Pergamon

Environment International, Vol. 22, Suppl. 1, pp. S699–S714, 1996
1996 Elsevier Science Ltd
Printed in the USA. All rights reserved
0160-4120/96 $15.00+.00

PII S0160-4120(96)00173-0

# JOINT ANALYSIS OF LONG- AND SHORT-TERM RADON MONITORING DATA FROM THE NORTHERN U.S.

P.N. Price and A.V. Nero
Indoor Environment Program, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

*EI 9510-355 M (Received 29 October 1995; accepted 12 December 1995)*

This paper analyzes data collected as part of two types of radon surveys of U.S. homes–the National Residential Radon Survey (NRRS) and the U.S. Environmental Protection Agency (EPA)/State Residential Radon Surveys (SRRS)–to determine the distribution of annual-average, living-area radon concentrations for ground-contact homes in the northern U.S. A statistical model is used to link the short-term SRRS measurement in each home with the home's annual-average, living-area radon concentration, although in no case are both a short- and long-term measurement available for the same home. This paper shows that, even though an individual short-term winter measurement from the SRRS is a poor predictor of the home's annual-average, living-area radon concentration, an aggregation of such measurements can be used, after adjusting for bias, to characterize the distribution of annual-average, living-area concentrations as determined by the NRRS. Different types of homes require different adjustment equations. This paper presents the adjustment equations and uses them to estimate parameters describing annual-average, living-area concentration distributions. Model approximations and validation are briefly discussed. The methods presented here could be applied to calibrate other radon data sets. *Copyright ©1996 Elsevier Science Ltd*

## INTRODUCTION

Residential radon measurements are commonly made following a variety of protocols. The most frequently used protocol in the U.S. has been the 'screening' measurement: a short-term (2-7 d) charcoal-canister measurement made on the lowest level of the home. Such measurements are often made by potential home buyers in an attempt to evaluate whether a particular home might have a radon 'problem', and by home-owners desiring a rapid evaluation of radon levels in their home. Winter-season screening measurements were also used in the U.S. Environmental Protection Agency (EPA)/State Residential Radon Surveys (SRRS), which were conducted with guidance from the U.S. EPA in many states of the U.S.

A radon measure that is far less common, but is believed to be much better for evaluating actual radon risk, is a 12-month integrated measurement of the radon concentration, averaged over living areas of the home ('annual-average, living-area radon concentration'). Alpha-track radon monitors, placed on every level of the home that is used as living space, can be used to estimate a living-area average concentration that is not subject to the biases and effects of day-to-day and seasonal variation that affect screening measurements.

In this paper, a description of a joint analysis of two sets of radon data is provided: data from the National Residential Radon Survey (NRRS), and data from the SRRS. Descriptions of these surveys can be found in the

references (Wirth et al. 1992; Alexander et al. 1993; Marcinowski et al. 1994). The present analysis was performed so as to address several issues:

1) To what extent can the SRRS measurements be used to predict the distribution of annual-average, living-area radon concentrations in different counties or regions? The ability to use SRRS measurements to predict distributions of annual-average, living-area radon concentrations within areas would facilitate identification of areas likely to contain large numbers of high radon homes.

2) What conversion procedure should be used to convert from short-term (SRRS) measurements to long-term, living-area average concentrations, such as those determined in the NRRS? Previous work by other researchers (Ronca-Battista and Chiles 1990; White et al. 1990; Klotz et al. 1993) has addressed the issue of the relationship between short- and long-term concentration measurements in homes, but such work has not fully investigated the effects of variables such as the presence or absence of basements, nor has it provided different relationships for different regions.

3) What estimates of parameters of interest–fraction of homes with annual-average, living-area concentrations over 150 Bq $m^{-3}$ (4 pCi/L), geometric mean (GM) radon concentrations by region, by state, and so on–are obtained from the joint analysis, and how much more precise are these values than those obtained from the NRRS alone?

Data from the NRRS are of high quality, in the sense that reported indoor radon concentrations are believed to closely reflect the actual, annual-average radon concentrations for each home. However, the NRRS sampled only 125 counties in the U.S., most of which had a relatively high population. This sparse and geographically uneven sampling is due to the goals of the initial design: the primary goal was to obtain a precise description of the radon exposure distribution of the U.S. population as a whole. Characterization of radon distributions within different regions or smaller areas was accorded a much lower priority.

The SRRS data set contains over 50 000 observations in 41 of the 50 states of the U.S., offering far greater geographic coverage than does the NRRS. In contrast to the NRRS, monitoring data from the SRRS are low quality screening measurements: short-term (a few days), winter charcoal-canister measurements, usually in the basement, if there was one. For any individual home, such a measurement is poorly correlated with the annual-average, living-area radon concentration of the home. Homes with the same annual-average, living-area concentrations frequently have screening measurements that differ by a factor of two or more (White et al. 1990). There are many reasons for this poor correlation, including temporal variation due to weather conditions and other factors, variation between basement radon concentrations and concentrations on higher floors, and the fact that in some homes the basement is a living area of the home (so that basement concentrations contribute directly to living-area averages), while in other homes the basement is not a living area. Although an individual SRRS measurement is a poor indicator of the home's annual-average, living-area radon concentration, this paper demonstrates that the SRRS data can be calibrated in such a way as to determine the statistical distribution of annual-average, living-area radon concentrations.

The relationship between the NRRS and the SRRS radon concentration measurements in the northern U.S., east of the Dakotas, is discussed here. These states are divided into four regions:

1) New England: Connecticut (CT), Massachusetts (MA), Maine (ME), Rhode Island (RI), and Vermont (VT).

2) Mid-Atlantic: Maryland (MD), Pennsylvania (PA), Virginia (VA), and West Virginia (WV).

3) Great Lakes: Illinois (IL), Indiana (IN), Michigan (MI), Minnesota (MN), Ohio (OH), and Wisconsin (WI).

4) Central Plains: Iowa (IA), Kansas (KS), Missouri (MO), and Nebraska (NE).

Although they are in the northern U.S., the states of New York, New Hampshire, New Jersey, and Delaware are excluded, as they did not participate in the SRRS. The states included in the present analysis are shown in Fig. 1, which also shows the counties that were selected for the NRRS. Except for the states that were excluded due to non-participation in the SRRS, these regions correspond with regions defined by the U.S. EPA for use in characterizing radon distributions.

## THE DATA

Nero et al. (1986) noted that the distribution of radon concentration measurements in much of the U.S. is approximately lognormal. For most counties with a large number of observations in the SRRS, the geometric standard deviation (GSD) of the observations in the county falls somewhere between 2.1 and 3.0. Since the within-county variation is so large, approximately 20 to 30 observations are needed in order to characterize the GM (or the GSD) within 20%.

