

Radon Exposure Dataset

Dakotah Maguire¹, Jeremy Logan¹, Heechan Lee², and Heidi Hanson¹

¹Oak Ridge National Laboratory

²Georgia Institute of Technology

1 Introduction

Exposure to elevated radon levels in the home is one of the leading causes of lung cancer in the world [1, 2, 3]. The following study describes the creation of a comprehensive, state-level dataset designed to enable the modeling and prediction of household radon concentrations at Zip Code Tabulation Area (ZCTA) and sub-kilometer scales. Details include the data collection and processing involved in compiling physical and demographic factors for Pennsylvania and Utah. Attempting to mitigate this risk requires identifying the underlying geological causes and the populations that might be at risk. This work focuses on identifying at-risk populations throughout Pennsylvania and Utah, where radon levels are some of the highest in the country [3].

The resulting dataset harmonizes geological and demographic factors from various sources and spatial resolutions, including temperature, geochemistry, and soil characteristics. Demographic variables such as the household heating fuel used, the age of building, and the housing type provide further insight into which populations could be most susceptible in areas with potentially high radon levels.

This dataset also serves as a foundational resource for two other studies conducted by the authors. The resolution of the data provides a novel approach to predicting potential radon exposure, and the data processing conducted for these states can be scaled up to larger spatial resolutions (e.g., the Contiguous United States [CONUS]) and allow for a broad reclassification of radon exposure potential in the United States.

2 Census Data

Data from the American Community Survey 5-Year (ACS5) and the Decennial Census (DEC) was collected for Pennsylvania and Utah at the block-group and ZCTA levels for the years 2000, 2013, 2015, and 2020. The variables collected are presented in Table 1. Next, block-group data for each variable was converted to H3 format by using the `area.interpolate()` function from the PySAL Tobler library [4].

Differences in census table definitions across surveys require adjustments before an analysis can be conducted. After the 2000 DEC, the long-form survey was replaced with the American Community Survey [5]. The US Census Bureau and other organizations such as IPUMS provide cross walks and documentation on how to correctly modify the data to ensure the variables are as close to each other as possible for analysis [6, 7].

Table 1: Census Variables Mapping and Descriptions

DEC Code	ACS Code	Description
P053001	B19013	Median Household Income in 1999
P053	B19013	Median Household Income in the Past 12 Months (Inflation-Adjusted Dollars)
P088	C17002	Ratio of Income to Poverty Level in the Past 12 Months
H001	B25001	Housing Units
H006	B25002	Occupancy Status
H008	B25004	Vacancy Status
H030	B25024	Units in Structure
H034	B25034	Year Structure Built
H023	B25017	Rooms
H024	B25018	Median Number of Rooms
H040	B25040	House Heating Fuel

Three tables in the collected data required modification.

- (1) The variables in table P088 from the 2000 DEC and table C17002 from the ACS5 needed slight adjustments due to changes in table structure made in the 2010 DEC. *H008003 Total Rented or sold, not occupied* was split into *Estimate Total Sold, not occupied* and *Estimate Total Rented, not occupied*. For analysis, these variables were summed under the new name, *B25004_003_005*.
- (2) The *H034: Year Structure Built* table in the 2000 DEC ends with the year 2000. The 1990–2000 decade was broken up into three date ranges—(1) *H034002: Built 1990 to 1994*, (2) *H034003: Built 1995 to 1998*, and (3) *H034004: Built 1999 to March 2000*. These three variables were summed to match the ACS5 variable *B25034_004: Estimate Total Built 1990 to 1999*.
- (3) Two changes were made for the variables in the P088 DEC table to reflect changes made in the 2010 DEC. *P08800: Total 0.50 to 0.74* and *P088004: Total 0.75 to 0.99* were combined and summed to match *C17002_003 Estimate Total 0.50 to 0.99*. Likewise, *P088007: Total 1.50 to 1.74* and *P088008: Total 1.75 to 1.84* were combined to align with *C17002_006: Estimate Total 1.50 to 1.84*.

2.1 Urban-Rural Classification

The DEC provides a classification system of geospatial layers at varying spatial resolutions and assigns an urban or rural designation to each component of the layer [8]. The data used was provided in a text file that contained urban block groups from the 2020 ACS5. This urban classification was appended to the other ACS5 data using the block-group geographic identifiers. Areas not classified as urban were considered rural.

