

**METHOD FOR PREDICTING THE OCCURRENCE  
OF SMALL RADON HOT SPOTS  
FROM RANDOMLY CHOSEN DATA**

W.E. Hobbs and L.Y. Maeda  
Department of Environmental Studies, University of California  
Santa Barbara, CA

**ABSTRACT**

Radon prone houses are often associated with uranium-rich geologic formations, but these formations are difficult to predict in the absence of indoor radon measurements. Even with data, it may be difficult to specify a hot spot if the area does not conform with established administrative boundaries or if there is uncertainty to the specific test location due to confidentiality. Radon in soil gas and in indoor air has been shown to follow lognormal distributions. If data from a random selection of homes are presented on lognormal graph paper, they will be represented by a straight line. Systematic deviation from the line at high radon concentrations is a sensitive indicator of the presence of a small geologically localized radon hot spot. The graphical procedure for finding a hot spot requires establishing the background distribution. The larger the sample size the more accurately the distribution is characterized in the high-radon tail. Sparsely populated hot spots remain difficult to find and the better strategy involves evaluating each geologic province separately.

**INTRODUCTION**

Since the discovery of the Reading Prong in Pennsylvania, many radon hot spots have been found in the United States. A radon hot spot is a region of unusually high indoor radon concentrations compared with neighboring areas and is usually associated with a specific geologic formation which contains anomalously high concentrations of uranium. The identification of hot spots represents a significant opportunity to reduce the potential exposure to ionizing radiation. We believe it may be beneficial to study Santa Barbara County as a model of a small geologic hot spot to aid in the future identification and assessment of high radon potential areas.

There are many outcroppings of Miocene Shale in Southern California. One of these, the Rincon Shale Formation, is associated with high concentrations of indoor radon. (For conciseness, we will henceforth refer to outcropping of the Rincon Shale geologic formation as the "Rincon.") The radon potential of the Rincon was discovered in 1989 by Carlisle and Azzouz (1993) through stratified sampling. In this paper we analyze a subsequent statewide survey of indoor radon levels which does not confirm this radon potential. This survey consists of 2319 measurements which are reported by county and postal zip code; 120 data were in Santa Barbara County. We report a statistical procedure to determine the presence of a radon hot spot. This analysis procedure is applied to the counties of California. Our basic hypothesis is that the Santa Barbara hot spot was not found in the state survey because random sampling is ill-suited for the evaluation of small subpopulations.

During 1990, the US Environmental Protection Agency (EPA) and the California Department of Health Services (DHS) made a survey of radon levels in California homes (DHS, 1993). The homes were chosen randomly within the counties, but the number of homes per county was chosen both to provide a reasonable estimate of the county distribution and to consider geologic and other factors. For example: 37 samples were taken from Siskiyou County with a population of about 44,000 for about one test for every 1200 people while 89 samples were taken from Los Angeles County with a population of about 9,000,000 for about one test for every 101,000 people. Elevated radon levels were not expected in these counties. In Santa Barbara and Ventura Counties, where higher radon levels were expected based on previous studies (Liu et al, 1991b; DHS, 1991), 120 and 159 samples were taken; with their populations of about 350,000 and 400,000, this provided about one test for approximately 2900 and

2500 people respectively. For 30 counties there were at least one test for every 6000 people; for 12 counties, there was one test for between 10,000 and 20,000 people; and for 6 densely populated coastal counties there was less than one test per 30,000 people. For 10 counties, generally of very low populations, there was not enough data to characterize the distributions.

Charcoal canisters were used to make the radon screening measurements. The test devices were distributed by the DHS. The tests were made using the testing protocols developed by the EPA. Because of the large uncertainties in relating collected radon to average residential radon concentrations at low levels, a minimum reportable level of 1.0 pCi/L (37 Bq/m<sup>3</sup>) was used. Any measurement showing a level less than this was simply tagged <1.0 pCi/L (<37 Bq/m<sup>3</sup>). Of the 2319 measurements taken, 1463 (~63%) were so designated.

