MAPPING INDOOR RADON-222 IN DENMARK: DESIGN AND TEST OF THE STATISTICAL MODEL USED IN THE SECOND NATION-WIDE SURVEY

Claus E. Andersen [1,*], Kaare Ulbak [2], Anders Damkjær [1], Peter Kirkegaard [1], Peter Gravesen [3]

[1] Risø National Laboratory, DK-4000 Roskilde, Denmark [2] National Institute of Radiation Hygiene, Knapholm 7, DK-2730 Herlev, Denmark [3] Geological Survey of Denmark and Greenland, Thoravej 8, DK-2400 Copenhagen NV, Denmark [*] Cooresponding author: Risø National Laboratory, Building NUK-125, DK-4000 Roskilde, Denmark, +45-4677 4912 (direct); Fax: +45-46774959; e-mail: claus.andersen@risoe.dk

In Denmark, a new survey of indoor radon-222 has been carried out. One-year alpha track measurements (CR-39) have been done in 3019 single-family houses. There is from 3 to 23 house measurements in each of the 275 municipalities. Within each municipality, houses have been selected randomly. One important outcome of the survey is the prediction of the fraction of houses in each municipality with an annual average radon concentration above 200 Bqm^{-3}. To obtain the most accurate estimate and to assess the associated uncertainties, a statistical model has been developed. The purpose of this paper is to describe the design of this model, and to report results of model tests. The model is based on a transformation of the data to normality and on analytical (conditionally) unbiased estimators of the quantities of interest. Bayesian statistics is used to minimize the effect of small sample size. In each municipality, the correction is dependent on the fraction of area where sand and gravel is a dominating surface geology. The uncertainty analysis is done with a Monte Carlo technique. It is demonstrated that the weighted sum of all municipality model estimates of fractions above 200 Bqm^{-3} (3.9 % with 95 %-confidence interval = [3.4,4.5]) is consistent with the weighted sum of the observations for Denmark taken as a whole (4.6 % with 95 %-confidence interval = [3.8,5.6]). The total number of single-family houses within each municipality is used as weight. Model estimates are also found to be consistent with observations at the level of individual counties. These typically include a few hundred house measurements. These tests indicate that the model is well suited for its purpose.

Keywords: Houses; Radon-222; Survey; Statistical model

INTRODUCTION

Radon is believed to cause an increased risk of lung cancer and it is therefore of interest to identify houses with high levels of indoor radon. It is important to know how many houses that have "high" levels (e.g. annual levels above 200 or 400 Bqm^{-3}) and it is important to know where these houses are located. Likewise, it is also of interest to know about the low-radon houses where there is no cause for alarm. This paper reports on a new Danish survey of indoor radon designed to tackle these problems. The survey is much larger than the first one from 1985/86 (Ulbak et al., 1988) and houses have been selected in a different way. The main objective in this paper is to describe the statistical model used in the new survey for prediction of fractions of houses above 200 and 400 Bqm^{-3}. The model is based on the work by Miles (1994, 1998) in the UK and Price and colleagues in the USA (1996). The task is to overcome what seems to be the main source of uncertainty in house-based radon surveys: The influence of small sampling sizes.

SURVEY DESIGN

Denmark is divided into 15 counties. Each county consists of a number of smaller municipalities. In total there are 275 municipalities. One-year alpha track measurements (CR-39) were done in 3019 single-family houses from December 1995 to December 1996. Detectors were placed in living

rooms. Within each municipality, houses were selected randomly by the Building and Dwelling Register (BBR). The median number of house measurements per municipality is 11. Nine municipalities have only 6 or less measurements, and nine municipalities have 18 or more measurements. The only geological information used directly in the model is the fraction of area (later referred to as g_k) in each municipality which is dominated by sand and gravel. These values are found by visual inspection of a map of the surface geology of Denmark (Pedersen et al., 1989). Except for Bornholm (where also granitic surface geology occurs) the fraction of area covered by clay (mainly glacial clayey till) is then $1 - g_k$.

MODEL

Transformations

We define the 'house concentration' c of a given house to be the average radon concentration of the living room and the bedroom: c_{Liv} and c_{Bed}. In this survey, we measured only the living-room radon concentration and we estimate the house concentration on the basis of the 1985/86 survey. We perform an unweighted linear regression analysis of the 1985/86 data: $\log(c) = a_0 + a_1 \log(c_{Liv})$ where log is the natural logarithmic function and c is calculated as the average of c_{Liv} and c_{Bed}. The fitted coefficients: a_0 = 0.227 (standard error 0.064) and a_1 = 0.922 (standard error 0.016) are used to convert the 3019 living-room measurements in the new survey.

