mining Science 2011 47:3 2011, 47 (3), 330–337. 10.1134/S1062739147030105.

(44). Sabbarese C; Ambrosino F; D'Onofrio A Development of Radon Transport Model in Different Types of Dwellings to Assess Indoor Activity Concentration. J Environ Radioact 2021, 227, 106501. 10.1016/J.JENVRAD.2020.106501. [PubMed: 33310392]

(45). National Research Council. Risk Assessment of Radon in Drinking Water; National Academies Press: Washington D.C, 1999.

(46). Stanley FKT; Irvine JL; Jacques WR; Salgia SR; Innes DG; Winquist BD; Torr D; Brenner DR; Goodarzi AA Radon Exposure Is Rising Steadily within the Modern North American Residential Environment, and Is Increasingly Uniform across Seasons. Sci Rep 2019, 9 (1), 18472. 10.1038/s41598-019-54891-8. [PubMed: 31796862]

(47). Şen GY; Içhedef M; Saç MM; Yener G Effect of Natural Gas Usage on Indoor Radon Levels. J Radioanal Nucl Chem 2013, 295 (1), 277–282. 10.1007/s10967-012-1841-8.

(48). Schubert M; Musolff A; Weiss H Influences of Meteorological Parameters on Indoor Radon Concentrations (222Rn) Excluding the Effects of Forced Ventilation and Radon Exhalation from Soil and Building Materials. J Environ Radioact 2018, 192, 81–85. 10.1016/j.jenvrad.2018.06.011. [PubMed: 29908412]

(49). Washington JW; Rose AW Regional and Temporal Relations of Radon in Soil Gas to Soil Temperature and Moisture. Geophys Res Lett 1990, 17 (6), 829–832. 10.1029/GL017i006p00829.

(50). Li L; Blomberg AJ; Lawrence J; Réquia WJ; Wei Y; Liu M; Peralta AA; Koutrakis P A Spatiotemporal Ensemble Model to Predict Gross Beta Particulate Radioactivity across the Contiguous United States. Environ Int 2021, 156, 106643. 10.1016/j.envint.2021.106643. [PubMed: 34020300]

(51). Liu M; Kang CM; Wolfson JM; Li L; Coull B; Schwartz J; Koutrakis P Measurements of Gross α- And β-Activities of Archived PM2.5and PM10Teflon Filter Samples. Environ Sci Technol 2020, 54 (19), 11780–11788. 10.1021/ACS.EST.0C02284/SUPPL_FILE/ES0C02284_SI_001.PDF. [PubMed: 32786555]

(52). Blomberg AJ; Li L; Schwartz JD; Coull BA; Koutrakis P Exposure to Particle Beta Radiation in Greater Massachusetts and Factors Influencing Its Spatial and Temporal Variability. Environ Sci Technol 2020, 54 (11), 6575–6583. 10.1021/acs.est.0c00454. [PubMed: 32363859]

(53). Breiman L Random Forests. Mach Learn 2001, 45, 5–32.

(54). Georganos S; Grippa T; Niang Gadiaga A; Linard C; Lennert M; Vanhuysse S; Mboga N; Wolff E; Kalogirou S Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling. 10.1080/10106049.2019.1595177 2019, 36 (2), 121–136. 10.1080/10106049.2019.1595177.

(55). Hengl T; Nussbaum M; Wright MN; Heuvelink GBM; Gräler B Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables. PeerJ 2018, 2018 (8), e5518. 10.7717/PEERJ.5518/SUPP-1.

(56). Goodchild MF; Li W Replication across Space and Time Must Be Weak in the Social and Environmental Sciences. Proc Natl Acad Sci U S A 2021, 118 (35), e2015759118. 10.1073/PNAS.2015759118/ASSET/9AFAE61C-E626-4060-B5D1-551D7E7D889C/ASSETS/IMAGES/LARGE/PNAS.2015759118FIG02.JPG. [PubMed: 34417345]

(57). Sui DZ Tobler's First Law of Geography: A Big Idea for a Small World? 10.1111/j.1467-8306.2004.09402003.x 2008, 94 (2), 269–277. 10.1111/J.1467-8306.2004.09402003.X.

