A comparative study of lognormal, gamma and beta modelling in radon mapping with recommendations regarding bias, sample sizes and the treatment of outliers

To cite this article: Patrick Murphy and Catherine Organo 2008 J. Radiol. Prot. 28 293

View the article online for updates and enhancements.

Related content
- The use of volunteer radon measurements for radon mapping purposes: an examination of sampling bias issues
  Orlaith Burke and Patrick Murphy

- The Castleisland Radon Survey—follow-up to the discovery of a house with extremely high radon concentrations in County Kerry (SW Ireland)
  Catherine Organo and Patrick Murphy

- Estimation of seasonal correction factors through Fourier decomposition analysis—a new model for indoor radon levels in Irish homes
  Orlaith Burke, Stephanie Long, Patrick Murphy et al.

Recent citations
- Variance of indoor radon concentration: Major influencing factors
  I. Yarmoshenko et al

- Log-normality of indoor radon data in the Walloon region of Belgium
  Giorgia Cinelli and François Tondeur

- Hierarchical modeling of indoor radon concentration: how much do geology and building factors matter?
  Riccardo Borgoni et al
A comparative study of lognormal, gamma and beta modelling in radon mapping with recommendations regarding bias, sample sizes and the treatment of outliers

Patrick Murphy$^{1,3}$ and Catherine Organo$^2$

1 School of Mathematical Sciences (Statistics), University College Dublin, Belfield, Dublin 4, Ireland
2 Radiological Protection Institute of Ireland, Advisory Services Division, 3 Clonskeagh Square, Dublin 14, Ireland

E-mail: Patrick.Murphy@ucd.ie

Received 18 March 2008, in final form 6 May 2008, accepted for publication 16 May 2008
Published 20 August 2008
Online at stacks.iop.org/JRP/28/293

Abstract

Radon risk maps have been produced in many countries using non-geologically based techniques utilising the lognormal and gamma distributions to display point estimates of the probability, $P_{RL}$, that indoor radon levels will exceed a reference level. Recent work on updating radon maps in Ireland prompted an examination of various statistical issues associated with non-geologically based radon mapping techniques. The purpose of this paper is to compare existing techniques with a new theoretically unbiased model which uses the beta distribution to estimate $P_{RL}$. We examine this new beta based model and four existing models for bias using simulated test data sets. Producing confidence intervals for $P_{RL}$, we also analyse the relationship between sample size and estimation error for each model and finally we consider the effect of extreme values on estimation procedures. We demonstrate that, under the assumption of lognormally distributed data with outliers, a two stage approach to testing which first detects and removes outliers from the data and then applies the beta based method is most satisfactory from both a theoretical and empirical perspective.

1. Introduction

There has been much progress in the field of radon mapping since Nero et al (1986) illustrated that indoor radon levels could be modelled well by lognormal distribution. It is known that indoor radon levels are significantly influenced by many factors related to house construction,
ventilation and heating (Arvela 2005). Indeed one of the interesting aspects of radon mapping work is the large house to house variability that is observed in measurements of radon levels. Two houses located beside each other may display quite different indoor radon levels. Gunby et al (1993) examined many of the factors that could be involved in producing this house to house variability. They hypothesised a mechanism by which the difference between observed indoor radon levels and the outdoor background level is the multiplicative product of a large set of factors. Thus they provided some statistical reasoning behind why the difference between indoor radon levels and the outdoor background level should follow a lognormal distribution. Other probability distributions have been considered, for example Janssen and Stebbings (1992) modelled radon measurements using the gamma distribution.

Acknowledging the influence of house effects, the presence of radon gas in a house is, in the vast majority of the cases related to the geological formations beneath that house. Recent radon mapping work has attempted to construct models that incorporate geological information. Constructing these models is not straightforward as it has long been recognised that the relationship between indoor radon levels and geology is complicated by many non-house related factors. The extent of these complications can be seen from an analysis performed by Gundersen and Schumann (1996). Despite this, work on integrating geological information using Bayesian and other methods has continued and recent work by for example Price et al (1996) and Miles and Appleton (2005) has shown much promise for this approach.