## STATISTICAL METHODS

Nero et al. (1986) have shown that radon levels in houses follow an apparent lognormal distribution. A simple approach to such a distribution is to note that the logarithms of the measurements follow a standard normal (Gaussian) distribution. We have fit the radon measurements for most counties in California to lognormal distributions. Since the sample is assumed to be chosen randomly, a normal-curve z-value is calculated for each point. The points with magnitudes <1.0 pCi/L (<37 Bq/m<sup>3</sup>) serve to establish the cumulative probability for the higher magnitude points, but are not used further (see Appendix for an example). The best-fit line correlating a county's non-negative radon logarithms and their z-values is found using least squares.

For a general normal curve the z-values are defined

$$z = [\ln(r) - \mu]/\sigma. \quad (1)$$

If z is chosen as the independent coordinate, the regression coefficients will provide the best-fit mean  $\mu$  and standard deviation s

$$\ln(r) = \mu + \sigma z. \quad (2)$$

For clarity, we have retained the logarithm notation for the radon variable  $\ln(r)$  but not for the mean and standard deviation.  $\mu$  is the natural logarithm of the geometric mean and s is the natural logarithm of the geometric standard deviation (g.s.d.).

## RESULTS: ANALYSIS OF RADON DATA

Fig. 1 shows the results of the analyses of 47 California counties. The ordinate has units of  $\ln(\text{pCi/L})$ : e.g., -1.0 corresponds to about 0.37 pCi/L (5 Bq/m<sup>3</sup>); 0 to 1.0 pCi/L (37 Bq/m<sup>3</sup>); and 1.0 to about 2.7 pCi/L (100 Bq/m<sup>3</sup>). Error bars show the standard deviations. For logarithms, the standard deviation corresponds to a multiplication factor. The legend shows factors of 2 and 3 and most of the standard deviations for the counties fall in this range. Thus, usually ~68% of the measurements for a county fall within a factor of 2 to 3 of the geometric mean. The geometric mean is approximately equal to the median or 50 percentile of the measurements.

In Fig. 1 the counties are ordered by the fraction of homes with radon levels above 4 pCi/L (150 Bq/m<sup>3</sup>), the US EPA action level, based on the lognormal fit. These fractions vary from less than 0.1% for Orange, Riverside, and San Francisco Counties (all with large populations) to greater than 10% for Madera, Merced, San Joaquin, Nevada, and San Luis Obispo. Note that the number of samples tends to decrease moving to the right in the figure. The counties with a lower number of samples have increased distribution spread.

All the data from the survey are weighted according to their county populations and the lognormal distribution for California was calculated: the log mean  $\mu$  is -0.43 (geometric mean ~0.65 pCi/L, 24 Bq/m<sup>3</sup>) and the log s.d. s is 0.87 (g.s.d. factor ~2.4). Fig. 1 also shows these parameters. Los Angeles County contains almost one-third of California's population and has parameters similar to California. Most of the large coastal population centers

are in counties with low radon concentrations. The county parameters are uniform; the county geometric means are within one California g.s.d. of the geometric mean except for Humboldt and Sacramento counties.

Fig. 2 shows a comparison of the California radon data with the approximate parametric lognormal representation for various radon concentration intervals. As can be seen, the agreement is excellent. The data conforms closely with a lognormal distribution. The parametric curve predicts the fraction of homes above 4.0 pCi/L (150 Bq/m<sup>3</sup>) as ~1.8%, while the data itself shows ~2.0%. Liu et al. (1991a) made a survey of California homes and found a similar radon profile. The geometric mean for their one-year alpha-track measurements was 0.85 pCi/L (30 Bq/m<sup>3</sup>) with a g.s.d. of 1.91. This leads to a prediction of 0.76% of the homes above 4.0 pCi/L (150 Bq/m<sup>3</sup>) although their survey found 6 homes or about 1.9% of the 312 homes above this level. When the lognormal parameters are recalculated using our method (Appendix), the geometric mean is 0.54 pCi/L (20 Bq/m<sup>3</sup>) and the g.s.d. is 2.64. This leads to a prediction of 1.9% of the homes statewide above 4.0 pCi/L (150 Bq/m<sup>3</sup>) which coincides with their data.

Although screening measurements are conservative and result in an overprediction of homes above the action level, the short-term screening measurements provide useful information. The state, as a whole, conforms closely to a lognormal distribution with generally low residential radon concentrations.