(58). Fotheringham AS; Crespo R; Yao J Geographical and Temporal Weighted Regression (GTWR). Geogr Anal 2015, 47 (4), 431–452. 10.1111/gean.12071.

(59). Wright MN; Ziegler A Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 2017, 77 (1), 1–17. 10.18637/JSS.V077.I01.

(60). Hagenauer J; Helbich M A Geographically Weighted Artificial Neural Network. International Journal of Geographical Information Science 2022, 36 (2), 215–235. 10.1080/13658816.2021.1871618/SUPPL_FILE/TGIS_A_1871618_SM7651.PDF.

(61). Field RW; Steck DJ; Smith BJ; Brus CP; Fisher EL; Neuberger JS; Platz CE; Robinson RA; Woolson RF; Lynch CF Residential Radon Gas Exposure and Lung Cancer: The Iowa Radon Lung Cancer Study. Am J Epidemiol 2000, 151 (11), 1091–1102. 10.1093/oxfordjournals.aje.a010153. [PubMed: 10873134]

(62). Fisher EL; Field RW; Smith BJ; Lynch CF; Steck DJ; Neuberger JS Spatial Variation of Residential Radon Concentrations: The Iowa Radon Lung Cancer Study. Health Phys 1998, 75 (5).

(63). Airgood-Obrycki W; Rieger S Defining Suburbs: How Definitions Shape the Suburban Landscape; Cambridge, MA., 2019.

(64). Luetzelschwab JW; Hastings L; Ellis SM Adsorption of 222Rn by Open-Faced and Diffusion-Barrier Canisters at Different Conditions of Temperature and Humidity. Health Phys 1994, 66 (1).

(65). Mitchell AL; Griffin WM; Casman EA Lung Cancer Risk from Radon in Marcellus Shale Gas in Northeast U.S. Homes. Risk Analysis 2016, 36 (11), 2105–2119. 10.1111/RISA.12570. [PubMed: 26882276]

(66). Field RW; Smith BJ; Brus CP; Lynch CF; Neuberger JS; Steck DJ Retrospective Temporal and Spatial Mobility of Adult Iowa Women. Risk Analysis 1998, 18 (5), 575–584. 10.1111/j.1539-6924.1998.tb00371.x. [PubMed: 9853393]

Synopsis:

We developed a geographical machine learning model based on over 2.8 million short-term radon measurements from independent laboratories to predict radon concentrations at unprecedented resolutions in the Northeast and Midwest United States. Our study demonstrates the potential of these measurements to accurately estimate longitudinal community-level radon exposures, and highlights the value of using them to enhance our understanding of the health effects of radon in future studies.