One of the reasons for the success shown in the UK has been the relatively large number of indoor radon measurements made. Indeed the amount of data available was deemed sufficient to allow radon maps to be produced at 1 km grid square level. This is in stark contrast to the situation in Ireland where the number of radon measurements possessing accurate geographical location identification is still quite small. It is worth noting that the main reason for the lack of accurate location information is that Ireland does not possess a postcode system.

The Radiological Protection Institute of Ireland (RPII) has recently conducted a localised radon survey of the area surrounding Castleisland in County Kerry. This survey, which was prompted by the discovery of a house with a seasonally adjusted annual average radon concentration of 49 000 Bq m\(^{-3}\) (Organo et al 2004) provided 383 new measurements in four 10 km \times 10 km grid squares. The aforementioned lack of accurate geographical identifiers on the houses that participated in the survey prevented any attempt to utilise a geological based modelling approach and required us to investigate other methods. Predictions for the proportion of houses with radon levels exceeding the 200 Bq m\(^{-3}\) reference level in the four grid squares under investigation had been previously made during the Irish National Radon Survey (Fennell et al 2002). The aim of our work was to produce updated predictions for each of the four studied grid squares and, if possible, to produce predictions for the sixteen 5 km\(^{2}\) grid squares in the reference area. The results of this analysis have been reported in Organo and Murphy (2007).

During the analysis of the Castleisland data, different approaches to modelling indoor radon levels were considered in an attempt to improve upon existing non-geologically based methods. The purpose of this paper is to report on the results of our work on model comparison which involved statistical analyses that were conducted on simulated data sets constructed to have similar statistical properties to the Castleisland data. Our work considers only estimation procedures which make use of sample data from a single area under investigation. Although we fully recognise their merits, we explicitly do not consider any methods which make use of neighbouring regions to enhance a given sample through smoothing or Bayesian methods. This paper consists of four elements:

(1) A description of five different methods for predicting \(P_{RL}\) (the proportion of measurements in a grid square above the reference level of 200 Bq m\(^{-3}\)). In particular this section presents
Comparison of statistical models in radon mapping

for the first time details of a new method used by the RPII for computing $P_{RL}$, which is theoretically superior to other methods described in the literature for estimating $P_{RL}$ from lognormally distributed data and is based on the beta distribution.

(2) Results of a simulation study to consider bias in each estimation procedure.

(3) An analysis of the relationship between sample size and the error on the predicted $P_{RL}$.

(4) An examination of the effect of extreme values on predictions from these models.

2. Description of the different modelling techniques

This section describes five different approaches that can be used to model indoor radon levels and predict the proportion of homes ($P_{RL}$) with radon levels above a reference threshold.

2.1. General linear model and lognormal model

Nero et al. (1986) demonstrated empirically that indoor radon measurements could be well modelled using the lognormal distribution. While not providing a full formal statistical explanation of this lognormal behaviour, Gunby et al. (1993) suggest some statistical justification for the lognormal behaviour by proposing the following general linear model for predicting $P_{RL}$:

$$\log(R_I - R_O) = \log(A) + \sum_{i=1}^{n} \log(\tau_i) + \epsilon',\quad (1)$$

where $R_I$ is the indoor radon concentration, $R_O$ is the background outdoor radon concentration, $A$ is an average difference between the indoor and outdoor concentrations, the $\tau$'s are factor effects related to the house which affect the indoor concentration, and $\epsilon'$ is an error term that is distributed as a normal random variable. Assuming that there are a large number of factor effects which are similarly distributed, then according to the central limit theorem we have

$$\log(R_I - R_O) = \epsilon,\quad (2)$$

where $\epsilon$ is approximately normally distributed with true mean $\mu$ and true standard deviation $\sigma$. With this simplification, the empirical lognormal behaviour finds a theoretical basis and the predicted proportion of homes with indoor radon levels above a given reference level RL can be computed from

$$P_{RL} = 1 - \Phi \left( \frac{\log(RL - R_O) - \mu}{\sigma} \right) = 1 - \int_{-\infty}^{\log(RL - R_O)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx,\quad (3)$$

where $\Phi$ is the cumulative distribution function of a standard normal random variable. As both $\mu$ and $\sigma$ are unknown, the standard practice is to replace these with the sample mean $\bar{x}$ and sample standard deviation $s$ of the log-transformed and background-subtracted radon levels and so compute the following estimate

$$\hat{P}_{RL} = 1 - \Phi \left( \frac{\log(RL - R_O) - \bar{x}}{s} \right).\quad (4)$$

We note that theoretically the nonlinearity of equation (3) means that the expression in equation (4) will not provide an unbiased estimator of $P_{RL}$. 