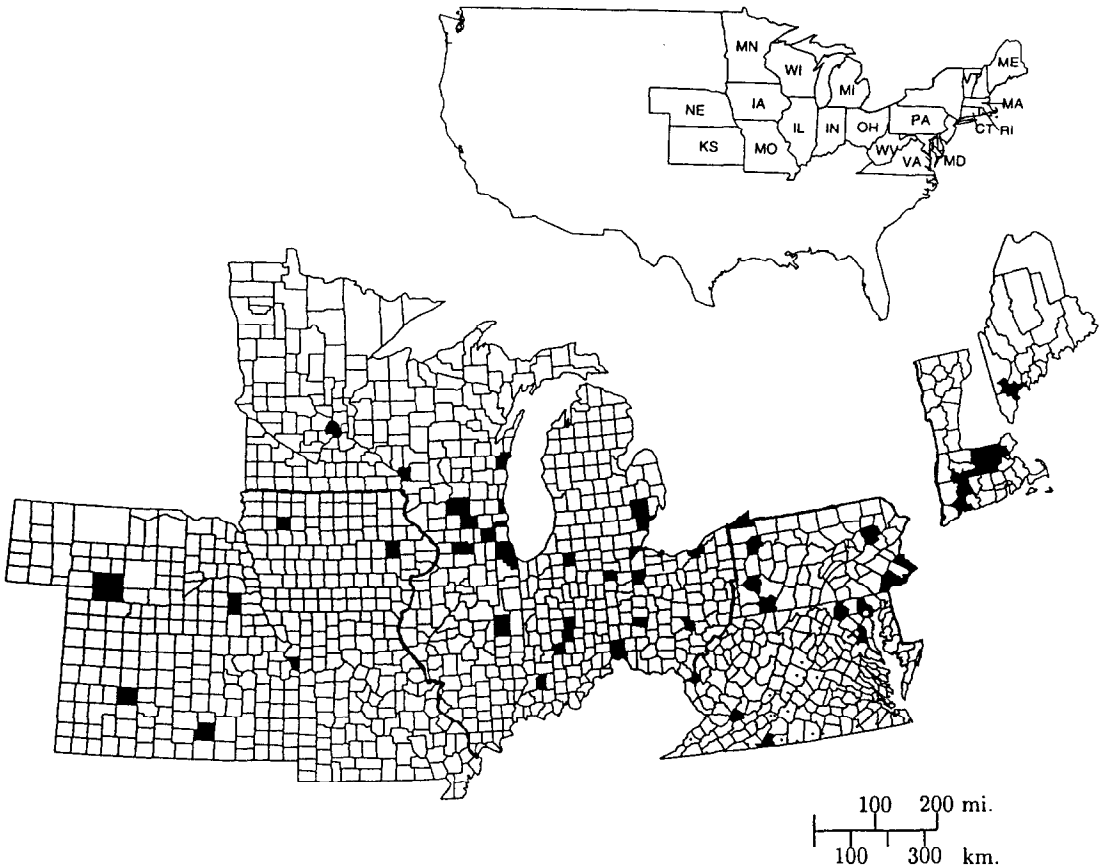


Fig. 1. Maps indicating the states included in the present analysis. The small map shows the states, the larger map shows the counties within those states. The larger map is an equal-area projection. Counties that were sampled in the NRRS are darkened. Almost all of the counties shown were sampled in the SRRS.

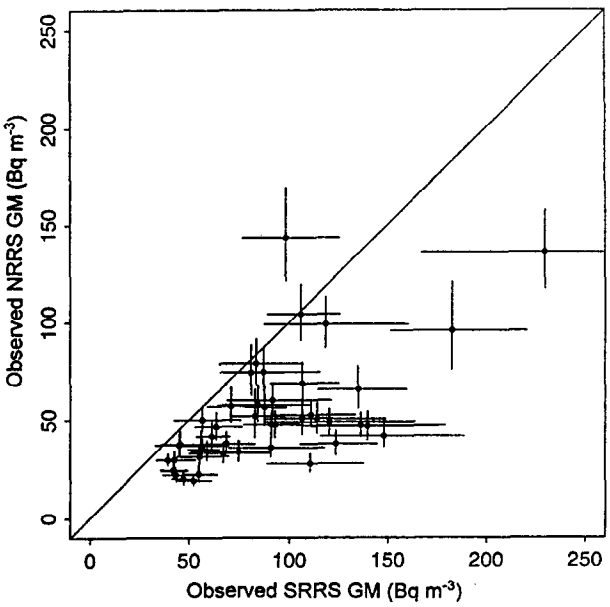


Fig. 2. Plot showing the GM of the annual-average, living-area radon measurements in the NRRS survey versus the GM of the SRRS measurements, by county. Only NRRS counties with more than 15 observations in the SRRS are shown, to avoid cluttering the plot.

Figure 2 shows the observed NRRS and SRRS GM radon concentrations for counties in the northern U.S. that were included in the NRRS. To reduce noise on the plot, only the 47 counties that have more than 15 observations in the SRRS data set are shown. However, all 59 NRRS counties in the regions, and all 1257 SRRS counties in the regions, were included in the full analysis described in the present paper.

The NRRS GM shown is the GM of annual-average, living-area concentration measurements for ground-contact homes. Error bars indicate one standard error in the mean. To calculate the size of the error bars, it is assumed that the observations are independent, identically-distributed selections from a lognormal distribution with a GSD equal to the observed county GSD. Note that in all but one of the counties, the SRRS GM is greater than the NRRS GM–frequently much greater. This plot does not seem encouraging from the perspective of making reliable predictions of annual-average, living-area distributions by county based on SRRS data; indeed, if the four or five counties with the highest NRRS GM were removed from this plot, one

might conclude that there is no significant relationship at all.

However, the superficial impression that there is little relationship between the two data sets is erroneous. As will be shown, after applying correction factors for different types of homes in different regions of the country, the SRRS data from a county can be used to predict the NRRS GM of the county fairly accurately. This becomes possible when one takes into account that the relationship between the screening measurement and the annual-average, living-area measurement depends on several factors. For example, the relationship varies from region- to-region, and depends on the type of house (basement vs. non-basement), and on whether the screening measurement was made in an unfinished basement. As a result, if a large number of SRRS observations are available for a county, they can be used to predict the GM of NRRS observations in the county fairly accurately. This is true in spite of the fact that any individual screening measurement is not very useful for determining the annual-average, living-area concentration in a home. In essence, each short-term measurement can be adjusted for systematic seasonal and housing-type effects to yield a predicted living area concentration that is subject to a great deal of random 'noise' due to temporal variability, detector errors, and so on. In aggregating many short-term predictions, these non-systematic errors tend to cancel out, yielding a valid estimate of the GM of annual-average concentrations.

In the sections that follow, a joint analysis of both data sets, the NRRS and the SRRS, is presented and discussed. The results of the analysis include estimates of parameters describing the distribution of annual-average, living-area radon concentrations in various parts of the country. These estimates apply only to homes included in the SRRS sampling frame: owner-occupied homes in which the lowest floor is in contact with the ground or with a crawlspace that is at ground level. The vast majority of such homes are single-family homes.

## THE STATISTICAL MODEL

Computationally (and perhaps conceptually), the easiest ways to model the relationship between the SRRS and the NRRS data sets would be based on direct comparisons of the county GMs and GSDs. For example, regressions of NRRS GMs on SRRS GMs, perhaps including some aggregated explanatory variables (such as, fraction of homes in the county that have basements, and so on), could be performed.

Unfortunately, simple models created on the aggregate level are inadequate for these purposes, for several reasons. First, the regression coefficients obtained through a conventional linear regression, as suggested above, are influenced by the size of the statistical errors (due to small sample sizes in the present case) on the x-axis, a phenomenon known as the 'regression effect' (Freedman et al. 1973; Price 1995). As a result, extracting the true, underlying relationship would require additional modeling.

Secondly, aggregating makes it difficult to construct a statistical model for the relationship between measurements in different types of homes. For example, consider Chester County, PA. The NRRS included 45 homes from this county, of which 31 had basements. The basement was a living area in 18 of these homes. The SRRS included 34 homes from Chester County, and the SRRS measurement was made in the basement in 29 of them, although in 21 of these homes the basement was not a living area. How should these facts be taken into account in modeling the relationship between the two types of data? Including such information is difficult in an aggregated model, but is relatively straightforward when the statistical model is constructed at the individual-house level, as shall be seen.

Thirdly, analysis at the aggregated level can lead to severe overfitting of the models. For example, the NRRS included only seven counties from the Central Plains states, so including even a few explanatory variables in a linear regression to predict county GMs would be problematic. Attempting to solve this problem by modeling all of the regions together would be unsatisfactory, since one would then be required to assume that the same relationship between the SRRS and the NRRS data prevails over the entire northern U.S.

Finally, fitting of parameters describing aggregated data yields results that are highly sensitive to slight differences between reality and the model being fit. For example, even if there were a good method for fitting county GMs and GSDs, the accuracy of the estimates of the fraction of homes with annual-average, living-area concentrations over 370 Bq $m^{-3}$ (10 pCi/L) would depend strongly on whether the within-county variation really is lognormal. The details of the distribution of actual observations are obviously not preserved when the distribution is summarized by two parameters.

For the reasons listed above, the statistical model is constructed at the individual-house level. Each home is assumed to have a true annual-average, living-area concentration which is determined by a combination of

factors, which are used as explanatory variables in the model: the county the home is in, whether the home has a basement, whether the basement is used as living space, and whether the home is in a single- or multi-family building. In addition, the presence of house-to-house variation that is not explained by any of the included variables is allowed.

