A Model of Radon and a New Radon Map: Predicting Household Indoor Radon among the universe of single-family homes in the US

Lachlan Watkins*

Wednesday 17th July, 2024

Abstract

I construct the most comprehensive dataset aimed at understanding indoor radon and generate a new predictive map of indoor radon concentrations across the contiguous United States. The dataset combines geological maps, elemental and mineralogical samples from the US Geological Survey, soil maps from the Natural Resource Conservation Service, household building characteristics from Zillow's database, short-term weather data from the PRISM Climate Group, and novel atmospheric measurements of bismuth-214 from the National Uranium Resource Evaluation (NURE). I match over 1.1 million short-term indoor radon tests from Alpha Energy Laboratories to this dataset and model the data-generating process of indoor radon using machine learning. I account for sample selection and endogenous building characteristics, enabling the estimation of yearly radon concentrations for the population of 75 million single-family homes in Zillow's database. Aggregating these predictions, I develop a new map of long-term indoor radon at the US Census tract level. Incorporating the full distribution of indoor radon exposure suggests yearly attributed lung cancer deaths may range between 30,000 to 35,000.

*Vanderbilt University
lachlan.j.watkins@vanderbilt.edu

1 Introduction

There are numerous challenges involved in developing a predictive model of indoor radon that can be applied to the entire housing population across the United States. The primary and deepest source of radon gas lies within the underlying geology, which comprises various minerals containing uranium and thorium. Ideally, a model of the data-generating process of indoor radon would incorporate accurate measurements of uranium, thorium, radium, and the geologic age. However, the current representation of geology across the US is either relatively generalized or was conducted in a fragmented manner on a state-by-state basis. Geological representations often employ geological units or indicators for rock lithology.[1] Thus, the first challenge I must overcome is constructing a seamless map of rock units and lithology for the contiguous US.

The presence of localized variation in deposits of radon-generating minerals poses a significant challenge. There is a natural correlation between the average concentration of uranium and the lithology of the rock. For example, granitoids typically have a mean uranium concentration of 3.3 ppm, with known ranges varying from 0.4 to 7.8 ppm (International Atomic Energy Agency, 2003). However, an important question arises: How does the uranium concentration in granite found in the Appalachian Mountains differ from that in granite located in Minnesota or the Rocky Mountains? This limitation arises from the use of geological

[1] Geological units are typically identified by the prehistoric period during which the rock was formed and its general classification, such as Quarternary volcanic rocks. On the other hand, lithology refers to the basic physical description of a rock unit and follows a hierarchical structure: Igneous $\Leftarrow$ Volcanic $\Leftarrow$ Mafic-volcanic $\Leftarrow$ Basalt.

rock formations and lithology as categorical indicators to predict concentrations of indoor radon. It becomes apparent that local variation cannot be adequately accounted for using such indicators. Hence, the second challenge revolves around obtaining the distribution of radon-forming minerals across the United States.

Moving forward, let's suppose that both the geological units and the distribution of radon-forming minerals across the contiguous US are observed. The next challenge in modeling the data-generating process of indoor radon pertains to the soil layer above the underlying geology. Given the extremely long half-lives of uranium and thorium, the decay of these elements into radon gas has reached an equilibrium or steady state (Kazkaz and Walsh, 2018). As a result, radon gas is being produced at a nearly constant rate within the geology and soil. However, this gas must migrate through the soil pores, which may be filled with water and other obstructing elements, in order to reach the surface and pose a potential risk of exposure. The shape, size, structure, and texture of soil particles, combined with the characteristics of the pore space and environmental conditions, influence the movement of fluids (Nazaroff (1992a), Chen et al. (1995), Klusman and Jaacks (1987), Pyngrope et al. (2022)). Thus, conditional on the underlying distribution of geology and radon-forming minerals, the interactions between environmental conditions and soil characteristics will mediate the flow of radon gas to the surface[2].

Let's proceed with the assumption that we have observed the amount of radon

[2]Additionally, there may be impervious obstructions, such as asphalt or concrete, that limit the amount of radon gas that makes it to the atmosphere.

that has reached the surface. The subsequent challenge in modeling the data-generating process of radon exposure involves considering the exterior diffusive barriers of the home. These barriers hinder the movement of radon from the outdoor and underground environments into indoor spaces. They encompass various factors, including basement indicators, foundation types, exterior wall materials, passive radon resistance measures[3], and the potential implementation of active radon mitigation systems, among others. In an ideal situation, the researcher would observe surface-level radon, weather, the complete set of building characteristics (including passive radon resistance and radon mitigation), and long-term indoor radon test results for a representative sample of the US housing stock. In practice, however, the challenges in modeling indoor radon in the current data environment arise from the potential consequences of not observing passive radon resistance or radon mitigation. Additionally, there may be sample selection biases introduced by conducting radon testing in homes that may not be representative of the overall US housing stock.

I have successfully overcome all of the aforementioned challenges and have constructed the most comprehensive dataset ever created to model the data-generating process of indoor radon. To accomplish this, I employ machine learning techniques to combine state-level geological maps with a conterminous geological map of the United States, allowing for the prediction of rock lithology in a seamless distribution. This predicted geological map, along with soil survey data and

[3]Passive radon resistance is a building technique that uses materials or methods that inherently make it more difficult for radon to enter the home. For example, the placement of a polyurethane sheet under the foundation to work as an impermeable layer.

long-term weather information, is utilized to predict the distribution of radon-forming elements and accessory minerals based on a recent geochemistry survey conducted across the conterminous US.

Furthermore, I integrate the predicted geology maps, soil data, predicted geochemistry, observed daily weather, and model historical measurements of atmospheric radon.[4]. Additionally, I develop models to account for radon test sample selection and endogenous building characteristics by incorporating various factors such as the EPA Map of Radon Zones, county testing rates, the complete distribution of exogenous housing characteristics, soil conditions, and other geographical variables. Finally, I combine the predicted values of atmospheric radon with housing characteristics, predicted endogenous building characteristics, radon testing propensity scores, and over 1.1 million short-term indoor radon tests.

By employing this comprehensive approach, I am able to construct an extensive model of the data-generating process of short-term indoor radon. This model enables me to predict and map long-term indoor radon at the census tract level. The combination of various data sources, including geological maps, soil data, geochemical surveys, weather information, atmospheric radon measurements, housing characteristics, radon testing rates, and other geographical variables, contributes to the development of the most expansive model in this field. The predictive map provides insights into the distribution of long-term indoor radon exposure across different regions and demonstrates significant radon exposure heterogene-

[4]I use bismuth-214 as a proxy for radon, a daughter isotope of radon-222, which has an extremely short half-life. More detailed information regarding this proxy is discussed in Section 2

ity.

2 Radioactivity

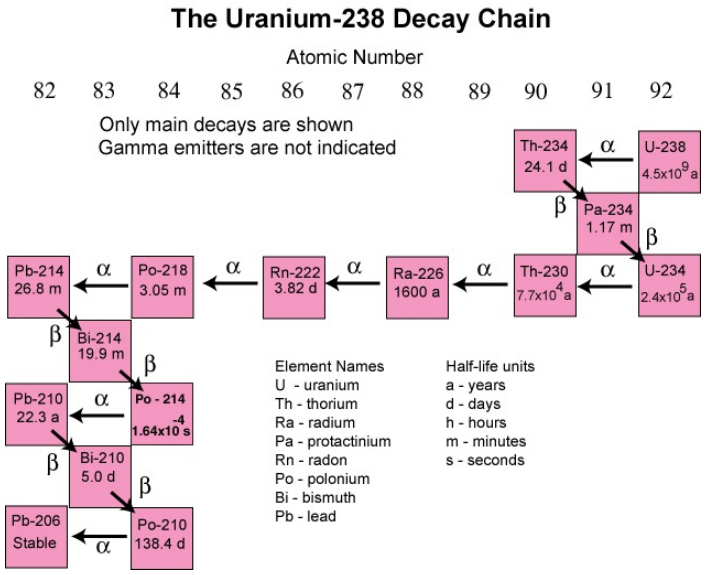
Indoor radon is a naturally occurring radioactive noble gas that is produced through the decay of radium isotopes. The longest-lived radium isotope, radium-226, has a half-life of approximately 1600 years, indicating that radon and radium originate from the decay of uranium and thorium that were deposited during the formation of the Earth. Figure 1 depicts the decay chain for uranium-238.[5] Radon-222 has a half-life of about 3.8 days, allowing it to diffuse from the ground into homes before emitting harmful alpha radiation and generating statically charged daughter particles that also emit alpha radiation.

It is worth noting the presence of the bismuth-214 isotope in Figure 1, which is part of the direct decay chain of radon-222. Bismuth-214 has a short half-life of about 20 minutes, and the two preceding isotopes (pb-214 and po-218) also have half-lives of less than 30 minutes. Due to the short half-lives of these daughter isotopes of radon-222, bismuth-214 is an effective measurement for estimating the concentration of radon gas in the immediate vicinity.[6] This observation is

[5]Figure 16 of Section 6 shows the decay chain for thorium-232, which produces radon-220. The short half-life of radon-220, just 55 seconds, implies that it is unlikely to provide a significant source of indoor radon exposure. Nonetheless, as noted in (Nazaroff, 1992b), studies have found that advection is the primary process by which radon enters the atmosphere through uncovered soil. Since the advection of radon-222 will be impacted by the amount of radon-220 that contributes to the advective velocities in the ground, thorium-232 should be included in any model of indoor radon (in other words, thorium-232 shoots out alpha particles, and generates radon-220 down the decay chain, which in turn increases the advective velocities in the soil that may co-determine the exhalation rates of radon-222).

[6]In fact, bismuth-214 is one of the isotopes used in gamma-ray counting to determine radon

Figure 1: The Radioactive Decay Chain of Uranium-238



critical and serves as the primary predictor of indoor radon levels.

Suppose we observe a deposit of radium-226 in a closed setting. When the observation period begins, the concentration of radon-222 will instantaneously increase as individual radium-226 nuclei randomly disintegrate into radon-222.[7] In a closed setting, radon gas will reach a steady state within 5-7 days, during which new radon-222 is produced from radium-226 at the same rate that current radon-222 decays into polonium-218 (Lehnert and Kearfott, 2010). However, in the underlying rocks and soil, radon-222 has already reached a steady state due to the 4.5 billion-year half-life of uranium-238. Therefore, as radon gas escapes from

concentrations from indoor activated charcoal radon tests.

[7]One gram of radium-226 decays at a rate of approximately 2.2 trillion decays per second. Radioactive measurements are commonly expressed in picocuries per liter (piC/L) or becquerels per cubic meter (Bq/m^3). One piC/L corresponds to 0.037 radioactive disintegrations per second in every liter of air. As a reference, the average human lung capacity is about 6 liters, and one cubic foot of air holds approximately 28.3 liters.

rocks and soil, it continuously and almost instantaneously produces bismuth-214 along the decay chain, which can be mathematically represented by the Bateman Equation.[8] To see why this is the case, consider the generalized form of the Bateman Equation:

$$X_n(t) = X_1(0) \prod_{k=1}^{n-1} \lambda_k \sum_{i=1}^{n} \prod_{j=1}^{n} \frac{e^{-\lambda_i t}}{\lambda_j - \lambda_i} \tag{1}$$

where $X_1(0)$ is the initial concentration of radon-222, t indexes time, and λ index the decay constants (for rn-222, po-218, pb-214, and bi-214). Given a positive amount of initial radon-222 in time period 0, the amount of bi-214 at time period 1 (say, 10 minutes later) is strictly positive ($X_n(1)$). For example, suppose radon-222 is constant at ground level at 0.1 becquerels/m^3. Then the Bateman equations estimate about 0.002 becquerels of bi-214 would have been produced after 10 minutes in every cubic meter of air. It is important to note that radon-222 is a gas and subject to flow. Indeed, it is possible for radon-222 nuclei that have yet to decay to travel a certain distance before bismuth-214 is produced along the decay chain.[9] This introduces a potential source of measurement error when using atmospheric bismuth-214 as a proxy for local atmospheric radon-222 since the bismuth-214 measurements represent a combination of locally generated and transported radon-222. However, despite this inherent limitation, atmospheric bismuth-214 remains an ideal proxy for atmospheric radon-222 and offers valu-

[8] The Bateman Equation describes the activities in a decay chain as a function of time and initial concentrations.

[9] In fact, radon-222 is the heaviest naturally occurring gas. As such, radon-222 at ground level will usually form low, flat clouds which spread based on their own density and surrounding terrain. At some point, the heavy gas cloud of radon will dilute, and reach a point where it is no longer heavy as its density becomes close to the surrounding air (Blackmore et al., 1982).

able insights into the data-generating process that brings radon-222 to the surface. This study represents the first attempt to leverage atmospheric bismuth-214 measurements in order to predict the distribution of indoor radon.

3 Empirical Methods

The objective of this paper is to enhance our comprehension of the distribution of indoor radon exposure, aiming to utilize the predictive model in research applications. As outlined earlier, numerous challenges must be tackled to develop a comprehensive model of indoor radon concentrations. This section provides a detailed account of each challenge and the corresponding steps taken to address them.

3.1 Part 1: Distribution of Geological Units

The primary sources of geological information are derived from Horton et al. (2017). Although the provided map represents a seamless spatial database of state-level geological maps, it is important to note that it is not a fully integrated geologic map, as the units have not been harmonized across state boundaries. The Horton et al. (2017) geological map includes rock unit lithology, which describes the physical characteristics of the geological units in a hierarchical structure. Typically, the maps include up to five lithologies for major, minor, and incidental rock formations within a specific geological unit. As an illustration, a sedimentary rock might have five lithologies listed as sedimentary, clastic, mudstone, Shale,

and black-shale. However, some geological units may not have such detailed lithology determinations and might only include up to three lithologies, such as sedimentary, clastic, and then mudstone. The varying precision in lithology categorization needs to be addressed in order to utilize a generalized map of unit lithologies in predictive models for the conterminous US.

To tackle this challenge, I leverage the hierarchical relationship of lithology and employ the Honest Random Forest (Athey et al., 2018) machine learning technique to estimate the distribution of lithologies for geological units with missing lithology categories. Figure 2 showcases a subset of lithology definitions for sedimentary rocks depicted in the Horton et al. (2017) map. For instance, let's consider a geological unit in one state where the 3rd lithology is missing but the presence of the first and second lithologies indicates sedimentary and carbonate, respectively. In such a case, we can infer that the 3rd lithology might be either limestone or marlstone. By utilizing information about the distribution of lithology combinations across the entire US, I impute missing lithology and generate a distribution of lithology indicators for each unique geological unit depicted in the Horton et al. (2017) map.