In line with many other surveys, it is found that $\log(c)$ is (relatively) well described by a normal distribution function. A Lilliefors test of normality can however be rejected since p <0.01. Further examination of the data shows that the transformed radon concentration x: $x = \log(c + b)$ where b =8.0 Bqm^{-3}, is closer to normality. All of the statistical analyses are therefore conducted for transformed radon concentrations x.

Distribution parameters

It is assumed that within each municipality k, the transformed radon concentration x is normally distributed with a (true) mean μ_k and a (true) standard deviation σ. We allow that μ_k can be different from one municipality to another, but require that all municipalities have the same σ. The latter requirement is supported by an analysis of the homogeneity of variances with a modified version of the Levene test based on absolute deviations from the municipality medians of transformed radon concentrations (Manly, 1997). That test could not be rejected (p =0.18).

The estimator $\hat{\sigma}$ of σ is found as follows: First, we calculate the simple mean $\bar{x}_k$ and standard deviation s_k of the N_k measurements in each municipality k:

$$\bar{x}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i \qquad (1)$$

and

$$s_k = \sqrt{\frac{1}{N_k - 1} \sum_{i=1}^{N_k} (x_i - \bar{x}_k)^2} \qquad (2)$$

Then unbiased estimates of the population standard deviations $\hat{\sigma}_k$ are obtained from (Sokal and Rohlf, 1995, p. 53):

$$\hat{\sigma}_k = C_k \, s_k \tag{3}$$

where

$$C_k = \sqrt{\frac{N_k - 1}{2}} \; \frac{\Gamma(\frac{N_k - 1}{2})}{\Gamma(\frac{N_k}{2})} \tag{4}$$

For example, if $N_k = 5$ we obtain $C_k = 1.064$. Finally, we pool the 275 $\hat{\sigma}_k$-values into a single weighted mean value: $\hat{\sigma}$. The number of house measurements (N_k) is used as weight. The value amounts to: $\hat{\sigma} = 0.59418$.

The estimators $\hat{\mu}_k$ of μ_k are found as follows: A simple estimate would be to let $\hat{\mu}_k = \bar{x}_k$. However, as demonstrated by Price et al. (1996), we can improve this estimate on the basis of Bayesian statistics. The essential trick is to look at the distribution of $\bar{x}_k$ for all municipalities. If $\bar{x}_k$ in a specific municipality deviates much from the typical value, we will adjust our $\hat{\mu}_k$ accordingly. As it is known that geology is an important factor of indoor radon (Ulbak et al., 1988), we will use geological information for the adjustment. First we conduct a linear regression:

$$x_k = \beta_0 + \beta_1 \, g_k + \varepsilon_k \tag{5}$$

where g_k is an estimate of the fraction of the total area of municipality k that has a surface geology dominated by sand and gravel. Based on all 275 municipalities, the regression coefficients amount to $\beta_0 = 4.54$ (standard error 0.0296) and $\beta_1 = -0.69$ (standard error 0.06). The R-squared value is 36 %. The variance σ_ε^2 of the residuals ε_k is 0.082. For each municipality, we calculate: $\theta_k = \beta_0 + \beta_1 \, g_k$ and use the following weighted average as the model estimate of μ_k:

$$\hat{\mu}_k = \frac{\omega_k \bar{x}_k + \omega_0 \theta_k}{\omega_k + \omega_0} \tag{6}$$

where the weights are: $\omega_k = N_k / \hat{\sigma}^2$ and $\omega_0 = 1/\sigma_\varepsilon^2$. Essentially, we estimate μ_k to be equal to the observed value $\bar{x}_k$ with some weigthed correction towards what on-the-averaged is found for municipalities with that type of surface geology. If there are few (or no) measurements in a municipality, then $\hat{\mu}_k \approx \theta_k$. If there are many measurements, then $\hat{\mu}_k \approx \bar{x}_k$. Essentially, the influence of θ_k equivalents about 4 extra measurements in each municipality. The main source of uncertainty in the survey is the small sample size. We apply equation (6) as a way to gently "stabilize" modelling results in all municipalities except those on the island Bornholm.

f_{200}-estimation

If x is normally distributed with a (true) mean μ_k and a (true) standard deviation σ, the (true) fraction of houses in municipality k with concentration above 200 Bqm^{-3} is:

$$f_{200,\text{true}}(k) = 1 - \Phi\left(\frac{\log(208) - \mu_k}{\sigma}\right) \tag{7}$$

where Φ is the cumulative distribution function for the normal distribution N(0,1). A straightforward estimator of $f_{200,\text{true}}$ for a given municipality would be to substitute μ_k and σ with $\hat{\mu}_k$ and $\hat{\sigma}$, respectively (as derived in the previous section). It can, however, be shown that on the average this does not give the correct result. The result is biased because of the non-linear nature of Φ. As shown in the appendix, an unbiased estimator of $f_{200,\text{true}}$ in a given municipality can be found as follows: We calculate u_k:

$$u_k = \frac{\log(208) - \hat{\mu}_k}{\hat{\sigma}} \tag{8}$$

and the bias term:

$$B_k = \frac{-1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u_k\right) \frac{u_k}{2 N_k} \tag{9}$$

and insert into:

$$f_{200,m}(k) = 1 - \Phi(u_k) + B_k \tag{10}$$

$f_{200,m}(k)$ is the (unbiased) model value for the fraction of houses in municipality k with radon concentrations above 200 Bqm^{-3}.

