Centralized analysis of local data, with dollars and lives on the line: Lessons from the home radon experience*

Phillip N. Price† Andrew Gelman‡

25 Nov 2013

We have been hearing a lot recently about how much can be done with statistical analysis of Big Data. But what happens after the numbers are crunched? What does it take for the results to make a difference in people's lives? We discuss here in the context of a story from our own research regarding decision making for health risks arising from exposure to radon gas in the home. The project was a research success that did not make its way into policy, and we think it provides some useful insights into the interplay between locally-collected data, statistical analysis, and individual decision making.

Big Data Compiled from Many Local Sources

Before getting to our story we set the stage with a brief discussion of general issues regarding data availability. Some Big Data problems involve data from a single source or collected through a single mechanism: a particle physics experiment might generate terabytes of data per day from just a few detectors, and the U.S. census generates data on hundreds of millions of people using just a few different survey instruments. More typically, though, a Big Data problem involves combining data from multiple sources. Moreover, although the input data might come from many sources and involve thousands or millions of people, at least some of the results of the analysis are often geared towards individuals. Some examples include:

1. In evidence-based medicine (e.g., Lau, Ioannidis, and Schmid, 1997), information from many separate experiments and observational studies are combined in a meta-analysis, with the goal being to produce recommendations that can be adapted to individual patients by doctors or regulatory boards;

2. Specialized online tools such as traffic analysis tools are used to gather and disseminate up-to-the-minute data so that people can get personalized estimates of travel time;

3. Some websites gather and analyze information on housing sales, house characteristics, and neighborhood characteristics, so that they can provide estimated prices for individual houses on the market.

In addition to coming from multiple sources, Big Data have multiple uses, often by design. Most obviously, Google and Facebook appear to users as tools for answering queries or sharing information with friends, but at the same time they analyze the queries and social media posts to give advertisers a means of targeting potential customers. Scholastic testing is used both to evaluate individual students and to evaluate schools or school systems.

From the perspective of the data analyst, or of the user of the analysis, having more data is always better. One might choose to ignore data, even entire categories of data, if it's not clear how to use them in a statistical model or if the computational cost of analyzing them is too great, but on average there should be no *harm* in having more data. To the analyst, privacy and confidentiality protections are nuisances, rendering some data inaccessible and other data accessible only under

*We thank the National Science Foundation for partial support of this work.

†Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory

‡Department of Statistics, Columbia University, New York

inconvenient restrictions. For instance, access to might be granted only if the researcher agrees that no raw data may be published. This may be acceptable inasmuch as it still allows publishing of summary statistics and derived quantities, but it might prevent publishing even exemplary plots or tables of raw data, and might make it hard for others to evaluate the validity of the work. Imagine the problems of verifying global temperature changes if the raw data could not be shared.

Although the researcher or data analyst would always prefer access to all data that can be had, and the ability to publish all data and related analyses, owners or controllers of data often have good reasons not to share information, or, if it is shared, to insist that the data be available only to a restricted group of researchers. Someone who is selling her house may not wish it known that the basement sometimes floods, and a political candidate might be reluctant to answer the question, "Have you ever had an affair'?'

Data privacy issues can lead to a sort of prisoner's dilemma in which a group of people would benefit if they were *all* to share their data, but no single person's expected benefit is great enough to overcome their privacy concerns. Employees at a company might be able to bargain more effectively if everyone knew everyone else's salary, but each individual employee might see the negative impact of revealing their salary as being greater than the positive. An employee might reasonably think "I already know my salary, so adding my data to the pool does me no good at all, whereas it could cause me embarrassment or make co-workers unhappy with me, therefore I will not share," but if everyone follows this approach then the employees as a group are at an informational disadvantage compared to their employer. At times, the desire to prevent the free flow of information can have important consequences.

Assessing Risks and Recommending Decisions Regarding for Indoor Radon Exposure

Much of our thinking in this area has been influenced by an example we worked on in the mid-1990s on evaluating risks of exposure to radon in the home (Lin et al., 1999). Radon is a naturally occurring radioactive gas that is drawn into houses from the surrounding soil due to wind- and temperature-driven pressure differences between the soil gas and the interior of the house. Radon has long been known to cause lung cancer if inhaled at high concentrations, an effect first recognized among miners. (To be technically correct, it is not radon that is dangerous, it is the decay products of radon, which are themselves radioactive. When we say "radon" in this article, we really mean "radon decay products.")