In addition to the variables that influence annual-average, living-area concentrations in a home, there are several factors that influence the SRRS screening measurement for the home:

1) The SRRS measurements were made in winter, when indoor radon concentrations tend to be higher than in other seasons.

2) The SRRS measurement is made on a single level of the home, whereas the NRRS measurement is an average of measurements from several levels.

3) The level on which the SRRS measurement is made is often not even a living area of the home. SRRS measurements were frequently made in unfinished basements.

4) The SRRS measurements were very short-term (a few days), and so are subject to substantial variation due to short-term temporal variation in indoor radon levels.

To include (as far as possible) all of the predictive variables available that affect the relationship between the NRRS and the SRRS observations in a county, a statistical model that summarizes this relationship is constructed. The explicit mathematical description of the model is as follows:

1) The actual value $\alpha_i$ of the logarithm of the living-area average concentration in home $i$ in county $j$ is assumed to have been drawn from the following distribution:

$$\alpha_i \sim N(\theta_j + X_i \cdot \beta, \sigma^2) \qquad (1)$$

where,

$N(a,b^2)$ indicates a normal distribution with mean $a$ and variance $b^2$;

$\theta_j$ is the 'county effect' for county $j$;

$X_i$ is a vector of explanatory variables for the home;

$\beta$ is a vector of coefficients corresponding to the explanatory variables in $X_i$; and,

$\sigma$ is the same for all homes.

The variables included in the $X$ matrix were indicator (dummy) variables indicating the following: the state of the U.S. in which the measurement was made; whether the home has a basement that is used as living space, a basement that is not used as living space, or does not have a basement; and whether the home is in a single-family or multi-family building.

2) In the NRRS homes, $\gamma_i$, the natural logarithm of the living-area average concentration measurement, is assumed equal to $\alpha_i$, the logarithm of the actual living-area concentration. This assumption, that there is no measurement error in the NRRS, is an approximation made for convenience of analysis. The NRRS 'mean over living areas' measurement in a home is an average of (usually several) alpha-track radon measurements, each of which is subject to error of about 8 to 20% (Lucas et al. 1992). The mean of several such measurements is, therefore, subject to error on the order of 5 to 15%. In this analysis, the presence of such error is ignored. Strictly speaking, the authors predict the distribution of the NRRS measurements for different regions and types of homes, and, in the text, refer to the NRRS measurements as if they represent the true concentration in the monitored homes.

3) In the SRRS homes $\gamma_i$, the logarithm of the reported radon concentration measurement, is drawn from the following distribution:

$$\gamma_i \sim N(\alpha_i + \Gamma_i \cdot \rho, \delta^2) \qquad (2)$$

where,

$\Gamma_i$ is a vector of explanatory variables;

$\rho$ is a set of coefficients corresponding to the variables in $\Gamma$; and,

$\delta$ is the same for all homes.

The matrix $\Gamma$ includes dummy variables indicating the following: whether the home has no basement, has a living-area basement and was monitored in the basement, has a basement that is not a living area and was monitored in the basement, or has a basement that was monitored on the first floor. Note that $\Gamma$ does not use the same variables as $X$. For example, whether or not the SRRS measurement was made on the first floor of a home, as opposed to the basement, may have only minor utility in predicting the annual-average, living-area concentration in the home, but certainly does influence the SRRS measurement in the home. Thus, the floor on which the SRRS measurement was performed is identified in $\Gamma$, but not in $X$. The dot product $\Gamma_i \cdot \rho$ is a conversion constant (which, when exponentiated, yields a conversion factor) for house $i$. The use of explanatory variables allows different conversion factors to be used in different types of homes (with, and

without, basements, for example). The full prediction of annual-average concentration from the SRRS measurement involves, not only the conversion factor, but also the variance estimates, as discussed below.

4) The county effects, $\{\theta\}$, are assumed to be drawn from a normal distribution, with variance $\tau^2$ which is estimated from the data. Recall that the model is constructed in log space, and that the assumption of normality of the county effects in log space is equivalent to assuming that the county effects are lognormally distributed in untransformed space. The assumption that the county effects are drawn from a common distribution helps to prevent overfitting of the model. For example, consider a county with only a single SRRS observation, and suppose this observation is much higher than is typical for similar types of homes in the rest of the state. In such a case, it would be possible to fit this home's measurement exactly (i.e., to make the predicted value equal to the observed value) by choosing the appropriate county effect. Such a procedure would allow the data to fit well, but would probably not be a very good reflection of reality (nor lead to helpful inference in counties in which fewer data are available), since, in fact, there are several ways the very high measurement could have occurred: the county effect may be high, the home might have a higher indoor radon concentration than typical homes in the county, or the measurement might have been made in a time period when the radon concentration in that home was particularly high. The present model accounts for the relative likelihoods of these different possibilities.

5) Each of the individual coefficients in $\{\beta\}$ and $\{\rho\}$ is assumed to be constant within a region. The sparse spatial coverage of the NRRS does not permit examination of spatial variation of coefficients on the scale of individual states (many states had only one or two counties in the NRRS), so there is little point in attempting to model coefficient variation on a scale smaller than a region.

Implementation of the model requires that the set of county effects $\{\theta\}$, the sets of coefficients $\{\beta\}$ and $\{\rho\}$, and the variance components $\sigma^2$, $\delta^2$, and $\tau^2$ are estimated.

Unfortunately, there is no direct way to obtain estimates for the coefficients $\{\rho\}$ or for the variance $\delta^2$: since there are no SRRS and NRRS measurements in the same homes, there is no way to do a regression of $\alpha_i$ on $\gamma_i$. However, modification of the model to rectify this problem is not as difficult as it might appear. The SRRS measurements are modeled with

$$\gamma_i \sim N(\theta_j + G_i \cdot \beta', \sigma_s^2), \tag{3}$$

where,

$G_i$ = a vector of explanatory variables from the SRRS dataset;

$\beta'$ = a vector of coefficients associated with those variables; and,

$\theta_j$ = same county effect as in Eq. 1.

Variables included in $G_i$ are the following: all of the variables in $\Gamma$, as well as dummy variables indicating the state in which the home is located and whether the home is in a single- or multi-family building. The model then implies that $\sigma_S^2 = \sigma^2 + \delta^2$ and that $\Gamma_i \cdot \rho = G_i \cdot \beta' + X_i \cdot \beta$. Specifically, the NRRS data allow one to estimate the coefficients $\{\beta\}$ that describe the influence of housing type on the annual-average, living-area concentrations, while the SRRS data allow one to estimate the coefficients $\{\beta'\}$ that describe the influence of housing type on the SRRS concentration measurements. Taking the difference $\beta' - \beta$ for the appropriate coefficients allows one to estimate the 'conversion constants' $\{\rho\}$ for the various types of homes.

The result gives two ways to predict the annual-average, living-area radon concentration in each SRRS home: the predicted annual-average, living-area concentration given by Eq. 1, or one obtained by applying the conversion constant to the SRRS measurement. The authors wish to generate a final prediction from the weighted average of these two predictions, with the weightings given by $1/\sigma^2$ and $1/\delta^2$, respectively, so that the predicted logarithm of the annual-average, living-area concentration in home $i$ from county $j$ is:

$$\alpha_i^{pred} = \frac{(\gamma_i - \Gamma_i \cdot \rho)/\delta^2 + (\theta_j X_i \cdot \beta)/\sigma^2}{1/\delta^2 + 1/\sigma^2} \tag{4}$$

Use of this weighted average has the effect of adjusting for the larger variance of the SRRS observations compared to the NRRS observations in the same county, as well as giving a more precise estimate for each home. The variance $v^2$ of true annual-average, living-area concentrations about the predicted values is given by

$$1/v^2 = 1/\sigma^2 + 1/\delta^2. \tag{5}$$

Unfortunately, the true values of the regression coefficients, the conversion constants$\{\rho\}$, the county effects $\{\theta\}$, and the variances $\delta^2$ and $\sigma^2$, are all unknown. However, they can all be estimated from the data through a linear mixed effects regression of the logarithm of the observed SRRS and NRRS radon concentrations on all of the explanatory variables. For an excellent introduction and discussion of mixed effects regressions, see Gelman et al. (1995). For an application to radon monitoring data, see Price et al. (1995). The random effects regression generates many different sets of estimates for all of these parameters, with the variation in each parameter estimate reflecting the uncertainty in the value of the parameter. Generally speaking, the regression coefficients $\{\beta\}$ and $\{\beta'\}$ are fairly well estimated (i.e., have fairly low uncertainties), while the county effects $\{\theta\}$ are highly uncertain due to small sample sizes in most counties. For each set of parameter predictions, a prediction and uncertainty of the annual-average, living-area radon concentration is generated for each SRRS home, which is used to predict distributional parameters such as the GM, GSD, etc. The distribution of predicted values then includes the uncertainty due to the uncertainty in the underlying parameters.