3 Soil and Geological Data

The soil dataset was created by using the gridded National Soil Survey Geographic Database (gNATSGO) provided by the US Department of Agriculture’s (USDA’s) Natural Resources Conservation Service [9]. Based on the available literature [10, 11], thirty variables that have either shown correlations or are thought to have correlations to elevated radon levels were chosen and are presented in Table 2. Various soil characteristics (e.g., density, composition) can influence radon levels by affecting fluid movement and gas permeability. Additionally, soil moisture and water retention capacity can both play a role in radon levels [12, 13]. Other factors that may influence radon accumulation or release, such as soil surface conditions or the presence of a basement, were also included in the analysis.

The ArcGIS Toolbox Soil Data Development Toolkit, designed and distributed by the US Geological Survey (USGS), and the 30-m CONUS gNATSGO grid were used to create this data. Variable processing criteria were standardized based on the data type (continuous or categorical). The data criteria included a soil depth of 0–200 cm, and the measure was represented as a weighted average or as a dominant condition for categorical data. Additional processing was required to format the

data to make it usable for analysis. This process is described in *How To Create an On Demand Soil Property or Interpretation Grid* from gNATSGO [14]. The definitions of the variables can be found in the *SSURGO Metadata Table Column Descriptions* document [15].

Table 2: Soil Variable Names and Their Dataset Abbreviations

Variable	Abbreviation
Available Water Capacity WTA, 0–200 cm	AWC
Available Water Storage WTA, 0–200 cm	AWS
Available Water Supply, 0–25 cm	AWSA
Available Water Supply, 0–50 cm	AWSB
Available Water Supply, 0–150 cm	AWSC
Available Water Supply, 0–100 cm	AWSD
Percent Clay WTA, 0–200 cm	CLAY
Depth to Any Soil Restrictive Layer WA	DARL
Bulk Density, One-Third Bar WTA, 0–200 cm	DB3R
Drainage Class DCD	DRCL
Depth to a Selected Soil Restrictive Layer WTA, Abrupt textural change	DSRL
Dwellings With Basements DCD	DWEL
Dwellings Without Basements DCD	DWLO
Hydric Rating by Map Unit PP	HYDR
Hydrologic Soil Group DCD	HYSG
Saturated Hydraulic Conductivity (Ksat) WTA, 0–200 cm	KSAT
Saturated Hydraulic Conductivity (Ksat), Standard Classes WTA, 0–200 cm	KSCL
Linear Extensibility WTA, 0–200 cm	LEP
Liquid Limit WTA, 0–200 cm	LQLM
Organic Matter WTA, 0–200 cm	OGMT
Plasticity Index WTA, 0–200 cm	PLSL
Percent Sand WTA, 0–200 cm	SAND
Percent Silt WTA, 0–200 cm	SILT
Soil Moisture Class DCD	SMCL
Soil Moisture Subclass DCD	SMSC
Surface Texture DCD, 0–1 cm	SRFT
Soil Temperature Regime DCD	STMP
Soil Taxonomy Classification DCD	STXC
Water Content, 15 Bar WTA, 0–200 cm	WC15
Water Content, One-Third Bar WTA, 0–200 cm	WC3R

After individual rasters were obtained for the gNATSGO variables, each raster was reprojected to EPSG:4326. Next, pixel values were extracted and reorganized into a GeoPandas DataFrame with each row containing the single pixel value along with a point geometry derived from the associated longitude and latitude of the pixel centroid. Using those values, zonal statistics for the target hexagons were computed using `geo_to_h3_aggregate` from the h3-pandas library [16]. To avoid memory overrun, the raster data was processed by using overlapping tiles. While processing each tile, the number of pixels used to produce the hexagon value for that tile was also tracked. This allowed the selection of hexagon values for hexagons in overlapping tile areas to be from a tile in which all contributing pixels were present by simply choosing the computed hex value with the largest number of pixels.

3.1 Lithology

Lithological data was obtained from the USGS [17]. The dataset contained 12 generalized lithological layers, which were reassigned based on the USGS state geologic map compilation for the CONUS. Various elements across different lithologic layers can contribute to increased radon levels. The data was converted to H3 format by using Python and the `polyfill` function from the h3-pandas library.