The second challenge arises from the fact that the Horton et al. (2017) geological map has not been reconciled consistently across state boundaries. While this would not pose a problem if the aim was to construct individual state-level models of indoor radon, it would not lead to a coherent model for the entire United States. To address this, I utilize another significant source of geological information from King and Beikman (1974), which provides a seamless map across state

Figure 2: A Subset of Geology Lithology

LITH1	LITH2	LITH3	LITH4
Sedimentary	Carbonate	Limestone	Coquina
Sedimentary	Carbonate	Marlstone	
Sedimentary	Chemical		
Sedimentary	Chemical	Banded-iron-formation	
Sedimentary	Chemical	Chert	
Sedimentary	Chemical	Diatomite	
Sedimentary	Chemical	Evaporite	
Sedimentary	Chemical	Evaporite	Anhydrite
Sedimentary	Chemical	Evaporite	Gypsum
Sedimentary	Chemical	Evaporite	Salt
Sedimentary	Chemical	Novaculite	
Sedimentary	Chemical	Phosphorite	
Sedimentary	Coal		
Sedimentary	Coal	Anthracite	
Sedimentary	Coal	Bituminous	
Sedimentary	Coal	Lignite	

Notes: This table shows a subset of the Lithology definitions from Horton et al. (2017)

boundaries and includes general geological classifications. In King and Beikman (1974)'s map, geological units are classified as Quarternary Volcanic Rocks or Miocene and Oligocene eugeosynclinal deposits, for instance.

To create a unified model, I utilize the King and Beikman (1974) map as the base map, which illustrates the discontinuities of geological units across the United States and state boundaries. Subsequently, I perform an intersection between the Horton et al. (2017) geological map and the King and Beikman (1974) map using ArcGIS. This process allows for the assignment of lithology distributions to each distinct map/geological unit represented in the seamless King and Beikman (1974) map[10]. The resulting map serves as the definitive source of geo-

[10]The foundation for the imputation procedure is based on the missForest R package (Stekhoven and Buehlmann, 2012). I adapted the missForest R package to use Honest Random Forests from

logical information used in the subsequent models. These final maps are available from the author upon request.

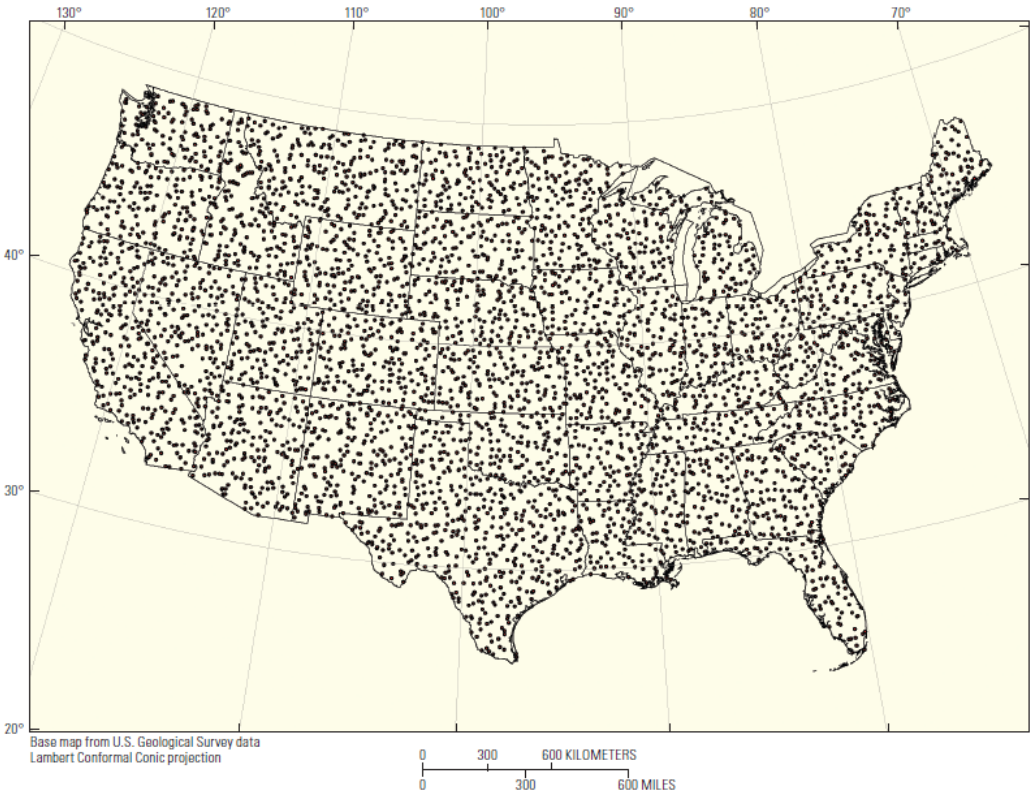
3.2 Part 2: Distribution of Radon-Generating Minerals

The second challenge stems from the limited variation in the categorical lithologies presented in the final map developed in Section 3.1. To overcome this limitation, I incorporate the geochemistry survey data obtained from Smith et al. (2013). This survey encompassed 4,857 sites across the United States, examining the concentrations of 45 major and trace elements that contribute to near-total elemental content. The sampling was conducted from 2010-2013, focusing on the A and C soil horizons, and utilized a generalized random tessellation design with a density of 1 site per 1,600 km^2. Figure 3 illustrates the distribution of these sampling locations.

It is worth noting that the sampling density of the geochemistry sites is relatively low. Each sample provides data for a small point in space, leaving the majority of the total land area of the conterminous US unsampled. If one were to utilize these observed point values in any model of atmospheric bi-214 or outdoor radon, it would require matching these points with the relevant outcome variable. However, these points were randomly selected to be representative of the overall population of elemental concentrations, excluding any obvious local anomalies

(Athey et al., 2018), which reduces bias through honest forest splitting and was significantly faster on the data sources with hundreds of thousands to millions of observations. See (Stekhoven and Buehlmann, 2012) for details on the missForest algorithm and (Athey et al., 2018) for a discussion on Honest Random Forests.

Figure 3: Geochemistry Sampling Locations



This map shows the sampling locations for the A and C soil horizons for the geochemistry elemental analysis. Source: (Smith et al., 2013)

such as uranium mine dust deposits. Based on this observation, I rely on a predictive approach by overlaying these points with the new map of underlying geology lithology, complete soil characteristics, long-term weather normals, and terrain features.

The soil characteristics are derived from the US General Soil Map (STATSGO2) sourced from (Soil Survey Staff, Natural Resources Conservation Service, 2016). This comprehensive dataset provides estimates of the physical and chemical prop-

erties at a 1:250,000 map scale. It includes information on horizon thickness, specific gravity, hydraulic conductivity, and numerous other variables[11]. The weather normals are obtained from the PRISM Climate Group (PRISM Climate Group, 2014), providing data on temperature, precipitation, dewpoint temperatures, solar radiation, and vapor pressure deficits at a 4km resolution. Additionally, digital elevation data from Danielson and Gesch (2011) is utilized, and I construct additional terrain features using the terra R package[12], and a measure of land imperviousness from Dewitz and U.S. Geological Survey (2021).

In theory, long-term weather patterns interact with the underlying lithology and the physical and chemical properties of the soil[13] to generate the observed concentrations of minerals and elements in the geochemistry samples. The selected minerals and elements are chosen based on their presence in uranium and thorium minerals and their classification as common accessory rock-forming minerals (International Atomic Energy Agency, 2003). To predict each element or mineral, I employ local nonparametric estimation that incorporates the latitude and longitude of each sampling location, along with the geology, soil, weather, and terrain features[14]. As a result, complete maps of quartz, aluminum, bismuth, calcium, cerium, iron, potassium, magnesium, niobium, phosphorous, lead, titanium, thallium, thorium, uranium, and yttrium for the conterminous US are gen-

[11] Some of which had missing values that were imputed using the modified missForest method mentioned above.

[12] I include elevation, slope, aspect, terrain ruggedness, and flow direction.

[13] Such as soil particle size, erodibility factor, pH, electrical conductivity, amount of readily exchangeable cations, and a measure of soil exchangeable hydrogen ions, among others.

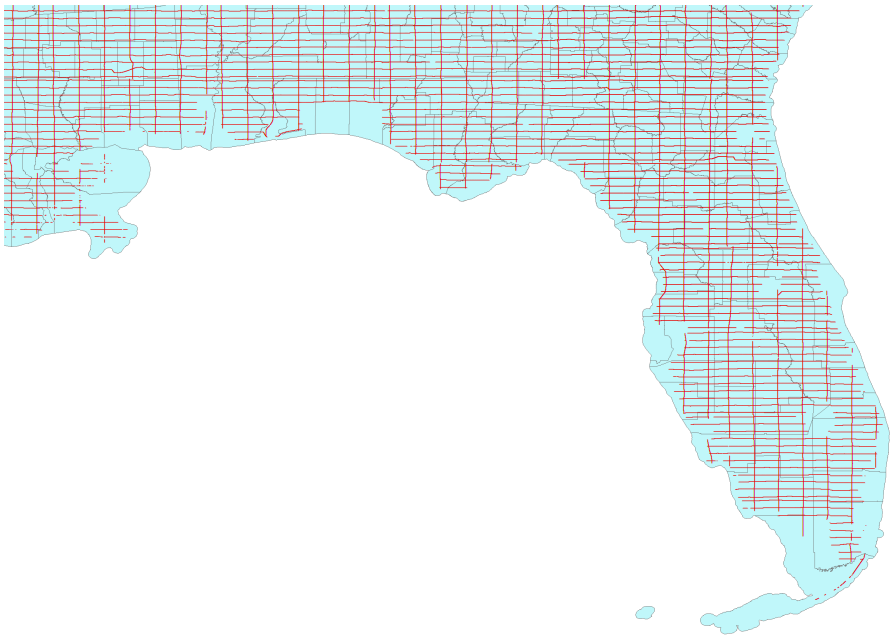
[14] This is accomplished using the Honest Random Forests with local linear corrections (Athey et al., 2018).

erated. These predicted maps are available upon request from the author.

3.3 Part 3: Modeling Atmospheric Bismuth-214

The third challenge revolves around the cleaning, leveling, and modeling of atmospheric bi-214 data obtained from the National Uranium Resource Evaluation (NURE). The NURE program was initiated by the United States government in the 1970s and 1980s as a comprehensive effort to identify and assess domestic uranium resources. It was a response to the energy crisis of the time, which underscored the importance of reliable domestic energy sources. Uranium, a crucial fuel for nuclear power plants, was recognized as a strategic resource that could help reduce dependence on foreign oil. The program aimed to systematically evaluate the uranium potential of the United States, covering both conventional and unconventional resources. The U.S. Geological Survey (USGS) led the program in collaboration with state and local governments, universities, and private industry, employing various activities such as geologic mapping, mineral resource assessments, and airborne geophysical surveys. The NURE program encompassed multiple phases, with the initial phase focusing on reconnaissance-level studies to identify areas with the highest potential for uranium deposits. Subsequent phases involved more detailed investigations of these areas, including ground-based geophysical surveys and drilling programs. The surveys generated approximately 36 million measurements of uranium, thorium, potassium, atmospheric bismuth-214, and magnetic properties, with flight lines covering a significant portion of the country. The flight surveys were conducted in “quadrangles”, defined as 2 degrees

Figure 4: NURE Survey Flight Lines of the State of Florida



Notes: This figure shows the location of the aerial flight line surveys conducted as part of the NURE program over the State of Florida.

longitude by 1 degrees latitude, with flight line spacing between 5 and 10 km. A total of 449 quadrangles cover the conterminous US. It is important to note that there are gaps in the measurements, particularly in Texas, and the flight lines only represent about 10 percent of the total land area of the United States. Figure 4 illustrates an example of the flight lines over Florida, where noticeable gaps in the measurements can be observed between the flight lines.

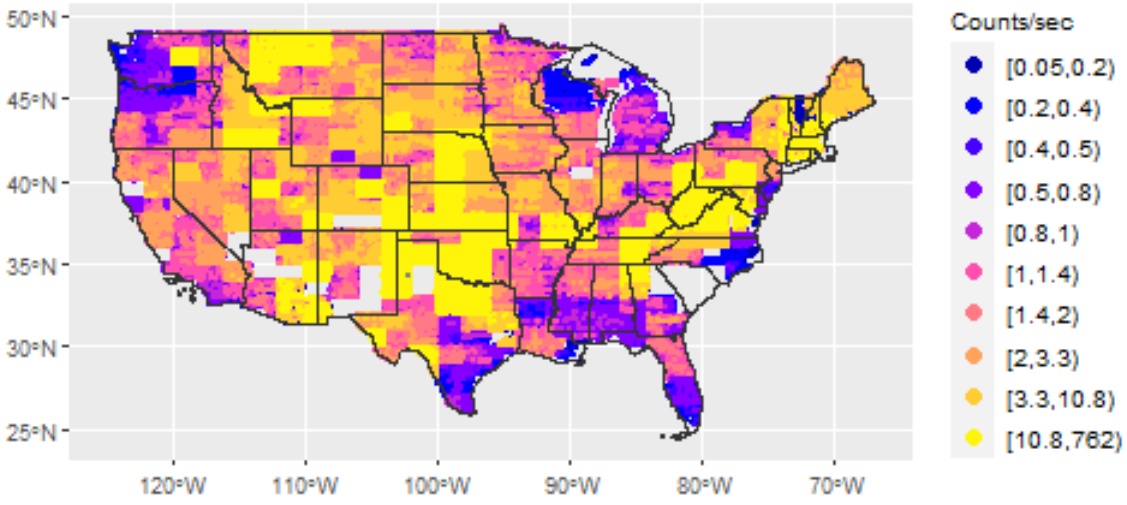
To obtain accurate measurements of uranium, thorium, and potassium in the upper 30 cm of the ground, the NURE surveys removed background radiation. Further details of the process can be found in (LKB Resources Inc, 1979), but in essence, the radiometric results are derived from counts per second of gamma radiation falling within specific energy windows[15]. Contractors responsible for the surveys conducted background corrections to eliminate cosmic and aircraft radiation, Compton scattering, and atmospheric bismuth interference. For the measurement of atmospheric bi-214, each surveyor typically employed a secondary radiospectrometer, which was extensively shielded from below with lead to ensure that the measurements exclusively captured gamma radiation from above (LKB Resources Inc, 1979). To the best of my knowledge, these background atmospheric bismuth measurements have been overlooked as a valuable source of information for assisting in indoor radon mapping.