Note that the relationship between $\alpha_i^{pred}$ and $\gamma_i$ in Eq. 4 is linear and has a slope of $1/(1+\delta^2/\sigma^2)$, which is less than unity. Transforming back from log space, this corresponds to a nonlinear relationship between the measured short-term concentration and the annual-average, living-area concentration. Such a nonlinear relationship has been found in previous studies (Ronca-Battista and Chiles 1990; White et al. 1990; Klotz et al. 1993) that investigated the relationship between long- and short-term measurements. For a discussion of one of the reasons behind this nonlinear relationship, see Price (1995).

Parameters that must be estimated for this model include all of the regression coefficients, as well as variance estimates for the house-to-house variation in radon levels, variance estimates for the county-to-county variation in radon levels, predicted county effects, and the measurement errors in the SRRS. It is assumed that the regression coefficients and variances are the same throughout a region.

## RESULTS

Results of the analysis include estimates of the coefficients and variance components in the model, as well as estimates of distributional parameters (GM, GSD, AM, etc.) concerning the distribution of annual-average, living-area concentrations within each state and within each region.

Coefficient estimates for the regression coefficients other than the county effects are presented in Table 1. Recall that the regression was performed in log space, and the coefficients thus represent the effect of each variable on the logarithm of the response. In addition to the coefficients and variances shown in the table, the model produced an estimate (with uncertainty) for every county for which at least one home was sampled in the SRRS. These county effect estimates are not shown, but they are approximately normally distributed (in log space) with the variance $\tau^2$ given in Table 1.

The coefficients in Table 1 show the average effect of various housing characteristics on measured radon concentrations relative to single-family non-basement NRRS homes. For example, the natural logarithm of the SRRS radon measurement in a New England home with a living-area basement is on average 1.14 higher than the logarithm of the NRRS annual-average, living-area measurement in a non-basement home in the same county.

The variance components are interpreted as follows: the SRRS and NRRS variances are estimates of the variation of the observations about the predicted radon levels for homes in the two surveys. The prediction is based on both the housing characteristics and the estimated 'county effect', which allows some counties to have generally higher or lower radon levels than other counties. Note that, in all regions, the SRRS variance is higher than the NRRS variance. This is probably due primarily to the fact that the SRRS observations are subject to temporal variation–any individual home may have been measured over a relatively high- or low-radon period. Recall that the analysis was performed on logarithmically transformed values: for example, the predicted logarithm of the observed SRRS radon concentration typically varies from the observed value by about $\pm\sqrt{0.92} = \pm 0.96$ in New England, corresponding to a multiplicative error of the untransformed values of exp (0.96) = 2.61. So, in New England, given the county in which the observation was made, and information about whether the home has a basement that is a living area, the home's SRRS observed radon concentration can be predicted only to within a factor of about 2.6. Similarly, the home's NRRS mean over living areas can be predicted to within a factor of about 2.3. As noted, in the model, it is assumed that the NRRS observation is made without error, so that it represents the home's true mean over living areas; however, it is assumed that the

Table 1. Coefficient and variance component estimates (in log space) for owner-occupied, ground-contact homes in the northern U.S. The notations 'bmt, liv' and 'bmt, nonliv' indicate the effect associated with basements that are and are not living areas, respectively. The notation 'bmt, meas 1st' indicates the effect associated with performing an SRRS measurement on the first floor, in a home that has a basement. In addition to the coefficients shown below, the model estimates a 'county effect' for each of the 1257 SRRS counties included in the analysis.

| Coefficient | New England | Mid-Atlantic | Great Lakes | Central Plains |
|---|---|---|---|---|
| CT | 2.92 | | | |
| MA | 3.23 | | | |
| ME | 3.03 | | | |
| RI | 3.09 | | | |
| VT | 2.96 | | | |
| MD | | 2.50 | | |
| PA | | 3.59 | | |
| VA | | 3.05 | | |
| WV | | 3.19 | | |
| IL | | | 3.82 | |
| IN | | | 3.92 | |
| MI | | | 3.31 | |
| MN | | | 4.29 | |
| OH | | | 3.92 | |
| WI | | | 3.82 | |
| KS | | | | 4.51 |
| IA | | | | 3.71 |
| MO | | | | 3.27 |
| NE | | | | 3.89 |
| multi-family | -0.03 | -0.34 | -0.33 | -0.22 |
| SRRS | | | | |
| no bmt | 0.77 | 0.47 | 0.15 | 0.39 |
| bmt, liv | 1.14 | 1.32 | 0.74 | 1.02 |
| bmt, nonliv | 1.33 | 1.40 | 0.80 | 1.08 |
| bmt, meas 1st | 0.60 | 0.64 | 0.24 | 0.40 |
| NRRS | | | | |
| no bmt | 0.00 | 0.00 | 0.00 | 0.00 |
| bmt, liv | 0.63 | 0.84 | 0.30 | 0.51 |
| bmt, nonliv | 0.08 | 0.38 | 0.00 | -0.03 |
| variance | | | | |
| SRRS ($\sigma^2_S$) | 0.92 | 1.01 | 0.77 | 0.65 |
| NRRS ($\sigma^2$) | 0.72 | 0.74 | 0.66 | 0.54 |
| btwn cnty ($\tau^2$) | 0.07 | 0.23 | 0.16 | 0.17 |

SRRS observation in a home is subject to both bias (requiring multiplication by a correction factor) and measurement error.

From the different variance estimates, the measurement error in the SRRS can be estimated: the SRRS variance is assumed to be equal to the true (NRRS) variance plus the error variance. This implies, for example, that in the New England region the error variance is about $\delta^2 = 0.92 - 0.72 = 0.20$, so that in log space the additive errors are about $\pm\sqrt{0.2} = \pm 0.45$, corresponding to a multiplicative error of about exp (0.45) = 1.56. If a set of similar homes (e.g., single-family homes without basements) from New England that have the same annual-average, living-area concentration were examined, one would expect the SRRS measurements for such homes to vary by a factor (one standard error) of about 1.56 in either direction.

As discussed above, two types of predictions are combined to obtain a final prediction for each home: the regression prediction of the home's annual-average, living-area concentration, and the prediction provided by the corrected SRRS measurement. This final prediction is still quite uncertain for an individual home. However, as previously noted, in a sufficiently large sample of SRRS homes, the non-systematic errors will tend to cancel out. Thus, a collection of SRRS observations can be used to predict, for example, the GM of annual-average, living-area concentrations for the

Table 2. Estimated factors for converting from NRRS to SRRS GM, for three classes of homes in the northern U.S.: those with no basement, with a living-area basement, and with a basement that is not used as living space.

| Type of home | New England | Mid-Atlantic | Great Lakes | Central Plains |
|---|---|---|---|---|
| no basement | 2.2 | 1.6 | 1.2 | 1.5 |
| living-area bmt | 1.7 | 1.6 | 1.6 | 1.7 |
| non-living-area bmt | 3.4 | 2.8 | 2.2 | 3.1 |

homes, although the value for any individual home will be quite uncertain.

The 'between county' variance component provides an estimate of the variation in county effects within a state. The county effects account for the variation in radon levels that is not explained by differences in housing types in different counties, and that is not due to sampling statistics or variation at the individual-house level. Presumably, most such variation is due to geologic factors. In New England, there is little unexplained between-county variation: in log space, additive county effects are about $\pm\sqrt{0.07} = \pm 0.26$, corresponding to multiplication or division by a factor of 1.3. In other regions, there is more between-county variability that is not accounted for by housing factors or sampling noise. One reason for the low between-county variability in New England is probably the small spatial size of the New England states, which one would naturally expect to be associated with less variation between counties than would occur in larger states.