Radon concentrations are often far higher in mines than in homes, and it was not until the mid-1980s that it was recognized that even some homes have indoor radon concentrations that expose occupants to dangerous levels of radiation. (A book from that era that is still useful scientifically but is now also an interesting historical document is Nazaroff and Nero, 1988). The most dramatic example is from 1984, when a Pennsylvanian who worked at a nuclear power plant kept triggering a radiation detector that was routinely used when workers left for the day. After some investigation, it was found that he was not being contaminated at work, but was carrying radon decay products to work with him on his clothes, and that his annual exposure from living in his house was far higher than the occupational safety limit for uranium miners. Within a few years of this highly publicized event—which led to the discovery of many other high-radon houses across the country—radon monitoring and mitigation companies sprung up across the country, and state and federal agencies had developed advice and guidelines. Radon monitoring is quite inexpensive, just \$15–\$30 depending on test type; mitigation, which usually involves using a fan to depressurize the soil beneath the house, typically costs \$800–\$2000 plus some energy cost to continuously run a fan.

Early on, the U.S. Environmental Protection Agency (EPA) established a recommendation that every house in the country should be tested for radon and that remediation actions should be per-

formed if a home's long-term living-area-average radioactivity concentration exceeded 4 picoCuries per liter of air (4 pCi/L), a threshold known as the "action level." In many places in the country, radon is one of many items specifically called out as potential risks in mandatory paperwork when a house changes hands, along with termites, mold, and so on.

It was clear from the outset was that some areas of the country have much higher average radon concentrations than others, and a much higher chance of having homes with extremely high radon concentrations. This was no surprise, given known geographical variation in soil types and home construction. Within a few years of radon becoming a national issue, the U.S. Geological Survey had begun working on mapping of "radon potential" (a somewhat ill-defined concept), and in the early 1990s they released maps for the coterminous U.S. (Schumann, 1988). However, these maps only identified areas of high, medium, and low "potential," and the current official policy was (and still is) that "Testing is the only way to know if you and your family are at risk from radon. EPA and the Surgeon General recommend testing all homes below the third floor for radon" (U.S. Environmental Protection Agency, 2012).

Contrary to federal policy, it would make sense to focus radon measurement and remediation efforts on homes with radon concentrations much higher than the recommended action level, first because the people in those houses are at greater individual risk and therefore stand to benefit most if their risks are identified and dealt with rapidly rather than slowly, but also because there is no question that residents of extremely high radon houses are at *some* risk. The dose-response relationship for radon decay products is not well known at typical residential concentrations, and many people question whether the EPA's recommended action level is too low: some people suggesting there may be no additional risk of lung cancer at 4 pCi/L compared to, say, 0.5 pCi/L, which is comparable to the outdoor radon exposure in some parts of the country—but there is no question that long-term exposure to a concentration of 20 pCi/L or more causes a substantially increased risk of lung cancer. We believe (but many people disagree) that there is fairly convincing evidence that long-term residence in a house with a radon concentration of 4 pCi/L does in fact increase the risk of lung cancer (see Field et al., 2000, for example).

To determine the nationwide statistical distribution of indoor radon concentrations and to begin to map the geography of the problem, in the late 1980s and early 1990s the EPA measured long-term living-area-average radon concentrations in a stratified random sample of about 5000 houses in the country, and worked with state agencies to measure short-term (2- to 3-day) radon concentrations in winter on the lowest level of a sample of tens out thousands of houses from most of the 50 states. A short-term measurement in winter on the lowest floor of the house is called a "screening" measurement and tends to over-estimate the long-term living-area-average concentration by a factor of 2 to 3.5, depending on details of climate and house construction, as well as being subject to considerable stochastic variability due to variation in the weather and physical constraints of the measuring device.

