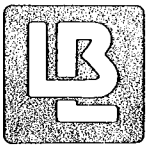
LBL-35818
UC-402
Preprint



Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

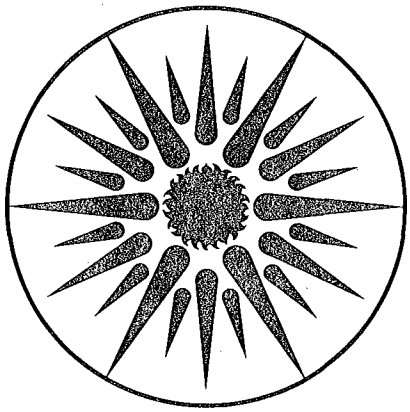
ENERGY & ENVIRONMENT DIVISION

Submitted to Health Physics

Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties

P.N. Price, A.V. Nero, and A. Gelman

August 1995



ENERGY
AND ENVIRONMENT
DIVISION

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

Submitted to *Health Physics*

LBL-35818
UC-402

Bayesian Predication of Mean Indoor Radon Concentrations for Minnesota Counties

P. N. Price
A.V. Nero
A. Gelman[1]

Energy & Environment Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California

[1]Department of Statistics
University of California Berkeley
Berkeley, CA -

August 1995

This work was supported by the Office of Energy Research, the Office of Health and Environmental Research and the Environmental Sciences Division of the U.S. Department of Energy (DOE) under Contract DE-AC03-76SF00098.

Abstract

Past efforts to identify areas having higher than average indoor radon concentrations by examining the statistical relationship between local mean concentrations and physical parameters such as the soil radium concentration have been hampered by the noise in local means caused by the small number of homes monitored in some or most areas. In the present paper, indoor radon data from a survey in Minnesota are analyzed in such a way as to minimize the effect of finite sample size within counties, in order to determine the true county-to-county variation of indoor radon concentrations in the state and the extent to which this variation is explained by the variation in surficial radium concentration among counties. The analysis uses hierarchical modeling, in which some parameters of interest (such as county geometric mean (GM) radon concentrations) are assumed to be drawn from a single population, for which the distributional parameters are estimated from the data. Extensions of this technique, known as a random effects regression and mixed effects regression, are used to determine the relationship between predictive variables and indoor radon concentrations; the results are used to refine the predictions of each county's radon levels, resulting in a great decrease in uncertainty. The true county-to-county variation of GM radon levels is found to be substantially less than the county-to-county variation of the observed GMs, much of which is due to the small sample size in each county. The variation in the logarithm of surficial radium content is shown to explain approximately 80% of the variation of the logarithm of GM radon concentration among counties. The influences of housing and measurement factors, such as whether the monitored home has a basement and whether the measurement was made in a basement, are also discussed. This approach offers a self-consistent statistical method for predicting the mean values of indoor radon concentrations or other geographically distributed environmental parameters.

Introduction

The ability to characterize the distribution of indoor radon concentrations within relatively small areas (such as counties) would greatly increase the efficiency of programs to identify individual homes having radon concentrations much greater than the norm. Examination of early data from the United States demonstrated substantial variability of concentration distributions from one area to another [Nero et al. 1986]. Later analysis suggested that most of the U.S. data are consistent with a picture wherein the overall U.S. distribution of indoor radon concentrations is a mixture of subsidiary local distributions that are approximately lognormal [Nero et al. 1990]. Within the U.S., the variation in the geometric means (GMs) among county-sized areas is generally much greater than the variation of geometric standard deviations (GSDs), so that most high-radon homes are located in areas with relatively high GMs. Identifying such areas is thus a useful step towards focusing efforts to locate individual homes with indoor radon levels much higher than average.

In 1987-88, the Minnesota Department of Health conducted a radon survey as part of U.S. Environmental Protection Agency's State/EPA Residential Radon Survey (SRRS) program [Tate et al. 1988, White et al. 1992, Alexander et al. 1994]. The results indicate that indoor radon levels in Minnesota tend to be higher than is typical in the U.S., and that there is significant variation of radon concentrations among the counties in the state. Earlier analysis [Nero et al. 1994] using ordinary regression techniques indicated that much of the variation in county GM indoor radon concentration could be predicted from soil radium data extracted from the National Uranium Resource Evaluation (NURE). That analysis, as well as accurate prediction of individual county GMs, was hampered by the statistical noise due to small sample sizes for most counties.

The present paper develops a more complete statistical approach, again using the SRRS survey data from Minnesota as a demonstration. The analysis is performed in several parts, with the goal of introducing the use of random effects regressions as a way of determining the underlying true distribution of county radon concentrations by removing (or at least minimizing) effects of finite sample size.

The Minnesota data include measurements made in a stratified random sample of 919 owner-occupied ground-contact homes in Minnesota, performed with a "screening" protocol: a two– to four–day, winter charcoal–canister measurement was taken, normally in the lowest level of each home. In addition to the measured radon concentration, data on each home include: the county in which the home sits, whether or not the home has a basement, whether the home was "single-family" (as opposed to a duplex, condominium, etc.), what room the measurement was made in (family room, dining room, etc), and on what floor of the home the measurement was made. The survey was conducted primarily with the goal of determining the overall screening radon distribution in the state. These data can also be used to investigate methodologies for predicting the radon distributions for smaller areas, such as counties, in order to locate particularly high-radon areas. We use the screening data because they are available and can be expected to exhibit roughly the same spatial distribution as would data from long-term monitoring in living areas. Locating areas that have generally high long-term living-area radon concentrations would require long-term measurement data, either to supplant or to normalize the screening data.

A population-based stratification scheme was used to choose the number of participants per county. Adjustments were made to increase the sampling rate of expected high-radon counties and of low-population counties [Wirth 1992], but the distribution of measurements

by county is extremely uneven: some counties had over 100 measurements, while other counties had few or none at all. Thus any attempt to use the data to determine parameters describing county radon concentrations—such as the geometric mean radon concentration for each county—must contend with the effects of finite sample size. The "noise" due to finite sample size also confounds analysis to find the relationship between county radon concentrations and physical factors such as geologic or soil information.

In this paper, we use the survey data to answer several questions:

1. What is the best estimate and uncertainty of each county's true geometric mean of radon screening measurements? By "true GM", we mean the GM that would have been obtained if every eligible home in the county had been measured with the survey's protocol, and if measurement error due to background subtraction (discussed below) were eliminated.

2. How much of the county-to-county variation can be explained by the variation in surficial radium concentration, as indicated by the uranium data from the National Uranium Resouce Evaluation?

3. Some of the observed variation between county radon concentrations is probably due to differences in known house construction parameters and differences in measurement procedures (such as whether the home has a basement, and whether the measurement was made in a basement). How can we discover the county-to-county variation that remains when these effects are removed?

We use regression techniques known as "random effects regression" and "mixed effects regression" to answer these questions. Although such techniques have been used in other fields for at least 15 years, we are not aware of their previous use in characterizing radon distributions or other environmental parameters. The procedures applied here are particularly useful when attempting to estimate parameters (such as county geometric means) based on sparse data.

A complete discussion of the mathematical details of hierarchical models and random effects regression is beyond the scope of this paper. Discussions of these techniques can be found in several of the references [Lindley et al. 1972, Rubin 1980, Bryk et al. 1992, Gelman et al. 1995]. Since these methods have not yet become as ubiquitous as more familiar tools such as conventional regressions, we discuss them briefly here in the context of the current problem, rather than simply presenting the results.

The Minnesota Screening Data.

Figure 1 shows a histogram of the radon concentrations reported from the state radon survey [Wirth 1992], weighted according to the sampling weights reported in the data set. A lognormal curve with GM = 132 Bq/m^3(3.6 pCi/L) and GSD= 2.18 has been superimposed on the data. The observed radon distribution in the state as a whole is nearly consistent with a lognormal distribution. The important exception for our purposes is the presence of a few too many extremely low radon concentrations (affecting the lowest bin in the linear plot of Fig. 1), which for example skews the calculation of geometric mean concentrations. Indeed, some of the reported radon concentrations are zero; in other state surveys, using

similar protocols, negative radon concentrations have been reported. The distribution above about 40 Bq/m^{3}appears almost perfectly lognormal.

The excess of low radon concentration measurements is consistent with being due to statistical errors in background subtraction: in determining radon levels, an expected number of background counts is subtracted from the observed number of total radioactive decays. The number of radioactive decays exceeding the assumed background count is ascribed to radon. Since the actual number of background counts varies statistically around the expected number, the number of counts attributed to radon (and thus the calculated radon concentration) can differ from the actual number by a small amount (typically equivalent to a few Bq/m^3). This effect is miniscule for moderate or high radon concentrations, but can have a large relative effect when the actual radon concentration is small; indeed, as noted, it can lead to negative reported concentrations.

When the reported value is extremely small, it is almost certain that the true value is higher than the reported value; however, the exact magnitude of this effect is unknown. We cannot simply discard the problematic points, since the low reported values really do indicate the presence of low-radon homes, although the exact values of the measurements are incorrect. If we were interested only in estimating distribution parameters for aggregated data, such as the geometric mean and geometric standard deviation, we could use a censored maximum likelihood estimate [Harter et al. 1966] with a censoring threshold set high enough to exclude the problematic points—that is, at 5 to 10 Bq/m^3. However, in the present paper we wish to perform analyses at the level of individual homes rather than county aggregates, so distribution estimates are not sufficient.

Since incorrect extremely low values can cause problems, especially when logarithmically transformed, we have adjusted all of the low values upwards slightly, with the extremely low values brought up the most and the values above 50 Bq/m^{3}essentially unaffected.

The empirical adjustment we used to convert the reported radon concentration $C_{\mathrm{Rn}}^{\mathrm{meas}}$ to a new value $C_{\mathrm{Rn}}^{\mathrm{new}}$ was

$$C_{\mathrm{Rn}}^{\mathrm{new}} = \frac{C_{\mathrm{Rn}}^{\mathrm{meas}}}{2} + \sqrt{\frac{(C_{\mathrm{Rn}}^{\mathrm{meas}})^2}{4} + D^2} \qquad (1)$$

with $D = 9.25$ Bq/m^3(0.25 pCi/L), which was found to make the entire distribution appear nearly lognormal even for low radon levels. We do not claim that this equation has any underlying physical validity—it is merely a convenient one-parameter correction that adjusts very low values upwards very slightly in absolute terms, while leaving higher values virtually unchanged: a measured value of 0.00 Bq/m^{3}is converted to 9.25 Bq/m^3, while a measured value of 20 Bq/m^{3}is converted to 23.6 Bq/m^3. Only a few measurements are affected substantially: of the 919 reported values, only 13 are below 20 Bq/m^3. In this paper all of our discussion of observed radon concentrations refers to the adjusted values $C_{\mathrm{Rn}}^{\mathrm{new}}$, hereafter referred to simply as C_{Rn}, rather than the measured values. The results presented here are quite insensitive to the exact value of D, as long as it is above about 5 Bq/m^3. We note that minimum values of 5 to 10 Bq/m^{3}are roughly consistent with observations of mean outdoor concentrations [Gesell 1983], and so are reasonable lower bounds for actual indoor radon concentrations. In addition, county GMs calculated with the adjusted values of C_{Rn} are in good agreement with the censored maximum likelihood estimates for the counties.