While no county distribution has dramatically elevated indoor-radon concentrations, several county distributions have large standard deviations. This may be accurate, perhaps resulting from insufficient sample size, or it may be evidence that the county data is not truly lognormal. It would not be lognormal if there is a geologically localized small radon hot spot located within the county and the hot spot has significantly higher radon levels. This may be difficult to determine if there is limited data and the fraction of houses in the hot spot is small.

## RESULTS: ANALYSIS OF HOME DISTRIBUTION STATISTICS

In Fig. 1, Santa Barbara County appears average for California. Its geometric mean is a little less than the California value, but its g.s.d. is somewhat larger. The fraction of homes which exceed the 4-pCi/L (150-Bq/m<sup>3</sup>) level is sensitive to the g.s.d. but only 6.3% are predicted to exceed this level. In their final report on radon in California Liu et al. (1990) make no mention of Santa Barbara County.

Carlisle and Azzouz (1993) did indoor radon tests of homes in the Santa Barbara region. In Santa Barbara County, 42 homes on the Rincon Shale and 34 homes on non-Rincon soil were tested for radon. In standardized screening tests, 74% of the homes on the Rincon Shale had measurements greater than 4 pCi/L and 26% were greater than 20 pCi/L. Homes on the non-Rincon Shale formations had low radon levels similar to those found for the general California population. Off-Rincon indoor air measurements had a geometric mean of 0.8 pCi/L with a geometric standard deviation of 2.3; a lognormal distribution with these parameters has about 3% of the population greater than 4 pCi/L.

From their data, Carlisle and Azzouz (1993) determined there are two distinct populations when measuring for radon in Santa Barbara County: homes on the Rincon Shale and homes on non-Rincon geologic formations. Simple random sampling may miss small geological radon-prone areas even for larger samples (>30 data). Deliberate geological exploration by soil-gas sampling is a more efficient means of determining radon prone regions.

Because of the uncertainties in the precise location of elevated indoor radon, the California DHS declared the southern part of Santa Barbara County below the crest of the Santa Inez Mountains from Summerland to Gaviota to be California's first radon hot spot (DHS, 1991). This region contains about half of Santa Barbara County's approximately 350,000 people.

Thus, a dramatic radon-prone region has been identified. The reason it was not seen by random sampling has two parts: (1) the lognormal distribution which is an accurate model for characterizing an indoor radon distribution has a long tail; and (2) the fraction of homes affected by the uranium rich source is small (<10% by

assumption). This combination results in the few measurements in the high-potential region appearing as chance data in the tail of the distribution. Fig. 3 shows a summation of two lognormal distributions where 5% of the population is Rincon Shale and the remaining 95% has the nominal radon potential of California. As you can see, although the distributions have more than an order of magnitude difference in geometric means, the Rincon dominated part of the combined distribution appears only as an extra long tail on the distribution.

The median for this bimodal distribution change very little from the California distribution. For instance, consider the expectation value  $R_s$  for the median of a sample from a population which consists of two lognormal distributions (note  $\mu_s = \ln R_s$ ). We assume a fraction  $f$  of the total is the Rincon radon population with median  $R_r$ , and the remainder of the distribution  $(1-f)$  is the California radon population with median  $R_c$ . The logarithms are the basic measure for these distributions and the mean of the sum  $R_s$  is given by

$$\ln(R_s) = f \cdot \ln(R_r) + (1 - f) \cdot \ln(R_c)$$

or

$$R_s = R_r^f \cdot R_c^{(1-f)} \tag{3}$$

As an example, if the California median is  $R_c=0.8$  pCi/L (30 Bq/m<sup>3</sup>) and the Rincon median is  $R_r=8.4$  pCi/L (310 Bq/m<sup>3</sup>) and if the Rincon fraction is  $f=5\%$ , then the Santa Barbara median would equal about  $R_s=0.9$  pCi/L (33 Bq/m<sup>3</sup>). If the Rincon fraction was  $f=10\%$ , the SB median would be about  $R_s=1.0$  pCi/L (37 Bq/m<sup>3</sup>). These small differences are not easily observed in random samples.