One apparent reason for the neglect of the bi-214 data is its availability exclusively in the raw NURE data files, not appearing in the widely-used processed

[15]Each radioactive isotope will release gamma rays with different energy intensities. These intensities are used to determine which isotope generated a particular measurement of gamma radiation

NURE data by (Duval et al., 2005).[16] Figure 5 illustrates the distribution of atmospheric bi-214 quantiles in the raw NURE data, clearly showing significant leveling and measurement errors. Some survey quadrangles lack measurements entirely. Leveling errors occur when adjacent surveys are conducted using different calibration settings and surveying instruments. The bright yellow regions, including northern Texas through Nebraska, Pennsylvania/West Virginia, and other parts of the United States, most likely represent these areas.

Figure 5: Atmospheric Bi-214 Measurements from NURE



Notes: This Figure shows the 10 quantiles of the raw atmospheric bismuth-214 measurements in radioactive counts/second in the raw NURE data.

It is worth noting the count/sec range represented by the highest quantile, spanning from 10 to over 700 counts/sec. Approximately 97 percent of the 25.2 million measurements of atmospheric bi-214 fall below 34 counts/sec. If every

[16] I downloaded the raw files from the USGS Online Publications directory https://pubs.usgs.gov/of/2009/1129/US/

measurement within a specific quadrangle surpasses 20 counts/sec (for example), it is highly likely that this quadrangle is ”out-of-level” compared to the rest of the population. This observation is evident from the distinct blocks displayed in Figure 5. However, it is essential to consider that the surveys were conducted under different environmental conditions. As a result, discrete boundaries observed across survey areas may genuinely reflect accurate measurements of atmospheric bi-214 influenced by variations in weather conditions. Therefore, the challenge lies in distinguishing between variations attributable to weather fluctuations and those arising from leveling errors across different regions of the United States. The process that generated these measurements involves the interplay between radon-producing elements, soil and geological characteristics, and environmental conditions. Conditional on observing these variables, any remaining differences between measurements are assumed to be leveling and measurement errors.

I integrate the predicted maps of lithology and geochemistry from Sections 3.1 and 3.2, along with observed soil characteristics, terrain features, land imperviousness, and daily weather data to model the data-generating process of daily atmospheric bi-214. The daily weather data used in this model is obtained from historical measurements at a 6 km resolution, sourced from Livneh et al. (2013), and matched to the NURE measurements based on survey coordinates and date.

To address leveling and measurement errors, I initially utilize k-means clustering with 5 clusters to identify surveys that require level adjustments. Each survey is assigned to a cluster based on the most common cluster assignment of the points within that survey. For instance, if the majority of points belong to

cluster 2, the entire survey is categorized as a cluster 2 survey, indicating that it is ”out of level” compared to the rest of the population. Using this approach, the major cluster is determined as the cluster with the highest number of points. The major cluster, representing approximately 95 percent of the 25 million measurements when using 5 clusters, generally exhibits atmospheric bi-214 levels below 35 counts/second.

In order to capture the underlying data-generating process, I utilize a semi-parametric linear modeling approach. The natural logarithm of bi-214 is modeled as a linear combination of the mean function of the major cluster, individual mean functions for each “out of level” quadrangle, a correction for flight line direction, and quadrangle fixed effects for the “out of level” surveys:

$$ln(B_i) = m(x)_m + \gamma' m(x)_q + \beta LD_i + \alpha_q I\{Quad_i = q\} + \varepsilon_i \tag{2}$$

where B_i represents the Bi-214 in radioactive counts/second, $m(x)_m$ represents the function of interest where x includes daily weather, geology, soil, geochemistry, and terrain features. LD is the flight line direction that is assumed to generate leveling/measurement error. I apply a Robinson (1988) style semi-parametric approach to identify $m(x)_m$. Note that I assume that conditional on x, the survey indicators contribute only to leveling errors. This assumption seems plausible given the local nature of the estimation and the fact that the survey grids represent exogenous discontinuities in the measurement across survey areas. The estimation procedure is given by:

1. Use honest random forests of $ln(B_i)$ on $m(x)_m$ using observations in the major cluster.

2. Use honest random forests to model $ln(B_i)$ on $m(x)_q$ for each separate survey that requires leveling.

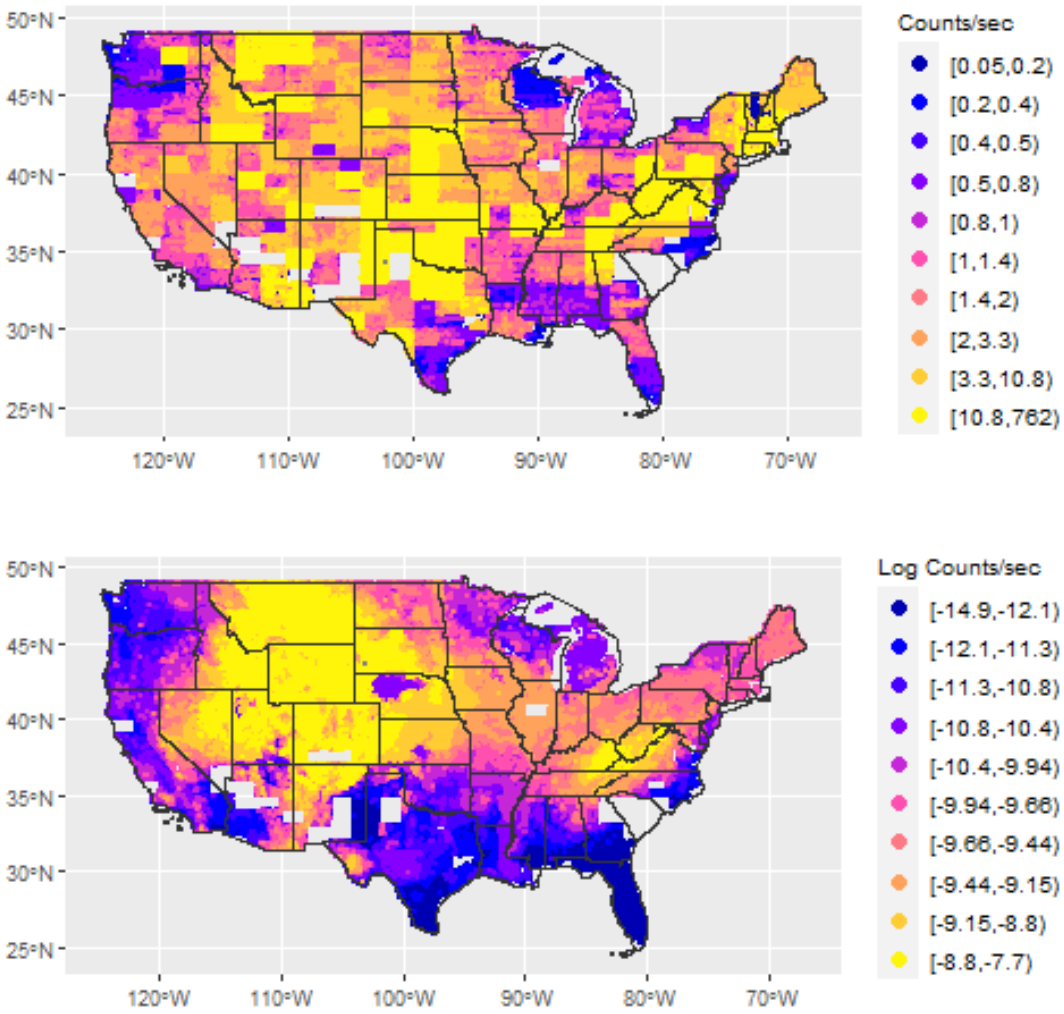
3. For all data, I assume $m(x)_m \approx m(x)$ so that $E[m(x)_m|X] = E[m(x)|X]$, and estimate $E[m(x)_q|X]$, $E[LD|X]$, and $E[I\{Quad_i = q\}|X]$ using honest random forests.

4. Run OLS of $ln(B_i) - m(x)_m$ on $\left[m(x)_q - E[m(x)_m|X]\right]$, $[LD_i - E[LD_i|X]]$, and
   $[I\{Quad_i = q\} - E[I\{Quad_i = q\}|X]]$ to obtain estimates of γ', β, and α_q.

5. Use honest random forest of $ln(B_i) - \hat{\gamma}' m(x)_q - \hat{\beta} LD_i - \hat{\alpha}_q I\{Quad_i = q\}$ to get a final estimate of $\hat{m(x)}_m$.

Figure 6 displays the conclusive predicted values from the $\hat{m(x)}_m$ model, assessed using the long-term weather normals. The figure illustrates evident geographic variation, with no noticeable arbitrary discontinuities emerging along the survey lines. Furthermore, the final model demonstrates exceptional accuracy in forecasting adjusted atmospheric bi-214 values beyond the sample, achieving an impressive out-of-sample R-squared value of 0.81.

Figure 6: Raw Atmospheric Bi-214 Compared to Their Predicted Values



This figure presents the orignal observed values of atmospheric bi-214 in the NURE data as a reference in the top map. The bottom map presents the predicted values after applying the semi-parametric modeling approach represented by Equation 2. The final model is evaluated at the long-term weather normals so that the point predictions can be compared based on the mean distribution of weather across the US.

3.4 Part 4: Indoor Radon

Let's consider the scenario where we have observed the presence of radon on the surface, specifically as bi-214. The subsequent challenge is to develop a model that captures the data-generating process for **indoor radon**, which is measured through short-term activated charcoal tests. The infiltration of surface radon into residential structures is influenced by various housing characteristics, some of which may not be directly observable to the econometrician. To create a comprehensive model for indoor radon applicable to the diverse housing stock across the contiguous United States, it is essential to address potential endogeneity issues associated with building characteristics.

For instance, factors such as passive radon resistance and radon mitigation, which are not captured in the housing data, are likely correlated with indoor radon levels and observable housing attributes like foundation type. Additionally, since radon tests are not conducted randomly, if the decision to perform a radon test is associated with unobservable characteristics that can influence the test results, sample selection becomes another important concern that needs to be addressed. Therefore, the subsequent steps involve constructing a model for endogenous building characteristics using exogenous variables and accounting for the sample selection bias inherent in radon testing.

3.4.1 Building Endogeneity

The housing data utilized in this project is sourced from The Zillow Group's ZTRAX database, which is the largest real estate database in the United States. It encompasses more than 400 million comprehensive public records from most U.S. counties. The dataset includes various property characteristics such as geographic location, number of bedrooms, square footage, number of bathrooms, foundation type, number of stories, roof material type, basement indicator, and more. These property details are primarily obtained from county property assessment records. For the analysis, only single-family residences are considered, although there is a plan to extend the analysis to other property types in future work.

One notable strength of the ZTRAX data is its extensive and detailed property records, similar to what is typically available during a Zillow property search. However, certain property characteristics had missing values and were imputed using the same imputation procedure employed in the previous sections. The imputation process involves generating a predicted distribution for categorical variables and conditional mean estimates for continuous characteristics. Table 1 provides summary statistics for some of the building property variables in ZTRAX for a random sample of homes, categorized by EPA zone.

Across the three EPA radon zones, there is a decrease in the number of ZTRAX homes with basements. This observation aligns with expectations, considering that the EPA zone construction is based on county-level foundation characteristics. The size of homes is relatively consistent across all EPA zones, with Zone 3 homes averaging 200 square feet larger. Additionally, homes in Zones 2 and 3

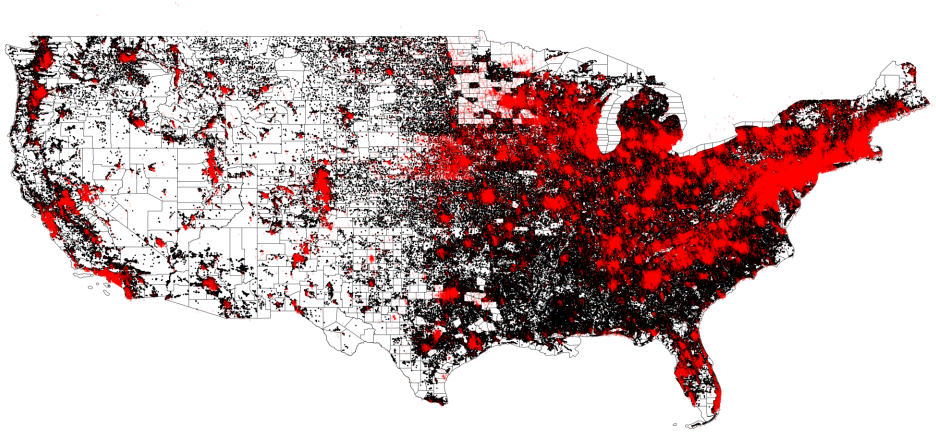
Table 1: ZTRAX 1% Descriptive Statistics