It might be recalled that the model assumed that the same regression coefficients and variance coefficients are applicable throughout a region, an approximation that is surely something of an oversimplification. Fortunately, mild spatial variation in the appropriate regression factor does not seriously affect the validity of the estimated distributional parameters for the region as a whole. In essence, the regression coefficients are those that are applicable for the 'average' home in the region. A particular regression coefficient will be higher than the true value in some of the states in a region, and lower in others, but for a random selection of homes drawn from the region as a whole, the coefficients (and the variation of true values about the predicted values) will be correctly determined by the model. However, applying the model to a non-random subset of homes from the region can lead to problems. For instance, it is possible that the estimates for basement homes in the state of MD will tend to be too high, while estimates for basement homes in PA may be too low, and so on. In practice, such errors are found to be fairly small, as discussed later.

The magnitude of the difference between regression coefficients for similar homes in the NRRS and SRRS can be exponentiated to determine 'conversion factors' for removing bias in the SRRS observations to yield unbiased predictions of NRRS annual-average, living-area concentrations. For any single home, the prediction from such a procedure is extremely uncertain: the estimate of $\exp(\delta)$ is 1.39 for the Great Lakes and Central Plains regions, 1.56 for New England, and 1.68 for the Mid-Atlantic. However, given a random selection of homes for the region, the over- and under-estimates for different homes tend to cancel out, so that the GM of the predicted annual-average, living-area concentrations approaches the true GM as the number of homes increases. In Table 2, conversion factors for various types of homes are presented, and for different regions of the northeastern U.S. These represent the factor by which the NNRS GM must be multiplied to predict the SSRS GM—of course division goes to either direction. The standard errors in the conversion factors are about 4% to 8% for homes with basements, which are common in the northeastern U.S., and are, thus, well represented in the database, and about 13% for homes without basements. However, it should be noted that the standard error estimates alone do not reflect the full uncertainty in these parameters, since they are conditional on the model. If the statistical model that has been applied were exactly correct, the uncertainties would be those given by the standard error estimates, but there is additional uncertainty due to differences between the model and reality.

The between-region variation in regression coefficients gives a rough impression of the within-region variation. For example, if adjacent regions have very different estimates, then what about the appropriate conversion factor for border states? In the present case, the estimates for most of the adjacent regions are not extremely different, although there is a fairly substantial difference for homes with basements that are not living

areas in the Great Lakes region compared to the Central Plains region. For that particular type of home, it is not exactly clear what the conversion factor should be for homes in, for example, eastern IA or western IL.

However, given the uncertainties in the conversion factors, the agreement in the estimates for the different regions is surprisingly good overall. Indeed, the remarkable agreement of conversion factors for living-area basements is surely partly coincidental, since one expects more variation on statistical grounds alone, even if the true conversion factors were identical. The estimated conversion factors for Great Lakes homes are lower for both non-basement and living-area-basement homes than in the adjacent Mid-Atlantic and Central Plains regions, but not by a large margin compared to the standard errors in the estimates. Although there is evidence of moderate between-region variation, the variation is not so large as to discredit the approximation that the factors are constant within a region.

To estimate distributional parameters for the annual-average, living-area concentrations in the various regions, the following method is used:

1) Use the mixed effects regression method (Gelman et al. 1995; Price et al. 1995) to generate distributions of likely values for the variance components, regression coefficients, and county effects;
2) Sample from the distribution determined in step 1 to obtain a particular 'possible' set of variances, regression coefficients, and county effect estimates;
3) Use the results from step 2 in Eq. 4 to generate a predicted annual-average, living-area concentration for every SRRS home;
4) Sample from the predictions in step 3, with variance $v^2$ determined from Eq. 5, to generate a predicted distribution of annual-average, living-area concentrations for the SRRS homes;
5) Calculate the GM, GSD, arithmetic mean (AM), and fraction of homes over 150 Bq m$^{-3}$ and over 370 Bq m$^{-3}$ for the predicted distribution, using the sampling weight reported in the SRRS data set to weight the measurement in each home;
6) Repeat steps 1-5 until an adequate number (i.e., 200) of simulations has been completed to incorporate the uncertainties in the variances, regression coefficients, and county effects;
7) Find the mean and standard deviation of the 200 predictions for each distributional parameter. These represent the prediction and uncertainty in the parameter.

Although estimates of fractions of homes in the extreme tails are somewhat sensitive to the assumptions of the model, the overall distributional parameters are much less sensitive. In Table 3, estimated parameters describing the distribution of annual-average, living-area concentrations in various regions are presented. These parameter estimates apply only to homes included in the SRRS data frame: owner-occupied homes in contact with the ground. In addition to the parameter estimates themselves, the analysis generates distributions of likely parameter values. To summarize the uncertainties in the parameter values, standard errors in the estimates under the approximation that the possible values are normally distributed were determined. The standard errors determined from the model are about ± 2 Bq m$^{-3}$ for the GM and the AM, and about ± 0.1 for the GSD. The standard error in the fraction of homes with annual-average, living-area concentrations over 150 Bq m$^{-3}$ is about 1 percentage point, and for homes over 370 Bq m$^{-3}$, it is about 0.2 percentage points.

The main source of uncertainty in the fraction of homes with annual-average, living-area concentrations over 150 Bq m$^{-3}$ and over 370 Bq m$^{-3}$ from the NRRS data alone is the 'noise' due to the relatively small number of counties and the relatively small number of high-radon homes included in the NRRS. In contrast, the main source of statistical uncertainty in the present analysis is the uncertainty in the values of the regression coefficients and variance components. Even though the SRRS certainly includes plenty of homes with annual-average, living-area concentrations over 370 Bq m$^{-3}$, the uncertainty in the conversion equations and the variation in the short-term measurements themselves are large enough that their homes (or even their proportion of all homes) cannot be identified with certainty. In consequence, the stated uncertainties in Table 3 for the fraction of homes over 150 Bq m$^{-3}$ and over 370 Bq m$^{-3}$ are only moderately smaller than the uncertainties based on the NRRS data alone (Lucas et al. 1992).

Although the present analysis does not provide regional parameter estimates that greatly improve on those from the NRRS analysis alone, it does provide markedly improved estimates for distributions of annual-average, living-area concentrations within individual states. The estimation of statewide distributional parameters is essentially impossible with the NRRS alone, since homes from only a few counties (one to four) were sampled in most states. Table 4 shows parameter estimates for the distribution of annual-average, living-area radon concentrations in individual states, grouped by region. Typical uncertainty estimates for the parameter estimates are shown as well, under the approximation that the distribution of possible values is normal.

Table 3. Parameter estimates for distributions of annual-average, living-area radon concentrations in owner-occupied, ground-contact homes in various regions. Uncertainty estimates (one standard error) are about ± 2 Bq m$^{-3}$ for the GM and AM, ± 1 percentage point in the fraction of homes over 150 Bq m$^{-3}$, and ± 0.2 percentage points in the fraction of homes over 370 Bq m$^{-3}$. True uncertainties are larger, as discussed in the 'model violations' section.

| | GM (Bq m$^{-3}$) | GSD | AM (Bq m$^{-3}$) | % >150 Bq m$^{-3}$ | % > 370 Bq m$^{-3}$ |
|---|---|---|---|---|---|
| New England | 26 | 2.5 | 37 | 3 | 0.3 |
| Mid-Atlantic | 37 | 3.1 | 74 | 12 | 2.9 |
| Great Lakes | 44 | 2.6 | 70 | 10 | 1.3 |
| Central | 48 | 2.7 | 78 | 13 | 1.5 |

Table 4. Parameter estimates for distributions of annual-average, living-area radon concentrations in owner-occupied, ground-contact homes, by state.