In addition to the statewide distribution of radon measurements being nearly lognormal

(except for the very low radon measurements, as discussed above), the observed distributions within the individual well-sampled counties are also approximately lognormal. Also, it has previously been noted that radon distributions in county-sized areas tend to be approximately lognormally distributed [Nero et al. 1986, Dudney et al. 1992]. For the present paper, therefore, we have chosen to model the within-county distribution of radon measurements as lognormal. Lognormal distributions do not always fit the extreme high tail of radon measurements perfectly [Cohen 1985; White et al. 1992, Janssen et al. 1992], so using the parameter estimates from the present paper to estimate the fraction of homes in a given county that would have extremely high screening measurements could conceivably lead to substantially incorrect inference. However, if such problems occur at all, they do so only for the very high tail of radon measurements, so that lognormal parameters do adequately summarize the distribution of the vast majority of homes [White et al. 1992, Janssen et al. 1992].

In order to characterize a lognormal distribution, both the GM of the distribution and the geometric standard deviation[1] (GSD) must be known. In Minnesota, all of the counties with more than 20 observations have observed GSDs between 1.8 and 2.35. The data as a whole are nearly consistent with the hypothesis that all of the counties have the same true GSD (equal to about 2.1). In this paper, we make the simplifying assumption that the true GSD for homes with a given set of explanatory variables is the same for all counties. In the following section, this is equivalent to the assumption that all counties have the same true GSD. In later sections, the assumption is modified to include different housing and measurement characteristics, but we still assume that the GSD conditional on those characteristics is the same for all counties. The predictions for the county GMs, the parameter that interests us most, are not strongly dependent on the true values of the GSDs. Since it does appear that the true county GMs are all close to one another, slight variations of the GSD among counties will have only minor influence on the predicted GMs.

Because the radon measurements within each county appear to be approximately lognormally distributed, it is often convenient to perform calculations with the natural logarithm of the radon measurement, C_{Rn} , rather than with the observed value itself. Although we often transform to log space to perform the calculations, we will usually report results in untransformed rather than log space.

Calculation of Posterior Estimates for the County GMs.

For simplicity, in this section we ignore the explanatory variables related to soil radium concentration and to housing type, and discuss only the county geometric mean radon concentrations; use of the explanatory variables will be discussed in the next section.

We wish to use the observed county GMs to try to predict the true county GMs (of "screening" radon concentrations). One simple and traditional approach is to use the observed GM as a direct prediction of the true GM (that is, to take $\mathrm{GM}^{\mathrm{pred}} = \mathrm{GM}^{\mathrm{obs}}$), but this approach has at least one serious drawback: it leads to a distribution of predicted county GMs that is far broader than the true distribution is likely to be. The distribution of observed GMs is almost certain to be much wider than the distribution of true GMs, because of the effect of finite sample size—given the small number of observations in most

[1]The logarithm of the geometric standard deviation of a set of measurements is equal to the standard deviation of the logarithms of the measurements.

counties, some high-radon counties will happen to yield observed GMs even higher than their true GMs, and some low-radon counties will happen to yield GMs even lower than their true GMs. Imagine, for example, the effect of finite sample size on a group of counties with exactly the same true GM: the measured GMs will be spread about the true GM, with the degree of spread depending on the number of observations in each county.

All of our questions about the true county GMs and the overall distribution of county GMs would be easily answered if a large amount of data were available for each county. In the present case, however, only a few of the counties are so heavily sampled that we already have a precise estimate of the true GM of radon concentrations in the county: about twenty observations are needed to determine a county GM with a standard error of twenty percent, while the median number of observations per county in Minnesota is only five. Much of the variation in observed county radon levels is certainly due to the effects of the small sample size in most counties. For example, consider Lac Qui Parle County: this county has only two observations, and the GM of the observations is about 500 Bq/m^3. This GM is considerably higher than the GMs of well–sampled counties (e.g. those with more than fifteen observations), all of which lie between 75 and 150 Bq/m^3. It seems likely that the true GM of Lac Qui Parle County is considerably lower than 500 Bq/m^3, and that the monitored homes from that county simply happened to have unusually high radon levels (at least over the days they were tested). How, then, can we obtain statistically well-founded predictions of the actual county GMs that adjust for the variation due to finite sample size?

A reasonable answer to this question is provided by a hierarchical model: we assume the true county GMs are drawn from some distribution of "possible" county GMs, and that the parameters of this distribution can be estimated from the data. For instance, suppose we knew the true GM for 86 counties, randomly selected from the 87 counties in Minnesota. Furthermore, suppose these 86 values were found to be approximately lognormally distributed with a geometric mean of 145 $Bq/m^3$and a geometric standard deviation of 1.4. Then, even if we had no observations from the missing county, it would be reasonable to guess that its true GM is likely to fall between 75 $Bq/m^3$and 285 Bq/m^3(two GSDs below and above the GM of the county GMs, respectively) with about 95% certainty.

In the hypothetical situation described here, we have substantial knowledge of the range in which the missing county's true GM is likely to fall even though we have *no measurements at all* from that county. This conclusion relies on the plausible assumption that the GM of the missing county is drawn from the same distribution as the GMs of the known counties; we would certainly be surprised if the GM of the missing county were later found to be, say, 800 Bq/m^3or 1 Bq/m^3. In conventional statistical notation, with θ representing the true value of the logarithm of the county's geometric mean radon concentration, such knowledge of the distribution from which the missing county's log(GM) is drawn would be written:

$$p(\theta) = N(\mu, \sigma^2), \tag{2}$$

indicating that the probability of obtaining a particular value of θ is normally distributed about μ [equal to log(145 Bq/m^3) in the current example] with standard deviation σ [equal to log(1.4) in the current example]. In such a case, in which the distribution from which the missing county's GM is drawn is known, $p(\theta)$ is known as an *informative prior distribution.*[2]

[2]The case of $\sigma^2 \to \infty$, corresponding to a distribution of county GMs that has infinite variance, would be a *noninformative* prior distribution, indicating total ignorance of the likely range containing the missing county's true log(GM).

We wish to avoid the misconception that the assumption of a distribution from which parameters are drawn is equivalent to the assumption that the variation between counties is "random" rather than having some physical explanation—in fact all it means is that explanatory variables useful to predict the exact values are unknown for purposes of the present analysis.

If we are now given some measurements from the missing county (in the form of a list y), Bayes's theorem [Bayes 1763] can be applied to determine a new estimate of the county's true GM. Bayes's theorem states that

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{3}$$

The notation $p(\theta|y)$ reads "the probability of θ *given* y "; in the current context it represents the probability that "the true mean is θ", given "the set of observations y." In order to apply this equation, we must have some way of calculating $p(y|\theta)$, known as the *likelihood.* The likelihood $p(y|\theta)$ is the probability that the values y would have been observed, if the true value of log(GM) is θ. In order to evaluate this likelihood, we require a statistical model for the distribution of observations *within* a county. The value of θ that maximizes Equation 3 can be thought of as a "best guess" at the true value of θ. (Note that the denominator of Eq. 3 is independent of θ; in practice we need not evaluate it, since it merely provides a normalization factor.) Generally, we are not interested only in the best guess but also in the uncertainty—the range of values of θ that are reasonably consistent with the observations y and with our prior knowledge of the possible values of θ.

In general, and in the case of the current Minnesota data set, we do not have direct knowledge of the true distribution of county GMs. There are only eight Minnesota counties for which more than twenty observations were made. Since the distribution within each county is approximately lognormal with a GSD of about 2.1, twenty observations are only enough to characterize the GM of a county within about 20%. For most Minnesota counties, then, the true GM is quite uncertain.

The observed county GMs, however, are approximately lognormally distributed, and the distribution of measurements *within* each county is also approximately lognormal. We therefore select the following statistical model for the distribution of radon measurements:

1. The true county GMs are lognormally distributed: the values of log(GM) are drawn from a normal distribution with unknown mean μ and unknown variance σ^2, as in Eq. 2.

2. The observations *within* a county are also lognormally distributed: the logarithms of the observations are drawn from a normal distribution with a mean equal to the true value of log(GM) and unknown variance κ^2. For the purposes of the present analysis, κ^2 is assumed to be the same for all counties. This is equivalent to the assumption that all of the counties have the same GSD.

The true value of log(GM) for each county is the main parameter of interest. With the lognormality assumptions mentioned above, the application of Bayes's theorem (Eq. 3) yields a particularly simple result for the prediction of log(GM) for county i: the most probable value of the true log(GM) is given by a weighted average between the observed value of log(GM) for the county and the 'grand mean' μ of the distribution from which all of

the county log(GM) values are drawn, where the relative weights depend upon the number of observations n_i in the county, and the variance estimates σ^2 and κ^2:

$$\log(\mathrm{GM}_i^{\text{point estimate}}) = \frac{(1/\sigma^2)\mu + (n_i/\kappa^2)\log(\mathrm{GM}_i^{\text{obs}})}{(1/\sigma^2) + (n_i/\kappa^2)}. \quad (4)$$

Equation 4 provides a point estimate of the true GM, but this estimate is uncertain: the probability distribution of the true value about this estimate, given μ, σ^2, and n_i/κ^2, is described by:

$$p(\theta|\mu, \sigma^2, n_i/\kappa^2) = \mathrm{N}(\log(\mathrm{GM}_i^{\text{point estimate}}), V_i^2) \quad (5)$$

where

$$V_i^2 \equiv (1/\sigma^2 + n_i/\kappa^2)^{-1} \quad (6)$$

Mathematically it's as though we already had some number (equal to κ^2/σ^2) of observations of $\log(C_{\mathrm{Rn}})$ in each county, with the mean of the observations being μ, before any actual observations were made.

In order to actually perform this adjustment, we need values for μ, σ, and κ. These parameters can be estimated from the data. For example, a point estimate for μ is provided by the mean of the observed county log(GM)s, yielding a value $\mu \approx 4.96$ (in units of $\log(\mathrm{Bq/m^3})$). Rough point estimates for the true within-county variance $\kappa^2 \approx 0.54$ and the true between-county variance $\sigma^2 \approx 0.11$ can be determined from an analysis of variance. In untransformed space, these correspond to a within-county GSD of $\exp(\sqrt{0.54}) = 2.1$, and a distribution of true county GMs that has a GM of $\exp(\mu) = 143$ $\mathrm{Bq/m^3}$and a GSD of $\exp(\sqrt{0.11}) = 1.4$. Obtaining point estimates of parameters from the data themselves, and then using those estimates through Bayes's theorem to obtain estimates for quantities of interest, is sometimes referred to as an "empirical Bayes" method.