Monte Carlo estimates of confidence intervals

We use a Monte-Carlo technique to assess confidence intervals for model output. The procedure works as follows: First, we calculate $\bar{x}_k$, $\hat{\sigma}$ and $f_{200,m}(k)$ for all 275 municipalities. Second, we generate a set of 3019 synthetic "measurement results" with a random generator corresponding to a new national survey. In municipality k, we draw N_k random x-values from a normal distribution function with mean $\bar{x}_k$ and standard deviation $\hat{\sigma}$. With this synthetic data set we calculate $f_{200,m}(k)$ and other statistical information. Only one thing is different from the treatment of the real measurements: the θ_k-values used in equation (6) are sampled randomly from a normal distribution with mean $\beta_0 + \beta_1 g_k$ and variance σ_ε^2. Hence the uncertainty of θ_k are taken into account as required in a full Bayesian analysis (Price et al., 1996). All results are stored. A new synthetic data set is generated and new results are calculated and stored. In total, 2000 Monte-Carlo realisations are generated in this way. The lists of synthetic $f_{200,m}$-values are sorted by size. The ranges that contain the middle 95 % of the elements in each list defines a biased 95 %-confidence interval for $f_{200,m}(k)$. The bias is caused by the following: The true "$f_{200,m}(k)$-value" in the Monte-Carlo simulations will be:

$$f_{\mathrm{MC}}(k) = 1 - \Phi\left(\frac{\log(208) - \bar{x}_k}{\hat{\sigma}}\right) \tag{11}$$

which is different from the observed value given by equation (7). Hence the bias can be removed by a subtraction of $f_{\mathrm{MC}}(k) - f_{200,\mathrm{true}}(k)$. It is assumed that the observed (list) confidence interval will relate to f_{MC} in the same way as the wanted confidence interval relates to $f_{200,\mathrm{true}}$.

RESULTS

In the survey, house radon levels (c) in the range from 2 to 590 Bqm^{-3} were observed. Results for all 15 counties are shown graphically in the top plot of Figure 1. Three counties have estimated f_{200} levels below 1 %. These counties have sand and gravel as a prevailing surface geology. The county with the highest level (f_{200}=12 %) partly has a granitic surface geology. Clayey till is a dominant surface geology in the remaining counties. The bottom plot in Figure 1 shows typical municipality results for one of the "sand-and-gravel" counties. Estimates of f_{200} range from 0.0 to 2.6 %. In total, 261 house measurements were conducted in the county. Only one house had a level above 200 Bqm^{-3}. It is assessed that 0.8 % of the single family houses in the county are above 200 Bqm^{-3}. The middle plot of Figure 1 shows results for a typical "clayey till" county. f_{200}-estimates range from 0.4 to 12 %. Out of 248 measurements, 11 were found to be above 200 Bqm^{-3}. It is assessed that 7.2 % of the single-family houses in the county are above 200 Bqm^{-3}.

DISCUSSION

Improved estimates by modelling?

The primary purpose of the statistical model is to provide estimates of the fraction of houses above 200 Bqm^{-3} at the level of individual municipalities. The idea is to make estimates that are better (i.e. more accurate and less variable) than estimates deduced from simple observations: $f_{200} = N_{200,k} / N_k$, where $N_{200,k}$ is the observed number of houses with c>200 Bqm^{-3} in municipality k, and N_k is the number of measurements. The main problem with such simple observations is that for the typical case of about 10 house measurements per municipality, the outcome will be in steps of 10 % (i.e. 0 %, 10 %, 20 % etc.). This can be illustrated with synthetic data. We draw 3019 synthetic (transformed radon concentrations) x from a normal distribution with mean 4.33 and standard deviation 0.5941. Subsequently we transform the data to ordinary radon concentrations (c-values) using the inverse of $x = \log(c + b)$. The true value of f_{200} in this case is 4.60 % (about the same as the national average). The data are grouped in municipalities and counties exactly as in the survey (this is important as the number of measurements determines the variability of parameter estimates). Also, we preserve the fraction of sand and gravel (g_k) which is needed in equation (6). In this case, however, the regression (see equation 5) will only be by chance. The model is applied exactly as with the real data set. To evaluate the importance of the Bayesian correction, we will also consider simplified-model estimates where ω_0 in equation 6 is set to 0 (such that $\hat{\mu}_k = \bar{x}_k$). The results for the 275 municipalities are shown in Figure 2. One curve (labelled observed) shows the observed fraction of houses with c>200 Bqm^{-3} ($N_{200,k} / N_k$). The