2.2. Median or sort technique

Miles (1994) adapts the lognormal model (3) by recognising that the sample mean and sample standard deviation could be influenced by the presence of extreme values in the sample data, and consequently an alternative so-called sort technique is recommended. In this model the following estimate is used:

\[ \hat{P}_{RL} = 1 - \frac{1}{\Phi_1 \left( \frac{\log(RL - RO) - M}{X_{84.15} - M} \right)} \]

(5)

where \( M \) is the median and \( X_{84.15} \) is the 84.15th percentile of the log-transformed sample radon levels. The logic here is that for a true normal distribution the median \( M \) and the mean \( \mu \) are identical and the standard deviation is given by

\[ \sigma = X_{84.15} - M. \]

(6)

Exponentiating equation (6) we find that the geometric standard deviation (GSD) of the original radon levels may be estimated from

\[ \text{GSD} = \frac{R_{84.15}}{m} \]

(7)

where \( R_{84.15} \) and \( m \) are the 84.15th percentile and median of the original radon measurements.

As both the median \( m \) and the percentile \( R_{84.15} \) are less influenced by the presence of outliers the estimator in equation (5) should be more robust than that in equation (4) and consequently a better estimator for use with real data. We do note however that, as was the case with equation (4), the nonlinearity of equation (3) implies that equation (5) is a theoretically biased estimator of \( P_{RL} \).

2.3. Lognormal model with bias correction

Andersen et al (2001) correctly observe that \( \hat{P}_{RL} \) estimated in (4) is not an unbiased estimator of \( P_{RL} \) and that a correction for bias should be made. In the situation where the standard deviation \( \sigma \) is known but the mean must be estimated from the sample they suggest the following revised estimator (cf equation (10) of Andersen et al 2001):

\[ \tilde{P}_{RL} = 1 - \frac{1}{\Phi_1 \left( \alpha \right)} - \frac{1}{\sqrt{2\pi}} \frac{\alpha}{\sigma} e^{-\frac{\alpha^2}{2}} \]

(8)

where

\[ \alpha = \frac{\log(RL - RO) - \bar{x}}{\sigma} \]

(9)

We note that the above expression from Andersen et al (2001) contains a typographical error. The correct expression, intended by Andersen et al (cf equation (26) of Andersen et al 2001) should read

\[ \tilde{P}_{RL} = 1 - \Phi(\alpha) - \frac{1}{\sqrt{2\pi}} \frac{\alpha}{2N} e^{-\frac{\alpha^2}{2}} \]

(10)

where \( \alpha \) is again given by equation (9).

This estimation equation (10) clearly still involves the unknown standard deviation, \( \sigma \), and as such is not complete. However Andersen et al observe empirically that the standard deviation is typically small and that the above correction is therefore somewhat exaggerated.

4 The 84.15th percentile is an approximation. To twelve decimal places the more accurate value is the 84.134474606854th percentile.

5 Readers may note that the procedure above does not replicate exactly that contained in Miles (1994). This is due to a misprint on page 208 of Miles (1994) not present in the original draft of that paper, cf Miles (1993). The sentence ‘...the GM and GSD of \( \ln(R_i - RO) \) are calculated...’ should read ‘the GM and GSD of \( (R_i - RO) \) are calculated.'
deviation of indoor radon levels in Denmark varies very little with geography. This fact has also been noted by Miles (1994) in the UK. Consequently they argue that the true standard deviation, $\sigma$, may be estimated essentially independently of the mean using an estimate based on an enhanced set of data composed of the sample from the region under investigation pooled with data from neighbouring regions.

We note that Andersen et al also suggest a Bayesian enhancement of their model in which the mean, $\mu$, is estimated using geological information. As our previously mentioned focus is upon comparing non-geologically based approaches we will examine the core model described in equation (10).