When considering only the geometric means (medians), a radon hot spot, even with very high levels, is essentially opaque if it subtends only a small fraction of the total number of homes. To assess the impact of the Rincon the fraction of homes which it contains are evaluated. From analysis of maps about 2% or 3,500 people live in homes built on the Rincon Shale and about 16% or 28,000 people live in homes in downwash contact regions.

Overall, about 11,000 people in Santa Barbara County (~3%) live in homes which exceed 4 pCi/L (150 Bq/m<sup>3</sup>) directly resulting from the Rincon. In addition, there may be ~3% of the people which live in homes with elevated radon unrelated to the Rincon. The data shows there are up to 1,500 people in the county which live in homes which exceed 20 pCi/L (740 Bq/m<sup>3</sup>) and essentially all of these are associated with the Rincon.

The standard deviation of the radon logarithms  $\sigma_s$ , for two lognormal distributions (e.g.,  $\sigma_r$  for the Rincon and  $\sigma_c$  for California), is given by

$$\sigma_s^2 = f \cdot \sigma_r^2 + (1 - f) \cdot \sigma_c^2 + f(1 - f) \cdot [\ln(R_r/R_c)]^2$$

and

$$\text{factor} = \exp(\sigma_s) \tag{4}$$

where "factor" is short for g.s.d. factor, a measure of the variation in the data distribution. Since the assumed composite distribution of logarithms of indoor radon levels is not normal, this factor does not necessarily enclose ~68% of the data. It is however, a measure of the spread of the observed radon levels. The California radon data has a factor of ~2.3 and the factor for the Rincon radon data in Table 1 has a factor of ~3.1. Using these values, when the Rincon fraction is  $f=5\%$ , the factor is ~2.8, and when the Rincon fraction is  $f=10\%$ , it is ~3.1. This shows that the geometric factor for the radon data is a more sensitive indicator of the presence of a hot spot than the geometric mean. The last term of Equation (4) involves the ratio of geometric means of the subpopulations  $R_r/R_c$  which is about an order of magnitude for California and Rincon indoor radon distributions. This term is the primary source of increased spread. If the two subpopulations have the same spread ( $\sigma_r=\sigma_c$ ), the increase results only from this final term. The maximum g.s.d. factor occurs when the two subpopulations have equal contributions ( $f=0.5$ ).

In our analysis of the data taken by the EPA and DHS (DHS, 1993), only the points with magnitude greater than 1.0 pCi/L (37 Bq/m<sup>3</sup>) were used and a single lognormal distribution was determined. For Santa Barbara a g.s.d. of 3.35 was found. The analysis of the previous paragraph applies to the distribution approximated as a sum of two distributions. The application of our curve-fitting scheme using truncated data (see Appendix) to the distribution shown in Fig. 3 would have resulted in a artificially large g.s.d. but an artificially small geometric mean. Thus, this statistical method acts to magnify the influence of a radon hot spot.

## DISCUSSION

Our approach to finding a statistical indication of a radon hot spot is as follows: First, the high-magnitude truncated data set is examined. In this paper the data with magnitude greater than 1.0 pCi/L (37 Bq/m<sup>3</sup>) is used. The lower magnitude data are not used because they have larger relative errors. This criteria generally eliminates about two-thirds of the data. Second, the background distribution must be known. The indoor radon levels for a general area are assumed to be distributed according to a lognormal distribution. In this study, the California distribution is taken as the background distribution. Third, the geometric mean and standard deviation for each of the subregions, the counties in this study, are calculated. Fourth, the geometric standard deviation for each county is examined. It is usually less than 3.0. If it is greater than 3.0, the county should be considered in more detail.

We are concerned about the occurrence of small localized regions with high radon levels within an otherwise low or moderate radon area. California, as a whole, is a low-radon area and is examined for possible radon hot spots. Based on the radon data from California counties, the counties seem to partition into groups. There are several candidates to be considered for potential radon hot spots.

**Average Counties**--Most (about 30) California counties appear to be well characterized by a simple lognormal distribution with a low median (< 1 pCi/L, <37 Bq/m<sup>3</sup>) and geometric standard deviation (< 3). Based on the lognormal fits, we predict that these counties will have less than 5 percent of their homes with levels above 4 pCi/L (150 Bq/m<sup>3</sup>) in standard screening tests. The data don't show any reason to expect serious elevated indoor radon problems in these counties. Fortunately, these counties contain over 85% of the total population of California (about 25 million people).