Although use of the point estimates for the model parameters would lead to reasonable estimates of the county GMs, the resulting uncertainty estimates would be too small, since they would not include the uncertainties in the model parameters themselves. For this reason, we do not restrict ourselves to point estimates of the values of the parameters; rather, we estimate the distribution of possible (or likely) values for the parameters, then draw randomly from that distribution and use the resulting parameters in Eq. 5 to obtain an estimate for each county's true log(GM). Details of this so-called "full Bayes" procedure can be found in Gelman et al. 1995. The sampling procedure is repeated many times (1000 in the current case), with each set of parameters yielding an estimate for each county's log(GM); the resulting distribution of 1000 GM estimates for each county is spread over a range due to both the uncertainty in the true values of the parameters in the hierarchical model *and* the uncertainty due to the finite number of measurements (which would remain even if we knew the exact parameters of the distribution from which the county GMs are drawn). We select the mean of the 1000 estimates for each county as our "best guess", or *posterior estimate*, of the county's GM. The procedure described here can be carried out directly, or as a special case of a random effects regression, described in the next section.

Results for counties with more than five observations are shown in Figure 2, in which the posterior estimates of county GM and uncertainty (an error bar containing the middle 50% of the posterior estimates for each county's GM when the sampling procedure described above is performed 1000 times) are plotted against the GM of the measurements in the county. The points are plotted as numbers, with the number being the number of observations in the

county. The distribution of posterior estimates of the GMs (shown on the abscissa) is much narrower than is the distribution of observed county GMs (on the ordinate), as expected. The mean estimate of σ^2 is 0.097, and the mean estimate of κ^2 is 0.570, corresponding to a county GSD of 2.13. The mean estimate of μ is 4.95.

Lac Qui Parle County, with only two observations yielding an observed GM of 498 Bq/m^3, has a mean posterior estimate of 196 Bq/m^3, although the true value may be as low as 122 Bq/m^{3}or as high as 303 Bq/m^3(the 5 and 95 percent posterior interval limits, respectively). The results appear reasonable, although it may be surprising how large the effect of finite sample size is estimated to be.

Interestingly, although Lac Qui Parle County had the highest observed GM (498 Bq/m^3), it does not have the highest "best guess" GM, losing out to Blue Earth County, which had an observed GM of 250 Bq/m^{3}and has a posterior estimate of 210 Bq/m^3. This is a consequence of the fact that Blue Earth County had many more observations than Lac Qui Parle County (14 as opposed to 2). Simply put, the distribution of observed county GMs suggests that most true county GMs fall in the range between 75 to 150 Bq/m^3, and there is more evidence that Blue Earth County falls beyond that range than there is evidence that Lac Qui Parle does so.

The predictions from the model seem reasonable, but that alone is not, of course, sufficient to give us confidence in them. Several validation checks have been carried out. One type of check concerns the degree of agreement between the model predictions and the observations. For example, given the posterior estimates of the county GMs, how often would we expect to see an observed GM as high as 500 Bq/m^3? To answer this question, we start with the posterior estimates for the county GMs, then simulate the sampling procedure (by selecting simulated "observations" from each county's assumed distribution) and examine the resulting "observations" to see how they compare statistically with the actual observations. For example, if the mean posterior estimate for each county is assumed to be the true value of the county's GM, repeated simulation of the sampling procedure yields at least one county with an "observed" county GM higher than 500 Bq/m^{3}about 30% of the time, so such a high observation clearly does not violate the conclusions based on the model.[3] In other words, even if the mean posterior estimates of the GMs (the highest of which is 210 Bq/m^3) are the true GMs, with the given distribution of sample sizes in each county there is about a 30% chance of getting an observed GM over 500 Bq/m^3.

Another type of validation check that we performed was to create a validation data set by discarding a random 90% of the data from the four counties with more than 50 observations. Complete data from all of the other counties, plus the reduced data from those four counties, were then used to fit the model again. The predictions for those four counties were then compared to the true GMs as known from the complete data for those counties. This sampling/predicting procedure was carried out many times. The model validated well, in the sense that the true values for the well-sampled counties fell within one standard error of the estimate in about 68% of the tests, and within two standard errors in about 95% of the tests.

Several other checks of model fit were made as well; we did not find any significant discrepancies between the model and the data. We conducted similar validation checks on

[3]This example was given to illustrate the basic idea of model checking, but in fact more sophisticated checks are appropriate; for instance, rather than assuming that the true GMs are given by the mean of the posterior estimates, one should simulate the testing procedure for *each set* of posterior estimates, to include the full range of likely values for each county's true GM.

the models discussed below.

Although the statistical model discussed above does validate well and does appear to provide better estimates for each county's GM, the estimates are still fairly uncertain, especially for the many poorly-sampled counties. In the next section, we discuss the use of predictive variables to improve the predictions for the county GMs.

Regression prediction of the county GMs.

It has been noted previously [Nero et al. 1994] that much of the county-to-county variation in Minnesota's indoor radon levels (as measured by the GM) can be explained by variation in surficial radium content as indicated by the aerometric survey conducted as part of the National Uranium Resource Evaluation (NURE). The NURE survey generated estimates of surficial uranium content along flight lines spaced every 6 to 12 miles across the U.S. These data were processed using various extrapolation and smoothing schemes [Duval et al. 1989] to produce a nationwide map of surface uranium concentration, which can be used to estimate the concentration of radium, which is a uranium decay product. Previous work [Moed et al. 1985, Revzan et al. 1988, Gundersen et al. 1991, Jackson 1992, Nero et al. 1994] has used aerial radiometric survey data to predict distributions of radon concentration measurements or to locate areas with high radon "potential". For purposes of the present research, we have aggregated the NURE data of Duval et al. 1989 to generate average surface uranium concentration (expressed in equivalent ppm of uranium) by county; in Minnesota, the resulting NURE values range from 0.14 ppm to 0.57 ppm, with a median of 0.39 ppm.

Figure 3 plots the GM of the radon concentration measurements in each county versus the predicted GM from a conventional weighted linear regression of the logarithm of each county's GM on the logarithm of the county-averaged NURE value; conventional error bars (1 standard error) are plotted for the observed GMs, based on the approximation that the true GSD of each county is 2.1. Only counties with more than five observations are shown, to avoid clutter. Note that over 60% of the error bars cross the line indicating agreement between prediction and reality—thus the data appear to be nearly (but not quite) consistent with the hypothesis that NURE is a perfect predictor of the GM of county radon levels, with the residuals being due to the finite sample size in each county. For the log-space conventional linear regression the value of R^2, a standard measure of model fit, is 0.58 for the counties shown here. However, this figure substantially underestimates the real predictive ability of NURE in this case, since much of the variation between predicted and observed GMs is certainly due to small-sample noise rather than differences between the true GMs and their predicted values.

In this section, we discuss a procedure to predict the true county GM using both the observations and the fitted results for the county. This procedure provides a method of using both observational data and explanatory variables *together* in a statistically consistent way in order to predict each county's true GM.

The statistical model we wish to apply is defined as follows:

1. The true values of log(GM) for each county are drawn from a normal distribution with a mean equal to the predicted value of log(GM) based on a regression, with unknown variance σ^2, so that for county i

$$\log(\mathrm{GM}_i) = \beta_0 + \beta_{\mathrm{NURE}} \log(\mathrm{NURE}_i) + \delta_i, \tag{7}$$

where $p(\delta_i) = \mathrm{N}(0, \sigma^2)$; or equivalently, with $\theta_i = \log(\mathrm{GM}_i)$,

$$p(\theta_i|\mathrm{NURE}_i) = N(\beta_0 + \beta_{\mathrm{NURE}} \log(\mathrm{NURE}_i), \sigma^2). \tag{8}$$

The parameter σ^2 does not have the same value as the σ^2 in the previous section unless $\beta_{\mathrm{NURE}} \equiv 0$, in which case the model reduces to the model in the previous section.

2. We number the 919 radon measurements with the index $\{j\}$, and assume that the logarithm of observation j within county i is drawn from a normal distribution with a mean equal to the county's true log(GM), and with an unknown variance κ^2 assumed the same for all counties, so that

$$p(\log(C_{\mathrm{Rn}})_j) = \mathrm{N}(\log(\mathrm{GM}_i), \kappa^2). \tag{9}$$

The parameters $\beta_0, \beta_{\mathrm{NURE}}, \sigma$, and κ are again to be estimated from the data. A large value of σ^2 would indicate that NURE is a poor predictor of true county GMs, while a small value of σ^2 would indicate that the true county GMs are closely grouped around their NURE predictions.

A simple but imperfect estimate of σ^2 can be obtained as follows: perform a regression of observed log(GM) on log(NURE), then apply the hierarchical model described in the previous section to the *residuals*. This procedure is correct in spirit and provides a quick estimate of the true distribution of the residuals. It has the drawback, however, that it is slightly over-optimistic, in the sense that it yields confidence intervals that are too narrow, since it does not include the uncertainty in the regression coefficients themselves. For the current data set, there are enough different counties that the regression coefficients are fairly well determined, so the overcertainty caused by using only the best-fit regression coefficients is fairly small. However, rather than present results of such an incomplete analysis, we will carry out a procedure, called "random effects regression", that takes into account all sources of uncertainty in the model parameters.

Before embarking on a description of random effects regressions, we first discuss the role of "dummy variables" in conventional linear regressions. In statistical regression, a dummy variable is used to indicate the presence or absence of a particular characteristic, or that the data are included or excluded from a particular class. For example, in the present case we create a dummy variable for each county in Minnesota (85 in all, if we include only the counties for which there is at least one measurement). Each of the 919 radon measurements C_{Rn} is therefore associated with 85 dummy explanatory variables, all but one of which takes the value of zero; the value unity is assigned to the variable that denotes the county in which the measurement was made.

A conventional linear regression of the values of $\log(C_{\mathrm{Rn}})$ on these 85 dummy explanatory variables alone yields 85 regression coefficients, each of which is the mean of the observations of $\log(C_{\mathrm{Rn}})$ in the indicated county. The hierarchical model introduced in section 3 can be reproduced by applying Bayes's theorem with the assumption that these regression coefficients are measurements with error of underlying "true" parameters, which are drawn from a normal distribution.

The hierarchical regression model introduced in the present section can be applied also, and the uncertainties properly estimated, by including another explanatory variable, in addition to the county dummy variables; for each of the 919 observations this variable takes

the value of NURE averaged over the county that contains the observation. With NURE included as an explanatory variable, the values of $\{\delta\}$ from Eq. 7, indicate the "true" residuals from the NURE regression (i.e. the difference between the true value of $\log(GM_i)$ and the NURE predictions).

In general, the difference between a county's true log(GM) and the regression prediction for the county is referred to the "county effect". Regression coefficients (such as county effects) that are assumed to be drawn from a common distribution are usually referred to as "random effects"; hence the name "random effects regression." When a model includes both conventional regression variables ("fixed effects") and random effects, it is called a "mixed effects model". All models discussed hereafter are mixed effects models. As before, the assumption that random effects are drawn from a common distribution in no way implies that there is no *reason* that some of the county effects are large while others are small, merely that we have no information that allows us to distinguish between them.

The mathematical details of performing a Bayesian random effects regression (or mixed effects regression) are rather involved; the reader is referred to the references for a complete discussion [Rubin 1980, Bryk et al. 1992, Zeger et al. 1991, Gelman et al. 1995]. The basic ideas, however, are those discussed in the previous section: the conditional distributions of the parameters (regression coefficients and variance components) are calculated, and parameters are drawn from the calculated distribution. Where appropriate, a hierarchical model is assumed (as for the county effects). The procedure is repeated many times in order to obtain posterior intervals (conceptually similar to confidence intervals, in that they reflect the range in which the true value is likely to fall) on the parameters.