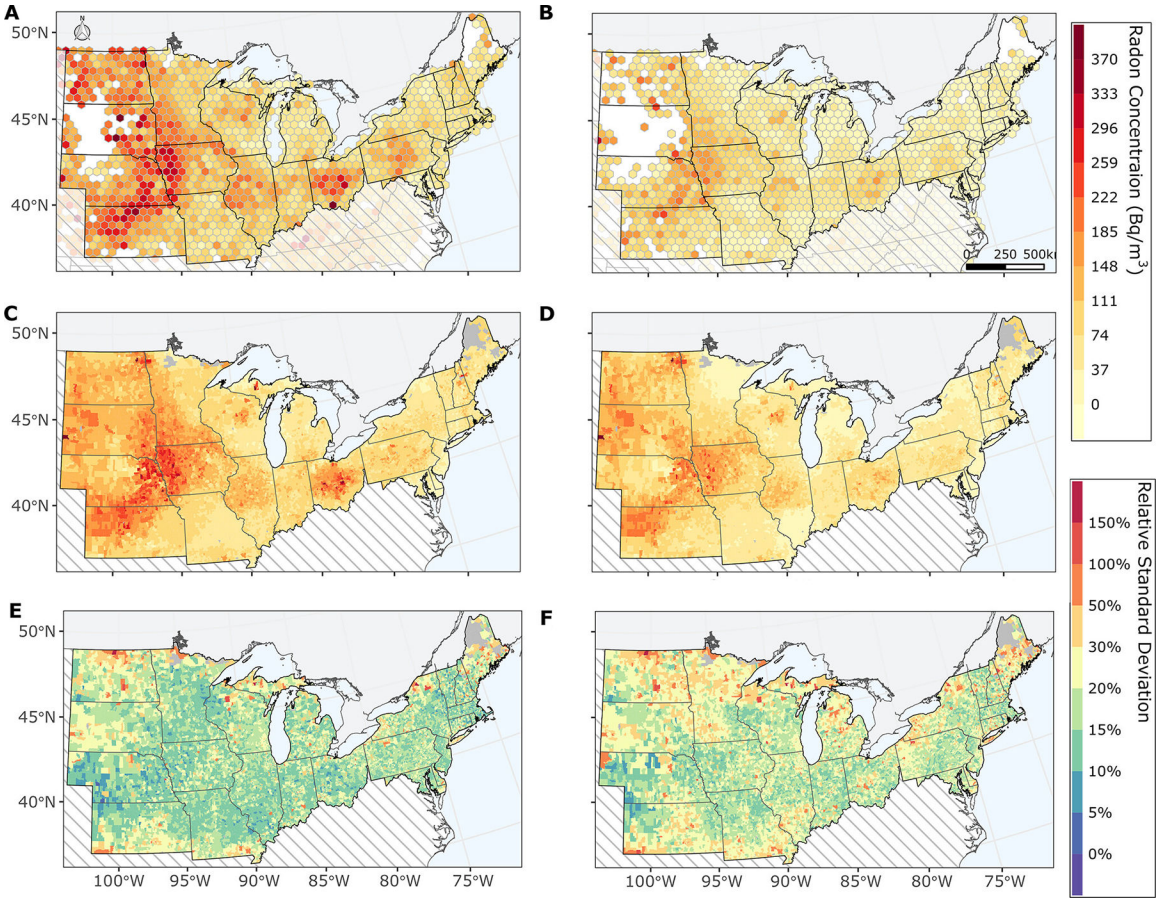


Figure 1.

The spatial distributions of observed radon concentrations, predicted radon concentrations, and the relative standard deviation of the prediction.

The 1,595,661 measurements were aggregated into hexagons due to the difficulty to visualize/print each individual point. The radius of hexagon is 25 km. On average, each hexagon contains over 1,000 radon measurements during 2001–2020. Each hexagon is colored based on the geometric mean of short-term measurements within the extent during the study period. The hexagon was not shown if fewer than 5 measurements exist within. Radon measurements conducted in states adjacent to the study region are covered by a semitransparent layer with diagonal lines. **Panel A** shows the radon concentrations measured in the basement; **Panel B** shows the radon concentrations measured in upper floors; **Panel C** shows the spatial distribution of the predicted ZCTA-level radon concentrations in the basement; **Panel D** shows the spatial distribution of the relative standard deviation of the predicted concentrations in the basements; **Panel E** shows the spatial distribution of predicted ZCTA-level radon concentrations in the aboveground floors; **Panel F** shows the spatial distribution of the relative standard deviation of the predicted concentrations in the aboveground floors.