mean and standard deviation of the results are 4.9 % and 7.2 %, respectively. In one case, f_{200} is found to be as high as 40 %. It is particularly problematic that about 60 % of the municipalities are without measured houses with concentrations above 200 Bqm^{-3}. It is little help that many of the remaining municipalities, have observed fractions above 10 %, such that on-the-average the correct result of about 4.6 % is observed. The curved labelled simplified model are the results of model estimates without the Bayesian correction ($\omega_0 = 0$). Compared with the first curve, these estimates are much better in the sense that the results are less variable (mean 4.9 % and standard deviation 3.7 %). The final curve labelled full model present by far the best estimates (mean 4.3 % and standard deviation 1.7 %). However, because the data in each municipality come from the same distribution, the variance of the regression residuals (ε in equation 5) is lower than in the real survey. This means that in this (synthetic) example, the Bayesian correction will correspond to about 9 extra measurements in each municipality (compared to 4 in the real situation). The confidence intervals of the simulations are not shown in the Figure 2. In summary, we can report, however, that the true value (4.6 %) is within the confidence intervals in all (275) cases of the simple observation model, in 95 % of the cases for simplified model estimates, and in all but 3 for the full model. These observations indicate, that the Monte-Carlo estimated confidence intervals are valid (or at least not too small).

Model versus measurements

The (weighted) national average of model predictions amounts to $f_{200,m}$ = 3.9 % with CI(95 %)=[3.4,4.5]. This is not significantly different from the observed fraction f_{200} = 4.6 % with CI(95 %)=[3.8,6.6]. Even if we apply the model for prediction of the fraction of houses above 400 Bqm^{-3}, it is found that there is an insignificant difference between the model estimate 0.21 % with CI(95 %)=[0.15,0.28] and the observed result 0.38 % with CI(95 %)=[0.17 %,0.66 %]. The latter agreement (that concerns the tail of the distribution) suggests that the assumption of normality is not greatly violated. As shown in the top plot of Figure 1, there is also good agreement between modelling results and observed values in the 15 individual counties. Even though model assumptions (such as those concerning normality and homogeneous variance) may not be perfect, the model does not seem to be strongly biased: On-the-average, the model accounts well for data at the level of individual counties and for Denmark as a whole. Therefore we believe that the model is reasonable also at the level of individual municipalities.

To illustrate how the model treats counties with different types of geology it is of interest to study Figure 1. The middle plot of the figure shows results for County no. 13. In this county, clayey till is the predominant geology. From the model, it is estimated that 7.2 % of the houses are above 200 Bqm^{-3} (CI(95 %)=[5.7,9.0]). This value is the third largest county value in Denmark (see the top plot). In comparison, the bottom plot of the figure shows results for County no. 3. Here the prevailing geology is sand and gravel. From the model, it is estimated that 0.8 % of the houses are above 200 Bqm^{-3} (CI(95 %)=[0.4,1.4]). This value is the third lowest county value in Denmark (see the top plot). In short, the model predicts that f_{200} for these counties differ by one order of magnitude, so at the county level the model can certainly resolve such geological differences. More importantly, this is not just because the model works correctly "on-the-average": All $f_{200,m}$ estimates for municipalities in the low-level county (county no. 3) are lower than all (except two) of the high level county (county no. 13). This results is not a consequence of the Bayesian correction (that uses geological information). If we remove the Bayesian correction (by setting

$\omega_0 = 0$) in equation 6, we obtain about the same result. Observe, that the relatively large scatter of values for the "clayey till" county could result from true differences among municipalities or from random fluctuations. The latter can be seen from the associated uncertainty predictions as well as the example in Figure 2.

Model elements

The model includes some special elements: (1) the offset b = 8 Bqm^{-3} in transformation $x = \log(c + b)$, (2) the C_k correction in equation 3, (3) the bias B correction in equation 10, and (4) the Bayesian correction in equation 6. These elements have been added to the model for the reasons given in the text. However, in retrospect it is interesting to investigate the importance of these elements. One benchmark is to compare model predictions with the observed fraction of houses above 200 Bqm^{-3} for Denmark taken as a whole. This is shown in Table 1. All results are weighted with the number of single-family houses in each municipality. The table shows result for 8 cases. Case 8 corresponds to the (full) model. In the other seven cases, one or more of the model elements have been turned off. Model estimates range from 3.9 to 5.5 % which is actually not significantly different from the observed value of 4.6 % when the associated confidence intervals are considered. This shows that the core of the model (assumption of normality and homogeneity of variance) gives the right answer on-the-average and that the four model elements tested in Table 1 provide relatively small refinements. For example, changing the offset from 8 to 7 Bqm^{-3} changes the (average) model estimate from 4.3 to 4.4 %. It is observed that the best agreement with the observed value is not for the full model (case 8): The best fit is for case 7 where the bias correction (B in equation 10) has been switched off. One explanation for this could be that the expression for B were derived on the assumption that the distribution parameters μ_k for each of the municipalities were found on the basis of only the measurements within each municipality. In the full model, the Bayesian correction makes the μ_k -estimates less variable, and therefore the bias B in equation 10 may overcorrect the problem. The Bayesian correction has an important impact on the uncertainty that we assign to the model predictions at the level of individual municipalities. Figure 3 shows the results for county no. 13 without the Bayesian correction. In comparison with the middle plot of Figure 1, it can be seen that the estimates in Figure 3 are more variable and that the associated uncertainties are much larger. The Bayesian correction is particularly important for the municipalities with very few (e.g. less that 6) measurements.