3.2 Soil Geochemistry

Increased levels of potassium, uranium, and thorium in the soil can correlate with higher concentrations of radon [18]. These variables were obtained from a USGS geochemical and mineralogical survey [19]. The available data was divided into three soil depths: O horizon (0–2 in.), A horizon (2–10 in.), and C horizon (30–48 in.) [20].

The data was processed using the USGS-provided shapefiles and converted into H3 hexagons for ZCTA-level aggregation. GeoTIFF files that contain mineralogical data included separate bands to represent RGB values rather than a single value per element. To interpret this data, the color codes were matched with the legends provided by the USGS, which indicated categorical values. A new raster band was then generated with these categorical values and used zonal statistics (specifically, the mode statistic with the `geo_to_h3_aggregate()` function from the `h3pandas` library) to assign representative values to each H3 level-8 hexagon. This process ensured that each hexagon in the study area had a consistent categorical representation of geochemical attributes, thereby enhancing the granularity and reliability of the radon risk predictions.

4 Elevation Data

Elevation data for Pennsylvania and Utah was collected from the USGS’s Global Multiresolution Terrain Elevation Data 2010 (GMTED2010) [21], which provides 30 arc-second resolution data for the United States. After obtaining 30 arc-second resolution raster format data, the elevation data was re-gridded to H3 level-8 hexagons using the `geo_to_h3_aggregate()` function from the `h3pandas` library [16]. The `geo_to_h3_aggregate()` function was passed raster data values, with the location of each raster pixel represented by its centroid. The function provides aggregated values for each hexagon that contains at least one raster pixel centroid by aggregating all such values according to a specified aggregation operation (mean, in this case) [16]. To fill in missing hexagons (which occurred due to grid misalignment), the authors applied a *ring smoothing* technique in which the values of missing hexagons are computed by averaging the values of adjacent hexagons [16].

5 Hydrologic Data

Hydrologic variables were acquired from the USGS Hydrologic Landscape Regions dataset [22]. Four variables most relevant to hydrological influences on radon mobility were selected from the dataset: *Aquifer Permeability Class*, *Minimum Elevation in Watershed*, *Relief of Watershed*, and *Percent Flat Land in Watershed* [23, 24]. The raw hydrologic data was provided as a vector dataset represented by a shapefile at a 1 km² scale. To align with the other datasets, the data was first read into a Geopandas GeoDataFrame, and then the `polyfill_resample()` method from the `h3-pandas` library was applied to obtain values for each target H3 level-8 hexagon [16].

5.1 The GLoBal HYdrogeology MaPS

The GLoBal HYdrogeology MaPS dataset provides permeability and porosity data as vector (non-gridded) data with an average polygon size of approximately 100 km² [25]. Soil porosity has been identified as a significant factor in radon levels [23, 24]. Conversion from the original shapefile to H3 format was done by using Python and the `polyfill` function from the `h3-pandas` library.

6 Meteorological Data

The Daymet dataset consists of daily values for each variable, and each day is given as a raster layer of modeled values estimated over a 1 km² grid covering North America. For the analysis, data from the CONUS was used for the following four variables during the 2008–2017 time period [26]: *Daily Total Precipitation*, *Snow Water Equivalent*, *Daily Minimum/Maximum 2-m Air Temperature*, and *Vapor Pressure*. These variables were chosen because they provide insight into the atmospheric and hydrological limitations to radon’s movement. These variables can influence the airflow of a building or the emission of radon from the ground, both of which directly affect radon levels in a structure [27, 28].

The daily Daymet grids were first aggregated into monthly averages by using NumPy, and the grid centroids were transformed from the provided Lambert Conformal Conic projection to EPSG:4326. Next, the monthly grids were re-gridded to H3 level-8 by using areal interpolation (specifically the `area_interpolate()` function from the PySAL Tobler library), and finally a ring-smoothing technique was used to fill in missing hexagon values caused by misalignment between raster cells and the H3 hexagons [4].