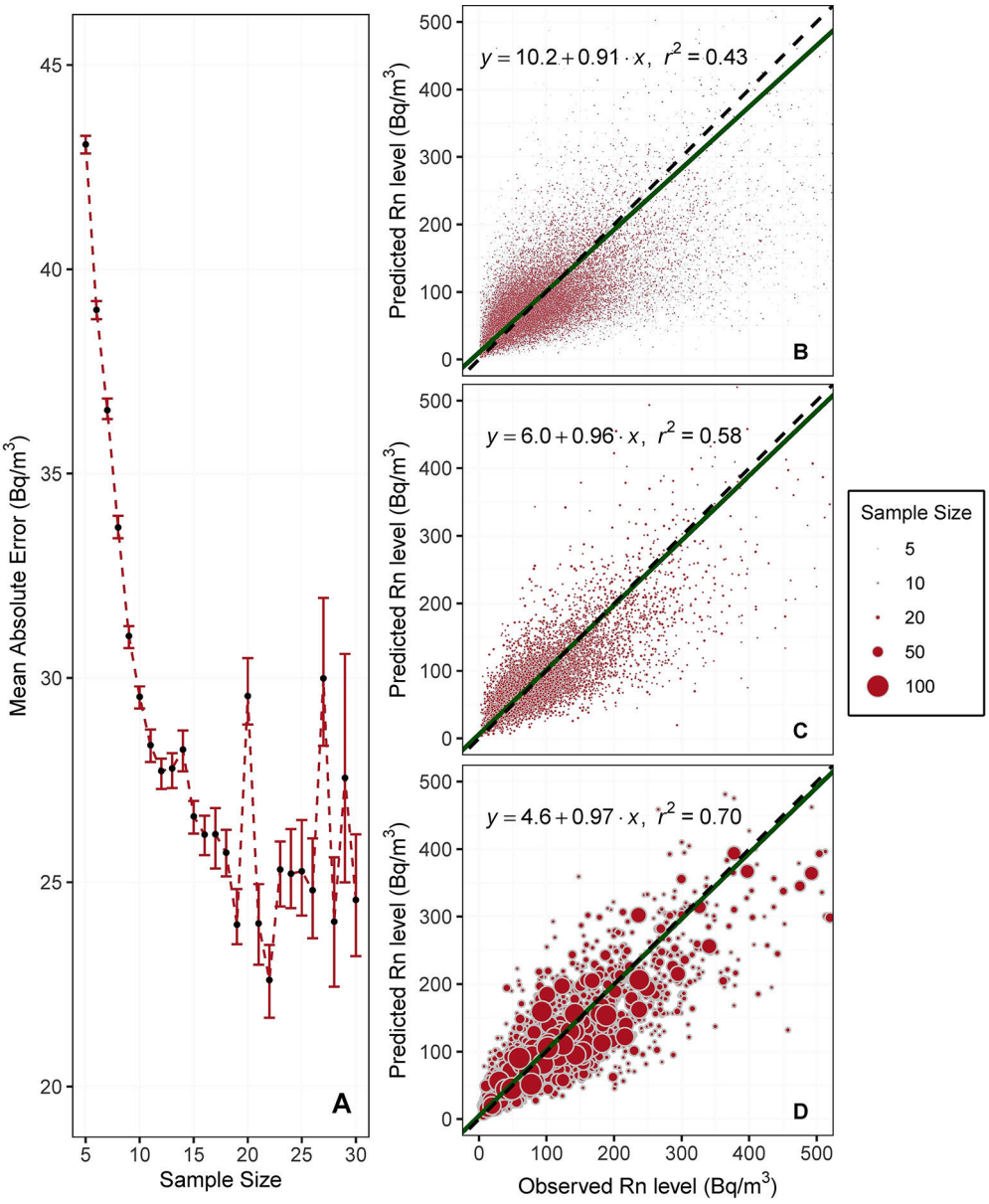


Figure 2.

The sample size-specific differences/correlations between the observed and predicted ZCTA-level monthly radon concentrations.

Panel A shows the mean absolute error of predicted ZCTA-level radon concentrations as a function of the sample size of observed ZCTA-level radon concentrations. **Panels B-D** show the correlations between observed and predicted ZCTA-level radon concentrations that are stratified by the sample size of ZCTA-level observation (Panel B for 5–9; Panel C for 10–14; Panel D for ≥15).

