Spatial Estimation of Radon Exposure for Epidemiologic Risk Assessment

Kyle J. Sorensen, Richard L. Smith, Eric A. Whitsel, Jason M. Collins

Department of Statistics and Operations Research & Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina @ Chapel Hill

Introduction

Problem

- Radon levels are rising across North America, linked to trends in climate change
- Radon exposure is associated with lung cancer, strokes and other cardiovascular events
- Current estimates of radon exposure are limited, classified into three levels at low spatial resolution

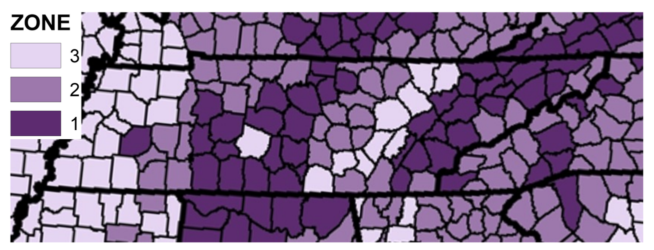


Figure 1: example of county-level resolution US EPA map of GRP radon zones in TN

Aim

- Create a spatial model for the geographic distribution of radon with some quantification of uncertainty
- Incorporate data accounting for geologic, atmospheric, and residential factors
- Provide improved, granular estimates of radon exposure

Solution

- Kriging, latent process modeling, alternative approaches
- Zip-code level model validation

Background

Radon

- Chemical element with symbol Rn, number 86
- Radioactive, colorless, odorless, and invisible gas
- Naturally occurring product of the decay of uranium

Exposure to radon

- Second leading cause of lung cancer
- Linked to strokes and other cardiovascular events
- Evidence of recent increases in North America

Radon and climate change - climate change may indirectly influence rises in radon exposure due to…

- Increased HVAC use
- Recycling of indoor air

Kriging

- Popular spatial modeling algorithm
- Model is a Gaussian process with…
  - Mean – function of covariates
  - Covariance – function of the spatial coordinates

Spatial dependence structure

- Nearby data is more similar than distant data
- Can cause artificially optimistic estimates of model performance

Spatial blocking cross-validation

- Folds from standard k-fold CV → geographically distinct regions
- Provides more realistic measure of model performance

Data

SRRS - EPA's State Residential Radon Survey

- Series of household-level short-term surveys
- 63,291 homes, 42 US states and six US territories
- Conducted between 1986 and 1992

GRP - USGS and EPA's Geologic Radon Potential

- Constructed from geologic, atmospheric and residential survey data
- Three levels:
  - "high" (estimated radon level > 4 picocuries per liter, or pCi/L) – zone 1
  - "moderate/variable" (2–4 pCi/L) – zone 2
  - "low" (< 2 pCi/L) – zone 3

Methods

Subset selection for example analysis

- 3 by 3 coordinate region in middle Tennessee
- 1247 homes across 256 zip-codes
- Relatively high spatial variability in GRP
  - Suggests that we may see sharper fluctuations in radon concentration across space

Kriging

- One of the most common methods for linear interpolation
  - Mean – linear function of covariates
  - Covariance – nonlinear function of spatial coordinates
- Two analyses: one for SRRS, one for integrated SRRS+GRP

Latent process modeling

- Highly flexible modeling approach that allows for robust integration of other data sources
  - Specified in a Bayesian hierarchical formulation
  - Allows us to condition the observed values on the so-called latent process values
- Two analyses: one for SRRS, one for integrated SRRS+GRP