	EPA Zone 1	EPA Zone 2	EPA Zone 3
Year Built	1968	1970	1978
Square Footage	2,085 (1,295)	2,090 (1,270)	2,261 (1,327)
Lot Size (Acres)	2.21 (9.83)	1.90 (8.22)	1.69 (7.50)
Number of Baths	2.10 (0.95)	2.14 (0.93)	2.19 (0.91)
Number of Stories	1.44 (0.46)	1.36 (0.43)	1.31 (0.42)
Basement Indicator	0.40 (0.47)	0.23 (0.40)	0.11 (0.28)
Uranium	2.13 (0.58)	1.95 (0.70)	1.48 (0.61)
Thorium	6.67 (1.77)	5.96 (2.47)	4.53 (2.32)
Foundation: Concrete Block	0.11 (0.22)	0.09 (0.17)	0.05 (0.14)
Foundation: Concrete	0.14 (0.30)	0.09 (0.21)	0.07 (0.20)
Foundation: Crawl Space	0.25 (0.27)	0.21 (0.22)	0.16 (0.25)
Foundation: Footing	0.07 (0.17)	0.08 (0.18)	0.08 (0.21)
Foundation: Slab	0.11 (0.15)	0.14 (0.21)	0.34 (0.41)
Exterior Wall: Brick	0.14 (0.31)	0.11 (0.25)	0.12 (0.28)
Exterior Wall: Siding	0.29 (0.44)	0.21 (0.39)	0.14 (0.33)
Exterior Wall: Wood	0.12 (0.28)	0.13 (0.27)	0.08 (0.22)
Exterior Wall: Stucco	0.04 (0.15)	0.08 (0.20)	0.13 (0.30)
Exterior Wall: Combination	0.05 (0.14)	0.07 (0.14)	0.07 (0.17)
Interior Floor: Combination	0.05 (0.11)	0.06 (0.12)	0.06 (0.16)
Interior Floor: Wood	0.12 (0.17	0.12) (0.16)	0.09 (0.18)
Interior Floor: Vinyl	0.05 (0.07)	0.05 (0.08)	0.04 (0.10)
Interior Floor: Carpet	0.25 (0.27)	0.23 (0.28)	0.33 (0.37)
Interior Floor: Concrete	0.08 (0.09)	0.08 (0.09)	0.08 (0.14)
Interior Wall: Drywall	0.43 (0.30)	0.40 (0.31)	0.51 (0.36)
Interior Wall: Wood	0.02 (0.06)	0.03 (0.06)	0.02 (0.07)
Interior Wall: Plaster	0.16 (0.19)	0.13 (0.14)	0.10 (0.17)
Interior Wall: Paneling	0.03 (0.09)	0.03 (0.09)	0.03 (0.11)
N	212,462	268,145	265,492

The building characteristics (foundation, exterior wall, interior walls and floors) that are shown four or five out of more than 15 possible values for each. This is a 1 percent sample of the buildings in the ZTRAX single-family home database.

tend to be newer compared to Zone 1 homes. Many other housing characteristics exhibit similar patterns across all three zones. However, there is a notable difference in the frequency of slab foundations, which is substantially higher in Zone 3. This finding is consistent with the EPA's use of slab foundations in the zone delineation. To geocode all properties in the ZTRAX database, ArcGIS was employed. Figure 7 displays the mapped locations of the properties, depicted in black. The ZTRAX database covers most of the United States, except for areas with limited habitation and certain counties in Minnesota. The red points represent the point locations of the indoor radon tests obtained for this project.

Figure 7: Locations of Single-Family Homes and Alpha Energy Radon Tests



This figure shows the geocoded locations of a 10 percent sample of ZTRAX single-family residences, represented by black points. The red points are the locations of the Alpha Energy indoor radon tests.

In the model of building endogeneity and sample selection, I utilize the EPA

Map of Radon Zones, along with pre-mitigation and post-mitigation county testing rates from the CDC's National Environmental Public Health Tracking of indoor radon (CDC, 2018), as the primary "instruments." It is important to note that while these variables are not technically instruments in the Two-Stage-Least-Squares approach due to the use of nonparametric probability forests for estimation, they do serve as valuable predictors of building characteristics and selection into radon testing.[17]

To elaborate further, the decision to build a basement is jointly determined by the designed foundation type of a home prior to construction. However, the home design is likely influenced by the EPA county designation, where builders are inclined to follow some of the radon-resistant guidelines outlined on the EPA website.[18] Moreover, given the observed housing characteristics, the assignment of a county as a high-risk area for indoor radon likely prompted certain homes to install radon mitigation systems. For instance, homes lacking passive radon resistance and featuring a basement are more likely to have installed mitigation systems. Hence, the joint distribution of basement and foundation type serves

[17]The probability forest models come from Athey et al. (2018) and can be represented by the following equation:

$$\sum_{i=1}^{n} \frac{1}{B} \sum_{b=1}^{B} \frac{1\{X_i \in L_b^*(x)\}}{|L_b^*(x)|} \left(\frac{1}{B} \sum_{b=1}^{B} \sum_{i:X_i \in L_b^*(x)} \frac{I\{Y_i = y_r\}}{|L_b^*(x)|} \right) \tag{3}$$

where b indexes the trees of a forest of size B, $L_b^*(x)$ is the set of observations falling in the same leaf (or terminal node) as x, and $I\{Y_i = y_r\}$ is an indicator function for outcome Y taking the value y_r. In the full probability forest, it averages across all indicator functions for each categorical value, thus generating a predicted distribution.

[18]https://19january2017snapshot.epa.gov/radon/building-codes-radon-resistant-new-construction-rrnc_.html#states

as a probable source of building endogeneity, as these factors correlate with unobserved radon resistance and mitigation. It is worth noting that the ZTRAX single-family home data underwent imputation, resulting in homes with missing foundation types having a predicted distribution rather than an indicator for a specific foundation type. Similarly, basements were also imputed, providing some homes with a probability of having a basement. Thus, the process to generate predictions of the joint distribution of basement and foundation type follows the following procedure on a random sample of 200,000 ZTRAX properties:

1. Use k-means clustering, with two clusters, to determine the cut point that assigns a home a basement based on their basement probability.

2. Using the predicted probability distribution of foundation type for individual observation, draw a random foundation type.

3. Construct the joint distribution of the imputed basement and foundation type from the previous steps.

4. Model the joint distribution using Probability Random Forests from (Athey et al., 2018) using long-term weather, long-term predicted atmospheric bi-214, NURE uranium and thorium, EPA region, the exogenous building characteristics, the EPA Zone, and testing rates.

5. Repeat this process 10 times

6. Average the 10 predicted distributions to generate the final predicted joint distribution of basement and foundation type.

7. Apply the final aggregate model to all 74 million single-family residences in Zillow.

3.4.2 Indoor Radon under Sample Selection

Suppose that all the building characteristics observed in the data-generating process of indoor radon are exogenous. Additionally, let's assume that surface-level radon continues to be observed. The next challenge is to account for sample selection, which could be influenced by unobservable characteristics. To tackle this challenge and develop the final model for indoor radon, I obtained indoor radon test results from Alpha Energy Laboratories. They provided 1.8 million indoor radon tests conducted between 2000 and 2021. Alpha Energy is a national radon testing lab based in Texas, and according to internal conversations with the company, they analyze approximately one-third of all radon tests conducted in the US. To comply with the confidentiality agreement, individual test results cannot be disclosed. After an initial cleaning process, I obtained 1.6 million usable test results, which were then matched to the ZTRAX population based on address, achieving a match rate of 92 percent. The geocoded test locations are displayed in Figure 7 as red points, covering a wide geographical distribution across the US.

Table 2 presents some summary statistics of the tests. On average, the test results for each EPA zone were approximately 6, 4, and 2.4 picocuries per liter (piC/L), respectively. Of particular interest are the standard deviations of the indoor radon tests. Zone 1 tests exhibit a standard deviation of 7 piC/L, while even Zone 3 tests show a relatively large standard deviation of 3.2 piC/L. The me-

dian year built for the Alpha Energy properties is remarkably similar across EPA zones, albeit somewhat newer than the ZTRAX population. This similarity is not surprising considering that radon tests were conducted only over the past 20 years, while ZTRAX represents the entire housing stock population. It is interesting to note the resemblance in building characteristics between the ZTRAX sample and the matched Alpha properties. The most apparent difference is the percentage of properties with a basement, which aligns with the understanding that basements are often associated with higher levels of radon.

There are some important limitations of the Alpha Energy radon test data. Data prior to 2012 did not collect the test address independently of the mailing address in a systematic manner. Thus, data prior to 2012 is assumed that the mailing and test addresses are the same. Furthermore, some of the intermediate calculations (such as the Lower Limit of Detection or LLD) prior to 2012 weren't stored in the database and were recalculated under a new methodology. Thus, the LLD isn't accurate in the data prior to 2012. Some test results, about 10 percent, resulted in radon measurements below the LLD. In these cases, the estimated radon measurements are confidently below the LLD, but are imprecise. At the risk of biasing the probability distribution of indoor radon, I use the reported results. Of the 1.6 million test results, only about 1100 tests reported an exact 0 test result. Exact 0 results represent an error result according to Alpha Energy Laboratories and were dropped from the analysis.

To adequately capture the complex data-generating process of indoor radon, I employ a non-parametric local linear approach that accounts for building endo-

Table 2: Alpha Energy Descriptive Statistics

	Zone 1	Zone 2	Zone 3
Radon Test Result (piC/L)	5.99 (7)	4.08 (5)	2.41 (3.2)
Radon Test Decay Time	5.2 (2)	5.29 (2.1)	5.27 (2.1)
Mitigation Test	0.06 (0.24)	0.04 (0.20)	0.02 (0.13)
Air Temperature	282 (10.44)	282 (10.41)	286 (9.90)
Accumulated Precipitation	2.52 (3.75)	2.50 (3.79)	2.75 (4.23)
Year Built	1979	1979	1979
Square Footage	2,419 (1,328)	2,490 (1,419)	2,489 (1,316)
Lot Size (Acres)	2.06 (9.21)	1.87 (7.14)	1.61 (6.57)
Number of Baths	2.43 (0.94)	2.45 (0.94)	2.41 (0.93)
Number of Stories	1.51 (0.46)	1.47 (0.45)	1.47 (0.49)
Basement Indicator	0.51 (0.48)	0.43 (0.47)	0.30 (0.43)
Uranium PPM	2.15 (0.57)	1.87 (0.66)	1.53 (0.63)
Thorium PPM	6.93 (1.77)	5.91 (2.23)	4.71 (2.30)
Foundation: Crawl Space	0.23 (0.24)	0.22 (0.20)	0.21 (0.26)
Foundation: Slab	0.11 (0.14)	0.13 (0.18)	0.22 (0.33)
Exterior Wall: Brick	0.13 (0.30)	0.13 (0.29)	0.14 (0.31)
Exterior Wall: Siding	0.30 (0.44)	0.26 (0.42)	0.19 (0.38)
Exterior Wall: Wood	0.12 (0.28)	0.10 (0.24)	0.07 (0.21)
Exterior Wall: Stucco	0.05 (0.17)	0.08 (0.23)	0.10 (0.28)
N	474,321	281,576	67,850

Notes: Table shows the descriptive statistics for the Alpha Energy radon tests that were matched to Zillow Single-Family Residences. Standard deviations are in parentheses.

geneity and sample selection:

$$R_i^* = m(x_1, x_2) + \omega \tag{4}$$

$$x_2 = \pi(z) + u \tag{5}$$

$$s_i = h(z) + v \tag{6}$$

where R_i^* is the observed short-term radon test result, x_2 represents the joint distribution of basement and foundation type (which is endogenous), and s_i represents the sample selection model (an indicator function that equals 1 when a radon test is performed). Sample selection is a potential problem if $Cov(\omega, v) \neq 0$. I model sample selection as a nonparametric function of various factors, including long-term weather, long-term predicted atmospheric bi-214, NURE uranium and thorium levels, EPA region, exogenous building characteristics, EPA Zone, and the CDC county-level testing rate (CDC, 2018). The analysis includes a random sample of 200,000 observations, consisting of an equal proportion of Alpha properties and non-matched ZTRAX properties. Following the approach outlined in Das et al. (2003), I replace x_2 with $\hat{x_2}$ and include a function of the propensity score in 4:

$$R_i^* = m(x_1, \hat{x_2}) + \lambda(\hat{s_i}) + \gamma \tag{7}$$

$$\Rightarrow R_i^* = g(x_1, \hat{x_2}, \hat{s_i}) + \gamma \tag{8}$$

where $g(x_1, \hat{x_2}, \hat{s_i}) \equiv m(x_1, \hat{x_2}) + \lambda(\hat{s_i})$. It is important to note that this paper does

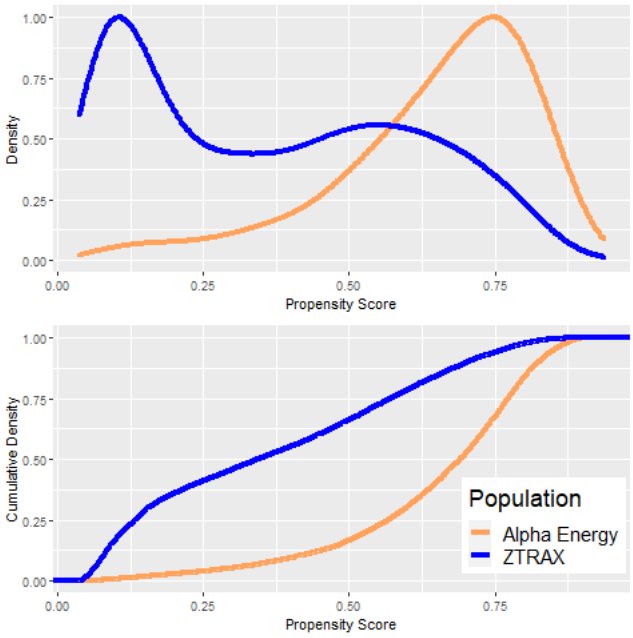
not aim to separately identify the $\lambda(\hat{s}_i)$ function or explore the factors influencing the decision to conduct a radon test. The primary objective is to develop a predictive model of indoor radon that controls for sample selection and building endogeneity, with estimates obtained from $g(\cdot)$. A comprehensive analysis of selection into radon testing is left for future research.

Nevertheless, it is worth providing a brief discussion of the sample selection model. I present the distribution of propensity scores among Alpha Energy properties and a random sample of the ZTRAX population. Figure 8 illustrates the density and cumulative density of the propensity scores, demonstrating sufficient overlap between the two populations. Notably, all predicted propensity scores fall strictly within the range of 0 and 1. Additionally, the cumulative density plot reveals that only 25 percent of the overall ZTRAX population has propensity scores below 0.125. Neither ZTRAX nor Alpha Energy properties exhibit significant counts of observations near the extremes of 0 or 1. Finally, both the EPA Zone designation and CDC county-level testing rates exhibit strong predictive power in determining individual selection for radon testing.[19]

Next, I present some descriptive results of the final model of indoor radon. First, it is important to note that although radon mitigation is unobserved in the Zillow data, the Alpha radon tests include information that can be used to identify if a test is a post-mitigation test. Controlling for this information increases the accuracy of the model. However, radon mitigation is not observed in the ZTRAX

[19]Since the estimation procedure is nonparametric, this was observed from a simple measure of "variable importance", which generates a weighted count of how many times in the forest a particular variable was chosen to determine a tree split.