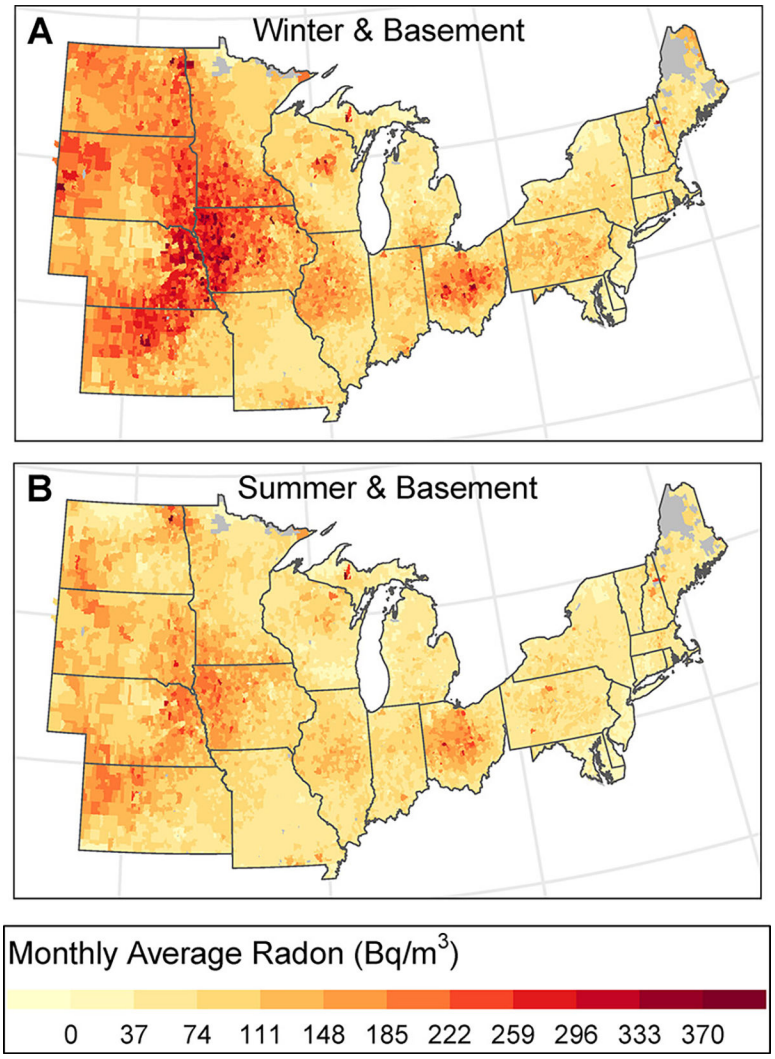


Figure 3.

Seasonal patterns of predicted ZCTA-level radon concentrations for basement.

Panel A shows the distribution of ZCTA-level basement radon predictions in the summer;

Panel B shows the distribution of ZCTA-level basement radon concentrations in the winter.