Although our analysis had many characteristics of a Big Data problem, it differed from such problems in at least one important way: we did not in fact have a lot of data, with random-sample radon measurements from only about 60,000 homes throughout the country. Many individual counties had fewer than 5 samples. At least in some areas of the country there were much larger datasets that were potentially available from radon testing and mitigation contractors, but these datasets had a variety of problems and, after examining some of the data, we ultimately decided not to pursue obtaining and using them. The biggest problems were inconsistent and often unrecorded measurement protocols, and including multiple confirmatory measurements from houses with high initial test measurements, but with no way of identifying when this occurred. We found only low correlations between the random sample radon measurements and those from private databases in the same zip code or county. This example illustrates the important point that simply finding a

way to get a larger dataset does not guarantee a more accurate or more useful analysis.

Local Data, Centralized Analysis, Local Decision Making

When we began to work on indoor radon mapping we immediately ran into problems with combining data from different sources. For example, the nationwide radon survey performed long-term living-area measurements on each floor of the home, whereas the surveys conducted by the individual states almost all made one or two short-term measurements on the lowest level of each home, which was often an unoccupied basement. Also, radon measurements are made in individual homes, but the only available data on soil uranium content (a useful predictive variable) were available only as spatial averages. Finding ways to jointly analyze all of the available data was a significant challenge (Price and Nero, 1996).

We made several time-consuming false starts in our analyses when trying to figure out ways to exploit various types of data. For example, due to confidentiality constraints some of our data provided house locations only as zip codes, which are large enough in rural areas that we were unable to determine the local geology except in very crude terms. We spent considerable time and resources investigating whether more detailed location information would lead to better predictive models—this required both performing a new radon survey whose participants allowed us to make use of their exact house location, and digitizing of old paper maps of local geology—only to find that this approach provided little benefit over the other variables we were already including in our models. Deciding what additional data are worth collecting is a part of many statistical modeling efforts.

In the end, having done the best we could to create a model that used the radon measurement data, along with other information on the geographic distribution of radiation risks, we constructed a hierarchical model that yielded a (probabilistic) prediction of the radon level in any house, given its geographic location, information about the surface-soil radioactivity in its area, and house-specific information such as whether the house has a basement. This portion of our analysis yielded predictions of, and uncertainties in, the statistical distribution of indoor radon concentrations in every county in the conterminous United States. Counties that were heavily sampled in the radon surveys had small uncertainties, whereas sparsely sampled or unsampled counties had statistical distributions based purely on the explanatory variables and therefore had large uncertainties. The goal of this government-funded effort was to identify areas of the country with the highest radon concentrations, either in terms of average levels or the fraction of homes whose radon concentration greatly exceeded the recommended action level, so that special attention could be focused on those areas by way of increased public outreach and perhaps government programs to perform radon testing.

The good news was that, after some effort, we were able to construct a model that fit the data well, gave insight into the geographic distribution of radon risks, gave informative predictions, and performed well under cross-validation. The bad news was that, even though our mid-90s efforts led to much more accurate maps of statistical distributions of radon concentrations than had previously been available, and in spite of the fact that much of our research had been federally funded, we found that our work had no effect on federal radon policy. Simply having radon maps that would have allowed specific counties around the country to be targeted for increased radon monitoring and mitigation did not lead to any such targeting, at least by federal agencies (some states do have targeted programs).

It was obvious to us that a targeted approach could be much cheaper than monitoring every home, and could presumably save more lives as well: most people were ignoring the monitoring recommendations whether they lived in a high-radon area or not, whereas targeting high-radon areas

would presumably improve compliance in those areas. Providing the ability to focus on high-radon areas did not promote change, but we thought, perhaps naively, that the government's radon policy would change if we illustrated the fact that a targeted policy could cost less money—and, depending on the policy and its reception by the public, could save more lives—than recommending testing in every house. Having fit the model, we were able to estimate the the potential cost in dollars and the potential savings in lives of various potential programs for measurement and remediation of radon (Lin et al., 1999). This allowed us to compare the current policy of monitoring every house with a short-term test to various more targeted programs, such as focusing on certain counties or only certain types of houses in certain counties. Similar work was done by others at the same time (Ford et al., 1999).