### 2.4. Beta model

The Irish National Radon Survey conducted by the RPII (Fennell et al 2002) estimate $P_{RL}$ using a method suggested by Daly (1994) and based on Lieberman and Resnikoff (1955). This method makes no approximation but instead derives an expression for

$$P_{RL} = 1 - \int_{-\infty}^{\log(RL - R_0)} \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{1}{2} \left( \frac{x - \bar{x}}{\sigma} \right)^2} dx,$$

when $\mu$ and $\sigma^2$ are unknown but are estimated by the sample mean $\bar{x}$ and sample variance $s^2$ respectively. In this situation the correct way to estimate $P_{RL}$ is to use

$$\tilde{P}_{RL} = 1 - \int_{-\infty}^{\log(RL - R_0)} f(x | \bar{x}, s^2) dx,$$

where $f(x | \bar{x}, s^2)$ is the density of $x$ conditional on knowing $\bar{x}$ and $s^2$. Lieberman and Resnikoff show that, after an appropriate change of variables, the density function $f(x | \bar{x}, s^2)$ is actually the density for a beta random variable. Thus the RPII use the following estimator

$$\tilde{P}_{RL} = 1 - \int_{-\infty}^{U} \frac{\Gamma(N - 2)}{\Gamma(N/2 - 1)\Gamma(N/2 - 1)} \zeta^{N/2 - 2} (1 - \zeta)^{N/2 - 1} d\zeta \quad (13)$$

where $\Gamma(x)$ is the Gamma function evaluated at $x$ and

$$U = \frac{1}{2} \left( 1 - \sqrt{\frac{N}{N-1} \left( \frac{\log(RL - R_0) - \bar{x}}{s} \right) } \right). \quad (14)$$

### 2.5. GammaLog model

All of the above models are based on the assumption of an underlying lognormal distribution with a background correction factor. If one does not correct for this background then radon levels are not well modelled by the lognormal distribution. Instead, as has been shown by Janssen and Stebbings (1992) such uncorrected, log-transformed indoor radon levels can instead be modelled well using the gamma distribution. As no correction is made in this model for outdoor background radon levels, the estimator for $P_{RL}$ is given by

$$P'_{RL} = 1 - \int_{-\infty}^{\log(RL)} \frac{1}{\theta \Gamma(\kappa)} x^{\kappa - 1} e^{-x/\theta} dx,$$

where $\theta$ and $\kappa$ are known as the scale and shape parameters respectively of the gamma distribution. As $\theta$ and $\kappa$ are unknown Janssen and Stebbings use the maximum likelihood estimates of these two parameters.
Table 1. Results of a simulation study to examine bias in different estimation procedures to compute the proportion of radon levels ($P_{RL}$) above the reference level of 200 Bq m$^{-3}$. Simulations produced 10 000 samples containing $N$ observations from an underlying lognormal model (equation (16)) with an assumed background outdoor radon level of 6 Bq m$^{-3}$. $P_{RL}$ was estimated for each sample and the tabulated values show the mean value of these 10 000 $P_{RL}$ values. In all cases the true proportion of observations above the reference level is 10.24%.

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimation equation</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal model</td>
<td>(4)</td>
<td>$N = 100$</td>
</tr>
<tr>
<td>Sort technique</td>
<td>(5)</td>
<td>$N = 50$</td>
</tr>
<tr>
<td>Bias corrected lognormal model</td>
<td>(10)</td>
<td>$N = 25$</td>
</tr>
<tr>
<td>Beta model</td>
<td>(13)</td>
<td>$N = 15$</td>
</tr>
<tr>
<td>Gamma model</td>
<td>(15)</td>
<td></td>
</tr>
</tbody>
</table>