7 ZCTA Aggregation

Non-census data was converted to h3 level-8 resolution. Because the radon measurements were conducted in residential buildings, hexagons that do not have residents were masked by using LandScan day and night population data. The remaining areas with population were then aggregated to ZCTA resolution. This was accomplished by using areal interpolation (specifically the `area_interpolate()` function from the PySAL Tobler library [4]). This data was then merged with the Pennsylvania ACS5 and DEC ZCTA data.

8 Conclusion

This effort produced a robust, high-resolution dataset that integrates area-based measures and demographic variables that are critical for understanding household radon concentrations in Pennsylvania and Utah. By harmonizing open-source data across multiple sources and spatial scales, the resulting dataset establishes a scalable framework for assessing radon risk at sub-kilometer resolution. This resource supports more precise modeling of exposure potential and insight into the spatial variability of radon at fine geographic scales.

Acknowledgement

Notice: This manuscript has been authored by UT-Battelle LLC under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<https://www.energy.gov/doe-public-access-plan>).

Oak Ridge National Laboratory’s work on the LUCID: Low-dose Understanding, Cellular Insights, and Molecular Discoveries program was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Contract UT-Battelle, LLC- ERKPA71

References

- [1] U.S. Department of Health and Human Services, “Surgeon general releases national health advisory on radon,” 2005, news release issued January 13, 2005. [Online]. Available: <https://www.adph.org/radon/assets/surgeon-general-radon.pdf>
- [2] D. González, Y. Wang, M. Tayarani, X. Jin, S. Mishra, H. Gao, L. Zhang, and P. Waddell, “Air pollution and urban form in 20 global cities,” *Environmental Impact Assessment Review*, vol. 85, p. 106444, 2020. [Online]. Available: <https://doi.org/10.1016/j.eiar.2020.106444>
- [3] L. Gundersen, R. Schumann, J. Otton, R. Dubiel, D. Owen, and K. Dickinson, “Geology of radon in the united states,” in *Geologic Controls on Radon*. Geological Society of America, 1992, vol. 271, pp. 1–16.
- [4] S. J. Rey and L. Anselin, “PySAL: A Python Library of Spatial Analytical Methods,” *The Review of Regional Studies*, vol. 37, no. 1, pp. 5–27, 2007.

- [5] U.S. Census Bureau, “Understanding and using american community survey data: What all data users need to know, chapter 9: Differences between the acs and the decennial census,” 2018, available online at https://www.census.gov/content/dam/Census/library/publications/2018/acs/acs_general_handbook_2018_ch09.pdf.
- [6] S. Manson, J. Schroeder, D. V. Riper, K. Knowles, T. Kugler, F. Roberts, and S. Ruggles, “IPUMS National Historical Geographic Information System: Version 19.0 [dataset],” 2024, minneapolis, MN: IPUMS. [Online]. Available: <http://doi.org/10.18128/D050.V19.0>
- [7] U.S. Census Bureau, “Comparing acs data,” available online at <https://www.census.gov/programs-surveys/acs/guidance/comparing-acs-data.html>.
- [8] —, “A List of 2020 Census Tabulation Blocks Classified as Urban in the 2020 Census with Associated 2020 Census Urban Area Census (UACE) Codes and Names for the U.S. and Puerto Rico [282 MB],” U.S. Census Bureau, 2020, available online at <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html>.
- [9] Soil Survey Staff, “Gridded National Soil Survey Geographic (gNATSGO) Database for the Conterminous United States,” United States Department of Agriculture, Natural Resources Conservation Service, December 19 2023, fY2024 official release. [Online]. Available: <https://nrcs.app.box.com/v/soils>
- [10] G. D. Guthrie, “Environmental toxicology of radon,” *Energy*, vol. 15, no. 3, pp. 527–534, 1990. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0360544290900454>
- [11] N. Barros, I. A. Fajardo, M. Veiga, and J. Cortes, “Study of radon exhalation in soil and building materials: Factors affecting radon exhalation rate,” *Journal of Radioanalytical and Nuclear Chemistry*, vol. 295, pp. 865–874, 2013. [Online]. Available: <https://link.springer.com/article/10.1007/s10967-012-1841-8>
- [12] J. Yang, H. Busen, H. Scherb, K. Hürkamp, Q. Guo, and J. Tschiersch, “Modeling of radon exhalation from soil influenced by environmental parameters,” *Science of The Total Environment*, vol. 656, pp. 1304–1311, 2019, editor: Pavlos Kassomenos.
- [13] K. Sun, Q. Guo, and J. Cheng, “The effect of some soil characteristics on soil radon concentration and radon exhalation from soil surface,” *Journal of Nuclear Science and Technology*, vol. 41, no. 11, pp. 1113–1117, 2004. [Online]. Available: <https://doi.org/10.1080/18811248.2004.9726337>
- [14] USDA Natural Resources Conservation Service, “Create soil property or interpretation grid from gnatsgo instructions,” 2022, accessed: [Insert access date here]. [Online]. Available: <https://www.nrcs.usda.gov/sites/default/files/2022-08/Create%20soil%20property%20or%20interpretation%20grid%20from%20gNATSGO%20Instructions.pdf>
- [15] —, “Ssurgo metadata table column descriptions report,” 2022, available online at <https://www.nrcs.usda.gov/sites/default/files/2022-08/SSURGO-Metadata-Table-Column-Descriptions-Report.pdf>.
- [16] D. Jahn, “H3-pandas: Pandas integration with uber h3 hexagons,” <https://github.com/DahnJ/H3-Pandas>, 2023, available online at <https://github.com/DahnJ/H3-Pandas>.
- [17] D. Anning and S. Ator, “Generalized lithology for the conterminous united states,” 2017, u.S. Geological Survey data release. [Online]. Available: <https://doi.org/10.5066/F7R78D4N>
- [18] U.S. Geological Survey, “Generalized lithology and lithologic units of the united states,” 2007, open-File Report. [Online]. Available: <https://pubs.usgs.gov/gip/7000018/report.pdf>
- [19] D. Smith, F. Solano, L. Woodruff, W. Cannon, and K. Ellefsen, “Geochemical and mineralogical maps, with interpretation, for soils of the conterminous united states,” U.S. Geological Survey, Scientific Investigations Report 2017-5118, 2019, available as HTML. [Online]. Available: <https://doi.org/10.3133/sir20175118>