Figure 8: Propensity Scores Densities and Cumulative Distributions



This Figure shows the densities and cumulative distributions of the predicted propensity scores from the sample selection model. Matched Alpha Energy properties are shown in orange, and a 10 percent random sample of ZTRAX properties is shown in blue.

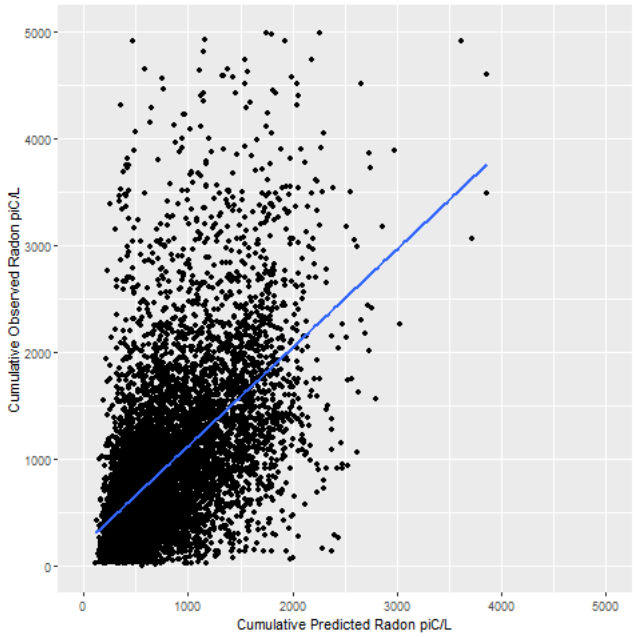
population, so the mitigation variable is omitted from the predictive model. If the mitigation rate in the Alpha properties is higher than in ZTRAX, then omitting mitigation from the model and predicting indoor radon for ZTRAX will underestimate true radon exposure. On the other hand, if the mitigation rate is lower in the Alpha properties, then this model will overestimate true radon exposure.

There are a couple of observations that suggest the model underestimates radon exposure in the ZTRAX population. Figure 9 shows that the out-of-sample predictions of the long-term alpha track test results underestimate observed cumulative exposures. This is confirmed by the linear projection of the predictions onto the observed exposures, represented by the blue line in Figure 9. The slope coefficient on the linear projection is 0.98. On average, increasing the predictions by 1 piC/L is associated with a 0.98 increase in the observed long-term test result. Furthermore, notice the significant portion of the long-term cumulative test results that appear above a 45-degree line. There is a clear pattern that the observed values towards the upper end of the cumulative exposure distribution have predictions that are much lower. In addition, it is important to note that for the short-term test results, more than 53 percent of the 1.1 million predictions are lower than the corresponding observed test result. This indicates that the majority of the predictions tend to underestimate the short-term indoor radon levels.

Finally, it is important to note that the predictive accuracy of the short-term model, which was used to forecast cumulative long-term exposures, exhibits an out-of-sample R-squared value of 0.30 among the alpha track test results.[20] This

[20]The long-term test results were held out of the model of indoor radon completely. The in-

Figure 9: Observed Long-Term Alpha-Track Radon Tests against Long-Term Predictions



The Figure illustrates the comparison between observed cumulative radon concentrations, obtained from long-term alpha-track radon tests conducted by Alpha Labs, and their corresponding predictions. The tests measure the total disintegrations over the testing period. To generate the long-term predictions displayed on the x-axis, each test was associated with the monthly weather data for every month during the test deployment. It is worth noting that long-term alpha track tests typically span between 90 and 365 days. For each month's weather values, short-term predictions were generated using the predictive model, and these individual predictions were then aggregated over the entire testing period.

indicates that the model explains approximately 30 percent of the variance in the observed long-term exposures, providing a reasonable level of predictive power. It suggests that the model captures a substantial portion of the underlying fac-

sample R-squared for the short-term radon test results is 0.84

tors influencing radon concentration levels in indoor environments. However, it also suggests the presence of other factors that are not currently considered in the model. These unaccounted factors contribute to the remaining variance in long-term exposures, highlighting the potential for further enhancements and refinements in future iterations of the model.

4 Discussion

The preceding sections have discussed the extensive measures undertaken to overcome significant challenges in order to develop a comprehensive model of indoor radon. This model can be utilized to estimate counterfactual concentrations for the entire single-family housing stock in the United States, as represented in Zillow's database. To summarize, the following steps were undertaken. Firstly, a predictive map of rock lithology for the contiguous United States was generated. This map was combined with geochemical surveys, long-term weather data, soil characteristics, and terrain features to model the distribution of radon-accessory elements. Next, the predicted lithology, distribution of radon-accessory elements, soil characteristics, terrain features, land imperviousness, and other relevant factors were used to model the data-generating process of atmospheric bi-214, a critical proxy variable for outdoor/surface-level radon. Subsequently, endogenous building characteristics and sample selection were modeled, and these adjustments were combined with the daily predicted atmospheric bi-214, weather data, and remaining building characteristics to model the data-generating process

of indoor radon.

Each step in this process involved making various assumptions, each with its own set of limitations. Currently, an important limitation of the final model is its inability to generate confidence and prediction intervals for predicted indoor radon in the ZTRAX population. Due to the high computational cost associated with generating predictions at each step, I am unable to provide estimates of uncertainty. This limitation is particularly significant considering the nonparametric estimation techniques employed, where bootstrapping the final indoor radon model would be the most straightforward approach. However, to generate monthly predictions of indoor radon for all 74 million single-family homes, it is necessary to match these homes with their monthly weather data, predict atmospheric bi-214 for each month, and subsequently predict indoor radon. With approximately 10.6 billion predictions required for a single year across all 74 million homes, one can imagine the additional computational burden that would arise from introducing bootstrapped indoor radon models.

Moreover, at each step, a model is developed and utilized to predict or impute geology, geochemistry, atmospheric bi-214, building characteristics, and ultimately indoor radon. Consequently, each set of predictions introduces some error into each subsequent model. It is possible that this cumulative error could result in large confidence and prediction intervals for long-term radon exposures, potentially limiting the usefulness of a predicted long-term map. Furthermore, these errors could lead to attenuation bias in any application that relies on indoor radon predictions, such as estimating the hedonic price of radon or assessing

causal health effects.

4.1 A New Radon Map and its Implications

The models developed in this paper represent significant contributions to the field and underscore the challenges associated with nationwide mapping of indoor radon. Based on the outlined procedures, modeling assumptions, and data limitations, a New Map of Indoor Radon at the census tract level is presented. Additionally, square kilometer maps of the State of Tennessee and Davidson County, TN, showcase the detailed mapping potential of the final model. Finally, radon-attributed lung cancer deaths are re-estimated using the EPA's radon dose-response, incorporating the full distribution of **predicted** indoor radon exposure across the population of single-family homes.

4.1.1 Census Tract Map

Figure 10 depicts an updated radon map for the lowest level of single-family homes (top map) at the census tract level, alongside the original 1993 EPA Map of Radon Zones (bottom map). A notable visual difference between these maps is the reduced prevalence of Zone 1 areas in the updated version. The new predictive approach indicates that approximately 10 percent of homes have long-term predicted radon values of 4 piC/L or higher, contrasting with the EPA's Zone 1 designations, which represented about 28 percent of homes. According to the new model, the mean long-term concentration in the first floor of US single-family homes is 1.8 piC/L, while in the lowest level of these homes, it is 2.2 piC/L.

The predictive map of indoor radon offers a markedly different risk profile compared to the 1993 EPA map, suggesting that the latter is outdated and potentially misrepresents the long-term distribution of indoor radon exposure. Several factors likely contribute to the disparities between the predictive and EPA maps.

First, let's consider the construction of the 1993 EPA map. The EPA Committee utilized five factors known to be significant indicators of radon potential in order to establish the Zone designations for each county in the US: (1) Indoor radon measurements; (2) Geology; (3) Aerial Radioactivity (NURE equivalent uranium data); (4) Soil parameters; and (5) Foundation types. The Committee assigned points on a scale of 3-17 based on their assessment of the individual contribution of each category to radon potential across approximately 360 distinct geological provinces. Once the total point count was determined, the final zone designation for each county was assigned based on the geological province(s) it overlapped, with counties overlapping multiple provinces being assigned based on the majority rule of the total overlapping land area.

It's worth noting that this method was developed at a time when computational and digital resources were still in their nascent stages. An evident limitation is how the EPA Committee assigned points based on the NURE uranium data, taking the "visual average" of the NURE points for each geological province. Additionally, the committee relied on uranium in the first 30 cm as represented by the NURE measurements. However, they failed to account for weather variations that influence the production and movement of radon from the ground into the atmosphere. The EPA attempted to address this issue by supplementing the analysis with ag-

gregated soil parameters and local geological survey information (where available), but this approach is susceptible to the same limitations discussed in Section 3. The categorical assignment of rock lithology and its association with indoor radon cannot accurately capture the spatial variation of radon-forming elements within geological units.

Furthermore, the EPA map was based on a relatively small sample of approximately 60,000 short-term radon tests, which provide limited approximations of long-term concentrations. Short-term tests are subject to significant variations influenced by weather conditions, and activated charcoal tests only capture the last 12-20 hours of exposure.

Lastly, it is essential to consider that new homes have been constructed with radon-resistant features, and existing homes have undergone remediation measures. The perceived lower risk in Zone 1 areas may be attributed to successful radon mitigation efforts in those counties. Exploring this possibility would be an intriguing subject for further research.

An additional intriguing disparity between the predictive and EPA maps emerges in the identification of counties initially categorized as EPA Zones 2 and 3, but now exhibit long-term predicted indoor radon levels exceeding 4 piC/L. Remarkably, among the 74 million single-family residences, approximately 2.2 percent of homes have been inaccurately classified (or about 1.6 million residences). It is unlikely that these households have recognized their true long-term risk, given that radon testing and mitigation are heavily influenced by the EPA map and are concentrated in Zone 1. Figure 11 depicts these areas and maps the percentage of

homes with long-term predicted indoor radon levels exceeding 4 piC/L, alongside the original 1993 EPA Map for comparison. Notably, these misclassified regions predominantly concentrate in NV/ID, KS, NE/SD, TN, WI, and NH. If the predictive model effectively captures long-term radon risks, a vital next step would involve collaborating with the EPA to efficiently educate individuals residing in these areas about their escalated radon risk.

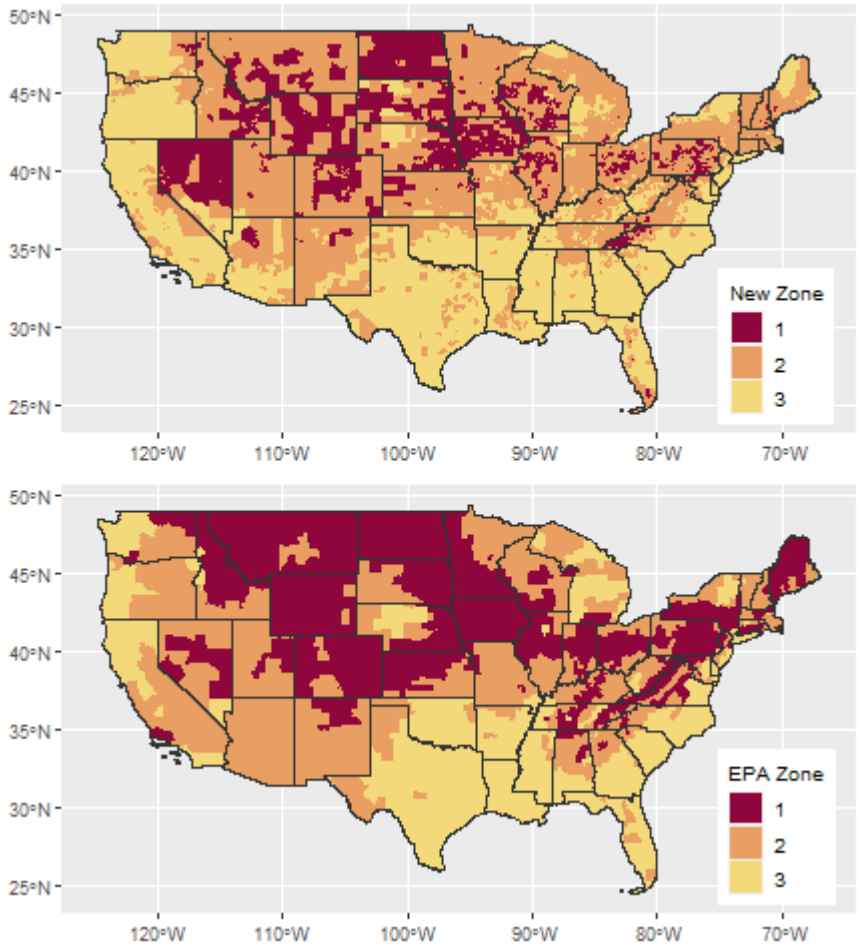
Next, the census tract map provides significant policy-relevant information. An analysis of the mean long-run indoor radon concentrations reveals some significant differences in household income at various levels of radon exposure. Among all census tracts with median incomes less than the first quartile, those located in the New Zone 1 (with predicted long-run indoor radon of 4 piC/L or more) have household incomes of $900 on average less than those located in the New Zone 3 (less than 2 piC/L). This is a descriptive result but demonstrates the importance of understanding the heterogeneity in the characteristics of consumers that are exposed to high levels of indoor radon. Additionally, it begs the question of whether indoor radon is driving the income differences observed between these census tracts. From a public policy perspective, exploring and understanding the implications of lower household incomes in high-radon areas compared to low-radon areas is crucial for addressing any disparities and encouraging equitable outcomes. This is an interesting future area of research and the models and maps developed in this dissertation can be used as tools in such an analysis. Figure 12 shows the New Zones for census tracts with median incomes less than the 25th percentile (top map), with the New Zone designations for the census tracts with

median incomes greater than the 75th percentile (bottom map). Future work could target homes that reside in the New Zone 1 depicted in the top map, which represent the poorest US residents. Compare this to the bottom map, which represent relatively wealthier areas. Given that income is most likely a critical factor that drives mitigation and the ability to avoid indoor radon exposure, understanding the distribution of indoor radon risk highlights the need for targeted interventions to improve housing conditions, raise awareness about radon risks, and provide resources for testing and mitigation in economically disadvantaged areas.