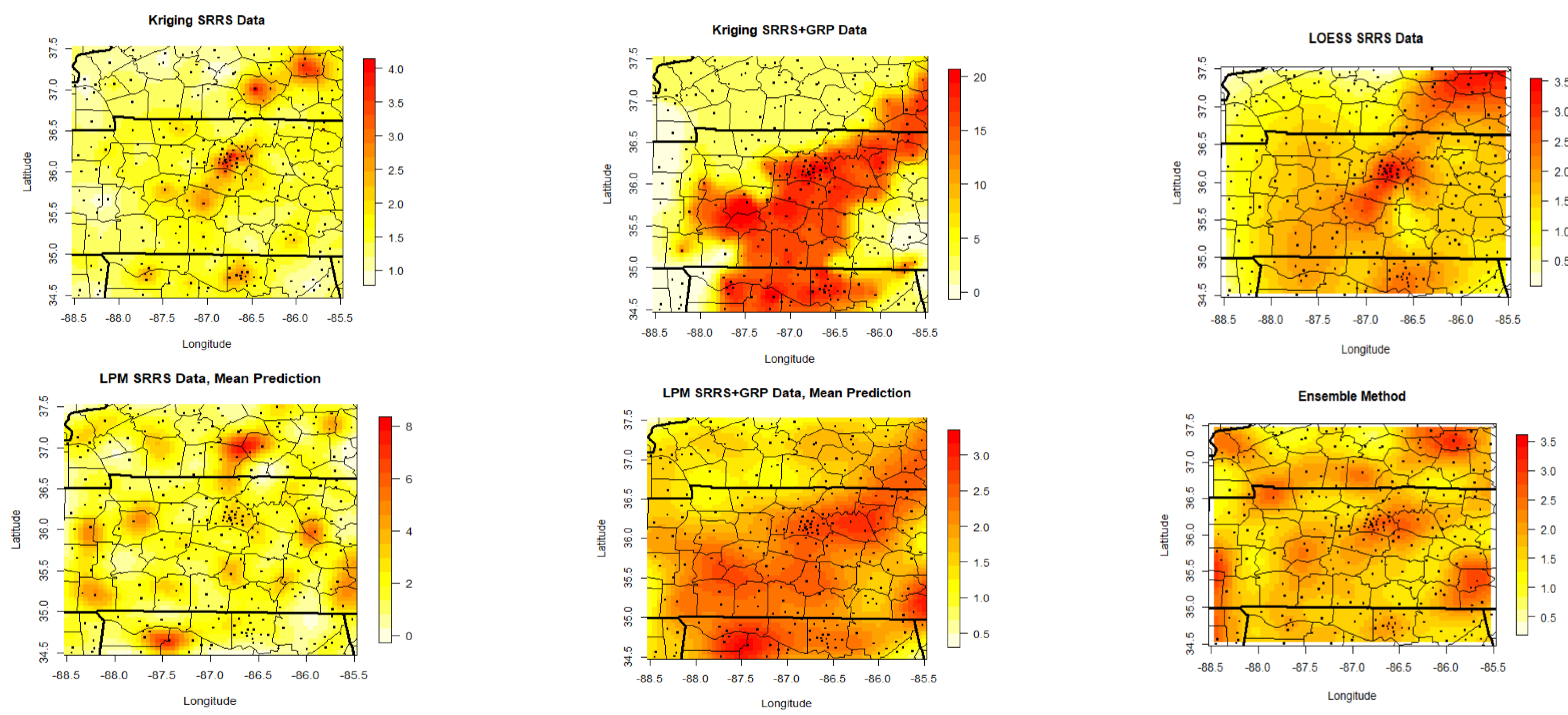
Alternative methods

- Locally estimated scatter plot smoothing (LOESS)
- Ensemble estimation
  - Simple average of predicted values across all models above

Model validation

- 80/20 training/test split on the level of zip-code, rather than individual observation
  - Ensures that our training and test sets are independent and eliminates concerns with spatial dependence
- Mean absolute error (MAE) used to compare accuracy and bias used to compare the direction of average error in predicted values

Results



Model	MAE (pCi/L)	Bias (pCi/L)
Kriging (SRRS)	2.112	-0.397
Kriging (SRRS+GRP)	9.941	+4.325
LPM (SRRS+GRP)	2.209	-0.416
LPM (SRRS+GRP)	2.071	-0.462
LOESS	2.329	-1.044
Ensemble estimation	2.059	-0.579