Using the model described by equations 7 and 9 above, we perform 1000 simulations to obtain 1000 estimates for each of the parameters: β_0, β_{NURE}, κ, σ, and each of the 85 values of δ_i. The estimated county effects $\{\delta\}$ do double duty: they allow us to predict the county GMs, and they also provide a way of measuring the extent to which the other explanatory variables allow prediction of indoor radon levels. The extent to which NURE is a good predictor of the true county GMs can be gauged from the likely values of σ: if σ is small (and thus the county effects are all near zero), then NURE alone is enough to predict the GM of radon concentrations in a county from Eq. 7. If σ is large, on the other hand, then at least some of the individual county effects are large, and NURE alone is not sufficient to obtain a good estimate of the county's true GM. Furthermore, σ can be used to define a measure of model fit for the county radon levels that is analogous to R^2: as in [Bryk et al. 1992], we wish to define an effective R^2_{eff} as

$$R^2_{\mathrm{eff}} \equiv 1 - \frac{\text{unexplained variance of true log(GM) values}}{\text{total variance of true log(GM) values}}. \tag{10}$$

We do not know either the actual unexplained variance or the total true variance, but we do have estimates of each: we obtain estimates of σ^2 from random effects regressions performed with and without NURE (or other explanatory variables), in both cases including the county dummy variables. The best estimate of σ^2 when *only* the county dummy variables are included provides us with an estimate of the true total variance of the county GMs, while the best estimate of σ^2 when both dummy variables and other variables are included provides us with an estimate of the unexplained variance. As long as only county-level variables are included, the estimate of R^2_{eff} obtained this way will behave similarly to the conventional measure of R^2, in the sense that it will always increase (or remain constant) as additional variables are added. However, if the model contains individual-house variables,

then the value of R^2_{eff} can actually *decrease* as additional variables are added. An interesting and informative example is considered in the next section.

For the current case, with NURE included as the only explanatory variable, the mean posterior estimate of σ^2 is 0.019. Combining this with the value of $\sigma^2 = 0.097$ obtained when only the county effects are included, we obtain the an estimated value of $R^2_{\text{eff}} = 0.80$; our best estimate is that log(NURE) explains 80% of the county–to–county variation in log(GM). In terms of determining a county's log(GM), knowledge of the county's NURE value is "worth" an extra $\kappa^2/\sigma^2 \approx 30$ observations in each county.

The ability of NURE to predict county GMs so well appears to be unique to the state of Minnesota—in the several other states of the U.S. that we have examined, NURE has lower predictive power, with log(NURE) typically explaining about 30% to 65% of the variation in the logarithm of the county GMs.

The coefficient of log(NURE) in Eq. 7 is estimated to be $b = 0.711$, with 90% posterior bounds of 0.561 and 0.849. Since a coefficient of log(NURE) different from unity implies (after transforming back from log space) a nonlinear relationship between county soil radium concentration and county indoor radon concentration, this result may seem peculiar: at least for individual homes, physical models suggest the indoor concentration should be approximately proportional to the radium concentration in the surrounding soil.

Several factors may contribute to a non-linear relationship between county-average NURE and the indoor radon concentration measurements. First, the use of county-average NURE would be completely appropriate only if homes in each county are uniformly distributed over the entire area of the county. In fact, in some counties homes are probably more heavily concentrated in high-radium areas, while in others they're more heavily concentrated in low-radium areas, simply by random chance. A consequence of such a phenomenon is a decreased coefficient on log(NURE), through the regression effect (cf. Price 1995 for another consequence of the regression effect in indoor radon studies). Furthermore, the NURE measurements are subject to errors, some of which are related to factors such as soil moisture content that are likely to affect indoor radon concentrations [Duval et al. 1989, Schumann et al. 1994]. Given these facts, a coefficient different from unity in the regression is certainly not a cause for concern.

Figure 4 shows the result of performing the mixed effects regression; as in Fig. 3, only counties with more than 5 observations have been plotted. The posterior predicted GM for each county has been plotted with a square, as a function of the prediction based on a conventional regression on log(NURE); thus, if the posterior prediction and the conventional regression prediction agreed perfectly, the square would be plotted on the 45-degree line on the figure. The GM of observations in each county has been plotted with a point (the same as Figure 3, except that that error bars are not shown). The position of each square represents a sort of weighted average between the observed GM and the GM predicted from a conventional regression on log(NURE), with the relative weighting determined from the data.

For counties with many observations the posterior estimate is always very close to the observed GM, while for counties with fewer observations the final estimate can be shifted fairly far. Most of the final estimates are very close to the regression line—there is strong evidence that NURE explains almost all of the county-to-county variation in radon levels in Minnesota. However, as noted previously the distribution *within* each county is quite broad: the best estimate of κ is 0.76, corresponding to a GSD of 2.1.

Table 1 presents results for each county in Minnesota (including the two counties with no observations): the number of observations in the county, the GM of the observations, the predicted GM from a conventional linear regression of county log(GM) on county log(NURE) alone, and the posterior estimate and uncertainty (one standard error) of the county's true GM, based on the mixed effects regression described above. The 'uncertainty' is an approximation, treating the posterior GM estimates as if they were normally distributed. In fact, the distribution of posterior estimates for a county is close to lognormal, but since the uncertainties are generally small compared to the estimates, the normal approximation is fairly good. If more accurate summaries of the uncertainties were desired, posterior intervals could be determined directly from the distribution of 1000 posterior estimates from each county.

Figure 5 displays histograms of the distribution of observed and estimated county GMs Each county is represented by a number, indicating the number of observations in the county. Each number is stacked in the column appropriate to the county GM radon concentration[4]. Note that all of the counties with observed GMs over 250 Bq/m^{3}have 5 or fewer observations. The distribution of predicted GMs is much tighter than the distribution of observed GMs—there is no convincing evidence that any of the true GMs are as high as 250 Bq/m^3, although some county predictions barely include 250 Bq/m^{3}within two standard errors. The distribution of true GMs is somewhat broader than the distribution of predicted GMs that is shown, since the true GMs are distributed about the predicted values, with standard errors given in Table 1. As noted previously, $\sigma^2 = 0.097$ is the estimated variance of the logarithms of the true county GMs, so that the distribution of true county GMs has a relatively small GSD, about $\exp(\sqrt{0.097}) = 1.37$.

Including additional explanatory variables.

In addition to the measured indoor radon concentration and the county NURE measurement, we have some information on each home in the survey: does the home have a basement, and, if so, was the measurement made in the basement. The presence of a basement might be expected to have some mild effect even on first-floor indoor radon measurements, and certainly measurements made in basements are expected to be considerably higher than measurements made on the first floor. There are two substantive reasons that we wish to take account of the basement and floor effects.

First, we are interested in the magnitude of the coefficients themselves: how much higher are measurements made in the basement than those made on the first floor?

Second, what are the county effects *after* controlling for the floor effect in the homes in each county? For example, do the low-radon counties have lower radon levels merely because they have more non-basement homes?

As an initial attempt to answer these questions, we introduce three individual-home explanatory dummy variables. One variable (γ) indicates homes that have basements and were monitored in the basement, one (ϕ) indicates homes that have basements but were monitored on the first floor, and one (ν) indicates homes without basements. Most homes—

[4]It is important to remember what these histograms represent: the "105" in the 120–140 Bq/m^{3}interval does *not* represent a county with 105 observations all of which fell in that interval; the observations from that county are spread over a very large range, (from 9.25 Bq/m^{3}to 888 Bq/m^3, as it happens), with a GM that falls in the range 120–140.

769 of the 919 homes tested—have a basement and were monitored in the basement.

The model is defined as follows. For a home j in county i, the probability of obtaining a given observation is given by

$$p(\log(C_{\mathrm{Rn}})_{\ j}) = \mathrm{N}(\beta_{\mathrm{NURE}} \log(\mathrm{NURE_i}) + \beta_{\mathrm{bb}}\gamma_j + \beta_{\mathrm{b1}}\phi_j + \beta_{\mathrm{nob}}\nu_j + \delta_i, \kappa^2). \qquad (11)$$

Here β_{bb} is the effect associated with a basement home that is monitored in the basement, β_{b1} is the effect for a basement home monitored on the first floor, and β_{nob} is the effect for a home without a basement. The county effects $\{\delta_i\}$ are assumed to be normally distributed, as before. As before, the county effects measure the extent to which the explanatory variables in the linear model fail to explain all of the county-to-county variation in radon concentrations. If NURE and the housing dummy variables were sufficient to predict the distribution of measured values in homes in different counties, with no remaining evidence of unexplained between-county variation, then the county effects would be near zero. Coefficient estimates and variance estimates are presented in Table 2, along with estimates from other models discussed below. The coefficient estimates all happen to have standard errors of about ± 0.1 or so, except for the coefficient associated with the fraction of homes that do not have basements (discussed below), which has a standard error of about ± 0.3.

The model including individual-house explanatory variables does not allow direct prediction of the county GMs, since Eq. 11 does not contain only county-level variables. Essentially, we obtain separate estimates for each county for homes in three different categories: homes with basements in which the Rn levels were measured in the basement, homes without basements, and homes with basements but in which the monitoring was nevertheless performed on the first floor. Use of the results to estimate the true county GMs would require knowledge of the distribution of housing types by county. For example, one might start (on one extreme) with the approximation that all counties have the same proportion of basement and non–basement homes, or (on the other extreme) with the approximation that all counties have exactly the observed proportion of basement and non-basement homes. We have not attempted to model the distribution of housing types. We perform the individual-house analysis only to illustrate that the techniques described in this paper can handle both individual and county–level data.

An interesting result of this regression is that the variance of the county effects goes *up* compared with the previous, NURE-only regression. How can this happen? Consider Roseau county. The value of NURE averaged over the county, when used in the conventional NURE–only regression, predicts the average value of $\log(C_{\mathrm{Rn}})$ for homes in the county should be about $\log(126\ \mathrm{Bq/m^3})$, in good agreement with the observed value of $\log(131\ \mathrm{Bq/m^3})$. However, in 5 of the 14 monitored houses in Roseau county, the measurement was made on the first floor of a home rather than in a basement. Since first-floor measurements are expected (based on the full regression) to be about half as high as basement levels, and since 5/14 represents a much larger fraction of non-basement homes than is typical in counties in Minnesota, the full regression prediction for the homes in Roseau county is now much too low, so the county effect estimate for this county must be made fairly large in order to bring the prediction into agreement with the observations.

The increase of the size of the county effects when additional data are included indicates some violation of the model. In this case, it indicates that there is some difference between counties with many non-basement homes and those with few non-basement homes— some difference that affects radon levels. For example, counties with generally high soil moisture

may have fewer basement homes, and the soil moisture may also influence indoor radon concentrations (and perhaps the NURE observations as well [Duval et al. 1989]).

To help resolve this issue, we add another county-level explanatory variable: observed fraction of non-basement homes. For all the homes in a county (whether or not they have a basement) this variable takes the value of the fraction of survey homes in the county that do not have a basement. (We would prefer to use *actual* fraction of non-basement homes in the county, rather than *observed* fraction, which is subject to significant noise due to the small number of observations in most counties. Unfortunately the actual fraction for each county is not available.) A sizeable coefficient for this variable would indicate that the fraction of non-basement homes is correlated with county radon levels, *over and above* the correlation due to the fact that levels in the measured homes depend upon whether the measurement was made in a basement or not.