4.1.2 Square Kilometer Maps

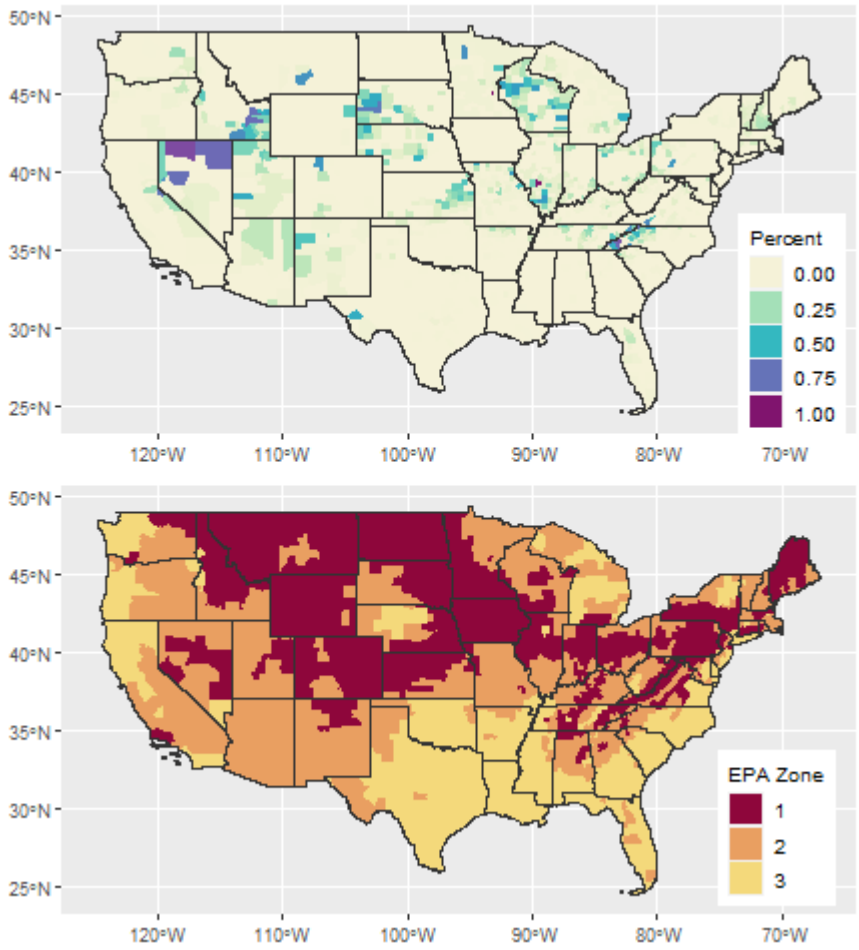
As part of the pilot feature, I have generated a predicted long-term map of indoor radon for the State of Tennessee at a 1 km resolution, as illustrated in Figure 13. The map showcases the average long-term radon levels in the lowest livable level of homes and reveals substantial variation across the entire state.

Figure 10: A New Radon Map at the Census Tract and the EPA Map of Radon Zones



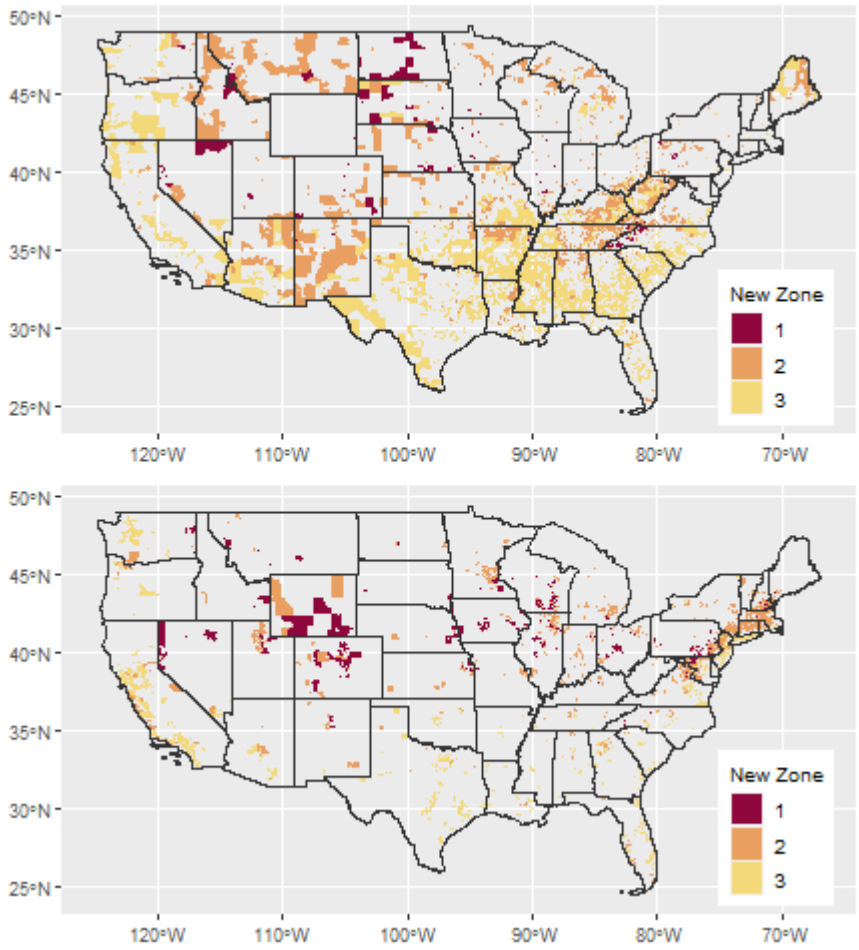
Notes: This figure provides a comparison between the new predictive map of long-term indoor radon and the original EPA Map of Radon Zones. In the top figure, the "New Zone" is assigned based on the average long-term prediction at the census tract level. Census tracts with average long-term predictions exceeding 4 piC/L are classified as New Zone 1, those with average long-term predictions ranging from 2 to 4 piC/L are designated as New Zone 2, and census tracts with average long-term predictions below 2 piC/L are labeled as New Zone 3. The EPA Zones 1-3 are defined similarly, except they represent "probable" screening levels of indoor radon for an area.

Figure 11: EPA Zone 2 and 3 Counties with the Percentage of Homes with more than 4 piC/L



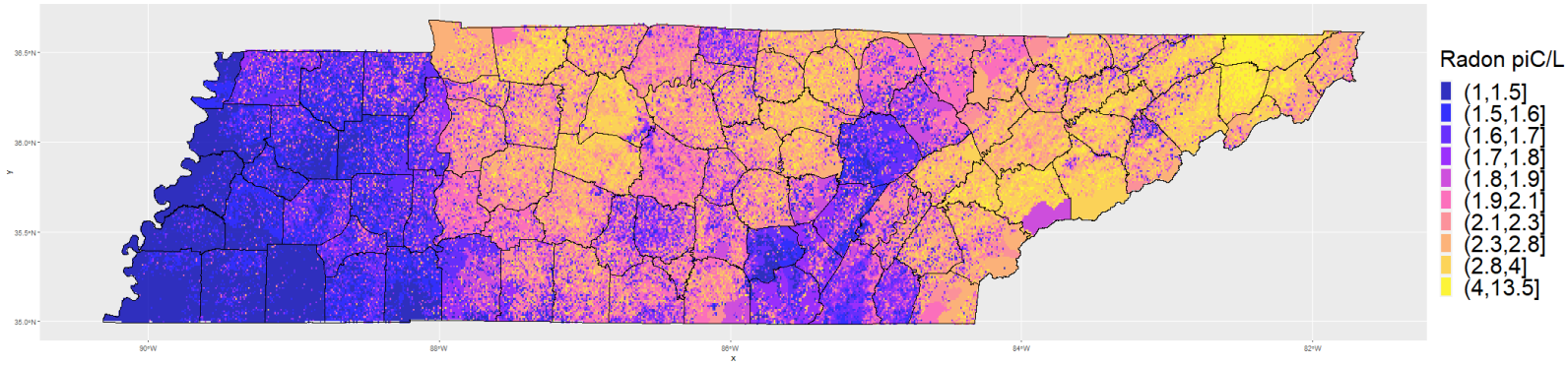
Notes: The top figure provides a visual representation of the percentage of homes situated in the EPA Zones 2 and 3, yet possessing predicted long-term indoor radon levels surpassing 4 piC/L in the lowest livable level of the home. The mapping is delineated by county, allowing for a comprehensive overview. To provide context, the bottom figure displays the original 1993 EPA map of radon zones, serving as a reference point for comparison.

Figure 12: New Zone Designations among Poor and Wealthy Census Tracts



Notes: This figure shows the New Zone designations for the poorest and wealthiest census tracts. The top map shows the New Zone designations based on the long-term predictions for census tracts with median household incomes less than the 25th percentile of the US distribution from the 2010 Census. The bottom map shows the New Zone designations for census tracts with median household incomes greater than the 75th percentile.

Figure 13: Predicted Long-Term Indoor Radon for the State of TN by Square Kilometer



The figure presented illustrates the average long-term indoor radon predictions per square kilometer in the State of Tennessee. It is important to note that certain square kilometers do not contain any single-family properties. In such cases, these square kilometers were assigned the mean long-term indoor radon value of the census tract to which they belong. However, if the census tract itself lacks single-family structures, the square kilometers were assigned the mean long-term indoor radon value of the respective county.

The next figure, Figure 14, provides a closer look at Davidson County (Nashville, TN) and highlights the ten quantiles of predicted indoor radon, offering a more detailed understanding of the distribution. The highest quantile, depicted by the bright yellow square kilometers, represents areas with long-term predicted radon values ranging from 4 to 13.5 piC/L. Notably, the original EPA map does not provide any insights into the distribution of indoor radon risk within counties or census tracts, making the new predictive approach a valuable resource for assessing radon risk at a more localized level. This advancement significantly contributes to the accurate and precise mapping of long-term indoor radon exposure levels across the United States.

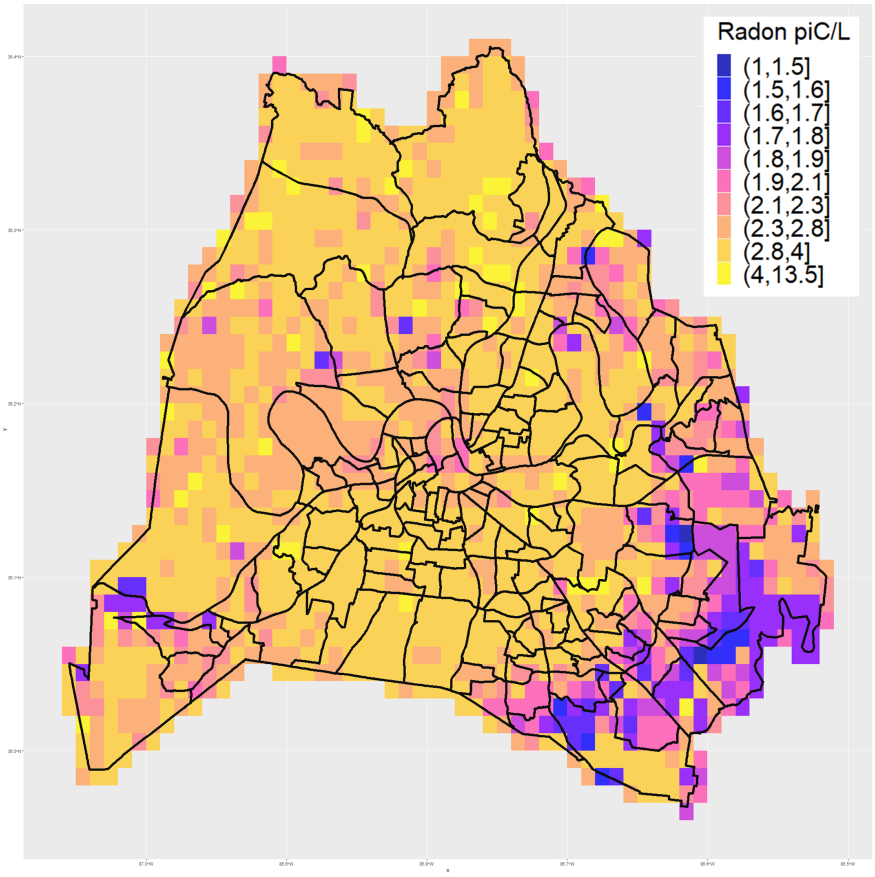
4.1.3 Radon-Attributed Lung Cancer Deaths

Finally, the household predictions can be used to provide updated estimates of radon-attributed lung cancer deaths by applying the EPA's linear dose-response model for lung cancer risk due to radon exposure.[21] Additionally, assuming that the long-term predictions consistently approximate true exposures,[22] I can estimate the probability of lung cancer originating from each single-family home by applying the parameter estimates of the linear dose-response to the *continuous* measures of predicted long-term exposures among all 74 million single-family homes.

[21] Refer to the EPA's table here: https://www.epa.gov/radon/health-risk-radon

[22] Furthermore, as discussed in Section 3.4.2, the long-term predictions tend to underestimate observed long-run indoor radon tests, suggesting a likely underestimation of lung cancer cases as well.

Figure 14: Predicted Long-Term Indoor Radon for Davidson County,TN



The figure presented illustrates the average long-term indoor radon predictions in the lowest level of a home, averaged by square kilometer in Davidson County, TN.

Several implicit assumptions are necessary to accomplish this estimation. First, a proportion of the US population smokes, and the EPA provides separate linear models for smokers and non-smokers. To reflect this, I randomly assign 12 percent of all single-family homes to the smoking linear model and the remaining 88 percent to the non-smoking linear model. Next, each home is assigned both a basement and first-floor prediction, regardless of the basement indicator, which is imputed for a certain proportion of homes. For simplicity, a home is given a basement value if its predicted value exceeds 0.5. Additionally, to assign exposure levels, I make assumptions based on the amount of time residents spend in the basement or first floor. Two exposure mechanisms are considered: the conservative approach uses only first-floor exposure, while the second approach assumes that 60 percent of the time individuals are exposed to first-floor radon and 40 percent to basement levels of radon.[23] Finally, to estimate the radon-attributable fraction, I define the reference group as 0-2 piC/L of exposure, medium risk as 2-4 piC/L, and high risk as greater than 4 piC/L, following the EPA's Zone designations.[24]

Based on these assumptions, new estimates of the radon attributable fraction range from 0.24 for 100 percent first-floor exposure to 0.28 for a split time between the basement and first floor. Applying these estimates to the 127,000 lung cancer deaths in 2023 implies that radon causes between 30,480 and 35,560 yearly lung cancer deaths, representing a substantial 50 percent increase compared to the

[23] Basement levels of radon generally are higher than first-floor levels.