Measurement uncertainty

Considerable measurement uncertainty is associated with the c -estimates (typically about 20 %). Part of this comes from the conversion from living room to house concentrations. Such uncertainties tend to have little impact on averages of quantities that relates linearly to the measurements (e.g. arithmetic means) as such random errors on the average will tend to cancel each other. Unfortunately, estimation of the fraction of houses above 200 Bqm^{-3} is an non-linear function of the individual radon concentration results, and random errors therefore will bias the estimation. This has previously been demonstrated by Miles (1994), and he uses a sort technique to overcome the problem. Because of the few measurements per municipality in this investigation, we have not adopted that method. To assess the importance of the problem, we took the original 3019 measurement results and added random (gaussian) noise with zero mean and standard deviation equal to the uncertainty of the measurement results. For Denmark as a whole (weighted by number of single-family houses in each municipality), this gave an observed fraction f_{200} $(= N_{200,k} / N_k)$

equal to 5.7 % (95 % confidence interval = [4.8,6.7]). For comparison, it is recalled that the value without added noise was 4.64 % (CI = [3.8,5.8]). For the model, we found that the results with the noise was: $f_{200,m}$ =4.7 % (CI = [4.1,5.4]). The value without noise was: 3.9 % (CI = [3.4,4.5]). These results indeed show that measurement errors can bias the survey results. For the national average, the bias amounts to about 0.8 %-points. At the level of individual municipalities, the main concern, however, relates to the effect of small sample sizes.

CONCLUSION

A statistical model has been developed. It predicts the fraction of single-family houses (in each municipality) with an annual radon level above 200 Bqm^{-3}. The investigation suggests that these estimates are better (more accurate and less variable) than simple observations based on direct observation of houses with levels above 200 Bqm^{-3}. Also, the model provides estimates of uncertainties associated with these predictions. The main source of uncertainty relates to the small sample size (typically only about 11 measurements in each municipality). Comparison between model predictions and measurements indicated that the model is well suited for mapping of indoor radon in Denmark.

ACKNOWLEDGEMENT

The project was supported financially by the Danish Ministry of Health.

APPENDIX: BIAS TERM

Let the stochastic variable X (the transformed radon concentration) come from a normal distribution $X \in N(\mu,\sigma^2)$. Then X has the density:

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) \tag{12}$$

and the distribution function:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{13}$$

where:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right) \tag{14}$$

and:

$$\Phi(z) = \int_{-\infty}^{z}\varphi(t)\mathrm{d}t \tag{15}$$

are the density and distribution functions of the standardized normal distribution $N(0,1)$. We are given a (random) sample $x_1,\ldots,x_N$ of size N (e.g. $N=11$) from the distribution, and we are told

the exact value of the (true) distribution parameter σ. In the survey, we do not know the true value of σ, but as outlined in the text, we apply all 3019 house measurements in the estimate $\hat{\sigma}$ = 0.59418. Hence, our σ estimate depend on many more degrees of freedom than N, and it is therefore essentially independent of the specific sample in question. From this information, we want to calculate the fraction f_L of X that exceeds a certain action level x_L (e.g. $x_L = \log(208)$).

First, we consider the fraction $p_{\mathrm{L}} = 1 - f_{\mathrm{L}}$ of houses below x_{L}. Given μ, the true answer is:

$$p_{\mathrm{L,true}} = F(x_{\mathrm{L}}) = \Phi\left(\frac{x_{\mathrm{L}} - \mu}{\sigma}\right) \tag{16}$$

As we do not know μ, we compute the sample average:

$$x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{17}$$

and estimate p_{L} as:

$$p_{\mathrm{L,b}} = \Phi\left(\frac{x_{\mathrm{L}} - \bar{x}}{\sigma}\right) \tag{18}$$