Including the county-level non-basement fraction variable does decrease the magnitude of the county effects when individual-house basement categories are included (Model 5), although the county effects are still slightly smaller in the models that do not include individual-house variables.

Spatial distribution of county effects.

Thus far, we have not included any spatial information in our analysis. This fact does not invalidate any of the analyses discussed above; specifically, the estimates of the county effects (and the estimates of σ) are valid even though spatial information has not been included. Given these facts, there might seem to be no need to delve into the spatial relationships in the data.

However, there are pitfalls to blindly applying the regression results without regard to spatial concerns. For example, suppose we wish to use the NURE-only regression to predict the mean radon level in some group of counties. If these counties are selected at random across the state, there is no problem with combining the regression predictions for the individual counties to predict the geometric mean of the entire group, and the more counties that are included in the group, the lower the error in the estimated GM is likely to be. If, on the other hand, the counties were all selected from a particular region of the state, then the presence of spatial correlations in the county effects would lead to problems: our estimated group GM would be overcertain, unless we account for such correlation.

Also, spatial correlation in the county effects presents an opportunity: if there are some areas that are higher or lower in radon than predicted, even after controlling for the available explanatory variables, then the locations of these areas might suggest avenues of exploration to improve the models. In principle, even if no explanatory variables can be found that explain the spatial correlations, the presence of the correlations themselves can allow improvements in the accuracy and precision of the models by creating an explicitly spatial model. However, such in-depth analysis of the spatial correlations is beyond the scope of the present paper.

Instead, we display the estimated county effects from Model 5 on a map of Minnesota (Figure 6). The estimated county effects have been multiplied by 100 to avoid printing unnecessary digits. Notice that there do seem to be patterns in the distribution of county effects; specifically, most of the large negative county effects occur in counties to the east of about 94 degrees longitude, while most of the large positive county effects occur to the west

of that line. "Large" is only relative in this context—the county effects with the largest magnitudes correspond to modifications from the ordinary regression predictions of only about 15%, and most of the effects are much lower.

To remove the obvious east-west spatial trend, we added a "longitude" variable to the model. This county-level variable assigns to each data point the scaled longitude of the center of the county in which the house sits; the variable was defined as (longitude $-$ 97)/7, which is zero at the western edge of the state, and unity near the eastern edge. The resulting models (numbers 6–9 in Table 2) show a barely improved fit, as indicated by the decrease in σ. In addition, examination of the spatial distribution of the associated county effects reveals no obvious large-scale trends, although non-random clumps of positive or negative county effects can still be found. The negative coefficient of the longitude variable indicates that county mean radon concentrations tend to be lower in the eastern part of the state than would be predicted based on the other explanatory variables alone, and higher on the western part of the state. However, the effect is quite small, changing most county posterior predicted GMs by a few Bq/m^3in spite of the fact that the coefficient of the longitude variable is substantial: the effect of the sizeable coefficient of the longitude variable is largely offset by the decrease in the coefficient of log(NURE), which is partially collinear with longitude.

In summary, although there is evidence for spatial variation in county GMs that is not explained by the included explanatory variables, the effect of such unexplained variation on the predictions for the true county GMs is very small.

Discussion and Conclusions.

The models discussed above contain four variables believed to be directly related to indoor radon concentration measurements: NURE, which is a measure of surficial radium concentration; and the three housing variables, which are related to the coupling between soil-gas radon concentrations and the indoor radon concentration.

We have included two additional county-level variables in the model: observed fraction of non-basement homes, and county longitude. These variables are *not* directly related to indoor concentration measurements; to the extent that they increase the predictive value of the models, they must be proxy for other (presently unknown) variables.

Inclusion of the individual-house basement categories improves the within-county fits, as indicated by the decrease in κ, although it does not decrease σ. Although including the basement categories does not result in lower county effects, it does lead to a slight decrease in the *uncertainty* of the individual county effect estimates—this is a small effect in most counties, but for a few counties the uncertainty (the width of the 68% posterior intervals) decreases by 15% or more.

Which of the models discussed above should be preferred? The answer depends on the purpose of the analysis. For purposes of estimating the true county GM's in Minnesota, using both the regression fits and the observations in each county (i.e., including the county effects estimates), models 2 or 6 are most convenient. Model 6 contains longitude, which obviously acts as a proxy for some other variable or variables; this fact does not affect its value in the prediction of radon levels in Minnesota counties, but does make it harder to compare the results of the current study to those from other states in which longitude does not act as a useful proxy.

Model 2, which contains only NURE as an explanatory variable (and which was used to generate the predicted GMs in Table 1), still does an excellent job at fitting the county means. The estimated county geometric means are slightly less certain than in model 6, but the fact that only one variable is included, and that it has direct physical interpretation, may be sufficient reason to prefer this model in some instances. Note, however, that it is possible that the NURE measurement is itself partly a proxy for other important variables, as illustrated perhaps by the overlap in explanatory power between NURE and soil classes observed in previous work [Nero et al. 1994].

The models that include individual-house explanatory variables are useful for understanding the factors that influence radon concentration measurements. All of the models agree that basement measurements in a county are about twice as high as first-floor measurements, and that there is no evidence that first-floor measurements are higher in basement homes than in non-basement homes in the same county. In addition to the models discussed in this paper, we also looked for variation of basement effect with latitude, and for a difference between measurements in finished and unfinished basements; the magnitudes of those effects were found to be very small, and to have no significant effect on the posterior estimates for the county GMs.

We remind the reader that the data used in the present work were short-term basement measurements made in winter, and so cannot be used directly to estimate radon risk or radon exposure for the living areas of the home.

Random and mixed effects regression modeling of the Minnesota radon data have proved to be extremely useful in obtaining predictions for the true county geometric mean indoor screening (i.e., short-term winter) radon concentrations, and in determining the explanatory value of NURE and of the housing parameters. The predictions use all of the available data—both measurements and explanatory variables—and take proper account of the varying number of measurements in each county. The techniques discussed in the present work allow investigation of the use of various explanatory variables to account for variations in radon measurements, while minimizing the effects of finite sample size in the various counties.

The models seem appropriate to the data, and we have confidence in their basic conclusions; specifically, we believe the county GM estimates presented in Table 1, and their posterior intervals, to be substantially correct, although admittedly we might wish to widen the posterior intervals a bit, just in case of some slight undetected model violation. We feel that the posterior estimates of the county GMs should be used rather than taking the observed GM as an estimate of the true GM: for example, it seems extremely unlikely that Lac Qui Parle county and Murray county have true GMs over 450 Bq/m^3, or even over 300 Bq/m^3.

The Bayesian techniques described in the current work promise more efficient use of data and more reliable prediction than the techniques currently in use in the radon characterization field, and we recommend their more widespread use. They are particularly important in identifying high-radon areas when only sparse monitoring data are available; even in cases in which counties are much better sampled than in the present case, attempts to predict radon levels at smaller spatial scales (such as zip code areas or census tracts) will inevitably need to cope with the effects of small sample sizes. Furthermore, the statistical techniques are obviously not specific to radon, and could profitably be applied to a wide variety of environmental problems.

Acknowledgements.

We are indebted to U.C. Berkeley graduate student John Boscardin, who wrote the random effects regression computer program used in this work, and graciously permitted its use. We would like to thank Randy Schumann of the U.S. Geological Survey, Richard Sextro and Kenneth Revzan of Lawrence Berkeley Laboratory, Deborah Nolan of the University of California, Berkeley, and Clarice Weinberg of the National Institute for Environmental Health Sciences, for their helpful comments and advice.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Health and Life Sciences Research Division of the U.S. Department of Energy (DOE) under contract DE-AC03-76SF00098.

References

Alexander, B.; Rodman, N.; White, S.B.; Phillips, J., Areas of the United States with elevated screening levels of Rn222, *Health Phys.* **66**, 50–54; 1994.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 330–418. Reprinted, with bibliographical note by G. A. Barnard, in *Biometrika* **45**, 293–315; 1958.

Bryk, A.S. and Raudenbush, S.W. (1992) Hierarchical Linear Models, Sage Publications, Newbury Park.

Cohen, B.L., Survey of one-year average Rn levels in Pittsburgh area homes. *Health Phys.* **49**, 1053–1059; 1985.

Dudney, C.S., Hawthorne, A.R., Wilson, D.L., Gammage, R.B., Indoor Rn222 in Tennessee Valley houses: seasonal, building, and geological factors. *Indoor Air* **2**, 32–39; 1992.

Duval, J.S., Jones, W.J., Riggle, F.R., and Pitkin, J.A., Equivalent Uranium Map of the Conterminous United States, *Open-File Report 89-478, U.S. Geological Survey* (1989).

Gelman, A; Carlin, J; Stern, H; Rubin, D. (1995) Bayesian Data Analysis, Chapman and Hall, New York.

Gesell, T.F., Background atmospheric Rn222 concentrations outdoors and indoors: A review, *Health Phys.* **45**, 289–302; 1983.

Gundersen, L.C.S., Schumann, R.R., Otton, D.E., Dickinson, K.A., Peake, R.T., and Wirth, S.J., 1991, Preliminary radon potential map of the United States, in Proceedings of the 1991 EPA International Symposium on Radon and Radon Reduction Technology, Volume 2, Symposium Oral Papers, Technical Sessions 6–10: U.S. Environmental Protection Agency Report EPA/600/9-91/037B, 9-13–9-32.

Harter, H. L. and Moore, A. H., Iterative maximum-likelihood estimation of the parameters of normal populations from singly and doubly censored samples. *Biometrika*, **53**, 205–213; 1966.

Jackson, S.A., Estimating radon potential from an aerial radiometric survey, *Health Phys.* **62**, 450–452; 1992.

Janssen, I. and Stebbings, J. H., Gamma distribution and house Rn222 measurements. *Health Phys.* **63**, 205–208; 1992.

Lindley, D. V. and Smith, A. F. M., Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* **34**, 1–4; 1972.

Moed, B.A.; Nazaroff, W.W.; Nero, A.V.; Schwehr, M.B.; Van Heuvelen, A., Identifying areas with potential for high radon levels: analysis of the National Airborne Radiometric Reconnaissance data for California and the Pacific Northwest. *Bonneville Power Administration* DOE/BP-00098-5 (1985).

Nero, A.V.; Schwehr, M.B.; Nazaroff, W.W.; Revzan, K.L. Distribution of airborne radon-222 concentrations in U.S. homes, *Science* **234**, 992–997; 1986.

Nero, A.V.; Gadgil, A.J.; Nazaroff, W.W.; Revzan, K.L., Indoor radon and decay products: Concentrations, causes, and control strategies; *Technical Report LBL-27798, Lawrence Berkeley Laboratory* (1990).

Nero, A. V.; Leiden, S. V.; Nolan, D. A.; Price, P. N.; Rein, S.; Revzan, K. L.; Wollenberg, H.R.; Gadgil, A. J., Statistically-based methodologies for mapping of radon "actual" concentrations: the case of Minnesota *Radiat. Protect. Dosim.* **56**, 215–219; 1994.