**Modestly elevated counties**--There are several counties, of medium population, running through the central part of the state with higher than average radon levels (greater than 2.0 pCi/L, 74 Bq/m<sup>3</sup>). These counties are also well characterized by a lognormal distribution, but the distributions are modestly elevated. These are Tulare, Stanislaus, Sutter, Madera, Merced, San Joaquin and Nevada. These counties probably don't have hot spots but 5-10% of their homes may have radon levels above 4 pCi/L (150 Bq/m<sup>3</sup>) in standardized screening tests.

**Anomalous Counties**--Sacramento and Humboldt Counties appear to be low-radon counties, but each has a single home with a particularly high radon measurement which does not conform to the others. If we knew, for sure, that the data was lognormal, we could preemptively discard this anomalous data point. We don't know this for sure because the high-radon house may be on a geologic radon hot spot. From our experience with radon tests and data, we believe there is a good chance that there may be some interference such as a uranium-rich building material which may cause these houses to test high. It is probably not prudent to expend significant resources tracking the precise cause.

**Broad-Distribution Counties**--There are three groups of counties which should be considered further because their radon distributions were found to have a wide spread (g.s.d. factor greater than 3.4). In all cases there is a systematic deviation from the lognormal distribution of data at the higher levels. The data is consistent with a subpopulation of anomalously high level indoor radon concentrations. In the north there are three counties with large g.s.d. factors: Lake, Solano, and Napa Counties. These counties are all on the eastern slope of the California Coastal Range, a region of geologic activity (e.g., geysers), but a specific geologic formation has not been identified. We do not know if the apparent high-radon subpopulation conforms with a geologic subregion in these counties.

Further north, Shasta and Tehama Counties have indoor radon concentration distributions with still larger spreads, geometric standard deviations of 3.8 and 4.0. This spread results from a pronounced upward trend for the higher magnitude data which is consistent with a subpopulation enhancing the high-magnitude tail of the distribution. The data of these counties is consistent with a sum of two lognormal distributions. Again, we do not know if the high-radon subpopulation conforms with a specific geologic subregion.

Monterey, and San Luis Obispo Counties also have very broad distributions but they also have few numbers of homes evaluated (22 and 21). These counties also have known outcroppings of the Rincon Shale, the same geologic formation associated with high radon levels in Santa Barbara County. We believe that the Rincon may be the cause of the radon-prone subpopulations in these counties. Table 1 summarizes the geometric standard deviations for these counties.

Carlisle and Azzouz (1993) found a radon prone geologic formation in Santa Barbara County. We have found that it is limited to a small portion of the housing population in the county which may explain why it was not identified from random testing. Health officials should be sensitive to the possibility of radon-prone subpopulations and subregions when they review radon survey data.

### ACKNOWLEDGMENTS

We thank Dr. Ed Keller and Dr. Don Carlisle of the Geology Departments of UCSB and UCLA, respectively, for sharing data and encouragement. Helmut Ehrenspeck of the Dibblee Foundation supplied detailed geologic maps of Santa Barbara County. Scott Hoskins of Santa Barbara County Public Works provided detailed street maps. Dr. Ron Churchill of the California Division of Mines and Geology, Dr. Kai-Shen Liu of the California Department of Health Services, and Peggy Jenkins of the California Air Resources Board also supplied data, valuable discussions, and encouragement.

**Table 1. Potential Radon Hot-Spot Counties in California**

County (no. samples)	Geometric Standard Deviation Factor	Fraction of homes expected above 4.0 pCi/l
<b>Anomalous counties</b>		
Humboldt (50)	5.4	1.7 %
Sacramento (68)	4.5	3.5 %
<b>Middle-north counties</b>		
Lake (20)	4.9	6.2 %
Napa (33)	3.5	6.2 %
Solano (59)	3.5	4.2 %
<b>Far-north counties</b>		
Shasta (96)	3.7	4.7 %
Tehama (19)	4.1	5.8 %
<b>Rincon Shale counties</b>		
Santa Barbara (120)	3.4	6.3 %
Monterey (22)	4.4	6.8 %
San Luis Obispo (21)	6.7	12.9 %

## Appendix Method for Estimating Lognormal Parameters from Radon Data

Most radon data is dominated by low-magnitude concentrations themselves dominated by errors. We have developed a scheme for fitting a lognormal curve to the data with magnitudes greater than 1.0 pCi/L (37 Bq/m<sup>3</sup>). It does not use the low-magnitude data and thereby transcends the low-magnitude error problem. We will briefly detail our procedure using the data obtained in the EPA/DHS survey of Santa Barbara County.