[24] For a discussion of the CDC's method for estimating the attributable fraction of alcohol, refer to https://www.cdc.gov/alcohol/ardi/methods.html.

EPA's estimates. To align with the EPA estimates, one must assume a 25 percent lifetime exposure to the first-floor long-term concentrations, averaging 1.3, 2.4, and 3.3 piC/L in the New Zones 3, 2, and 1, respectively.

The discrepancies between these estimates and the EPA estimates can primarily be attributed to the implicit assumptions made in the EPA approach. In their calculations of the lifetime risks of radon-induced lung cancer death, the EPA relied on life tables and assumed a constant mean exposure of 1.25 piC/L for the entire US population, relative to the baseline lung cancer risk for the country (US Environmental Protection Agency, 2003). The mean exposure assumption was based on estimates from a study conducted by Marcinowski et al. (1994), which sampled 5694 homes across the US. This study found long-term averages of 1.25 piC/L and a standard deviation of 0.08 piC/L. However, it is important to note that this sample size represents less than 0.008% of the US housing stock, making it unlikely that extrapolating this assumed distribution accurately captures the true extent of exposure. On the contrary, the data from 1.1 million single-family radon tests reveal significantly higher means (with expected higher standard deviations for short-term tests) compared to the EPA assumptions, even within EPA Zone 3 areas. Additionally, the long-term radon tests conducted by Alpha Energy indicate a mean concentration of 5.4 piC/L and a sample standard deviation of 4.5 piC/L.

Furthermore, it is crucial to account for the continuous nature of radon exposure when attempting to estimate risks, particularly in EPA Zones 2 and 3. Given that there is no level of indoor radon deemed safe and the dose-response is assumed to be linear, significant cases of lung cancer can result from exposures to

radon levels above 2 piC/L. In fact, under the predicted New Zones, nearly 35 percent of all homes have indoor radon levels between 2-4 piC/L. Therefore, accurately estimating the true impact of indoor radon exposure on the prevalence of lung cancer and death in the United States necessitates accounting for the full distribution of radon exposure. These findings highlight the limitations of relying solely on small samples and underscore the need to consider a broader and more comprehensive dataset when assessing the risks associated with radon-induced lung cancer.

Finally, I estimate the cost-effectiveness of targeted remediation at the household level. Assuming a remediation threshold of 2.7 piC/L, remediating homes accurately predicted above this limit could potentially reduce yearly radon-attributed lung cancer deaths by approximately 8,800, considering the split-time assumption of radon exposure between the basement and first-floor levels. The total costs, assuming $2,000 per home for testing and remediation for accurately identified homes[25] and $500 for those inaccurately predicted, would amount to approximately $3.02 billion. This translates to approximately $354,545 per avoided lung cancer death. Similarly, at the 4 piC/L remediation level, targeted remediation could save 992 yearly lung cancer deaths at a cost of $333,000 per death averted. Assuming a 20/80 radon exposure split between the lowest level and first floor, and a remediation threshold of 3 piC/L, targeted remediation saves 4,010 yearly lung cancer deaths at a cost of about $400,000 per lung cancer death. These findings

[25]The short-term predictive model accurately identifies 72 percent of the homes with long-term average indoor radon values above 2.7 piC/L in the out-of-sample alpha track tests.

suggest potentially significant increases in the cost-effectiveness of remediation using the predictions from this model.

5 Conclusions

This paper has aimed to address the significant challenges associated with accurately modeling and mapping indoor radon exposure levels across the United States. Through a series of comprehensive steps and models, this work has made notable contributions to the field and shed light on the limitations of existing approaches.

The initial focus of this study involved generating a complete map of rock lithology for the contiguous United States. This map, combined with geochemical surveys, long-term weather data, soil characteristics, and terrain features, allowed me to model the distribution of radon-accessory elements. Subsequently, I employed nonparametric estimation techniques to predict daily atmospheric bi-214, a critical proxy variable for outdoor/surface-level radon, based on the imputed lithology, predicted radon-accessory element distributions, soil characteristics, terrain features, land imperviousness, daily weather, and other relevant factors.

To further refine our understanding of indoor radon exposure, I developed models to account for endogenous building characteristics, sample selection, and the daily predicted atmospheric bi-214, weather data, and remaining building characteristics. This enabled me to model the data-generating process of radon

that migrates indoors.

Throughout this work, it has become evident that the modeling process involves a series of assumptions, each with its own limitations. However, it is important to note that the current model lacks the ability to generate confidence and prediction intervals for predicted indoor radon in the population represented in the ZTRAX database. The computational cost associated with generating predictions at each step has posed restrictions on providing estimates of uncertainty. Moreover, the use of nonparametric estimation techniques and the extensive volume of predictions required for all single-family homes across the United States present additional challenges in terms of computational resources.

Considering the lack of confidence and prediction intervals, it is crucial to acknowledge that the predictive map, radon-attributable fraction, and cost-effective mitigation estimates presented in this study are still in the process of development. These estimates are expected to undergo further refinement and enhancement as additional data and iterations are incorporated, allowing for a more robust assessment of uncertainty.

Nevertheless, the findings have revealed significant differences between a predictive map of long-term indoor radon and the original 1993 EPA Map of Radon Zones. The prevalence of Zone 1 areas, representing higher radon levels, has reduced substantially in the predictive map. Approximately 10 percent of homes have long-term predicted radon values of 4 piC/L or more, compared to about 28 percent represented by the EPA Zone 1 designations. This indicates that the 1993 EPA map may be outdated and potentially misrepresents the long-term distribu-

tion of indoor radon exposure.

The new map has also highlighted the misclassification of counties originally labeled as EPA Zones 2 and 3, but with long-term predicted indoor radon levels exceeding 4 piC/L. Approximately 2.2 percent of homes among the 74 million single-family residences have been misclassified, raising concerns about the accurate identification of long-term radon risks. Given the influence of the EPA map on testing and mitigation efforts, it is crucial to collaborate with the EPA to educate individuals residing in these areas about their elevated radon risk.

The development of a predicted long-term indoor radon map for the State of Tennessee at a high-resolution level has further demonstrated the significant spatial variation in radon levels within counties and across the state. By providing information on the distribution of indoor radon risk within small geographical areas, the predictive approach offers valuable insights that were previously absent from the original EPA map.

Furthermore, this research has provided new estimates of the radon-attributable fraction, indicating that radon causes between 30,480 to 35,560 yearly lung cancer deaths, representing a 50 percent increase relative to the EPA's estimates. These estimates consider different exposure scenarios, and it becomes evident that assuming a 25 percent lifetime exposure to first-floor long-term concentrations would align radon-attributed lung cancer deaths with the EPA estimates.

I have also evaluated the cost-effectiveness of targeted remediation at the household level. Based on the predictions, remediation of homes accurately identified above the WHO 2.7 piC/L threshold would potentially reduce yearly radon-

attributed lung cancer deaths by approximately 8,800. The estimated cost of testing and remediation for accurately identified homes, along with the associated costs for testing inaccurately predicted homes, amounts to approximately $3.02 billion. This translates to a cost of about $354,545 per avoided lung cancer death. Similar exercises conducted at different remediation levels indicate potentially significant increases in the cost-effectiveness of remediation using the predictions from our model.

In conclusion, this research has contributed to the advancement of accurately and precisely mapping long-term indoor radon exposure levels across the United States. By addressing the limitations of existing models and providing more nuanced insights, this research paves the way for improved risk assessment, mitigation strategies, and public awareness regarding indoor radon exposure. Future studies should focus on refining the models, incorporating additional data sources, and collaborating with relevant partners to further enhance our understanding of this important public health concern.

6 Appendix

6.1 The EPA Map

The Environmental Protection Agency (EPA) published the Map of Radon Zones in 1993, which is widely used to provide information about indoor radon risks. The EPA map estimates that one-third of all continental U.S. counties are high-risk zones for indoor radon. The Map of Radon Zones was developed to assist national, state, and local governments and organizations in targeting their radon program activities and resources. Each of the three Zone designations represents an estimate of radon potential in the lowest livable area of a structure, but it cannot estimate actual exposures to radon.

The EPA Committee used five factors that were known to be important indicators of radon potential to construct the Zone designations for each county across the US: (1) Indoor radon measurements; (2) Geology; (3); Aerial Radioactivity (NURE equivalent uranium data); (4) Soil parameters; and (5) Foundation types. The Committee assigned numerical values based on the assessment of each category's individual contribution to radon potential for about 360 separate geological provinces.

The Committee acknowledges that this extrapolation may hide significant variation within counties. Figure 15 shows the index approach used by the EPA Committee to construct the Zones.

Figure 15: EPA Index Approach for Radon Mapping

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

INCREASING RADON POTENTIAL →

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

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

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

SCORING:

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

POSSIBLE RANGE OF POINTS = 3 to 17

6.2 Figures

Figure 16: The Decay Chain of Thorium-232

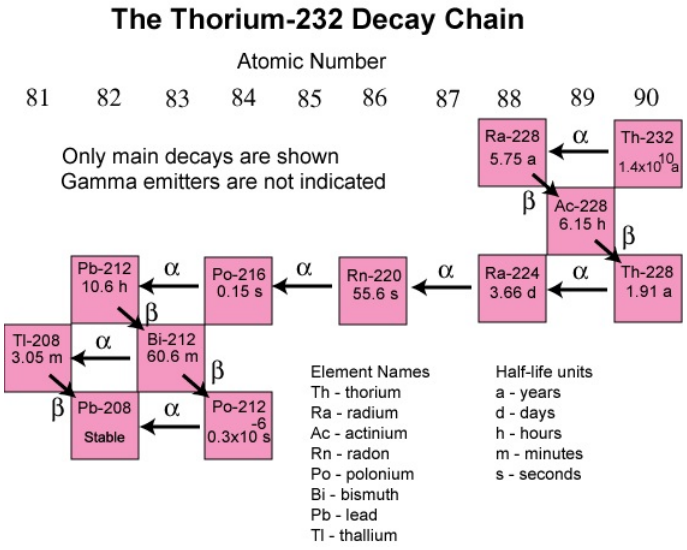
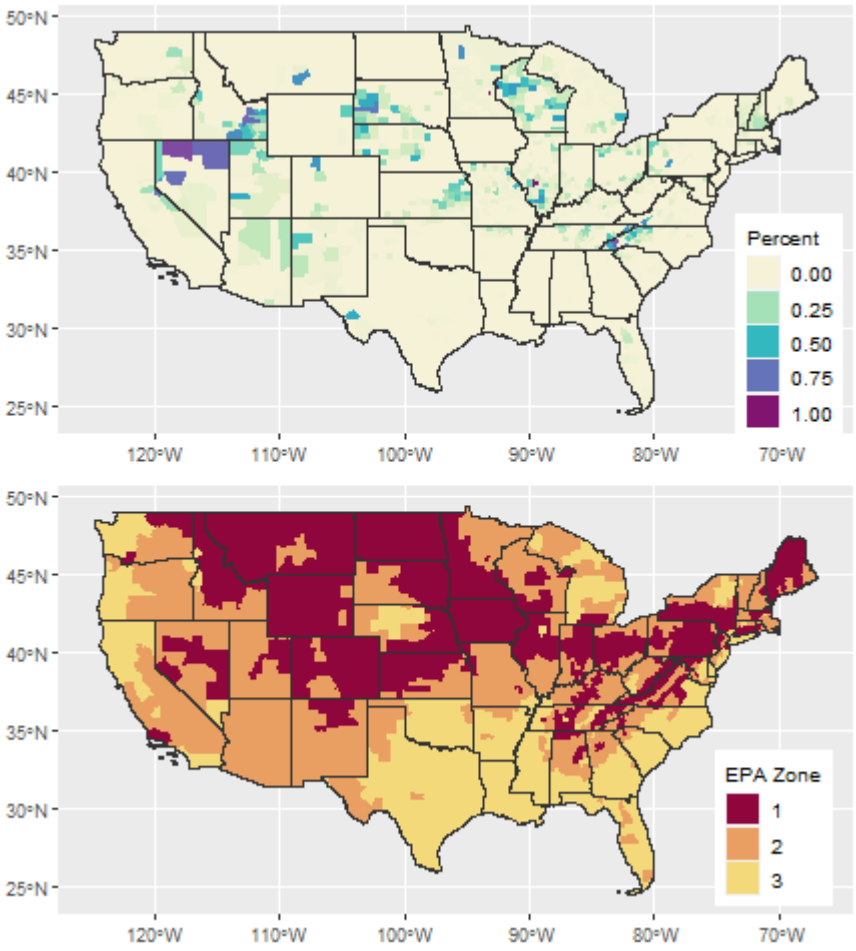
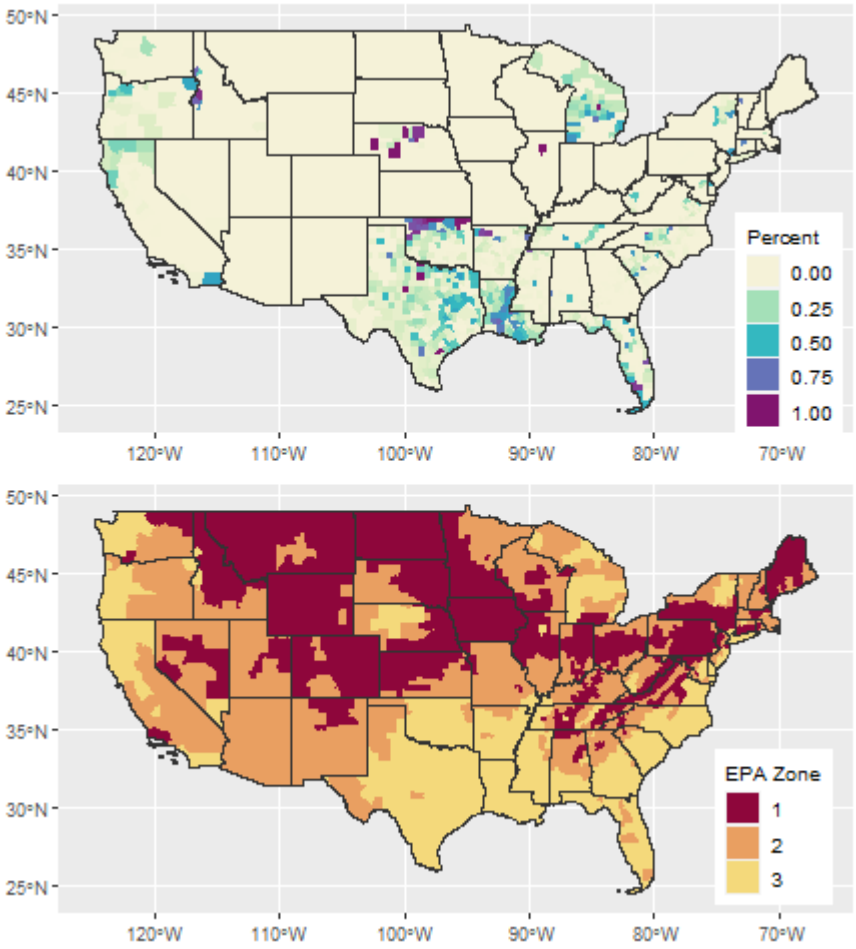