Price, P. N., The regression effect as a cause of the nonlinear relationship between short- and long-term radon concentration measurements. *Health Phys.* **69**, 111–114; 1995.

Revzan, K. L.; Nero, A. V.; Sextro, R. G., Mapping surficial radium content as a partial indicator of radon concentration in U.S. homes. *Radiat. Protect. Dosim.* **24**, 179–184; 1988.

Rubin, D. B. Using empirical Bayes techniques in the law school validity studies (with discussion). *Journal of the American Statistical Association* **75**, 801–827; 1980.

Schumann, R.R., Gundersen, L.C.S., Tanner, A.B., "Geology and Occurrence of Radon", *Radon: Prevalence, Measurements, Health Risks and Control*, Niren L. Nagda, Ed., ASTM Manual Series MNL 15, Philadelphia, PA, 1994.

Tate, E.E.; Lori, A.; Pedersen, D.; Schowalter, D.; Jachim, J.; Jaros, J.; Morin, S.; Fuoss, S., Survey of radon in Minnesota homes, *Minnesota Department of Health, Division of Environmental Health*, Minneapolis (1988).

White, S.B.; Bergsten, J.W.; Alexander, B. V.; Rodman, N. F., Phillips, J.L., Indoor Rn222 concentrations in a probability sample of 43000 houses across 30 states, *Health Phys.* **62**, 41–50; 1992.

Wirth, S. National Radon Database Documentation: The EPA/State Residential Radon Surveys, *Sanford Cohen and Associates, for the U.S. Environmental Protection Agency* (1992).

Zeger, S. L. and Karim, M. R., Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–89; 1991.

Table 1: Comparison of GM of observations, GM predicted by ordinary regression on NURE, and posterior prediction as discussed in the text. Absolute uncertainties tend to be larger for high-radon counties than for low-radon counties. All posterior predictions and uncertainties are subject to small errors due to the finite number of simulation runs.

county fips code	county name	number of obs.	observed GM (Bq/m^3)	NURE predicted GM (Bq/m^3)	posterior GM (Bq/m^3)
1	AITKIN	4	73	90	87 ± 12
3	ANOKA	52	88	80	84 ± 8
5	BECKER	3	107	135	131 ± 18
7	BELTRAMI	7	121	96	100 ± 13
9	BENTON	4	130	133	131 ± 18
11	BIGSTONE	3	169	193	188 ± 24
13	BLUEEARTH	14	250	178	194 ± 24
15	BROWN	4	189	179	179 ± 25
17	CARLTON	10	96	116	109 ± 12
19	CARVER	6	144	157	153 ± 18
21	CASS	5	151	95	100 ± 14
23	CHIPPEWA	4	210	178	179 ± 24
25	CHISAGO	6	107	87	88 ± 12
27	CLAY	14	222	187	195 ± 23
29	CLEARWATER	4	100	140	134 ± 18
31	COOK	2	73	103	100 ± 14
33	COTTONWOOD	4	97	186	172 ± 24
35	CROWWING	12	97	94	95 ± 11
37	DAKOTA	63	137	144	139 ± 10
39	DODGE	3	224	177	179 ± 24
41	DOUGLAS	9	194	164	168 ± 21
43	FARIBAULT	6	75	181	156 ± 22
45	FILLMORE	2	105	197	187 ± 25
47	FREEBORN	9	259	172	185 ± 24
49	GOODHUE	14	235	169	184 ± 23
51	GRANT	0	NA	190	191 ± 27
53	HENNEPIN	105	136	137	136 ± 9
55	HOUSTON	6	172	210	201 ± 27
57	HUBBARD	5	85	110	105 ± 13
59	ISANTI	3	107	86	87 ± 12
61	ITASCA	11	95	92	93 ± 11
63	JACKSON	5	280	183	191 ± 27
65	KANABEC	4	128	141	137 ± 19
67	KANDIYOHI	4	291	158	168 ± 24

county fips code	county name	number of obs.	observed GM (Bq/m^3)	NURE predicted GM (Bq/m^3)	posterior GM (Bq/m^3)
69	KITTSON	3	115	145	141 ± 18
71	KOOCHICHING	7	57	78	73 ± 10
73	LACQUIPARLE	2	498	183	192 ± 29
75	LAKE	9	54	90	80 ± 10
77	LAKEOFTHEWOODS	4	168	91	95 ± 14
79	LESUEUR	5	185	168	169 ± 23
81	LINCOLN	4	314	201	207 ± 28
83	LYON	8	242	194	201 ± 26
85	MCLEOD	13	118	162	147 ± 18
87	MAHNOMEN	1	145	163	162 ± 22
89	MARSHALL	9	127	148	141 ± 17
91	MARTIN	7	100	165	150 ± 19
93	MEEKER	5	126	149	145 ± 19
95	MILLELACS	2	69	127	119 ± 17
97	MORRISON	9	109	137	130 ± 15
99	MOWER	13	183	176	177 ± 21
101	MURRAY	1	448	195	196 ± 29
103	NICOLLET	4	323	175	185 ± 26
105	NOBLES	3	255	195	198 ± 28
107	NORMAN	3	103	177	166 ± 22
109	OLMSTED	23	126	174	152 ± 17
111	OTTERTAIL	8	144	127	129 ± 17
113	PENNINGTON	3	78	139	130 ± 17
115	PINE	6	72	131	118 ± 16
117	PIPESTONE	4	200	206	205 ± 27
119	POLK	4	146	177	171 ± 23
121	POPE	2	134	179	175 ± 24
123	RAMSEY	32	112	110	110 ± 10
125	REDLAKE	0	NA	146	148 ± 21
127	REDWOOD	5	234	190	193 ± 26
129	RENVILLE	3	156	192	186 ± 25
131	RICE	11	221	168	177 ± 22
133	ROCK	2	137	213	205 ± 29
135	ROSEAU	14	131	126	127 ± 14
137	STLOUIS	116	83	105	88 ± 6
139	SCOTT	13	181	153	159 ± 19

county fips code	county name	number of obs.	observed GM (Bq/m^3)	NURE predicted GM (Bq/m^3)	posterior GM (Bq/m^3)
141	SHERBURNE	8	111	90	92 ± 13
143	SIBLEY	4	129	173	166 ± 21
145	STEARNS	25	148	159	153 ± 15
147	STEELE	10	181	177	178 ± 20
149	STEVENS	2	222	205	205 ± 29
151	SWIFT	4	100	183	170 ± 24
153	TODD	3	164	142	142 ± 19
155	TRAVERSE	4	231	209	209 ± 28
157	WABASHA	7	208	165	170 ± 22
159	WADENA	5	103	91	92 ± 12
161	WASECA	4	62	170	151 ± 21
163	WASHINGTON	46	131	132	131 ± 11
165	WATONWAN	3	344	167	177 ± 26
167	WILKIN	1	344	173	175 ± 25
169	WINONA	13	163	205	190 ± 23
171	WRIGHT	13	182	138	147 ± 18
173	YELLOWMEDICINE	2	122	189	181 ± 24

Table 2: Coefficient estimates and measures of model fit for models discussed in the text. Recall that coefficients apply in transformed (natural log) space. Each row includes all of the coefficients estimated for a given model, except that each model also included county dummy variables which were treated as random effects (assumed drawn from a normal distribution with mean 0 and variance σ^2), as discussed in the text. "Const" refers to the constant term in the models (where appropriate), and "long." refers to the scaled longitude variable (longitude–90)/7.

	coefficient estimates							variances	
		county-level			individual house				
	const	log of NURE	fraction w/o bmt	long.	has bmt meas bmt	has bmt meas 1st	no bmt	κ	σ
1	4.86							0.76	0.31
2	5.73	0.71						0.76	0.14
3		0.70			5.85	5.22	5.22	0.72	0.17
4	5.74	0.71	−0.13					0.76	0.14
5		0.75	0.54		5.87	5.24	5.19	0.72	0.15
6	5.83	0.67		−0.32				0.76	0.13
7	5.86	0.63	−0.40	−0.41				0.76	0.13
8		0.67	0.22	−0.52	6.03	5.38	5.34	0.72	0.13
9		0.64		−0.55	6.03	5.38	5.37	0.72	0.13

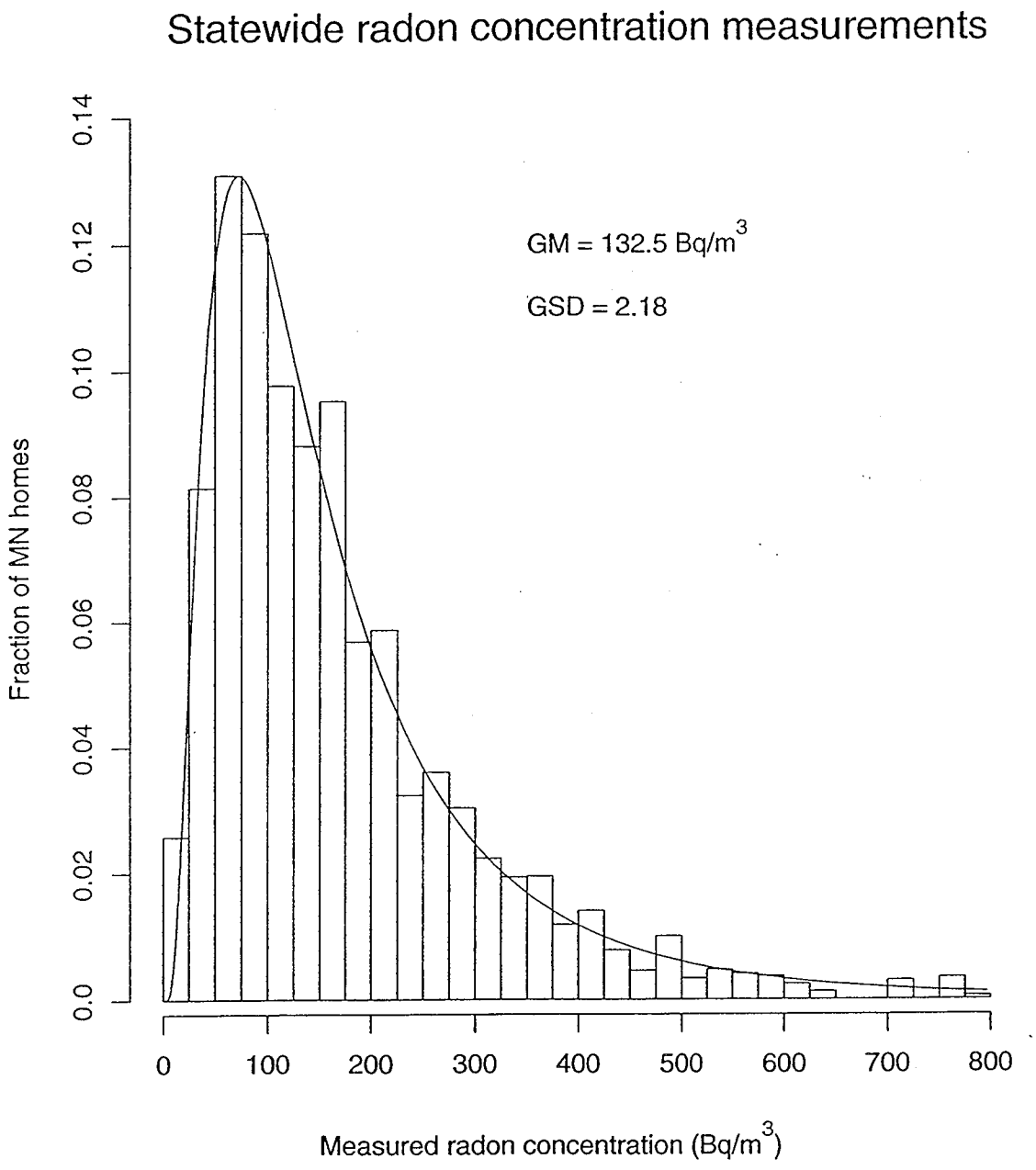


Figure 1: Histogram showing the distribution of screening radon concentration measurements in Minnesota, weighted by sampling weight reported in the SRRS data set. A lognormal distribution with GM=132.5 Bq/m^3 and GSD=2.18 has been superimposed on the data.