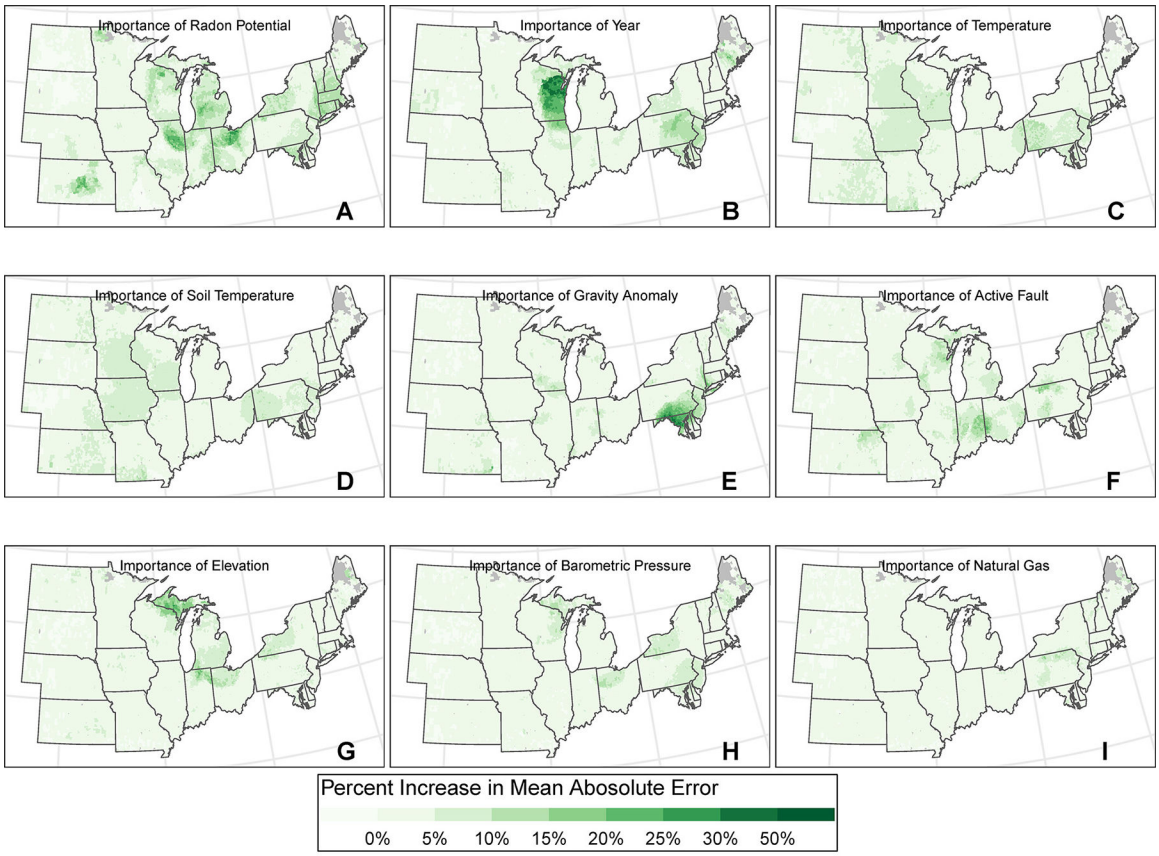


Figure 4.

The spatial distributions of the importance of nine radon predictors.

Panel A shows the spatial patterns of the importance of geological radon potential; **Panel B** shows the spatial patterns of the importance of calendar year; **Panel C** shows the spatial patterns of the importance of atmospheric temperature; **Panel D** shows the spatial patterns of the importance of soil temperature; **Panel E** shows the spatial patterns of the importance of gravity anomalies; **Panel F** shows the spatial patterns of the importance of distance to the closest active geological fault; **Panel G** shows the spatial patterns of the importance of elevation; **Panel H** shows the spatial patterns of the importance of barometric pressure; **Panel I** shows the spatial patterns of the importance of proportion of buildings fueled by natural gas.

Table 1.

Summary of collected short-term radon concentrations

Categories	Basement		Aboveground	
	Count	Median (IQR) in Bq/m^3	Count	Median (IQR) in Bq/m^3
All	1,242,375	92.5 (37.0, 196.1)	353,286	59.2 (25.9, 129.5)
Regions				
New England[1] (Northeast)	277,868	70.3 (37.0, 144.3)	32,958	37.0 (14.8, 77.7)
Mid Atlantic[2] (Northeast)	217,013	66.6 (25.9, 162.8)	45,206	40.7 (14.8, 107.3)
East North Central[3] (Midwest)	288,215	92.5 (37.0, 199.8)	82,803	55.5 (22.2, 125.8)
West North Central[4] (Midwest)	288,858	133.2 (59.2, 255.3)	75,968	88.8 (40.7, 173.9)
Neighboring states	170,421	99.9 (40.7, 218.3)	116,351	55.5 (25.9, 125.8)
Seasons				
Winter	362,683	111.0 (48.1, 229.4)	123,634	74.0 (29.6, 159.1)
Spring	368,359	92.5 (40.7, 203.5)	105,875	51.8 (22.2, 122.1)
Summer	253,703	66.6 (29.6, 144.3)	57,592	37.0 (12.9, 85.1)
Autumn	257,630	88.8 (37.0, 199.8)	66,185	59.2 (25.9, 133.2)
Years				
2001–2005	162,262	107.3 (48.1, 222.0)	57,743	59.2 (22.2, 136.9)
2006–2010	308,325	99.9 (44.4, 207.2)	88,276	55.5 (24.0, 129.5)
2011–2015	371,921	90.7 (37.0, 199.8)	96,635	59.2 (25.9, 136.9)
2016–2020	399,867	81.4 (33.3, 177.6)	110,632	55.5 (22.2, 122.1)
Type of device				
AccuStar-Liquid scintillation	269,357	77.7 (37.0, 162.8)	29,258	44.4 (22.2, 99.9)
AccuStar-Canister	144,021	55.5 (25.9, 125.8)	21,474	37.0 (18.5, 81.4)
AirChek-Foil bag	828,997	103.6 (44.4, 222.0)	302,554	61.0 (25.9, 136.9)

[1] New England region consists of Maine, Vermont, New Hampshire, Massachusetts, Connecticut, and Rhode Island.

[2] Mid Atlantic region consists of Delaware, Maryland, New Jersey, New York, and Pennsylvania.

[3] East North Central region consists of Illinois, Indiana, Michigan, Ohio, and Wisconsin.

[4] West North Central region consists of Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota.