Probabilistic estimate of indoor radon distribution in Abruzzo (central Italy): comparison of different statistical methods

Journal of Physics: Conference Series **1391** (2019) 012001 doi:10.1088/1742-6596/1391/1/012001

Probabilistic estimate of indoor radon distribution in Abruzzo (central Italy): comparison of different statistical methods

A Pasculli[1], F Rizzo[1], P Zazzini[1]

[1]Department of Engineering and Geology (InGeo), University of G. D'Annunzio, Chieti-Pescara, Via dei Vestini 31 (Italy)

antonio.pasculli@unich.it

Abstract. Since some years ago in Abruzzo (Central Italy), through a number of monitoring campaigns, a set of more than 1900 indoor radon measures has been acquired by the Regional Agency for the environmental protection. Thus, on the basis of these public experimental data, different statistical approaches, aimed to estimate the probability to exceed the level of 200 Bq/m^3 (lower than 300 Bq/m^3, threshold value currently recommended by the Euratom commission for indoor radon risk acceptability), taken just as a working reference value, have been selected and discussed in this paper. Essentially, 'Monte Carlo Empirical Bayesian approach', 'Bootstrap' and 'Gibbs samplers' methods have been applied and the results have been partially compared. Moreover, some insights on the minimum number of samples, needed to assess the probability distribution as reasonable as possible, are provided.

1. Introduction

Radon-222 is a noble natural gas belonging to the radioactive chain starting from Uranium-238, an ubiquitous trace component of the terrestrial crust. The ground on which the buildings are located is the primary source of this kind of radioactive gas that, by transport and diffusion through pore space and fractures, may reach the internal environment of buildings. Its inhalation exposes bronchial epithelium to alpha radiation and, by consequence, may be a possible cause of lung cancer. Minor roles are played by building materials, well water supplies and combustion of natural gas for cooking/heating. The data were obtained by one year exposure of solid state detectors located in selected dwellings. Then, the measures had to undergo some manipulation before analysis, including subtraction of mean outdoor concentration, normalization for seasonal variations (if necessary) and standardization to a virtual ground level condition. Many related consideration have been included and discussed in an previous paper [1]. Due to the limited number of available data, the identification of possible areas that may be at Radon risk requires, necessarily, the application of statistical methods. To this purpose, an important tool is the use of the Monte Carlo approach, widely exploited in many different fields ([2], [3]). Also, the spectral representation of a random variable or polynomial chaos expansion was employed in common research practice ([4], [5]). Other approaches, like morphometric index and fractals, found interesting applications in spatial parameters correlation [6] and in time series analysis [7] respectively. In our paper, however, we assumed the hypothesis of time-stationarity of the variables. Moreover, our paper is focused on the comparison among selected statistical approaches and the analyses of radon potential based on environmental geology and geographically weighted regression, already proposed in [8], [9] and [10], were not included in the present paper. In order to perform the statistical simulations

discussed in the following sections, a computer code was developed using the Fortran Compiler, beside other free software like *R* and *WinBugs*. Useful insights have been shared with PhD dissertations [11] and [12].

2. Multilevel modeling for indoor radon

A simple two-level model was selected, in which a categorical variable identifies the connection of the single *j-th* R_{ij} measure to a certain municipal territory *i* (level I), while the upper level of the model is represented by the averages α_i (level II) of the logarithms of concentrations at the *i*-th municipal territory level:

$$R_{ij}=\alpha_i+\varepsilon_{ij} \qquad \alpha_i \sim G\left(\mu_\alpha,\sigma_\alpha^2\right) \qquad \varepsilon_{ij} \sim N\left(0,\sigma_R^2\right) \tag{1}$$

The averages α_i and the residual error ε_{ij} were assumed to belong to the Gaussian and to the Normal Gaussian distribution (as common practise), respectively. The former was characterized by mean μ_α and variance σ_α^2, while the latter by variance σ_R^2, whose values should be inferred by means of experimental data. This α_i parameter (or regression coefficient at the lowest level of the model) is usually called 'intercept', as it represents a typical value (average) of the variable *R* within a group (in our case, of a municipal territory), while μ_α, σ_α^2 and σ_R^2 are known as the 'hyperparameters' of the model and concern the hierarchically highest level of the model itself [13]. The model can be rewritten in a more compact formulation, in which the index-variable *i* [*j*] codifies the belonging of the single datum to the municipal territory (in practice, *i* [23] = 3 means that the 23rd unit (measure) in the data belongs to the group, municipality, 3):