Although $\bar{x}$ is an unbiased estimator of μ, $p_{\mathrm{L,b}}$ is not an unbiased estimator of $p_{\mathrm{L,true}}$. This is because of the non-linear nature of Φ. This means that if we were provided with many samples of size N, then the average of our $p_{\mathrm{L,b}}$-estimates = $\langle p_{\mathrm{L,b}} \rangle$ would not converge to $p_{\mathrm{L,true}}$. The bias $B = \langle p_{\mathrm{L,b}} \rangle - p_{\mathrm{L,true}}$ will depend on the sample size N and on the action level x_L. The purpose of the following is to derive an expression for $B = B(N, x_{\mathrm{L}})$ such that we can make an improved estimate of p_{L}:

$$p_{\mathrm{L,ub}} = p_{\mathrm{L,b}} - B \tag{19}$$

with the property that $\langle p_{\mathrm{L,ub}} \rangle \approx p_{\mathrm{L,true}}$. We find $\langle p_{\mathrm{L,b}} \rangle$ by integration over all values of $\bar{x} \in N(\mu, \sigma^2/N)$:

$$\langle p_{\mathrm{L,b}} \rangle = \int_{-\infty}^{+\infty} \Phi\left(\frac{x_{\mathrm{L}} - \bar{x}}{\sigma}\right) \varphi\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{\mathrm{N}}}\right) \mathrm{d}z \tag{20}$$

We then introduce: $u = (x_{\mathrm{L}} - \mu)/\sigma$ and $t = (\bar{x} - \mu)/\sigma$ such that:

$$\langle p_{\mathrm{L,b}} \rangle = \sqrt{N} \int_{-\infty}^{+\infty} \Phi(u - t)\, \varphi\left(\sqrt{N}\, t\right) \mathrm{d}z \tag{21}$$

Our evaluation of equation 21 involves the detour of first taking the derivative of the integrand with respect to u:

$$\frac{\mathrm{d}\langle p_{\mathrm{L,b}} \rangle}{\mathrm{d}u} = \sqrt{N} \int_{-\infty}^{+\infty} \varphi(u - t)\varphi\left(\sqrt{N}\, t\right) \mathrm{d}t = A \int_{-\infty}^{+\infty} \varphi\left(\frac{t - t_0}{s_0}\right) \mathrm{d}t = A \tag{22}$$

where we (after some simple manipulations) find the constants:

$$t_0 = \frac{u}{1+N}, \quad s_0 = \frac{1}{\sqrt{1+N}}, \text{ and } \quad A = \varphi\left(\sqrt{\frac{N}{N+1}}\, u\right)$$

We therefore have that:

$$\left\langle p_{\mathrm{L,b}} \right\rangle = \Phi\left(\sqrt{\frac{N}{1+N}}\, u\right) + C \tag{23}$$

where C is an integration constant. Evaluation of equation 21 at $u \to -\infty$ shows that $C = 0$. With Taylor expansion of Φ in mind we write:

$$\varepsilon = \left(\sqrt{\frac{N}{1+N}} - 1\right) u \approx -\frac{1}{2N} u \tag{24}$$

such that:

$$\left\langle p_{\mathrm{L,b}} \right\rangle = \Phi(u+\varepsilon) = \Phi(u) + \varphi(u)\varepsilon + \ldots \tag{25}$$

Finally, we find the bias B:

$$B = \left\langle p_{\mathrm{L,b}} \right\rangle - \left\langle p_{\mathrm{L,true}} \right\rangle \approx -\varphi(u)\frac{u}{2N} \tag{26}$$

and with the approximation: $u \approx (x_L - \bar{x})/\sigma$ we have the result used in equation 10.

REFERENCES

[1] Manly BFJ. Randomization, bootstrap and Monte Carlo methods in biology. Second ed. Chapman & Hall (1997).

[2] Miles JCH. Mapping the proportion of the housing stock exceeding a radon reference level. Radiat Prot Dosim 1994;56(1-4);207-210.

[3] Miles J. Mapping radon-prone areas by lognormal modeling of house radon data. Health Physics 1998;74(3);370-378.

[4] Pedersen SAS, Rasmussen LAa, Petersen KS, Salinas I. Surface geology map of Denmark. 1:200000 (4 map sheets). Geological Survey of Denmark and Greenland; 1989.

[5] Price PN, Nero AV, Gelman A. Bayesian prediction of mean indoor radon concentrations for Minnesota counties. Health Physics 1996;71(6);922-936.

[6] Sokal RR, Rohlf FJ. Biometry: The principles and practice of statistics in biological research, third edition. W.H. Freeman and Company, New York, 1995.

[7] Ulbak K, Stenum B, Sørensen A, Majborn B, Bøtter-Jensen L, Nielsen SP. Results from the Danish Indoor Radiation Survey. Radiat Prot Dosim 1988;24(1/4);401-405.