Although certainly not a Big Data problem by today's standards, or even the standards of the time, our analysis had some of the characteristics typical of such problems: we were putting together locally-collected data in an aggregate context; our data came from disparate sources; and, as with most real-world problems but not textbook problems, it was up to us to decide whether it was worth seeking out and trying to use other data sources (for example, maps of local geology). Our analysis also shared an unnoticed feature of many real-world data-based problems, which is that they are *dispersed*. In a classical decision problem there is a single decision tree, a single utility function, a single probability distribution, and a single decision to make. Here we are considering problems in which many different people each have to make decisions based in part on analysis of a large dataset and in part on information that is specific to them. You can use a website to get summary statistics (or even raw data) about house prices, or the reliability of different ages and models of used cars, but you then apply that information to deciding whether to buy a specific house or a specific car. This is related to the idea of Hand (2009) that statistics is the science of the individual as well as the aggregate.

Our radon analysis was centralized, and the government offers general recommendations, but the ultimate decisions, as well as the financial and health risks, were borne locally by individuals. Each homeowner needs to make his or her decision of what to do about home radon, as the costs of measurement and remediation, as well as the risk of cancer, vary locally, and each person has her own risk tolerance and willingness to pay to reduce risk. We recommended that the government's simple decision rule, which was the same everywhere in the U.S. and which did not vary according to house-specific information, should be replaced by a more complicated rule that takes account of local knowledge of the statistical distribution of indoor radon concentrations, as well as information specific to the house (such as whether it has a basement). And if a homeowner measured their indoor radon concentration, our rule considered whether the measurement was short- or long-term, and, if short-term, in what season the measurement was made. Additionally, our rule allowed for variation of household size and makeup: smokers are at higher risk from radon than non-smokers; the risk is different for men, women, and children; and all else equal a larger household has more chance of experiencing a radon-induced lung cancer than a smaller household.

A homeowner's risk tolerance and household makeup, combined with the estimated dose-response relationship for men, women, and children exposed to radon decay products, determines a "perfect-information action level": if the homeowner knew for sure that their home's long-term living-area radon concentration exceeded this level, they should remediate their home for reduce the radon concentration. Given the perfect-information action level, the decision of whether or not to measure the radon concentration depends on the statistical distribution of radon concentrations for the home, and some other parameters such as the cost of testing and the uncertainty in the resulting measurement. For one set of assumptions about these matters, the results are shown in Figure 1.

What made this analysis work at a statistical level was hierarchical modeling. We had data

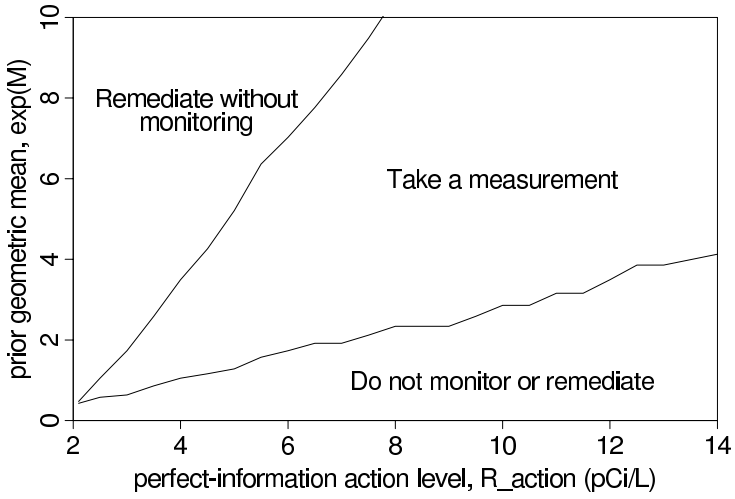

Figure 1: *Recommended radon remediation/measurement decision as a function of two inputs at the individual level: (a) the perfect-information action level R_{action}, which represents an individual risk assessment, and (b) the prior geometric mean radon level e^M, which captures information about individual exposure. A homeowner can read off his or her recommended decision from this graph. (Wiggles in the lines are due to simulation variability in the calculations.) As can be seen from this figure, the output from our statistical and decision analysis is not a single decision but rather is a decision function allowing different individual decisions under different conditions. From Gelman et al. (2013).*

at the level of measurements, houses, counties, and states. For example, two key predictors were surface geologic type and surface uranium concentration, and both of these were given at somewhat course geographic divisions which we then matched to counties, as these political divisions were the best geographic information we had on house locations. A hierarchical model (as opposed to a simple linear regression) allowed us to get reasonable predictions of distributions of radon level and of uncertainties in these distributions in every county, even those for which very few measurements were available (see Price, Nero, and Gelman, 1996). This allowed us to make decision recommendations that differ from county to county.