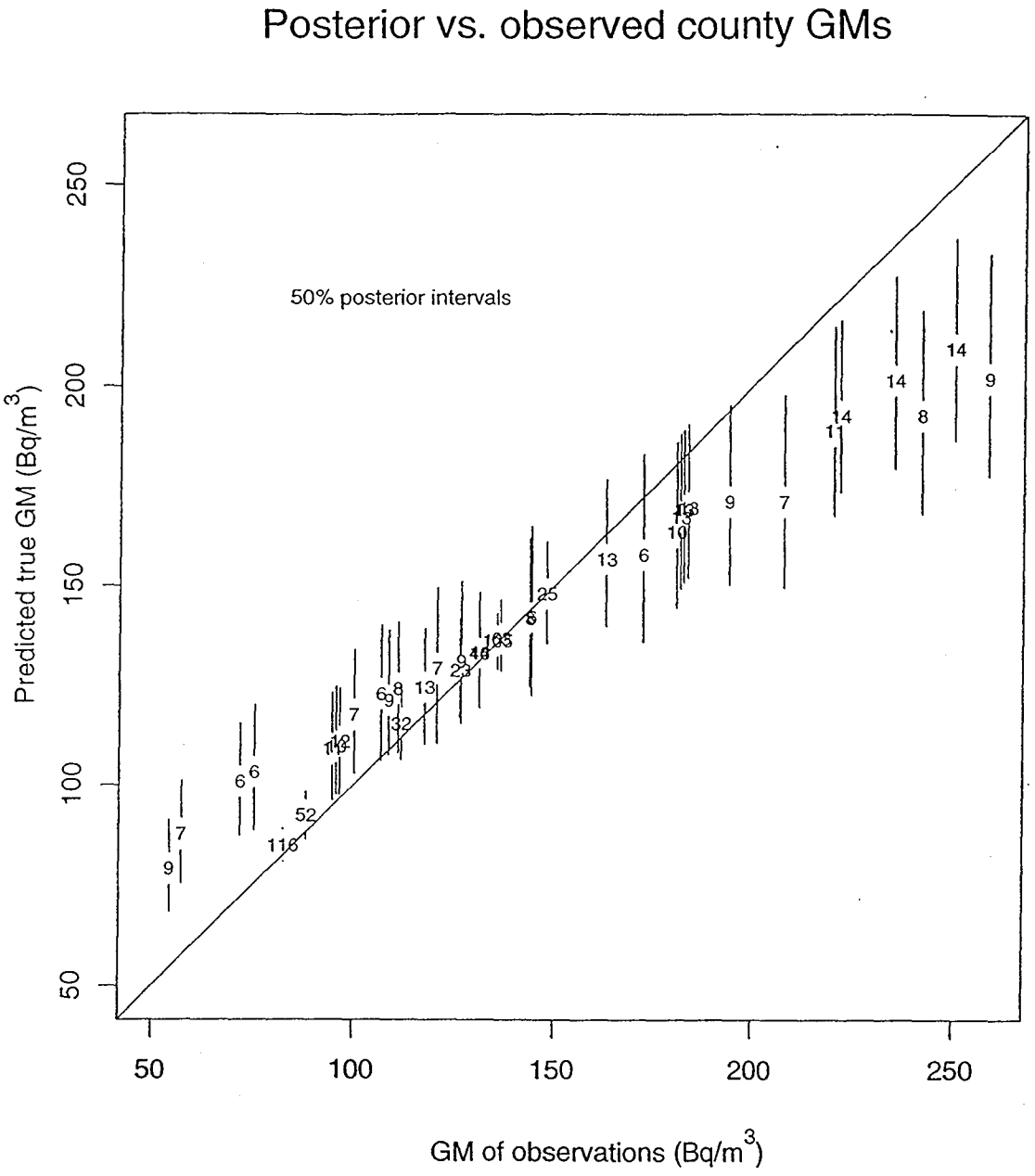


Figure 2: Posterior predictions of county GM values vs. GM of observations, for counties with more than 5 observations. Bars indicate the range that includes the middle 50% of posterior predictions based on 1000 simulation draws from the probability distribution of parameter values, as discussed in the text. The posterior prediction for each county is a compromise between the observed GM in the county and the grand mean of all of the county GMs, with the relative weighting determined from the data as described in the text.

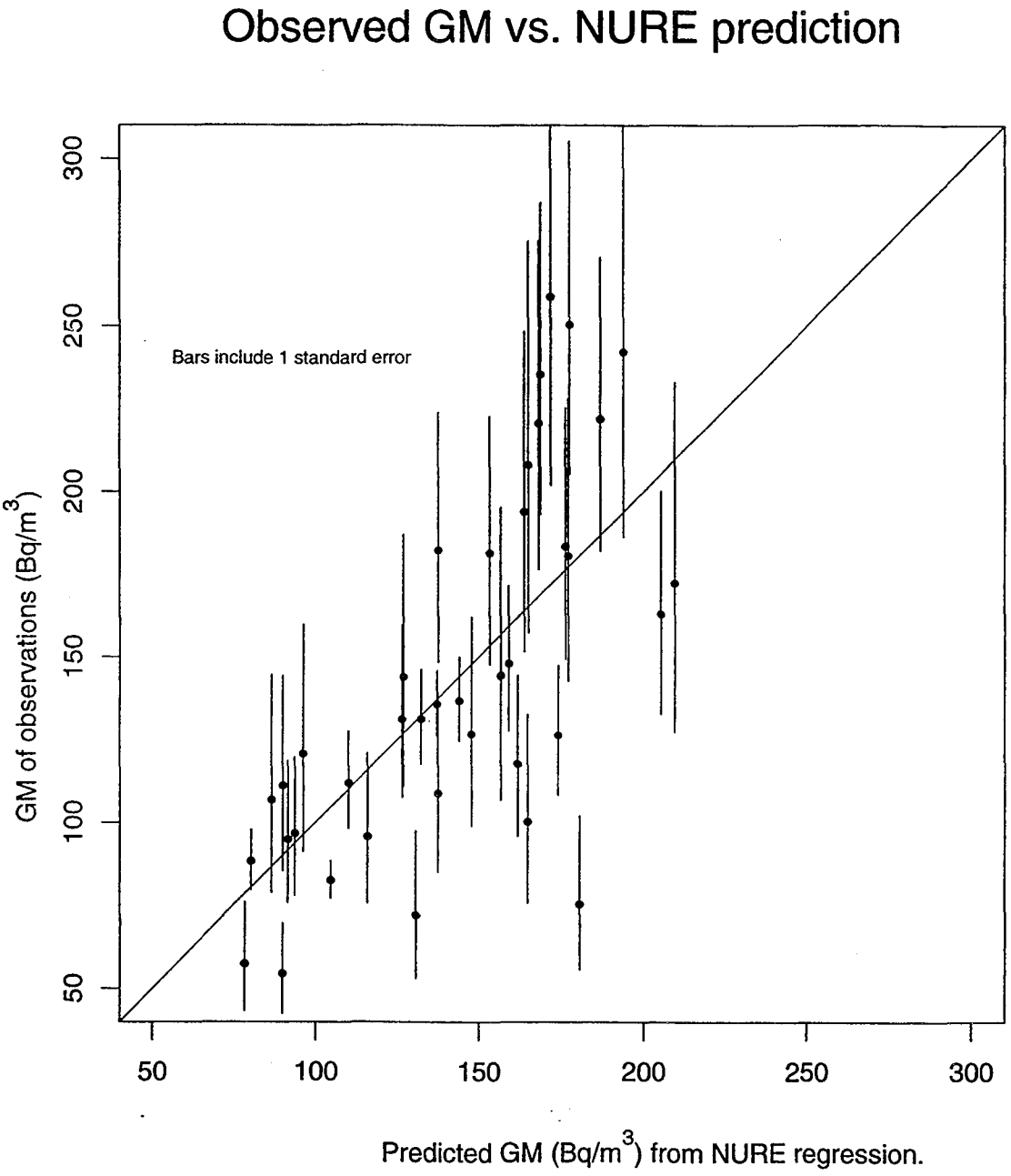


Figure 3: Results of a conventional regression of log(GM) on log(NURE), for counties with more than 5 observations. Note that more than 60% of the error bars cross the line representing perfect agreement, and that predictions for well-sampled counties (those with small error bars) tend to fall close to the line representing perfect agreement.