| | State | GM Bq m$^{-3}$ | GSD | AM Bq m$^{-3}$ | %>150 Bq m$^{-3}$ | % >370 Bq m$^{-3}$ |
|---|---|---|---|---|---|---|
| CT | Connecticut | 21 | 2.51 | 32 | 1.8 | 0.1 |
| MA | Massachussets | 30 | 2.52 | 48 | 5.0 | 0.5 |
| ME | Maine | 27 | 2.48 | 41 | 4.0 | 0.4 |
| RI | Rhode Island | 25 | 2.34 | 37 | 2.6 | 0.3 |
| VT | Vermont | 21 | 2.54 | 33 | 2.2 | 0.2 |
| | standard error | 2 | 0.09 | 3 | 0.7 | 0.2 |
| MD | Maryland | 23 | 3.20 | 46 | 5.8 | 0.9 |
| PA | Pennsylvania | 55 | 2.92 | 103 | 17.9 | 4.9 |
| VA | Virginia | 24 | 2.76 | 41 | 3.6 | 0.5 |
| WV | West Virginia | 31 | 2.40 | 46 | 3.9 | 0.5 |
| | standard error | 2 | 0.08 | 3 | 0.8 | 0.3 |
| IL | Illinois | 36 | 2.80 | 58 | 7.4 | 0.8 |
| IN | Indiana | 53 | 2.44 | 80 | 12.9 | 1.6 |
| MI | Michigan | 29 | 2.24 | 42 | 3.3 | 0.3 |
| MN | Minnesota | 74 | 2.14 | 98 | 17.6 | 1.6 |
| OH | Ohio | 48 | 2.71 | 82 | 13.6 | 2.5 |
| WI | Wisconsin | 47 | 2.39 | 70 | 9.4 | 0.8 |
| | standard error | 3 | 0.04 | 3 | 0.8 | 0.2 |
| IA | Iowa | 95 | 2.36 | 135 | 31.0 | 4.5 |
| KS | Kansas | 37 | 2.36 | 54 | 6.0 | 0.4 |
| MO | Missouri | 30 | 2.28 | 49 | 3.5 | 0.3 |
| NE | Nebraska | 74 | 2.24 | 100 | 18.9 | 1.4 |
| | standard error | 4 | 0.07 | 4 | 0.6 | 0.4 |

In a few of the states with particularly variable or elevated radon concentrations, uncertainties differ substantially from uncertainties in other states in the region. The parameters for which the uncertainty is markedly different from the uncertainty given in Table 4 are as follows. In PA, the AM has a standard error of 0.18 Bq m$^{-3}$ and the fraction of homes over 150 Bq m$^{-3}$ has a standard error of 1.3 percentage points. In IA, the AM has a standard error of 0.27 Bq m$^{-3}$, the GM has a standard error of 0.19 Bq m$^{-3}$, and the fraction of homes with annual-average, living-area concentration over 150 Bq m$^{-3}$ has a standard error of 4.0 percentage points. In NE, the AM has a standard error of 0.17 Bq m$^{-3}$, and the fraction of homes with annual-average, living-area concentration over 150 Bq m$^{-3}$ has a standard error of 3.0 percentage points.

The stated uncertainties for the parameters in Tables 3 and 4 and in the list above are conditional on the statistical model, and do not include the possibility of additional error due to model violations. A discussion

of such model violations and their importance is included next.

Figure 3 shows the same subset of counties as those shown in Fig. 2: counties from the NRRS with more than 15 observations in the NRRS set. For each type of home, the appropriate conversion factor is applied to the SRRS observations; results are then grouped by county to obtain a predicted NRRS GM for each county. Again, error bars represent one standard error. On the x-axis, error bars include error due to both sampling variation and due to uncertainty in conversion factors. To calculate the error bars for the NRRS, it was assumed that the NRRS homes in a county were independently drawn from all eligible homes in the county. In fact, a stratification scheme was used to select homes for the NRRS, so that spatial clusters of homes were selected within each county. In some counties, such clustering leads to uncertainty in the true county GM beyond the uncertainty shown in the plot.

Note that the predicted NRRS GMs fall far closer to the 45-degree line (which indicates perfect agreement between the predicted and observed values) than did the original data shown in Fig. 1, a strong indication that the statistical model is behaving well. The predicted NRRS GMs shown in Fig. 3 were not generated from a fit of SRRS GM to NRRS GM by county, but rather were generated by fitting the statistical model at the individual-house level and aggregating the results by county.

Figure 4 shows a further reduced set of counties: those with more than 35 observations in each data set. This set contains 21 counties, from 12 states. Most of the counties fall close to the line indicating agreement between prediction and NRRS observed GM, the sole exception being Frederick County, MD, which will be discussed later. The good agreement for the vast majority of counties suggests that within-region variation of appropriate conversion factors is not likely to introduce large errors in the predicted GM based on SRRS data, even if a spatial subset of observations is selected. If such variation were large, individual county predictions would not be this good.

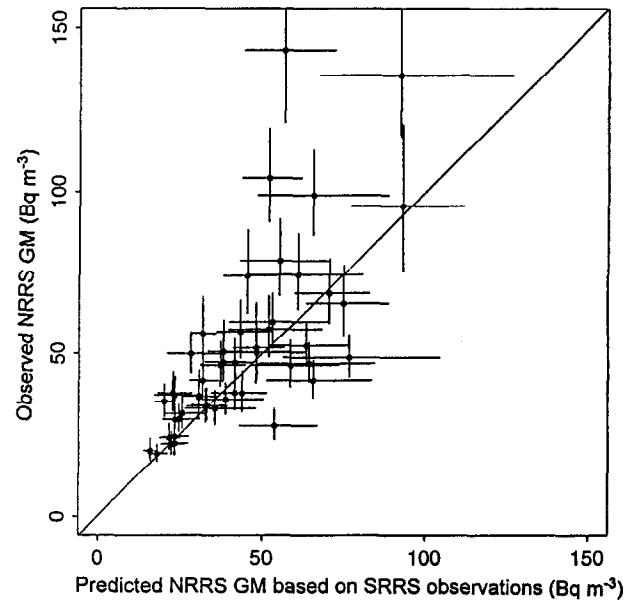


Fig. 3. Plot showing the GM of the annual-average, living-area measurements in the NRRS survey vs. the prediction based on the SRRS measurements after applying the suitable conversion factors, as discussed in the text. Only NRRS counties with more than 15 observations in the SRRS are shown.

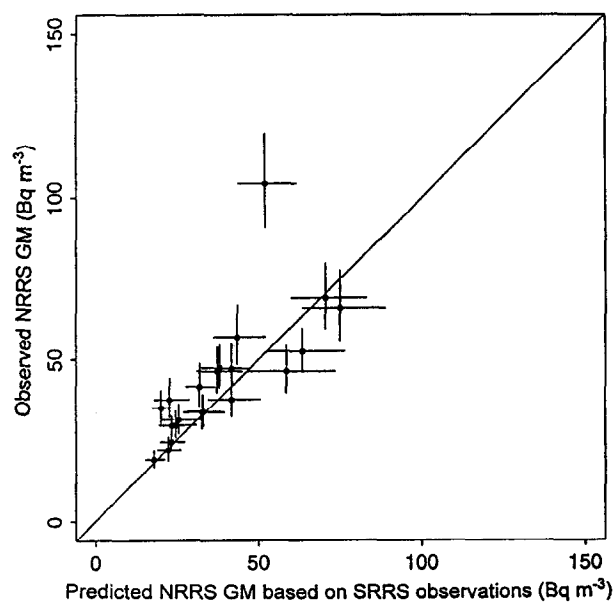


Fig. 4. Same as Fig. 3, but only counties with more than 35 observations in each survey are shown.

## UNCERTAINTIES AND MODEL VIOLATIONS

Two types of problems arise in estimating parameters as discussed in the present work. First, there are problems related to the construction and appropriateness of the statistical model itself–no statistical model is ever perfect. Second, there is the statistical noise due to small sample sizes. The effect of statistical noise on the estimates is already included in the error estimates presented above, but it is harder to estimate the uncertainties due to model violations. A few general comments are:

1) The overall distributional parameters (GM, GSD, and AM) and their uncertainties are the most reliable. They are relatively insensitive to the type of model violations

(such as within-region variation in the regression coefficients) that are present in this analysis.

2) Estimates pertaining to the tails of the distribution are more sensitive to model violations than are estimates of distributional parameters, and the effect becomes larger for higher radon concentrations. Estimates of uncertainty in the fraction of homes over 150 Bq $m^{-3}$ (4 pCi/L) are likely to be fairly good, while uncertainty estimates for the more extreme tail are likely to underestimate the true uncertainties.