- [20] Natural Resources Conservation Service, “A soil profile,” n.d. [Online]. Available: <https://www.nrcs.usda.gov/resources/education-and-teaching-materials/a-soil-profile>
- [21] U.S. Geological Survey, “Geographic information systems (gis) data for the national atlas of the united states,” United States Geological Survey, 2011, open-File Report 2011-1073. [Online]. Available: <https://pubs.usgs.gov/publication/ofr20111073>
- [22] —, “Attributes for NHDPlus Catchments (Version 1.1) for the Conterminous United States: Hydrologic Landscape Regions [Data set],” U.S. Geological Survey, 2023. [Online]. Available: <https://doi.org/10.5066/P9142BM0>
- [23] A. B. Tanner, *Measurement and Determination of Radon Source Potential*. Gaithersburg, MD: US Department of Commerce, 1994.
- [24] D. A. Sprinkel and B. J. Solomon, “Radon hazards in utah,” Salt Lake City, UT, 1990, circular 81. [Online]. Available: <https://ugspub.nr.utah.gov/publications/circular/C-81.pdf>
- [25] T. Gleeson, N. Moosdorf, J. Hartmann, and L. van Beek, “A glimpse beneath earth’s surface: Global hydrogeology maps (glhymps) of permeability and porosity,” *Geophysical Research Letters*, vol. 41, p. 2014GL059856, 2014.
- [26] Daymet, “Daily surface weather data on a 1-km grid for north america, version 4 r1,” 2024, available online at <https://doi.org/10.3334/ORNLDAAAC/2129>.
- [27] J. Yang and K. H. Q. G. J. T. H. Busen, H. Scherb, “Modeling of radon exhalation from soil influenced by environmental parameters,” *Science of The Total Environment*, vol. 656, pp. 1304–1311, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0048969718348198>
- [28] A. R. J. Porstendorfer, G. Butterweck, “Daily variation of the radon concentration indoors and outdoors and the influence of meteorological parameters,” *Health Physics*, vol. 67(3), pp. 283–287, 1994. [Online]. Available: https://journals.lww.com/health-physics/abstract/1994/09000/daily_variation_of_the_radon_concentration_indoors.11.aspx