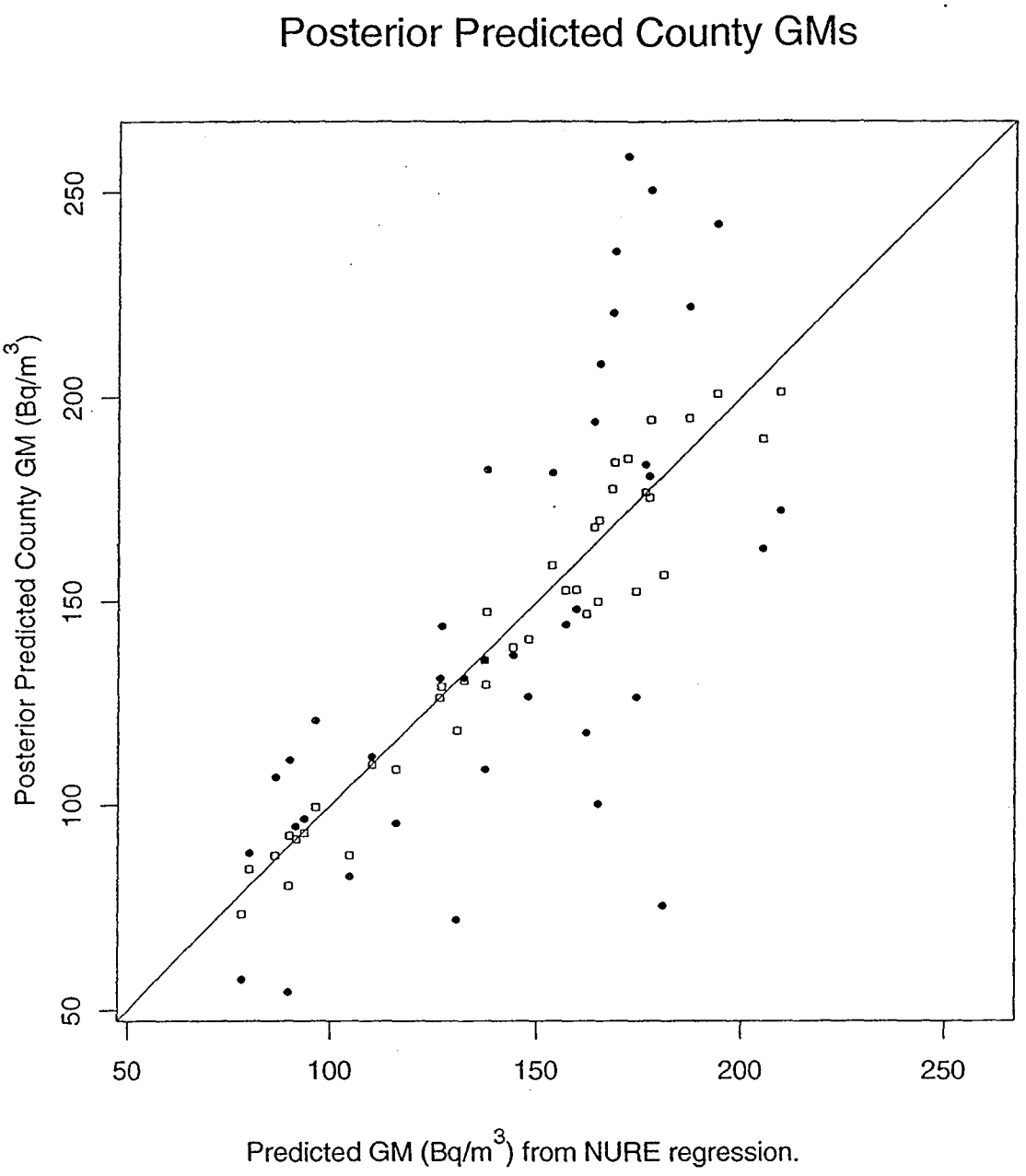


Figure 4: Results of a mixed-effects regression of logarithm of individual-house radon concentration on logarithm of county NURE and county random effects. Only counties with more than 5 observations are shown, although all counties were used in the analysis. Prediction from a conventional NURE regression is on the x-axis. Posterior predicted GMs are indicated by squares, while the GM of observations for each county is plotted as a point. Note that the posterior prediction for each county lies between the observed GM (point) and the conventional NURE prediction (45-degree line). For highly sampled counties, the posterior prediction always lies near the observed value, while for poorly sampled counties the posterior prediction can be very different from the observed value. Posterior predictions are subject to slight variation due to the finite number of simulation draws.

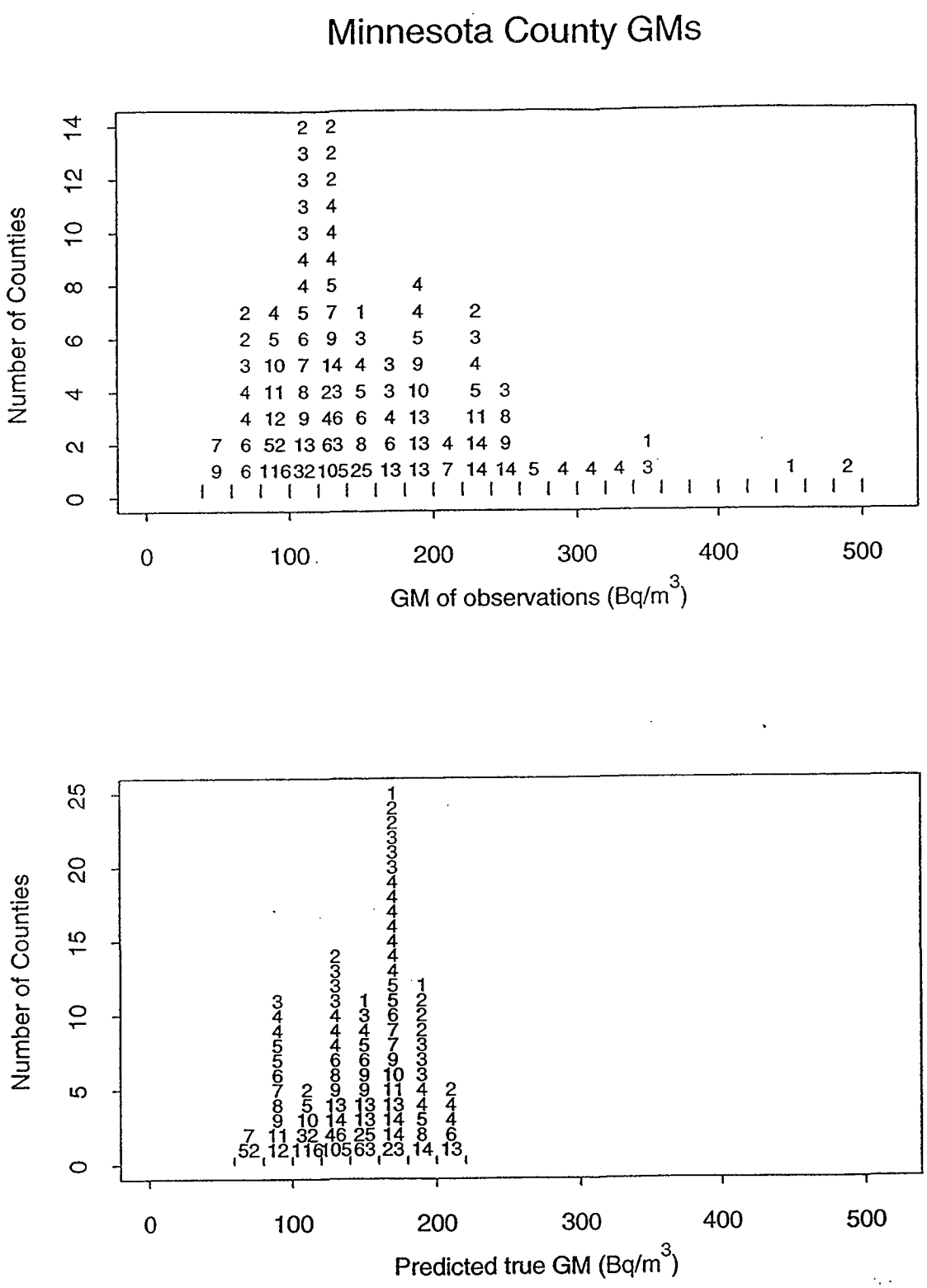


Figure 5: Stacked-number histograms showing the distribution of observed county GMs (top) and the distribution of posterior predicted GMs (bottom). Each county is represented by digit(s) indicating the number of observations in the county. For example, in the upper figure the value "105" in the 120-140 Bq/m^3 interval represents a county with an observed GM between 120 and 140 Bq/m^3, based on 105 observations. Note that the distribution of predicted true GMs is much tighter than the distribution of observed GMs. However, recall that the predicted GMs are themselves uncertain (by about 15 to 30 Bq/m^3 ---about one bin---for most counties).

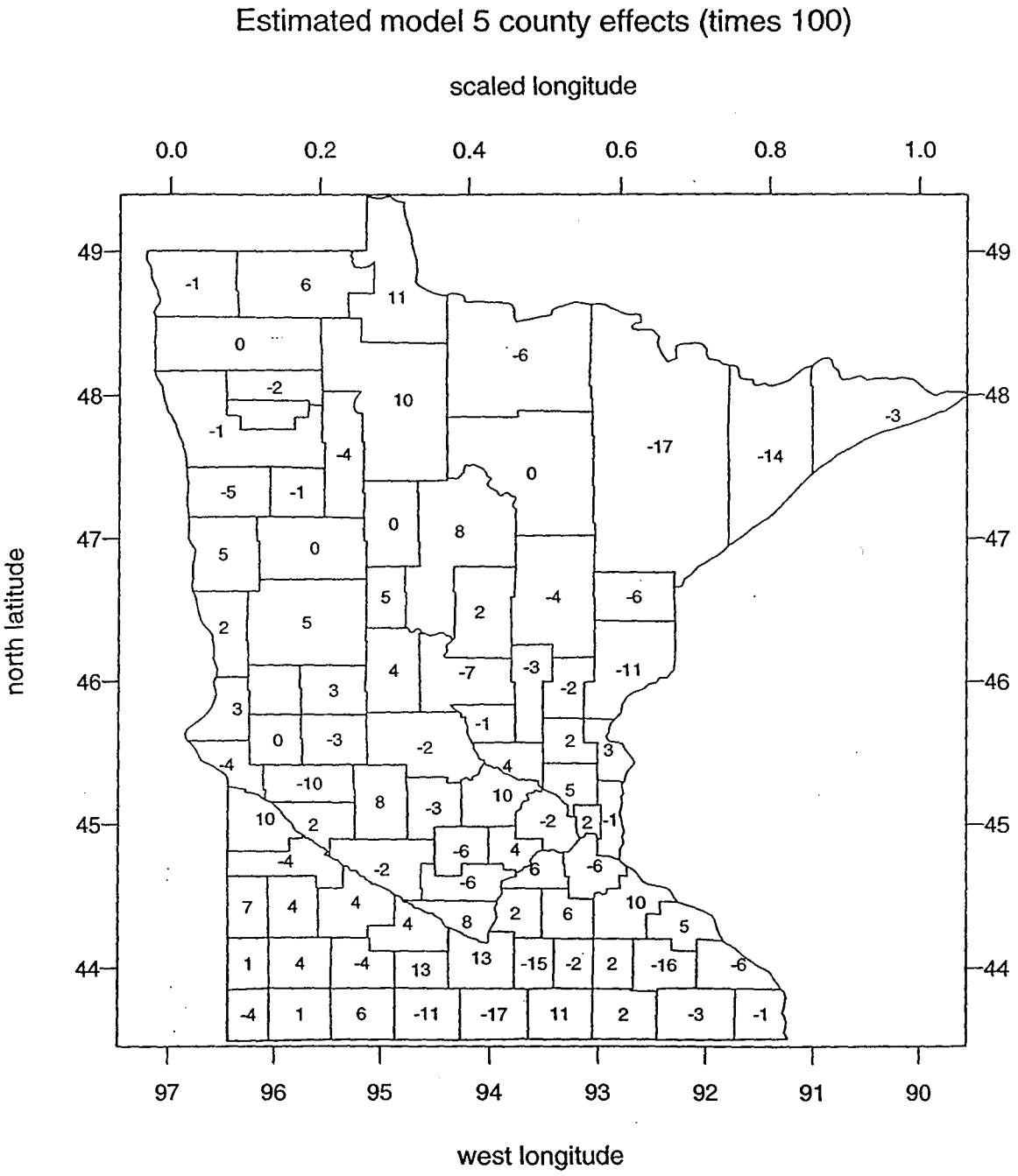


Figure 6: Map showing estimated county effects from model 5, multiplied by 100. Note the spatial grouping of negative county effects in the northeastern portion of the state, and the sparseness of negative county effects in the western half of the state.