Figure 17: Counties with Predicted Radon 4 piC/L or More



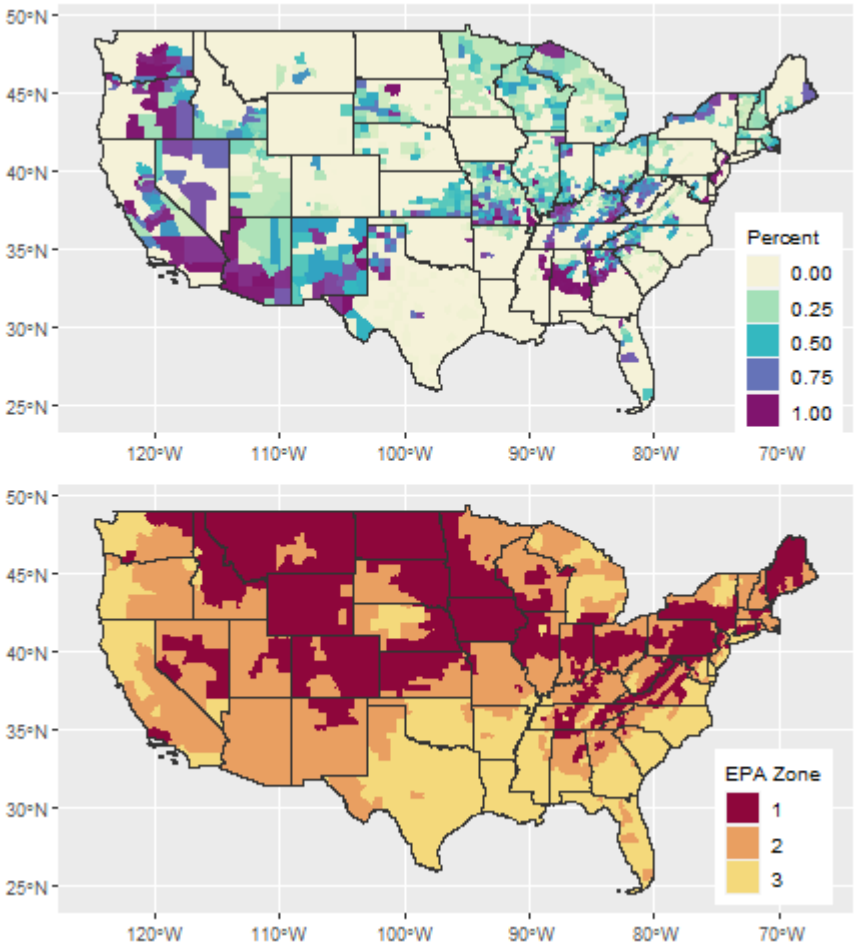
Notes: The top figure shows the percentage of homes that are located in the EPA Zones 2 and 3 with predicted lowest-level indoor radon at 4 piC/L or more. The bottom map shows the 1993 EPA map of Zones for reference.

Figure 18: EPA Zone 3 Counties Missclassified



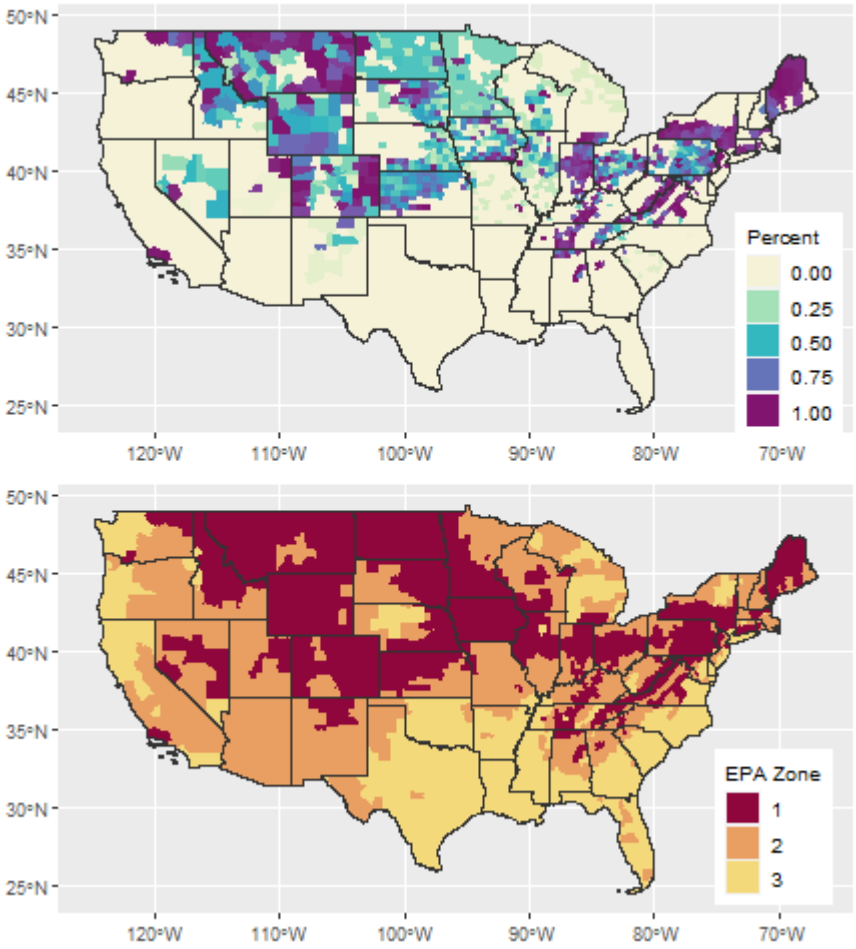
Notes: The top figure shows the percentage of homes that are located in the EPA Zone 3 that are misclassified under the new map. A home is defined as misclassified if its long-term predicted radon is not less than 2 piC/L, the definition of EPA Zone 3. The bottom map shows the 1993 EPA map of Zones for reference.

Figure 19: EPA Zone 2 Counties Missclassified



Notes: The top figure shows the percentage of homes that are located in the EPA Zone 2 that are misclassified under the new map. A home is defined as misclassified if its long-term predicted radon does not fall within 2-4 piC/L, the definition of EPA Zone 2. The bottom map shows the 1993 EPA map of Zones for reference.

Figure 20: EPA Zone 1 Counties Missclassified



Notes: The top figure shows the percentage of homes that are located in the EPA Zones 1 that are misclassified under the new map. A home is defined as misclassified if its long-term predicted radon is not greater than 4 piC/L, the definition of EPA Zone 1. The bottom map shows the 1993 EPA map of Zones for reference.

6.3 Tables

Table 3: Descriptive Statistics of Predicted Radon

	Zone 1	Zone 2	Zone 3	Zone 1*	Zone 2*	Zone 3*
Lowest-Level Mean	3.28	2.16	1.38	4.86	2.85	1.36
Lowest-Level SD	1.27	0.98	0.49	0.83	0.56	0.29
First-Floor Mean	2.68	1.85	1.27	3.33	2.36	1.33
First-Floor SD	0.83	0.63	0.30	0.77	0.61	0.28
N	2,083,376	2,639,091	2,617,623	722,722	2,464,913	4,152,455

Table reports the descriptive statistics of the radon predictions among a 10 percent sample of the ZTRAX population of single-family homes. The columns Zone 1 - Zone 3 are the original EPA Zone designations. The columns Zone 1* - Zone 3* are the new zone designations under the predictive map.

Table 4: OLS of New Zone on Median Income at the Census Tract

	Census Tract Medium Income				
	Incomes $<$ 25th	25th $<$ Incomes $<$ 50th	50th $<$ Incomes $<$ 75th	Incomes $>$ 75th	All Census Tracts
	(1)	(2)	(3)	(4)	(5)
Intercept	22,610.29***	37,288.43***	43,891.13***	−84,676.77***	−41,991.38***
	(437.10)	(324.59)	(610.30)	(3,662.48)	(918.01)
New Zone 2	−356.18***	73.98	124.50	3,036.99***	1,196.86***
	(121.37)	(73.52)	(112.35)	(488.57)	(213.12)
New Zone 1	−860.94***	332.66***	408.38**	143.63	891.72**
	(256.08)	(128.39)	(180.16)	(896.61)	(376.42)
Percent High School	106.55***	84.66***	186.63***	2,034.13***	1,342.24***
	(3.89)	(3.04)	(5.38)	(33.23)	(7.01)
Percent White	12.50***	−10.20***	−35.79***	−224.00***	−169.82***
	(3.47)	(2.51)	(3.87)	(15.44)	(6.74)
Percent Black	−54.49***	−27.50***	−52.48***	−413.32***	−394.25***
	(3.56)	(2.92)	(4.79)	(22.19)	(7.49)
State fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	17,907	17,916	17,916	17,917	71,666

*p<0.1; **p<0.05; ***p<0.01

Notes: This shows a simple linear regression of the New Zone categories on the median income at the census tract level. The first four columns represent separate models for census tracts with incomes below the first quartile, between the first and second, between the second and third, and above the third quartile. Column (5) shows the results for the US overall. These results are descriptive and are not meant to be interpreted as causal. They represent average differences in income across the New Zone designations, while contolling for demographics and State fixed effects.

Tables

Figures

References

Athey, S., Tibshirani, J., and Wager, S. (2018). Generalized random forests.

Blackmore, D., Herman, M., and Woodward, J. (1982). Heavy gas dispersion models. *Journal of Hazardous Materials*, 6:107–128.

CDC (2018). Cdc national environmental public health tracking network. Accessed From: https://ephtracking.cdc.gov/DataExplorer.

Chen, C., Thomas, D., and Green, R. (1995). Modeling of radon transport in unsaturated soil. *Journal of Geophysical Research*, 100(B8):15517–15525.

Danielson, J. and Gesch, D. (2011). Global multi-resolution terrain elevation data 2010 (gmted2010). US Geological Survey Open-File Report 2011–1073.

Das, M., Newey, W., and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70:33–58.

Dewitz, J. and U.S. Geological Survey (2021). National land cover database (nlcd) 2019 products. U.S. Geological Survey data release.

Duval, J., Carson, J., Holman, P., and Darnley, A. (2005). Terrestrial radioactivity and gamma-ray exposure in the united states and canada. US Geological Survey Open-File Report 2005–1413.

Horton, J., Juan, C. S., and Stoeser, D. (2017). The state geologic map compilation (sgmc) geodatabase of conterminous united states. US Geological Survey Data Series 1052.

International Atomic Energy Agency (2003). Guidelines for radioelement mapping using gamma ray spectrometry data.

Kazkaz, K. and Walsh, N. (2018). Combining stochastics and analytics for a fast monte carlo decay chain generator. *arXiv preprint arXiv:1104.2834v1*.

King, P. and Beikman, H. (1974). The geologic map of the united states. US Geological Survey Professional Paper 901.

Klusman, R. and Jaacks, J. (1987). Environmental influences upon mercury, radon, and helium concentration in soil gases at a site near denver, colorado. *Journal of Geochemical Exploration*, 27:259–280.

Lehnert, A. and Kearfott, K. (2010). An equilibrium-based model for measuring environmental radon using charcoal canisters. *Health Phys*, 99(2):S154–S163.

Livneh, B., Rosenberg, E., Lin, C., Nijssen, B., Mishra, V., Andreadis, K., Maurer, E., and Lettenmaier, D. (2013). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states: Update and extensions. *J. Climate*, 26.

LKB Resources Inc (1979). Nure aerial gamma-ray and magnetic reconnaissance survey: Ne washington area. Prepared for the Department of Energy.

Marcinowski, F., Lucas, R. M., and Yeager, W. M. (1994). National and regional distribution of airborne radon concentrations in u.s. homes. *Health Physics*, 66(6):699–706.

Nazaroff, W. (1992a). Radon transport from soil to air. *Reviews of Geophysics*, 30(2):137–160.

Nazaroff, W. W. (1992b). Radon transport from soil to air. *Review of Geophysics*, 30(2):137–160.

PRISM Climate Group (2014). Oregon State University.

Pyngrope, A., Saxena, A., Khardewsaw, A., Sharma, Y., and Sahoo, B. (2022). Effect of soil's porosity and moisture content on radon and thoron exhalation rates. *Journal of Radioanalytical and Nuclear Chemistry*, 331:1975–1984.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Smith, D., Cannon, W., Woodruff, L., Solano, F., Kilburn, J., and Fey, D. (2013). Geochemical and mineralogical data for soils of the conterminous united states. US Geological Survey Data Series 801.

Soil Survey Staff, Natural Resources Conservation Service (2016). Digital general soil map of u.s.: General soil map (statsgo2). U.S. Department of Agriculture, Natural Resources Conservation Service.

Stekhoven, D. and Buehlmann, P. (2012). Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

US Environmental Protection Agency (2003). Epa assessment of risks from radon in homes. Technical report, Air and Radiation, US EPA.