3) Attempts to use the regression coefficients or conversion factors to draw conclusions about spatial subsets of houses (a single state, for example) are subject to considerably larger errors than application to a selection of homes drawn from the region as a whole, although the results may be satisfactory in many cases.

The model used in this analysis, although reasonable, has some obvious shortcomings. The most questionable assumption is that the coefficients for various explanatory variables are constant within a region. Does the presence of a basement have the same influence on radon levels in southern IL that it does in northern MN? The answer is, of course, 'no'. The coefficients, and the appropriate conversion factors for various house types, surely cannot be exactly the same everywhere within a region. At best, the approximation that the coefficients are constant within a region is good enough that the substantive estimates produced by the model are fairly accurate.

One signature of large within-region variation in regression coefficients and conversion factors would be substantial scatter on Figs. 3 and 4. And indeed, there are two (though only two) obvious outliers on the plots–outliers in the sense of being more than one or two standard errors away from the line representing agreement between prediction and observations. The most extreme, with a predicted NRRS GM of about 56 Bq $m^{-3}$ in contrast with the observed NRRS GM of 140 Bq $m^{-3}$, is Butler County, OH. This county is clearly anomalous. Indeed, it is the only county on Fig. 2 in which the GM of the NRRS observations is greater than the GM of the SRRS observations in the county. None of the seven other OH counties in the NRRS exhibits such behavior.

The other outlier is Frederick County, MD. This county stands out more in Fig. 4, which shows only counties with more than 35 observations in each data set (the ones that are subject to relatively little noise due to small sample sizes). Again, the observed NRRS GM is substantially higher than would be predicted from its SRRS observations, although the other NRRS counties in the state of MD do not exhibit such behavior.

In the case of both Butler and Frederick Counties, the stratification scheme used to select homes for the NRRS may be important. In the NRRS, census tracts were selected from within each county, and 'secondary sampling units' were selected within each census tract. Homes in the selected unit were then sampled. In most counties, the variation between secondary sampling units is not very large, and the within-county variation is dominated by house-to-house variation. In such cases, the uncertainty in the county NRRS GM can be determined by assuming that the NRRS homes were independent selections from all eligible homes in the county. In a few cases, however, variation between secondary sampling units is substantial. Two such cases are Butler County, OH, and Frederick County, MD. In both of these counties, a single one of the selected secondary sampling units has much higher radon levels than other units in the county, suggesting the possibility that the observed NRRS GM may be higher than the true GM of the county due to the chance inclusion of a disproportionate fraction of high-radon homes. For counties with considerable variation between sampling units, the approximation that the NRRS homes are independent selections from all eligible homes in the county is not a good one, and significantly underestimates the true uncertainty in the county's GM. A 'bootstrap' method (Efron and Gong 1983) was used to estimate the uncertainty in the county GM for both Butler and Frederick Counties, and find that the true standard errors for both counties are approximately double the uncertainties shown in the figures. Thus, the anomalous SRRS/NRRS relationship for these counties does not necessarily indicate a serious problem with the conversion factors or the methodology used in the present work.

However, the fact that the model behaves fairly well should not be interpreted to mean that the model is perfect. There is some within-region variation of regression coefficients (and conversion factors) that is not accounted for in the model, so use of the conversion factors for a spatial subset of SRRS counties will lead to errors larger than implied from the standard errors of the conversion factors alone. The size of the errors can be estimated by assuming that the difference between predicted and observed NRRS GM has two components: a noise component due to finite sample sizes in both surveys, and a component due to true variation between prediction and reality, even if one had large sample sizes. The result of an analysis of variance indicates that about 70% of the variance between the

observed and predicted NRRS GM (Figs. 2 and 3) is attributable to small-sample noise in the NRRS and the SRRS. The remaining variation is due to the discrepancy between the statistical model and reality: even if we had many SRRS observations in a county, it would be impossible to predict the NRRS GM perfectly using the conversion factors determined here. The analysis of variance indicates that using the conversion factor to estimate the NRRS GM for a county from a large number of SRRS observations in the county will tend to yield an estimate that is subject to a standard error of about 15%.

In addition to errors that affect the prediction of the GM of annual-average, living-area concentrations from a set of short-term measurements (such as, violation of the assumption of constancy of the regression coefficients throughout a region), there is the potential for errors affecting other parameters of interest. Several types of model violation would lead to incorrect inference about the tails of the distribution such as the fraction of homes with annual-average, living-area concentrations over 150 Bq m$^{-3}$. One way to check for such violations is to use the model to simulate a draw of the SRRS data, and compare the results to the actual data observed. This was done by following steps 1-4 in the previous section to obtain a predicted distribution of annual-average, living-area concentrations for the SRRS homes, then simulating the SRRS measurement process by adding the appropriate bias ($\Gamma_i \cdot \rho$) and random noise [$N(0,\sigma^2_S)$] to the prediction for each home to generate a predicted distribution of SRRS measurements. This procedure was followed 200 times, to take into account the uncertainties in the variance, coefficients, and county effects. By comparing the distribution of 'observations' in the simulations with the distribution of actual observations, we can search for model violations. For example, if the simulated observations were much more (or less) widely spread than the actual observations, that could indicate a problem with the assumption that true concentrations are lognormally distributed about their predicted values.

In Table 5, the results of one type of comparison between the predicted tail of SRRS observations and the actual tail of SRRS observations are shown for two of the regions: New England (where the tail of the distribution of SRRS measurements was fit very well) and the Mid-Atlantic (where the fit to the high tail was relatively poor). The 'predicted' values shown in the table indicate the GM predicted value for the 50th-highest, 250th-highest, etc. observed SRRS radon concentrations, based on the 200 simulation draws from the individual-house predictions. For a discussion of this type of posterior predictive check, see Gelman et al. (1995).

As shown in the table, the statistical model for New England predicts that the 50th-highest SRRS measurement (from the 5035 SRRS homes in the region) should be around 792 Bq m$^{-3}$, in good agreement with the actual 50th-highest measurement of 803 Bq m$^{-3}$. The p-value of 0.40 indicates that in 40% of the simulation draws, the predicted 50th-highest value exceeded the 50th-highest measured value, while in the other 60%, the predicted value was lower than the measured value. P-values close to 1 or 0 indicate potential model violations, since they indicate cases in which the observed measurements are unlikely given the model.

As the table shows, the predicted distribution of SRRS measurements in New England is in excellent agreement with the observed distribution of SRRS measurements, at least through the prediction for the 1250th-highest measurement. However, the predicted 2000th-highest measurement exceeded the actual 2000th-highest measurement in 94% of the simulations (the p-value is 0.94). In other words, the model consistently predicts that the 2000th-highest observation should be higher than the observed 2000th-highest concentration actually was. Note, however, that both the absolute and relative difference between the prediction and the actual observation are very small. The observed 2000th-highest measurement was 85 Bq m$^{-3}$, while the predicted 2000th-highest measurement was 89 Bq m$^{-3}$ (with an uncertainty of a few Bq m$^{-3}$). This is a good example of a discrepancy that is statistically significant but not practically significant–the model may systematically overpredict the measured concentration for homes near the middle of the distribution of radon measurements, but only by about 4 Bq m$^{-3}$. In fact, the magnitude of the overestimate is comparable to the error introduced in the reporting of the radon measurements in the first place, which were rounded to the nearest 0.1 pCi/L (3.7 Bq m$^{-3}$), so even the statistical significance of this small discrepancy is in question.