3. Examination of bias in the different modelling approaches

The existence in the radon mapping literature of these five different estimation approaches poses the obvious question as to whether all of these procedures are equivalent or whether any one of the models is to be preferred. The first four estimation approaches described in the previous section all assume that indoor radon levels are correctly modelled by the lognormal process described in equations (1) and (2). Each of these approaches attempts to accurately estimate $P_{RL}$ in equation (3). Three of these approaches are based on approximations while the RPII approach, which utilises the beta distribution, makes no approximation but provides the theoretically correct solution to the problem of estimating $P_{RL}$ based on lognormally distributed data. To compare the performance of each estimation approach we constructed 10 000 data sets each consisting of $N$ observations simulated from a lognormal distribution $X = R_0 + \exp(\varepsilon)$, where the background level $R_0$ is set to 6 Bq m$^{-3}$ on the basis of observations made by the RPII (Fennell et al 2002) and $\varepsilon$ is normal with mean 4 and variance 1 (these values were chosen to produce simulated data which match closely the actual data from Castleisland). In this model the true proportion of observations above the reference level, $P_{RL}$, is 10.24%. We estimated $P_{RL}$ using each of the approaches for each of the 10 000 data sets and then computed the mean of the 10 000 estimates for each approach. The results of this exercise are contained in table 1.

What is immediately evident from the results in table 1 is that the beta model provides the most accurate unbiased estimates of $P_{RL}$. Recalling that we have simulated data from a lognormal distribution it is to be expected that the beta model performs best. This is because it addresses the problem of estimating $P_{RL}$ under the assumption that the underlying data are from a lognormal distribution from a correct theoretical standpoint.

In applying the method of Andersen et al (2001) we have estimated $\sigma$ using $s$ the standard deviation of the sample data. This is clearly not what was suggested by Andersen et al (2001) but we reiterate that our aim is to consider estimation procedures based on a single sample of data which do not make use of external data sources. All of the methods described here can provide better estimates if they make use of data external to the sample. Indeed the assumption of a constant standard deviation, which was used by Andersen et al (2001) has also been used by others including Miles (1994). The RPII itself makes use of a likelihood based smoothing technique (Fennell et al 1999) to enhance small sample data.
Comparison of statistical models in radon mapping

Table 2. 90% Confidence intervals for the proportion of observations ($P_{RL}$) above the reference level of 200 Bq m$^{-3}$. Simulations produced 10000 samples containing $N$ observations from an underlying lognormal model (equation (16)) with an assumed background outdoor radon level of 6 Bq m$^{-3}$. $P_{RL}$ was estimated for each sample and the tabulated values show the mean value of these 10000 $P_{RL}$ values. In all cases the true proportion of observations above the reference level is 10.24%.

<table>
<thead>
<tr>
<th>Sample size, $N$</th>
<th>Lognormal model</th>
<th>Sort technique</th>
<th>Beta model</th>
<th>Gamma model</th>
<th>Bias corrected lognormal model equation (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2.05%–21.48%</td>
<td>0.19%–20.87%</td>
<td>1.40%–21.69%</td>
<td>1.86%–19.93%</td>
<td>1.71%–20.71%</td>
</tr>
<tr>
<td>25</td>
<td>3.40%–18.70%</td>
<td>1.11%–18.87%</td>
<td>3.04%–18.77%</td>
<td>3.29%–17.69%</td>
<td>3.12%–18.22%</td>
</tr>
<tr>
<td>50</td>
<td>5.12%–16.16%</td>
<td>3.15%–16.87%</td>
<td>4.97%–16.16%</td>
<td>5.06%–15.52%</td>
<td>4.95%–15.92%</td>
</tr>
<tr>
<td>100</td>
<td>6.53%–14.28%</td>
<td>4.97%–15.04%</td>
<td>6.47%–14.27%</td>
<td>6.50%–13.82%</td>
<td>6.44%–14.16%</td>
</tr>
<tr>
<td>200</td>
<td>7.59%–13.11%</td>
<td>6.47%–13.71%</td>
<td>7.56%–13.11%</td>
<td>7.53%–12.77%</td>
<td>7.54%–13.06%</td>
</tr>
<tr>
<td>500</td>
<td>8.52%–12.09%</td>
<td>7.88%–12.54%</td>
<td>8.52%–12.08%</td>
<td>8.46%–11.83%</td>
<td>8.51%–12.06%</td>
</tr>
</tbody>
</table>

4. The relationship between sample size and error on estimation of $P_{RL}$

In the previous section we considered how best to estimate $P_{RL}$ with minimum bias. While Gunby et al (1993) discuss how to compute a confidence interval for an individual radon concentration level we note, with some notable exceptions (e.g. Price et al 1996, Andersen et al 2001), the general absence within the radon mapping literature of any computations of confidence intervals for the proportion of radon levels exceeding a reference level, $P_{RL}$. Using the simulated data sets described earlier we can compute such confidence intervals for different sample sizes for each of the estimation procedures. By considering these confidence intervals we can investigate the error associated with the estimates discussed in section 3 and we can better understand how sample sizes influence the accuracy of estimates of $P_{RL}$.