Research on residential radon in support of government policy required communication among several groups: the people whose homes were measured, the EPA, us, and many other players, including commercial suppliers of radon test kids and home remediation, state-level regulatory officials, and public health officials. Some aspects of this communication went better than others. The EPA got excellent compliance in their radon survey, and they made the data available so that many different research groups could work on the radon prediction and remediation problem. But our own interactions with regulatory agencies and end-users were not so effective. The people we spoke with at the EPA resisted our efforts to create a calibrated decision analysis with different recommendations for different counties and house types; they wanted to stick with a uniform recommendation (measure all houses, then remediate all houses where the measured radon level exceeds some threshold) which we estimated would be a much less cost-effective way to reduce risk.

What Went Wrong: Why is There Still a Single Nationwide Recommendation?

We thought at first that the resistance to a targeted radon monitoring approach was the belief that it would be too hard to implement. To try to demonstrate that such an approach was feasible, we created a high-tech (as of 2000!) website where anyone could click and find a map showing his or her county, along with estimated costs and benefits of radon measurement and remediation for a typical home in the county. Optionally, the user could fill in additional information such as measurements on neighbors' houses that could inform the decision of whether to make a radon measurement, what type of measurement to make, and what they should use as their personal "action level" for remediation. We provided defaults for household size and makeup, and for risk tolerance, but a user could change these as well if he wished. The user could even change the assumed relationship between radon exposure and health risk.

We did not expect to reach a large number of individual homeowners through the website: as we stated in a companion article to our decision analysis paper (Lin et al., 1999b), by performing our analysis (and making the website), "we are hoping to influence government policy; we do not expect our recommendations to reach a substantial number of individual homeowners."

We had some small successes publicizing our maps, website, and research project (including, at one point, an article with maps in the *New York Times*; Fairfield, 2005) but it never became wildly popular: in about ten years fewer than forty thousand visitors used our site to get a recommendation. Over the years, different pieces of the webpage became nonfunctional—victims of software updates and the like—and we eventually abandoned it almost ten years after it launched. The modest popularity of the site was neither disappointing nor surprising to us, since the public was not really the intended audience of the site: what we were trying to demonstrate was that the federal government could, if it chose, create and promote a targeted radon policy. The disappointment was that we never saw evidence that we had helped shift national policies towards targeted radon testing.

We believe one of the reasons targeted testing didn't catch on is that there is pressure from homeowners and realtors not to identify specific areas as having elevated risk from radon. If you're

a homeowner, you want to know if you're in a high-radon area so that you know to measure your home's radon concentration, but at the same time you may not want *other* people to know that you are in a high-radon area because this might decrease the value of your home. Concern about property values was evident even, or perhaps especially, in the first few years after high residential radon concentrations were discovered. As early as 1985 the state of Florida, which has very high radon concentrations on or near areas where phosphate mining has occurred, was considering requiring a formal warning to homeowners in those areas. A lobbyist said "If they pass this notification thing, it's war with a capital W." A few years later a 1989 Chicago Tribune article reported: "Lawyers and environmental specialists warned corporate relocation specialists gathering in Chicago last week that liability for environmental hazards in homes is likely to emerge soon as a major problem for sellers and real estate brokers," with radon specifically mentioned (Allen, J.L., Chicago Tribune, 1989). Viewed in terms of a single home, the desire that a high radon test result should not be made public is just a routine data privacy and confidentiality issue. But because high-radon homes tend to occur in spatial clusters, many homeowners (or at least home sellers) in high-radon areas would strongly prefer that even the *statistical distribution* of radon should not be accurately mapped.

Interestingly, there is also a constituency—radon testing and mitigation companies—that does not want specific areas to be identified as having *low* risk from radon, since this will decrease the number of tests and mitigation installations. In fact, because of seasonal and short-term variability in indoor radon concentrations, and peoples' insistence on taking short-term rather than long-term tests, it is likely that in some relatively low-radon areas of the country a majority of radon mitigations are unnecessary, in the sense that the house did not have a long-term living-area radon concentration in excess of the EPA's recommended action level (Lin et al., 1999).