$$R_j \sim G\left(\alpha_{i[j]},\sigma_R^2\right) \text{ for } j = 1,\ldots, n_i; \qquad \alpha_i \sim G\left(\mu_\alpha,\sigma_\alpha^2\right) \text{ for } i = 1,\ldots, I \tag{2}$$

namely, the *j*-th measure, among the total *n* measures which were carried out in a site located within the *i*-th municipality (the total number of which was *I*), belongs to a Gaussian distribution $G\left(\alpha_{i[j]},\sigma_R^2\right)$, whose average in turn belongs to another Gaussian distribution $G\left(\mu_\alpha,\sigma_\alpha^2\right)$ as described by the expressions (1). The mean reason to select a Gaussian distribution for the possible value of α_i was to bring the intercepts estimates closer to the global average μ_α in such a way that, in each *i*-th municipality, a compromise (pooling) was obtained between μ_α and the sample mean of the available n_i experimental data: $\bar{R}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}$. This kind of compromise may be theoretically supported by Bayes' theorem, selected in our paper. The following assumptions were adopted: the natural logarithms $ln(R_{ij})$ were supposed to be normally distributed within each geographic unit *i*. Accordingly, the R_{ij} values were assumed to follow a log-normal statistic, while $ln(GM_i)$ ($\equiv \alpha_i$ in (1) and (2)) and k^2 ($\equiv \sigma_R^2$ in (1) and (2)), were, respectively, the mean and variance of associated distributions. These parameters were calculated for each unit, considering the Geometric Mean (*GM*) and the Geometric Standard Deviation (*GSD*) of the measured ensemble data. whose size was very different from case to case. Of course, in those units where very few data were available, the resulting uncertainty related to the estimated *GM*s and *GSD*s can be very high. Also for this reason, the selection of a well-tested Bayesian inference method was further appropriate [14]. Moreover, this kind of approach, including also the measures acquired in the other units, improved the evaluation of *GM* and *GSD* related to each municipality. It worth to notice that k^2 ($\equiv \sigma_R^2$ the variance of the data, logarithmically transformed, within each group of measures associated with each *i*-th individual municipality) is assumed to be the same for all municipalities. The assumption that the indoor radon concentrations, within each selected unit, follow a log-normal distribution has been discussed in [1] and [8]. From the application of Bayes' theorem, the "true" value of the variable $ln(GM_j)$ can be inferred:

$$p\left[ln\left(GM_i\right)|ln\left(R_{ij}\right)\right]=\frac{p\left[ln\left(GM_i\right)\right]\cdot p\left[ln\left(R_{ij}\right)|ln\left(GM_i\right)\right]}{p\left[ln\left(R_{ij}\right)\right]} \tag{3}$$

where R_{ij} is the *j*-th measure of radon concentration, minus the outdoor concentration and after the normalization, within the *i*-th unit (municipality area). The $p\left[ln\left(GM_i\right)|ln\left(R_{ij}\right)\right]$ term in (3) is the conditional probability (posterior probability) that the expected value of log-transformed radon concentrations, within the i^{th} unit, is $ln\left(GM_i\right)$, given the set of experimental data $ln\left(R_{ij}\right)$; in the right-hand side, $p\left[ln\left(GM_i\right)\right]=N(\mu,\sigma^2)$ is an informative (normal) prior distribution and the remaining expression is the normalized likelihood. After some manipulations (see [13]), a particularly simple result for the estimate of $ln\left(GM_i\right)$ (known as Bayesian point or empirical estimate EB) is given:

$$ln\left(GM_i^{estimate}\right)=\frac{\frac{\mu}{\sigma^2}+\frac{n_i}{k^2}ln\left(GM_i^{obs}\right)}{\frac{1}{\sigma^2}+\frac{n_i}{k^2}} \tag{4}$$

$$V_i^2=\left(\frac{1}{\sigma^2}+\frac{n_i}{k^2}\right)^{-1} \tag{5}$$

Where V_i^2 is the variance. In the previous expressions n_i is the sample size for the geographic unit (municipality) *i* from which the sample mean $ln\left(GM_i^{estimate}\right)$ is estimated. Accordingly to (4), which is actually a weighted mean between the sample local mean and μ, the higher the number of measures n_i, the closer the estimated value $ln(GM_i^{estimate})$ to the sample value $ln(GM_i^{obs})$. The parameters σ^2 and k^2 are the variance components, respectively, between and within geographic units, and can be determined from an analysis of variance (ANOVA, [15]). The uncertainties of the point estimates (4) and (5) underestimate the true variability because they do not incorporate the variability due to the estimation of the hyperparameters. Hence, the point estimates were processed by means of a *Bootstrap* procedure [16] in order to get reliable confidence intervals for the α_i'*s* and the hyperparameters in the model. The key concept of the bootstrap is that the distribution of the statistic of interest can be approximated by estimates from repeated samples, hence, from an approximation of the unknown population [17]. Basically, a set of synthetic R_{ij} values is generated and treated as a set of responses from which a new set of bootstrap hyperameter estimates $\hat{\mu}'_\alpha$, $\hat{\sigma}'^2_\alpha$, $\hat{\sigma}'^2_R$ is obtained. In our paper, a random sample of 10000 elements was extracted from these empirical distributions for each hyperparameter. Once the set of 10000 bootstrap values was available we estimated the parameter standard errors using the usual sample procedures. Confidence intervals for the original parameter estimates or functions of them can be also constructed nonparametrically from the percentiles of the set of empirical bootstrap values.

2.1. Gibbs sampling approach

The underlying idea of Gibbs sampling is to partition the set of unknown parameters and then estimate them one at a time, or one group at a time, with each parameter or group of parameters estimated conditional on all the others and on the data. After a suitable number of iterations, we obtain a sample of values from the distribution of any component which we can then use to derive any desired characteristic such as the covariance matrix, the mean, etc. In general (see, e.g., [13]) the algorithm requires the choice of some number n_{chains} of parallel simulation runs. After completing the iterations (hundreds if not thousands may be required to ensure that convergence has been achieved) and assessing the convergence of the four chains (using proper convergence diagnostics), the resulting iterates are used for obtain summaries of the desired posterior distribution using the traditional sample estimates (mean,

median, standard deviation, quantiles, correlations). The simulations were performed by the free open source code *WinBugs*.

2.2. Multilevel modelling by R.
The model expressed by equations (2) was applied to experimental data set using the functions available in the *R* environment (a very popular free software for statistical computing [18]). In particular, the *lmer* function available in the *lme4* package [18] was used.

3. Identification of radon prone areas

The estimate of the variable $ln(GM_i^{estimate})$ and of its variance V_i^2, related to each i^{th} areal unit, (eq.s (4) and (5)) is deduced applying the Bayesian inference to the available data set according to the two different areal subdivisions (Fig. 3).
The variance of the data within each unit is given by $V_i^2 + k^2$. Then, accordingly to the previous discussion, a normal distribution is assumed for the variable $ln(R_i)$:

$$ln(R_i) \cong N\left[ln\left(GM_i^{estimate}\right), ln\left(GSD_i^{estimate}\right)^2 \right] \tag{6}$$

Where $ln\left(GSD_i^{estimate}\right)^2 = V_i^2 + k^2$ and $GSD_i^{estimate}$ is the geometric standard deviation of the R_i, estimated by Bayesian procedure. Then, in order to estimate the percentage of ground floor dwellings with indoor radon levels above a given reference level X, the standard normal variable $Z_i(x)$ is introduced by

$$Z_i(x) = \frac{x - ln(GM_i)}{ln(GSD_i)} \tag{7}$$

where $x = ln(C_{Rn} - Rn_{out})$ and C_{Rn} = radon concentration; thus $-\infty < x \le ln(X - Rn_{out})$ and $X \ge Rn_{out}$ is the selected reference level of the *Rn* concentration. Consequently, the area $Q(Z_i)$ under the curve for which the *Rn* concentration logarithm exceeds the value $ln(X - Rn_{out})$ can be calculated as follows:

$$f_i(X) = Q[Z_i(X)] = 1 - \Phi[Z_i(X)] = 1 - \frac{1}{\sqrt{2\pi}} \int_{\infty}^{Z_i(X)} e^{-\frac{Z_i^2}{2}} dZ_i \tag{8}$$

where Φ is the cumulative distribution function for the normal distribution $N(0,1)$.
Changing the variable, from Z_i in x, the expression (8) yields:

$$f_i(X) = Q(X) = 1 - \Phi(X) = 1 - \int_{-\infty}^{ln(X - Rn_{out})} \frac{1}{ln(GSD_i)\sqrt{2\pi}} e^{-\frac{(x - ln(GM_i))^2}{2\,ln(GSD_i)^2}} dx \tag{9}$$

In order to efficiently implement the expression (8) we consider a new variable change: $t = \frac{Z_i(x)}{\sqrt{2}}$. Thus:

$$f_i(X) = 1 - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{Z_i(X)}{\sqrt{2}}} e^{-t^2} dt = 1 - \left[\frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{Z_i(X)}{\sqrt{2}}} e^{-t^2} dt \right] \tag{10}$$

Then, introducing the *Error Function*: $Erf\left(\frac{Z_i}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{Z_i}{\sqrt{2}}} e^{-t^2} dt$ the probability distribution within the entire variability range of the standard normal Z_i variable assumes the following expression:

$$f_i(X)=\begin{cases}\frac{1}{2}\left[1-Erf\left(\frac{Z_i}{\sqrt{2}}\right)\right] & if \quad Z_i>0\\ \frac{1}{2} & if \quad Z_i=0\\ \frac{1}{2}\left[1+Erf\left(\frac{|Z_i|}{\sqrt{2}}\right)\right] & if \quad Z_i<0\end{cases} \tag{11}$$

As shown by [19], the simple application of (8) gives a biased result because of the non-linear nature of Φ ; the same authors, suggest the following unbiased estimator of $f_i(X)$:

$$f_i(X)=1-\Phi\left[Z_i(X)\right]-\frac{Z_i(X)}{2n_i\sqrt{2\pi}}e^{-\frac{Z_i(X)^2}{2}} \tag{12}$$

where n_i is the number of measurement in the i^{th} unit. Fig. 1 shows the plot of expression (12), while Figure 2 highlights how the unbiased probability values could be unreliable for small size number n_i and relatively high value of Z. Negative probability can occur even for n_i=10, a value commonly occurring also in literature. In this paper X=200 Bq/m^3 was selected as the reference value. A proportion of 20% for $Q(Z_i)$ (i.e. the expected percentage of ground floor dwellings with radon concentration above 200 Bq/m^3) could be one of the possible values to identify radon areas presenting a higher potential hazard, respect to the selected threshold.

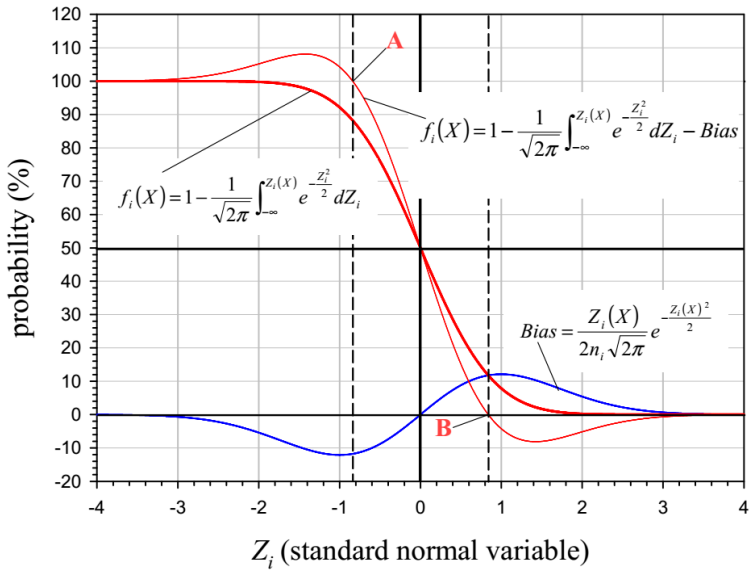

Figure 1. Biased and unbiased estimators

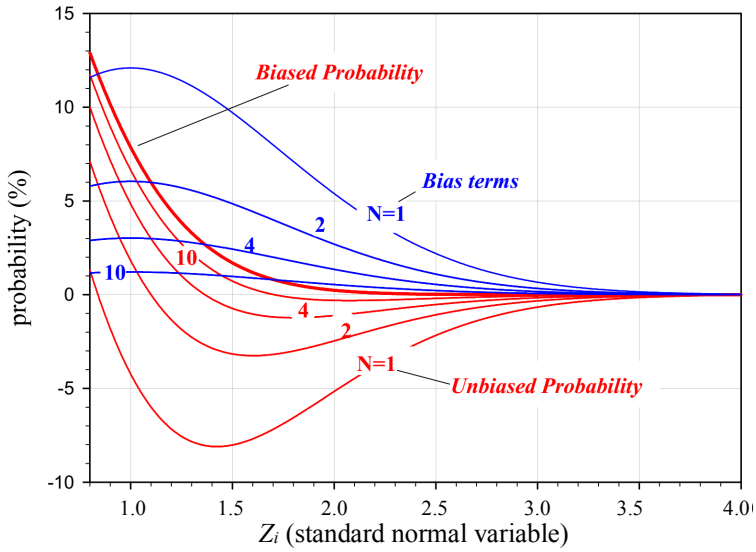

Figure 2. Probability parameterised by size sample n_i

4. Discussion

Figure 3 shows the sites (black dots) where radon measurements were taken during multiple campaigns. Furthermore the subdivision in the two zones was highlighted: coastal (zone 1) and mountain (zone 2)

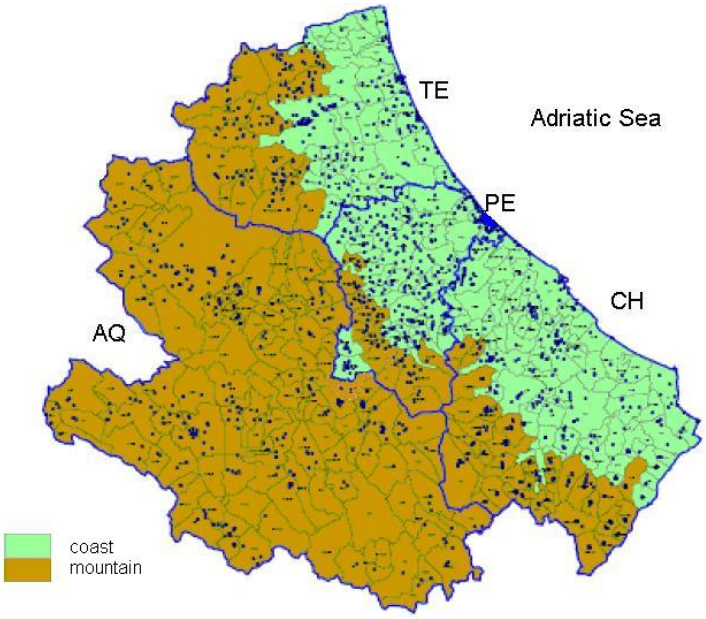

Figure 3. Subdivision of the Abruzzo's territory into two macro-areas: hilly-coastal area (green, zone 1) and the montane-piedmont area (brown, zone 2)

In figure 4, comparison among observed mean and estimated values by different methods are reported, while figure 5 shows how the statistical methodologies is able to lower the standard error related to the observed means.

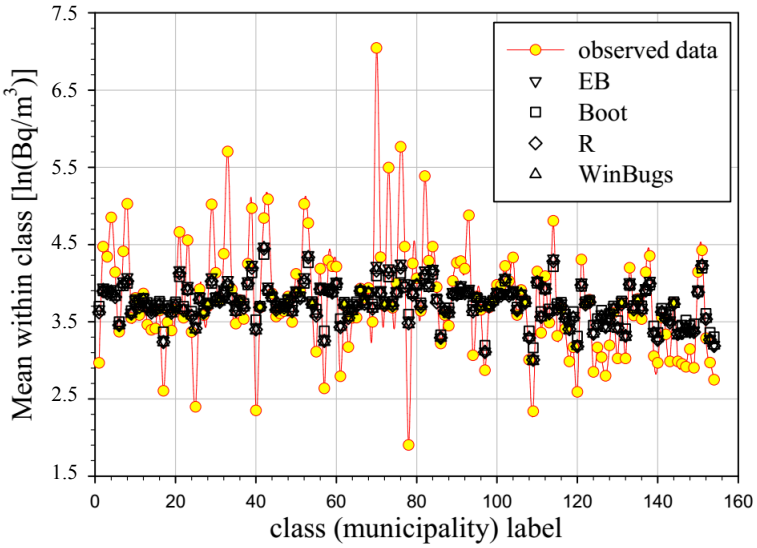

Figure 4. Comparison among observed and estimated mean (zone 1)

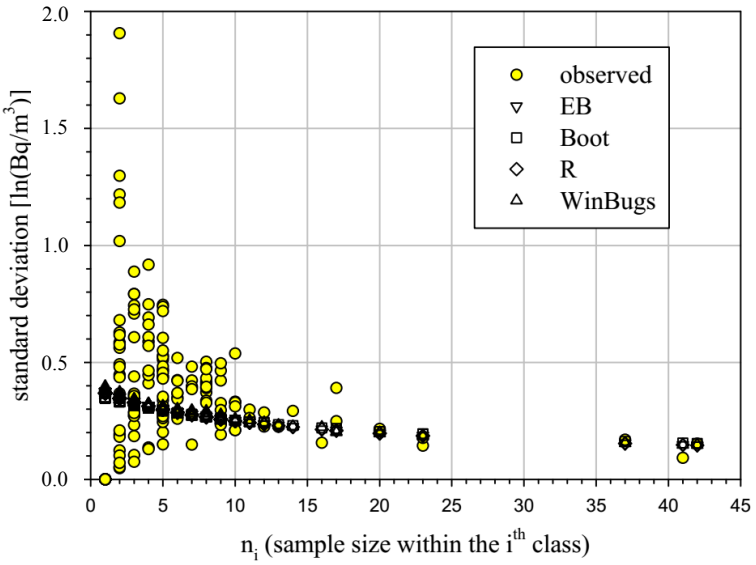

Figure 5. Standard deviations related to classes mean (zone 1)

Moreover, figure 6 highlights the difference between, in particular, Gibbs Sampling (open source code *WinBugs*) and *Bootstrap* (our research code) results.

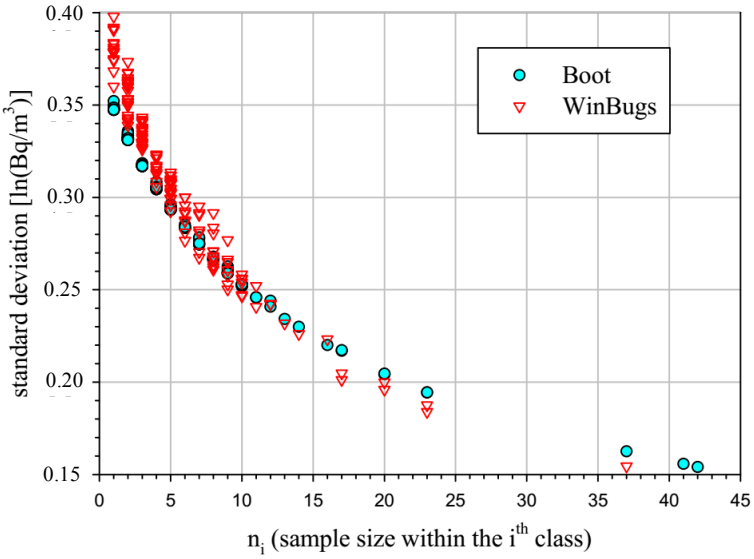

Figure 6. Standard deviation comparison (zone 1)

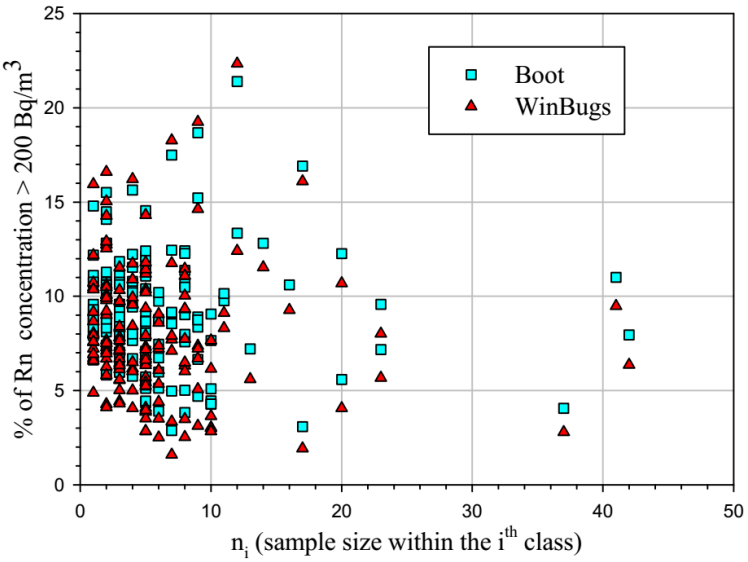

Figure 7. Gibbs Sampling and Bootstrap probability (zone 1)

Gibbs Sampling approach seems to be more suitable than Bootstrap approach to catch statistical differences among each class for low size samples (figure 6). In figure 8 the probabilities, expressed by equation (12), that the indoor-radon concentration exceed the selected threshold value of 200 Bq/m^3, for each municipality area in which the region was subdivided, are plotted.