Table 1: Influence of four model elements: (1) the offset b, (2) the C_k correction, (3) the bias correction B, and (4) the Bayesian correction on the weighted mean $f_{200,m}$ for Denmark as a whole. For comparison the observed value of f_{200} is also given. The distribution parameter $\hat{\sigma}$ (see equation 3) is listed in the last column.

Case	Model elements				$f_{200,m}$	95% Conf. interval	$\hat{\sigma}$
	b	C_k	B	Bayesian			
	Bqm^{-3}				%	%	
0	0	off	off	off	5.52	[4.9,6.2]	0.682
1	0	on	on	off	5.23	[4.6,5.9]	0.699
2	7	on	on	off	4.42	[3.9,5.0]	0.604
3	8	on	on	off	4.34	[3.8,4.9]	0.594
4	9	on	on	off	4.26	[3.7,4.9]	0.584
5	8	on	off	off	4.93	[4.3,5.5]	0.594
6	8	off	off	off	4.03	[3.5,4.6]	0.579
7	8	on	off	on	4.49	[3.9,5.1]	0.594
8	8	on	on	on	3.89	[3.4,4.5]	0.594
Observed					4.64	[3.8,5.6]	

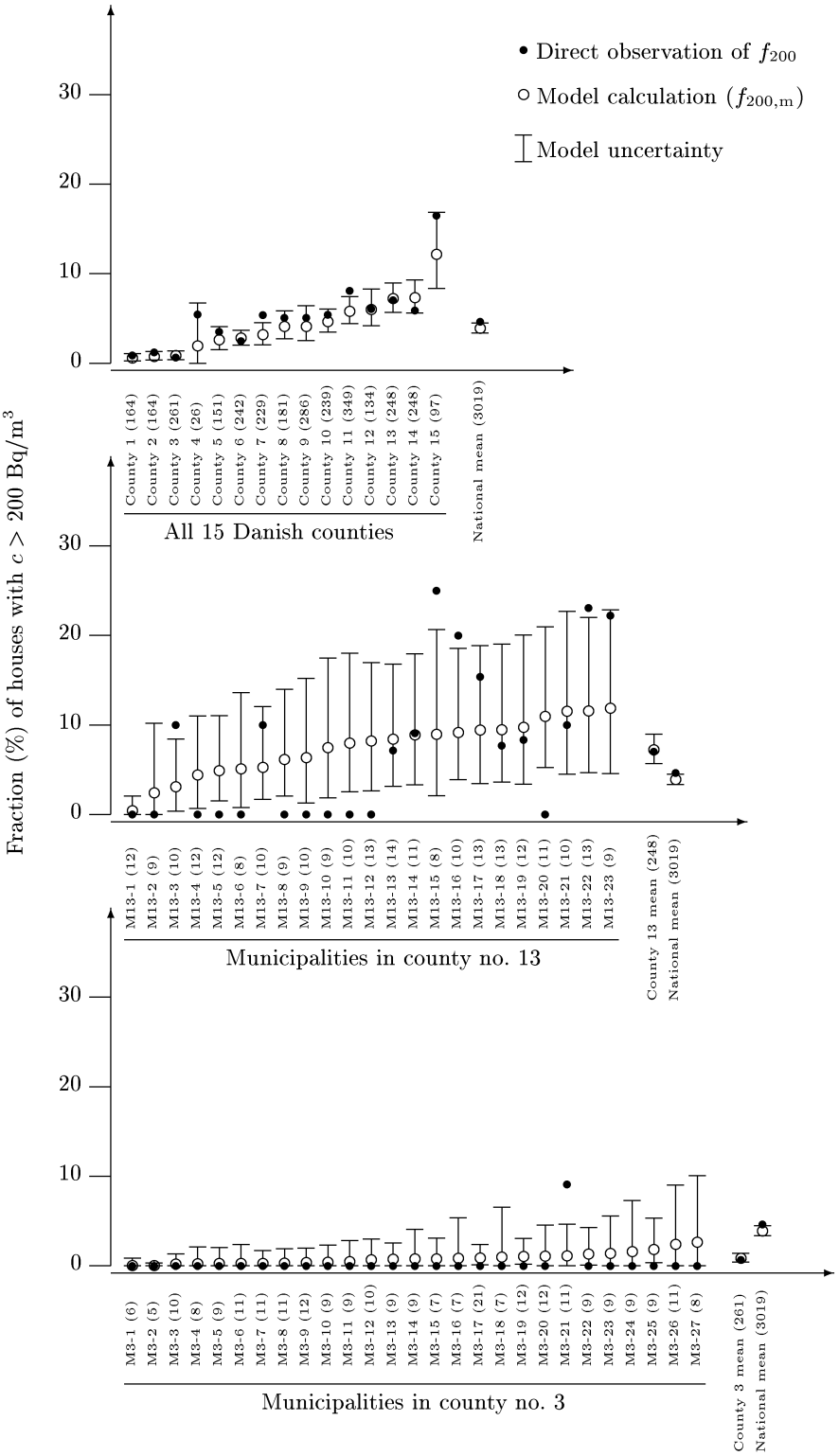

Figure 1: Survey results ($N_{200,k} / N_k$) and model estimates ($f_{200,m}$) of the fraction of houses above 200 Bq m^{-3}. TOP: Results for all 15 Danish counties. MIDDLE: Results for the municipalities in a high-concentration county (county no. 13). BOTTOM: Results for the municipalities in a low-concentration county (county no. 3). The numbers of house measurements are shown in parentheses.