Table 2 contains 90% confidence intervals for $P_{RL}$ for various sample sizes from $N = 15$ to 500. Even with a confidence level of 90% it is clear that the errors on the estimates of $P_{RL}$ produced by each of the estimation procedures are unacceptably large for small sample sizes. Indeed the error bounds decrease very slowly with increasing sample sizes. We note that a popular approach to dealing with small sample sizes is to attempt to enhance the data in a given region by weighting those data with information from neighbouring regions through a smoothing algorithm (e.g. Fennell et al 1999). The results in table 2 demonstrate the drawbacks of using small sample sizes and reinforce the need to supplement small sample sizes through smoothing methods or other alternatives before deriving estimates for $P_{RL}$.

5. The effect of outliers on predictions of $P_{RL}$

We now consider how to best estimate $P_{RL}$ given the presence of extreme observations within a data set. The sort technique of Miles (1994) assumes an underlying lognormal distribution for indoor radon levels but recognises that the sample mean and standard deviation do not act as robust estimators in the presence of extreme values in the data. Table 1 demonstrated that this method suffers from a certain bias when estimating $P_{RL}$ in a true lognormal model but its real advantage is when dealing with more realistic data where deviations from the lognormal model are present.

The other models as previously described contain no mechanism for dealing with extreme values that deviate from the lognormal model. However, it is possible to adapt these methods by firstly screening the raw data for the presence of extreme values, removing these values and then
Table 3. Results of a simulation study to examine the effect of outliers on different estimation procedures to compute the proportion of radon levels ($\text{P}_{RL}$) above the reference level of $200 \text{ Bq m}^{-3}$. Table displays means and 90% confidence intervals from simulations that produced $10\,000$ samples containing $N$ observations from an underlying lognormal model (equation (16)) with an assumed background outdoor radon level of $6 \text{ Bq m}^{-3}$. Two extreme values of $6184 \text{ Bq m}^{-3}$ and $1263 \text{ Bq m}^{-3}$ were added to each simulated sample. $\text{P}_{RL}$ was estimated for each sample and the tabulated values show the mean value of these $10\,000 \text{P}_{RL}$ values. In all cases the true proportion of observations above the reference level in the underlying lognormal model is $10.24\%$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$N = 200$</th>
<th>$N = 100$</th>
<th>$N = 50$</th>
<th>$N = 25$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sort technique</td>
<td>10.8% (7.11%, 14.50%)</td>
<td>11.36% (6.11%, 16.7%)</td>
<td>12.49% (5.19%, 20.23%)</td>
<td>15.03% (4.11%, 25.77%)</td>
</tr>
<tr>
<td>Lognormal model without truncation</td>
<td>12.60% (9.96%, 15.34%)</td>
<td>14.78% (11.19%, 18.54%)</td>
<td>18.68% (13.97%, 23.81%)</td>
<td>25.00% (19.12%, 31.44%)</td>
</tr>
<tr>
<td>Lognormal model with truncation</td>
<td>9.79% (6.90%, 12.92%)</td>
<td>9.97% (5.77%, 14.79%)</td>
<td>10.49% (4.37%, 18.16%)</td>
<td>12.07% (2.70%, 23.33%)</td>
</tr>
<tr>
<td>Beta model with truncation</td>
<td>9.77% (6.87%, 12.92%)</td>
<td>9.93% (5.69%, 14.79%)</td>
<td>10.42% (4.21%, 18.19%)</td>
<td>11.96% (2.34%, 23.46%)</td>
</tr>
</tbody>
</table>

basing one’s estimation on the truncated data set. Our suggestion for determining which values should be identified as outliers and removed is based on the box-plot device that is commonly used in statistical analyses. According to this methodology values which lie further that 1.5 interquartile ranges from the upper and lower quartiles of the data are considered as outliers. Organo and Murphy (2007) utilised this new procedure in modelling radon measurements from Castleisland.