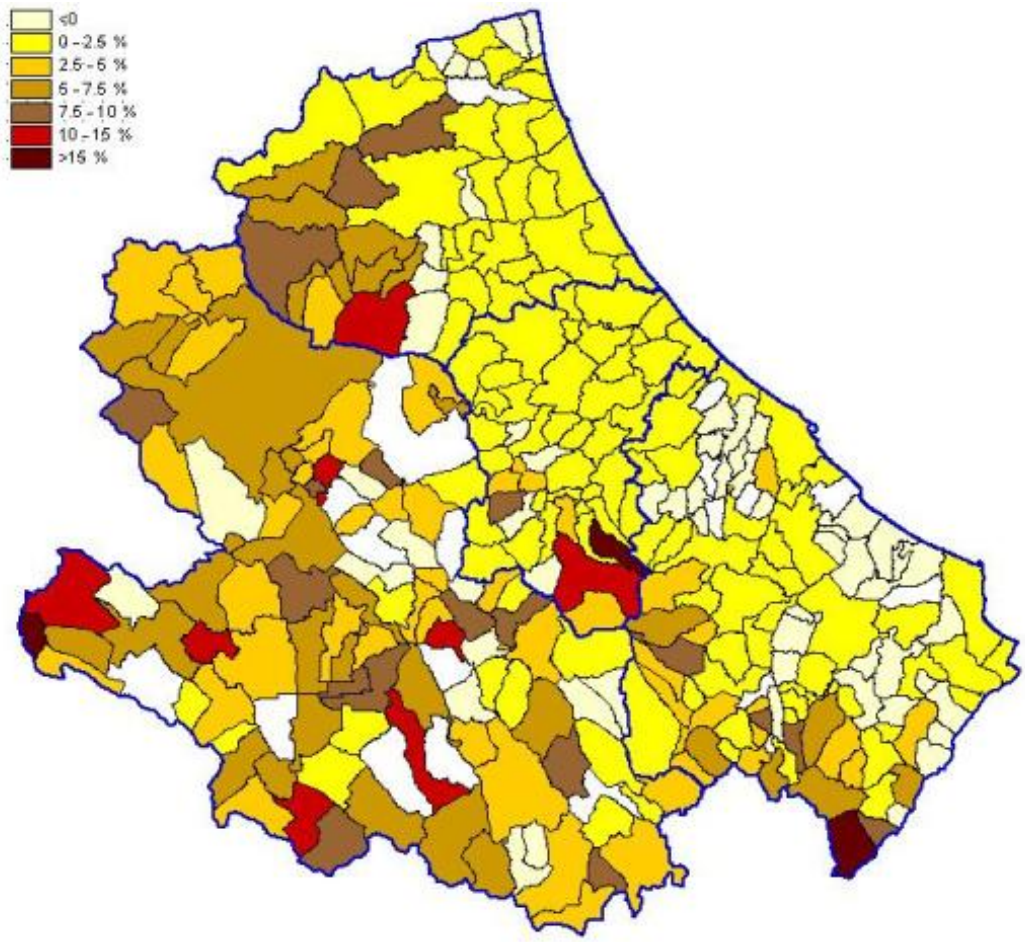

Figure 8. Percentage of ground floors with radon level exceeding 200 Bq/m^3 (eq. (12))

Conclusions

The content of the paper was the identification of areas affected by high radon concentration applying different statistical approach and comparing each results. The statistical analysis, based in particular on Monte Carlo realization of Full Bayesian Bootstrap and Gibbs sampling inference techniques, shows a substantial similarity between the two methods, at least for the problem explored in this paper. Nevertheless, Gibbs sampling seems to be more suitable to evidence statistical differences for classes characterized by a small size sample. Another important point has been the warning emerged on the application of unbiased expression to evaluate the probability of exceeding a reference level. The coastal area is generally characterized by pretty smooth values around or below 50 Bq/m^3. Besides, the radon levels in the montane-piedmont area are substantially higher, showing also a marked variability from zone to zone. Some areas, located on western and southern side, are affected by relatively high radon level. Different choices comparison of reference levels, evaluation of the representativeness of data respect to the population and geostatistical analysis of data improvement, will be the contents of next efforts. The currently available dataset, actually, lacks of a necessary number of data just in many of the potentially hazardous areas, in particular in the province of L'Aquila, the most mountainous region of the selected territory.

Acknowledgment

Authors wish to acknowledge fruitful discussion with dr. Sergio Palermi (Arta Abruzzo).

References

[1] S. Palermi, A. Pasculli, 2008. Radon Mapping In Abruzzo, Italy. Proceedings of 4th Canadian Conference on Geohazards Québec City Canada, May 20-24th 2008.

[2] A. Pasculli, M. Calista, N., Sciarra, 2018. Variability of local stress states resulting from the application of Monte Carlo and finite difference methods to the stability study of a selected slope. *Engineering Geology*. Vol. 245, pp. 370-389. WOS:000448494600031; SCOPUS id=2-s2.0-85054313956.

[3] M. Calista, A. Pasculli, N. Sciarra, 2015. Reconstruction of the geotechnical model considering random parameters distributions. *Engineering Geology for Society and Territory* Vol.2: Landslide processes pp. 1347-1351. SCOPUS id: 2-s2.0-84944628353.

[4] F. Rizzo, L. Caracoglia, S. Montelpare, 2018. Predicting the flutter speed of a pedestrian suspension bridge through examination of laboratory experimental errors. *Engineering Structures*, 172, 589-613. doi: https://doi.org/10.1016/j.engstruct.2018.06.042.

[5] F. Rizzo, L. Caracoglia, 2018. Examining wind tunnel errors in Scanlan derivatives and flutter speed of a closed-box. Journal of Wind and Structures. Vol. 26 (4), 231-251. doi: https://doi.org/10.12989/was.2018.26.4.231.

[6] M. Calista, E. Miccadei, A. Pasculli, T. Piacentini, M. Sciarra, N. Sciarra, 2016. Geomorphological features of the Montebello sul Sangro large landslide (Abruzzo, Central Italy). *Journal of Maps*. Vol. 12 (5), 882-891. WOS:000389542600021; SCOPUS id: 2-s2.0-84944909158.

[7] A. Chiaudani, D. Di Curzio, W. Palmucci, A. Pasculli, M**.** Polemio, S. Rusi, 2017. Statistical and Fractal Approaches on Long Time-Series to Surface-Water/Groundwater Relationship Assessment: A Central Italy Alluvial Plain Case Study. *Water***.** Vol. 9, pp. 1-28. WOS:000416798300036; SCOPUS id=2-s2.0-85033560093.

[8] S. De Novellis, A. Pasculli, S. Palermi, 2014. Innovative modeling methodology for mapping of radon potential based on local relationships between indoor radon measurements and environmental geology factors. *Wit Transaction on Information and Communication Technologies; 9th International Conference on Computer Simulation in Risk Analysis and Hazard Mitigation, RISK 2014*; Vol. 47, pp. 109-119. ISSN: 17433517; ISBN: 978-184564792-6; SCOPUS id: 2-s2.0-84903171519.

[9] A. Pasculli, S. Palermi, A. Sarra, T. Piacentini, E. Miccadei, 2014. A modelling methodology for the analysis of radon potential based on environmental geology and geographically weighted regression. *Environmental Modelling and Software*, Vol. 54, 165-181. WOS:000332267300013; SCOPUS id: 2-s2.0-84893040001.

[10] G. Ciotoli, M. Voltaggio, P. Tuccimei, M. Soligo, A. Pasculli**,** S. E. Beaubien, S. Bigi, 2017. Geographically weighted regression and geostatistical techniques to construct the geogenic radon potential map of the Lazio region: A methodological proposal for the European Atlas of Natural Radiation*. Journal of Environmental Radioactivity*. Vol. 166 (1) 355-375. WOS:000390073700014; SCOPUS id: 2-s2.0-84970046044**.**

[11] E. Ponzetti, 2004. Elaborazione statistica dei dati di concentrazione di radon per l'individuazione delle radon-prone areas. *PhD dissertation in Healt Physics*.University of Florence (Italy) (in Italian).

[12] S. Palermi, 2009. A measurement survey aimed to assess the exposure to radon of the population of Abruzzo region (Italy). *Phd dissertation in Healt Physics*, University La Sapienza Rome (Italy).

[13] A. Gelman and J. Hill, 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models *Cambridge University Press*.

[14] P.N. Price, A.V. Nero, A. Gelman, 1996. Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties, *Health Physics* 71(6), 922-936.

[15] M. Kutner, C.J. Nachtsheim, J. Neter, W. Li 2005. Applied Linear Statistical Models*, McGraw-Hill*.

[16] B. Efron, 1979. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* Vol. 7 (1), 1-26.
[17] B.N. Dixon, 2001. Bootstrap resampling, in Encyclopedia of Environmetrics, *A.H. El-Shaarawi*, W.W. Piegorsch (eds), Wiley.
[18] D. Bates, 2010. Computational methods for mixed models http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf
[19] C.E. Andersen, K. Ulbak, A. Damkjær, P. Kirkegaard and P. Gravesen, 2001. Mapping indoor radon-222 in Denmark: design and test of the statistical model used in the second nationwide survey *Sci. Total Environ.* 272 231–41.