The most enduring effect of our research is probably not in environmental policy but as an example of dispersed decision analysis that has appeared in two textbooks written by one of us (Gelman). Having a good example is not nothing—indeed, it was one of our original motivations in pushing through advanced statistical methods for this problem, to explore the application of hierarchical modeling for decision analysis.

We hope that the technical success of our work will motivate future uses of hierarchical modeling to enable effective localized decision recommendations. And we hope people can learn from our social failures so as to better integrate future projects involving statistical analysis, policy makers, and local actors in real-world decision problems.

Denouement: Information Wants To Be Free

We spent much of the past fifteen years feeling disappointed that our work, of which we are proud, had no influence on radon policy. In addition to our personal disappointment, we have been unhappy about the inefficiency of having a single nationwide policy, and even a bit angry about the loss of life and the waste of money compared to what a targeted radon policy could achieve.

But a funny thing has happened over the years: although the federal government still has a blanket recommendation, state policies and economic pressures have gradually led to a de facto targeting of high-radon areas. Some state governments, such as New York, have carried out radon mapping programs and information campaigns. Regions of the country with generally high radon concentrations tend to have more companies that perform testing and mitigation, and those companies advertise. Those regions have more news items about radon; a typical example is Minnick (2009), a short news article that reports, "Newlyweds Mark and Karen Hite learned about the high level of radon [in their area] months after they bought a north Raleigh home. A radon test revealed

levels twice as high as the recommended EPA limit." Word of mouth in high-radon areas also leads to higher awareness.

Overall, the situation concerning radon measurement and mitigation decisions is far less efficient than it could be, but most radon measurements do seem to be made in high-radon areas, and although there are plenty of unnecessary mitigations, and plenty of high-radon houses that escape remediation, many high-radon houses are found and fixed. The situation could be a lot better, but at least it's better than it used to be.

References

Allen, J. L. (1989). Radon new threat to home sales. *Chicago Tribune*, 8 Apr.

Dunkelberger, L. (1985). Realtors, builders fight radon warning bid. *Ocala Star-Banner*, 9 Oct.

Fairfield, H. (2005). In a new map, radon looks less risky for many. *New York Times*, 11 Jan.

Field, R. W., Steck, D. J., Smith, B. J., Brus, C. P., Neuberger, J. S., Fisher, E. F., Platz, C. E., Robinson, R. A., Woolson, R. F., Lynch, C. F. (2000). Residential radon gas exposure and lung cancer: The Iowa radon lung cancer study. *American Journal of Epidemiology* **151**, 1091–1102.

Ford, E. S., Kelly, A. E., Teutsch, S. M., Thacker, S. B., and Garbe, P. L. (1999). Radon and lung cancer: A cost-effectiveness analysis. *American Journal of Public Health* **89**, 351–357.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*, third edition. London: Chapman and Hall.

Hand, D. J. (2009). Modern statistics: the myth and the magic. *Journal of the Royal Statistical Society A* **172**, 287–306.

Lau, J., Ioannidis, J. P. A., and Schmid, C. H. (1997). Quantitative synthesis in systematic reviews. *Annals of internal medicine* **127**, 820–826.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. (1999). Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation (with discussion). *Statistical Science* **14**, 305–337.

Minnick, B. (2009). High radon levels can be a problem in Triangle homes. WRAL.com, Raleigh-Durham-Fayetteville, 22 Jun.

Nazaroff, W. W., and Nero, A. V., eds. (1988). *Radon and its Decay Products in Indoor Air*. New York: Wiley.

Price, P. N., Nero, A. V., and Gelman, A. (1996). Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Health Physics* **71**, 922–936.

Schumann, R. R., and Owen, D. E. (1988). Relationships between geology, equivalent uranium concentration, and radon in soil gas. Fairfax County, Virginia: U.S. Geological Survey Open-File Report 88-18, 27 pp.

U.S. Environmental Protection Agency (2012). A citizen's guide to radon. EPA report 402/C-12/001.