The other columns in the table show similar statistics for the Mid-Atlantic states. The fit to the extreme high tail in the Mid-Atlantic states is not as good as the fit in the other regions, in which the agreement was similar to that for the New England region. The predicted 50th-highest measurement in the Mid-Atlantic is almost always substantially lower than is the observed 50th-highest measurement, with a typical discrepancy of about 396 Bq m$^{-3}$ (about 20%). Thus, the model does not fit the highest 1% of the measurements in the 5677

Table 5. Measures of fit between the predicted and observed distributions of SRRS measurements.

| | New England | | | Mid-Atlantic | | |
|---|---|---|---|---|---|---|
| | | predicted meas. | actual meas. | | predicted meas. | actual meas. |
| position | p-value | (Bq m$^{-3}$) | (Bq m$^{-3}$) | p-value | (Bq m$^{-3}$) | (Bq m$^{-3}$) |
| 50 | 0.40 | 792 | 803 | 0.01 | 1628 | 2024 |
| 250 | 0.25 | 340 | 352 | 0.06 | 644 | 703 |
| 500 | 0.22 | 255 | 259 | 0.39 | 396 | 418 |
| 1250 | 0.78 | 137 | 133 | 0.91 | 181 | 185 |
| 2000 | 0.94 | 89 | 85 | 1.00 | 109 | 107 |

monitored homes in the region very well. The situation for the highest 5% of measurements is considerably better: the observed 250th-highest measurement is still unlikely under the model, but the typical under-prediction is only around 59 Bq m$^{-3}$, or 8%. As was the case in the New England region, there are discrepancies at lower measured values that are statistically significant but not practically significant. For example, in 200 simulated samples the 2000th-highest predicted measurement was never as low as the observed value of 107 Bq m$^{-3}$; however, the predicted measurement was always within 6 Bq m$^{-3}$ of the observed value and the typical discrepancy was only 2 Bq m$^{-3}$, less than the roundoff error in the reported measurements.

Unfortunately, there is no easy way to distinguish between two contradictory causes for the problems with the high-measurement tail in the Mid-Atlantic states: it is possible that the SRRS measurements are subject to variation about the annual-average, living-area concentration that is heavier-tailed than the lognormal, after adjusting by a correction factor. It is also possible that the distribution of actual annual-average, living-area concentrations is heavier-tailed than lognormal. In the former case, the slight excess of very high observations would be due to the measurement procedure rather than to the presence of extra high-radon homes, while in the latter case, the extra weight in the high tail would be due to an excess of high-radon homes over the number predicted from the lognormal model. The latter case may be the more likely of the two, since radon distributions with heavier than lognormal tails have been noted previously (Nero et al. 1986; Janssen and Stebbings 1992; Hobbs and Maeda 1996), although that is by no means certain. There is no obvious way to modify the model so that the tail is included correctly: is the problem caused by assumption 1 of the model (see Section: The Statistical Model), by assumption 4, or by something else? Resolving this issue would require collecting long-term, living-area measurements and housing types (as was done in the NRRS) for many more homes in many counties, and investigating the within- and between-county variation in detail. Such an investigation is not possible with the present data.

In any event, the discrepancy is apparent only in the extreme high tail (the highest 1 to 5% of homes in the region), and does not seem likely to cause problems in the estimates of fraction of homes with true concentrations over 150 Bq m$^{-3}$. Use of the current model to predict the fraction of homes with true concentrations over 370 Bq m$^{-3}$ is subject to much more uncertainty—many of the homes with SRRS observed concentrations in the highest few percent of homes have predicted true concentrations that are near this threshold, and this validation check indicates that (at least in the Mid-Atlantic region) there is a problem with the fit for these homes. Depending on the cause of the slight excess weight in the high tail of actual observations, estimates of the fraction of homes with true concentrations over 370 Bq m$^{-3}$ could be either systematic slight overestimates or slight underestimates. The noisy SRRS data are simply too crude a tool to allow accurate characterization of the high tail of annual-average, living-area concentrations. The standard errors of the model imply that the estimates given in Table 3 for fraction of homes over 370 Bq m$^{-3}$ should be accurate to within a fraction of a percentage point. The possibility of model violations affecting the high tail suggests that these uncertainty estimates should be increased substantially, perhaps by a factor of two or so. Exact quantification is difficult, since there is no way to use currently available data to determine what is causing the problem in the extreme high-radon limit.

## CONCLUSIONS

Results of the present analysis are appropriate only for homes in the northern U.S. in which screening measurements were made following the SRRS protocol, i.e., short-term charcoal-canister measurements made in winter on the lowest level of ground-contact homes. Conversion factors for non-winter measurements would certainly be different, since short-term radon concentrations vary seasonally.

Although a single short-term winter screening measurement in a home is a poor predictor of the home's annual-average, living-area radon concentration, the present work shows that a collection of such measurements (as from the SRRS) can be used to characterize the distribution of annual-average, living-area concentrations in an area. Thus, the SRRS, which provides a fairly large database of geographically dispersed, short-term monitoring data, can be used to predict annual-average, living-area radon distributions for entire regions (Table 3), for individual states (Table 4), and for individual counties (Figs. 3 and 4). The ability to use the existing base of short-term data to predict long-term concentration distributions has the potential of substantially improving the efficiency of government programs geared towards locating high-radon homes or high-radon areas of the U.S.

This analysis has determined conversion factors appropriate to different types of homes in different regions. If the parameter of interest is simply the GM of annual-average, living-area radon concentrations, an estimate can be obtained simply by multiplying each SRRS observation by a conversion factor appropriate to the type of home, and aggregating the results by county. If more detailed distributional information is desired, such as the GSD of annual-average, living-area concentrations, the full statistical procedure discussed in the present paper must be used.

In addition to providing novel results concerning the U.S annual-average indoor radon distribution by region, the present analysis illustrates the feasibility of using long-term concentration measurements to 'calibrate' short-term data, even if the long- and short-term measurements are from different homes. The approach could certainly be used on data other than those from the NRRS and the SRRS.

*Acknowledgment*—The authors would like to thank graduate student John Boscardin from the University of California, Berkeley, for the use of his mixed-effects statistical modeling program. Graduate student Lan Zhou performed additional programming. Statistics professor Andrew Gelman of U.C., Berkeley, offered many helpful suggestions, as did Rich Sextro of the Lawrence Berkeley National Laboratory. This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Environmental Services Division of the U.S. Department of Energy (DOE) under contract DE-AC03-76SF00098, and by the U.S. Environmental Protection Agency (EPA). Although the research was partially funded by the EPA, the report may not necessarily reflect the views of the EPA and no official endorsement should be inferred.

## REFERENCES

Alexander, B.; Rodman, N.; White, S.B.; Phillips, J. Areas of the United States with elevated screening levels of Rn-222. Health Phys. 66: 50-54; 1993.

Efron, B.; Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37: 36-48; 1983.

Freedman, D.; Pisani, R.; Purves, R.; Adhikari, A. Statistics, second edition. New York, NY: Wiley & Sons; 1973.

Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. Bayesian data analysis. New York, NY: Chapman and Hall; 1995.

Hobbs, W.E.; Maeda, L. Identification and assessment of a small, geologically localized radon hot spot. Environ. Int. 22 (Suppl. 1): S809-S817; 1996.

Janssen, I.; Stebbings, J.H. Gamma distribution and house Rn-222 measurements. Health Phys. 63: 205-208; 1992.

Klotz, J.B.; Schoenberg, J.B.; Wilcox, H.B. Relationship among short- and long-term radon measurements within dwellings: Influence of radon concentrations. Health Phys. 65: 367-374; 1993.

Lucas, R.M.; Grillo, R.B.; Perez-Michael, A.; Kemp, S.S. Final report (for USEPA): National residential radon survey statistical analysis. Research Triangle Institute, Research Triangle Park, NC; 1992

Marcinowski, F.; Lucas, R.; Yeager, W. National and regional distributions of airborne radon concentrations in U.S. homes. Health Phys. 66: 699-706; 1994.

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.

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.

Price, P.N.; Nero, A.V.; Gelman, A. Bayesian prediction of mean indoor radon concentrations for Minnesota counties. Report LBL-35818. Berkeley, CA: Lawrence Berkeley National Laboratory; 1995.

Ronca-Battista, M.; Chiles, B. The relationship between winter screening and annual average radon concentrations in U.S. homes. In: Proc. 1990 international symposium on radon and radon reduction technology. Vol. 1: 2,3- 2,14; 1990. Washington, D.C.: U.S. Environmental Protection Agency.

White, S.B.; Clayton, G.A.; Alexander, B.V.; Clifford, M.A. A statistical analysis: Predicting annual Rn-222 concentrations from 2-day screening tests. In: Proc. 1990 symposium on radon and radon reduction technology. Vol. 1: 3,117-3,118; 1990. U.S. Environmetal Protection Agency, Washington, D.C.

Wirth, S. et al. National radon database documentation: The EPA state/residential radon surveys. 1992. U.S. Environmental Protection Agency, Washington, D.C.