Future Work and Recommendations

Integration of additional data sets

- Base has been established for integrating additional data sets
- Could extend the latent process modeling approach to include other data sets accounting for the individual factors used to construct the GRP

Temporal component

- More recent radon measurement data exists, including the National Residential Radon Survey (NRRS)
- May allow for forecasting of changes in radon concentration across time

Bias in the sampling design, sampling weights

- We have not yet corrected for is the effect of the biased sampling design of SRRS
- Could leverage sampling weights to limit the effects of strong outliers

Other modeling approaches

- Nearest neighboring measure
- Inverse distance weighted mean

Further development of second-stage model

- Ensemble approach currently uses a simple mean as the second stage
- More complex modeling techniques have been applied in multi-stage models for spatial data with promising results

References

[1] United States Environmental Protection Agency. Air and Radiation (6604J). data retrieved from National Radon Database, volume 1-5 State/EPA Residential Radon Survey. 1993.
[2] United States Environmental Protection Agency. "National Residential Radon Survey Summary Report". In: Air and Radiation 6604J.EPA-402-R-92-0011 (1993), pp. 1–36.
[3] A. Gelman et al. "Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation". In: Statistical Science 14.3 (1999), pp. 305–337. doi: 10.1111/ecog.02881.
[4] D.R. Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". In: Ecography 40.1 (2017), pp. 913–929. doi: 10.1111/ecog.02881.
[5] F.K.T. Stanley et al. "Radon exposure is rising steadily within the modern North American residential environment, and is increasingly uniform across seasons". In: Scientific Reports 9.18472 (2019). doi: 10.1111/ecog.02881.
[6] K.M. Manson et al. "Radon is associated with clonal hematopoiesis of indeterminate potential in the Women's Health Initiative". In: Circulation 141.S1 (2020).
[7] L. Li et al. "Predicting monthly community-level domestic radon concentrations in the greater Boston area with an ensemble learning model". In: Environmental Science and Technology 55 (2021), pp. 7157–7166.
[8] P.N. Price et al. "Bayesian prediction of mean indoor radon concentrations for Minnesota counties". In: Health Phys 71.6 (1996), pp. 922–936.
[9] R.L. Smith et al. "Spatial models for radon data analysis". In: NIH Grant Proposal (2021), pp. 1–14.
[10] S.H. Kim et al. "The prevalence of stroke according to indoor radon concentration in South Koreans: Nationwide cross section study". In: Medicine (Baltimore) 99.4 (2020). doi: 10.1097/MD.0000000000018859.
[11] W.S. Cleveland. "Robust locally weighted regression and smoothing scatterplots". In: Journal of the American Statistical Association 74.368 (1979), pp. 829–836.
[12] N. Cressie. "Spatial prediction and ordinary kriging". In: Mathematical Geology 74.368 (1979), pp. 829–836.
[13] W. Field. "Climate change and indoor air quality". In: Contractor Report (2010), pp. 1–15.
[14] L. Gundersen and R. Schumann. "Mapping the radon potential of the United States: Examples from the Appalachians". In: Environment International 22.1 (1996), pp. 829–837.
[15] F.J. Massey. "The Kolmogorov-Smirnov test for goodness of fit". In: Journal of the American Statistical Association 46.253 (1951), pp. 68–78.
[16] R.L. Smith. Environmental Statistics. Lecture Notes, University of North Carolina at Chapel Hill. http://rls.sites.oasis.unc.edu/postscript/rs/envnotes.pdf, 2001.
[17] Stan Development Team. Stan User's Guide. English. Version 2.31. Stan Development Team. 2022. 468 pp. url: https://mc-stan.org/users/documentation/.
[18] H. Zhang and Y. Wang. "Kriging and cross-validation for massive spatial data". In: Environmetrics 21.1 (2010), pp. 290–304. doi: 10.1002/env.1023.