Fig. A1 shows the raw data generated by the survey. The data have been ordered in increasing size. The number of points below a given magnitude defines the cumulative probability for that magnitude. There were 120 measurements taken of which five eighths (75 data) are too small to accurately quantify. The low-magnitude points were simply designated <1.0 pCi/L (<37 Bq/m<sup>3</sup>). (Santa Barbara has low average outdoor levels of radon <0.2 pCi/L, <7 Bq/m<sup>3</sup>.) The low magnitude data serve to establish the probability magnitude for the points above 1.0 pCi/L (37 Bq/m<sup>3</sup>).

Both axes of Fig. A1 are modified. The natural logarithm of the radon magnitude and the z-value for the cumulative probability are evaluated. The z-value is implicitly defined using the normal probability curve,

$$P(r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp(-x^2/2) dx \quad (A)$$

where P(r) is the cumulative probability for radon level r. The resulting points in this lognormal space approximate a straight line and this shows the data conform to a lognormal probability distribution. Using linear least-squares, the curve parameters are estimated. The logarithm of the geometric mean is equal to the y-intercept and the logarithm of the geometric standard deviation is equal to the slope. In Fig. A2 shows the processed data and the best-fit linear relation; the geometric mean is specified. The legend shows the regression analysis with the coefficient of correlation R. R<sup>2</sup>=98.3%, showing an excellent fit for the data.

There is a subtlety in these manipulations. In this paper we use the logarithms of the indoor radon levels as the basic data and they are usually closely approximated by a normal curve. The value on the curve gives the expected portion of the distribution per differential unit logarithm of the radon levels, not per differential unit of radon level. It is convenient to think of the radon intervals as being defined by a multiplicative factor (>1), not an additive amount. Suppose we are interested in 2 pCi/L (74 Bq/m<sup>3</sup>) within a factor of 2; this would be the interval from 1 to 4 pCi/L (37 to 150 Bq/m<sup>3</sup>). The difference is proportional to the radon level. For example, if we are interested in 20 pCi/L (740 Bq/m<sup>3</sup>) within a factor of 2, the interval would be from 10 to 40 pCi/L (370 to 1480 Bq/m<sup>3</sup>).

### REFERENCES

1. Carlisle, D. and Azzouz, H. Discovery of Radon Potential in the Rincon Shale California - a Case History of Deliberate Exploration. *Indoor Air* 3:131-142; 1993.
2. DHS (California Department of Health Services). Santa Barbara County Radon Survey February - March 1991; California Department of Health Services, Sacramento, 1991.
3. DHS (California Department of Health Services). California Statewide Radon Survey Screening Results; California Department of Health Services, Sacramento, 1993.
4. Liu, K.S.; Hayward, S.B.; Girman, J.R.; Moed, B.A.; and Huang, F.Y. Survey of Residential Indoor and Outdoor Radon Concentrations in California (Final Report CA/DOH/AIHL/SP-53). California Department of Health Services, Berkeley, CA. 1990.

5. Liu, K.S.; Hayward, S.B.; Girman, J.R.; Moed, B.A.; and Huang, F.Y. Annual Average Radon Concentrations in California Residences. *J Air Waste Management Assoc* 41: 1207-1212; September 1991a.
6. Liu, K.S.; Chang, Y.L.; Hayward, S.B.; and Huang, F.Y. Survey of Residential Radon Levels in Ventura County and Northwestern Los Angeles County, (Final Report CA/DOH/AIHL). California Department of Health Services, Berkeley, CA, 1991b.
7. Nero, A.V.; Schwehr, M.B.; Nazaroff, W.W.; and Revzan, K.L. Distribution of Airborne Radon-222 Concentrations in U.S. Homes. *Science* 234: 992-997; 1986.