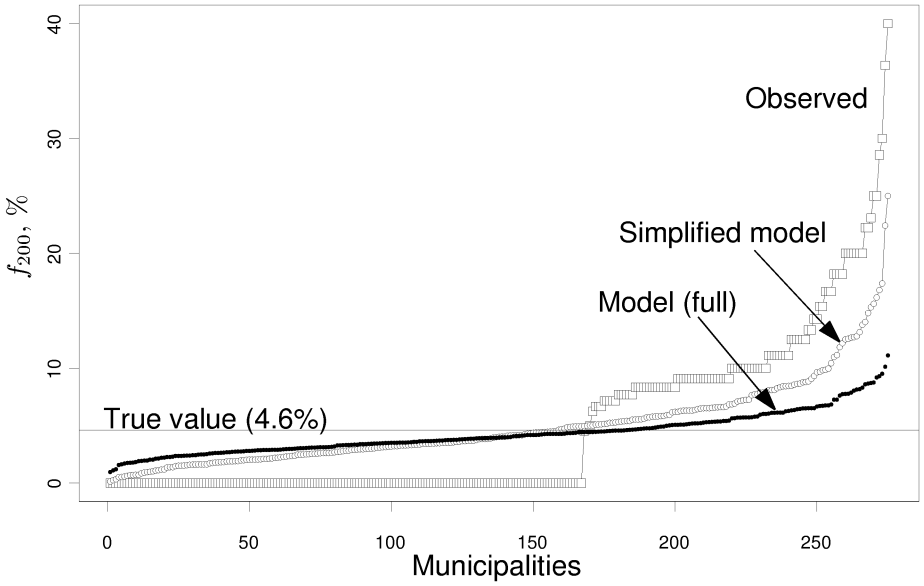

Figure 2: Test with synthetic data: Comparison between model estimates ($f_{200,m}$) and observed values for f_{200} in 275 municipalities when the true fraction above 200 Bqm^{-3} is 4.6 %. The curve labelled simplified model corresponds to the situation without the Bayesian correction.

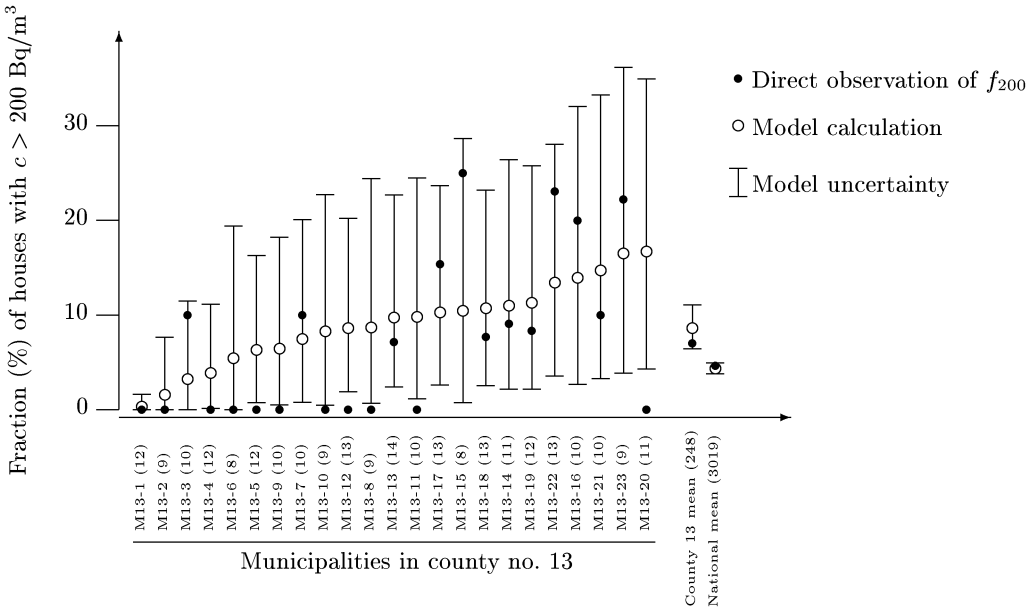

Figure 3: Modelling results for County no. 13 without the Bayesian correction. The numbering of municipalities is identical to that of Figure 1. The numbers of house measurements are shown in parentheses.