We will now demonstrate why it is important to account for extreme values in radon measurement data and investigate the relative performance of this new methodology compared to the sort technique when dealing with such extreme values. Following Miles (1994) we assume data are produced by an underlying lognormal model but with deviations from this model due to the presence of a small number of extreme observations above $400 \text{ Bq m}^{-3}$.

We simulated $10\,000$ data sets consisting of $N$ observations from a lognormal model as described by equation (16) and to each of these data sets we added two extreme values$^6$. We then estimated the proportion of homes above the reference level in the underlying lognormal model using four different methods. The first two methods involved applying Miles’ sort technique and the standard lognormal estimation method (equations (4)) to the raw data. For the next two methods we screened the data for outliers, removed any outliers which were found and then applied the lognormal and beta methods to the truncated data sets. The results of these estimation procedures are contained in table 3.

It is clear from these results that ignoring the presence of outliers within the data can introduce substantial bias into estimates of $\text{P}_{RL}$. Miles’ sort technique is very successful in reducing this bias and performs extremely well with large samples. However, it is also very clear from table 3 that by systematically removing any outliers from the data and then applying either the lognormal or beta methods to the truncated data set biases in $\text{P}_{RL}$ can be reduced even further. We note that the methodology we propose for identifying outliers removed more observations from each simulated data set than just the two outliers which we introduced making the truncated data set smaller than was absolutely necessary. However, despite this

---

$^6$ Two extreme values which were actually observed in the Castleisland survey ($6184$ and $1263 \text{ Bq m}^{-3}$) were chosen for the purpose of this simulation.
extra reduction in sample size the new procedure produced the least biased estimates when dealing with sample sizes up to \( N = 100 \).

6. Conclusions

A detailed description of five different models for estimating the proportion of homes, \( P_{RL} \), with indoor radon levels exceeding a reference level is presented. It is shown that the theoretically correct method for estimating \( P_{RL} \) is based on the beta distribution. Based on a true underlying lognormal model, the beta distribution produces the least biased estimates of \( P_{RL} \). However, the simple lognormal method also provides extremely good estimates. Confidence intervals are provided which display the relationship between sample size and the error on estimates of \( P_{RL} \). These results demonstrate clearly that small sample sizes should not be used to produce estimates for \( P_{RL} \) using any of the methodologies described in this paper. In such cases small samples may be supplemented with data from nearby regions through the use of smoothing algorithms.

The presence of outliers in a data set can introduce significant bias into estimation of \( P_{RL} \) and this should not be ignored. Using a robust estimation technique such as Miles’ sort technique can reduce this bias. However bias can be reduced even further by first detecting and removing outliers from the data and then applying the beta or lognormal method. This assumes that the sample data to be analysed contain observations which are generated by an underlying lognormal model together with outliers in the form of extreme observations above 400 Bq m\(^{-3}\). There is a need for improved estimation techniques for use in the case when the deviations from the lognormal model above 400 Bq m\(^{-3}\) are more substantial than the presence of a relatively small number of outliers.

References


Daly L 1994 personal communication


Gundersen L C S and Schumann R R 1996 Mapping the radon potential of the United States: examples from the Appalachian Environ. Int. 22 829–37

Janssen I and Stebbings J H 1992 Gamma distribution and house \(^{222}\)Rn measurements Health Phys. 63 205–8


Miles J C H 1993 Mapping the proportion of the housing stock exceeding a radon reference level 1st Int. Workshop on Indoor Radon Remedial Action (Rimini, June 1993)


Miles J C H and Appleton J D 2005 Mapping variation in radon potential both between and within geological units J. Radiol. Prot. 25 257–76


Organo C and Murphy P 2007 The Castleisland Radon Survey—follow-up to the discovery of a house with extremely high radon concentrations in County Kerry (SW Ireland) J. Radiol. Prot. 27 275–85

Price P N, Nero A V and Gelman A 1996 Bayesian prediction of mean indoor radon concentrations for Minnesota Counties Health Phys. 71 922–36