Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-01643-1

ORIGINAL PAPER

Analysing radon accumulation in the home by flexible M-quantile mixed effect regression

R. Borgoni[1] · A. Carcagní[1] · N. Salvati[2] · T. Schmid[3]

Abstract

Radon is a noble gas that occurs in nature as a decay product of uranium. Radon is the principal contributor to natural background radiation and is considered to be one of the major leading causes of lung cancer. The main concern revolves around indoor environments where radon accumulates and reaches high concentrations. In this paper, a semiparametric random-effect M-quantile model is introduced to model radon concentration inside a building, and a way to estimate the model within the framework of robust maximum likelihood is presented. Using data collected in a monitoring survey carried out in the Lombardy Region (Italy) in 2003–2004, we investigate the impact of a number of factors, such as geological typologies of the soil and building characteristics, on indoor concentration. The proposed methodology permits the identification of building typologies prone to a high concentration of the pollutant. It is shown how these effects are largely not constant across the entire distribution of indoor radon concentration, making the suggested approach preferable to ordinary regression techniques since high concentrations are usually of concern. Furthermore, we demonstrate how our model provides a natural way of identifying those areas more prone to high concentration, displaying them by thematic maps. Understanding how buildings' characteristics affect indoor concentration is fundamental both for preventing the gas from accumulating in new buildings and for mitigating those situations where the amount of radon detected inside a building is too high and has to be reduced.

Keywords Environmental radioactivity · Building factors · Radon-prone areas · Hierarchical mixed models · Penalised splines · Lombardy region

1 Introduction

Radon (the ^{222}Rn isotope) is a noble, radioactive gas naturally occurring as a decay product of uranium. It is a gas without colour or smell that is detectable only by specialised measurement devices and represents the main contributor to natural background radiation. The becquerel (Bq) is the standard international unit for radon activity i.e. the amounts of radioactive material, and radon activity concentration is measured in Bq/m^3.

Evidence since the sixteenth century suggests that exposure to elevated concentrations of radon and radon progeny is a potential cause of the high prevalence of lung cancer mortality among miners (Jacobi 1993). However, it is only since the 1970s that human exposure to radon became of general concern. Radon is present both indoors and outdoors. While outdoors, its concentration in the air is diluted to a low level and therefore does not pose any significant health problems. Indoor concentrations are far higher simply because the gas enters into a smaller space and accumulates there. Hence, this paper focusses on indoor radon concentration (IRC), since it is in the indoor environment that radon becomes a serious health concern.

Extensive epidemiological studies (Lubin and Boice 1997; Kreienbrock et al. 2001; Darby et al. 2005; Krewski et al. 2005; Tiefelsdorf 2007) point out that long-term radon exposure in homes determines a remarkable increase in the risk of lung cancer. Nowadays, the International

✉ R. Borgoni
riccardo.borgoni@unimib.it

[1] Università degli studi di Milano - Bicocca, Milan, Italy

[2] Università di Pisa, Pisa, Italy

[3] Freie Universität Berlin, Berlin, Germany

Agency for Research on Cancer (IARC) and the US Environmental Protection Agency (EPA) have categorised radon as a Group 1 and Group A human carcinogen, respectively. Other papers have investigated the impact of radon on the increase in the risk of other cancer typologies (Smith et al. 2007).

To estimate the exposure of people to radon and to identify building typologies and geographical areas more prone to high IRC, monitoring surveys have been implemented in many countries such as in the UK (Green et al. 2002) USA (USEPA 1992) (Smith and Field 2007), Canada (Shi et al. 2006), and Belgium (Cinelli et al. 2011), to mention but a few. These surveys provide geocoded data that are often fundamental for planning remediation activities. Collecting information at several locations, data coming from these monitoring campaigns allow us to account for the multifactorial dependencies of IRC through statistical models that combine a number of explanatory variables. Raised IRC levels can often be traced back to the radon content in the underlying rocks and soils and are detected in dwellings close to the ground. Hence, the geological and lithologic nature of the soil as well as other soil characteristics, such as porosity and permeability, can influence indoor accumulation. Radon exhalation from building materials is another relevant source of radon since many building materials, such as concrete containing alum shale or volcanic tuffs and pozzolana, may have high radium content. The relationship between IRC and geological indicators of high radon potential as well as other structural building factors and materials has long been documented: see Gunby et al. (1993), Gates and Gundersen (1992), Price et al. (1996), Apte et al. (1999), Levesque et al. (1997), Sundal et al. (2004), Shi et al. (2006), Smith and Field (2007), Hunter et al. (2009), Cinelli et al. (2011) amongst others. All these types of effects will be considered further in the paper.

All the papers mentioned so far aimed to assess the influence of various characteristics of buildings on the average IRC. However, modelling a central tendency measure of the conditional distribution, typically the mean, may provide a rather incomplete or even inappropriate picture if the actual interest is in the tail of the distribution as is the case in the presence of outliers, asymmetry or reference concentration values endorsed by law or international recommendations. The conventional regression approach, modelling the conditional expectation of IRC, does not permit this kind of analysis. Instead, this can be obtained using quantile regression, i.e. by fitting a family of robust regression models, each summarising the behaviour at different levels of the IRC conditional distribution. Quantile regression was introduced in the econometrics literature by Koenker and Bassett (1978) and has been extended towards many directions. For instance, Chaudhuri (1991) proposed locally polynomial quantile regression whereas Koenker et al. (1994) and Bosch et al. (1995) discussed penalty methods for smoothing quantile regression. Geraci and Bottai (2014) investigated linear quantile regression models for clustered/hierarchical data. Since quantile regression enables us to fully describe the conditional distribution, the method has been used in many applications in recent decades. We refer to Yu et al. (2003) and Koenker (2005) for some examples. However, quantile regression has seldom been applied in the context of radon mapping. Exceptions are represented by Borgoni (2011) who adopted a spatial semiparametric model to define a conditional quantile regression model. Furthermore, Fontanella et al. (2015) and Sarra et al. (2016) proposed a spatial quantile hierarchical Bayesian model in this context.

An alternative to quantile regression is the M-quantile regression introduced by Breckling and Chambers (1988) to integrate expectile (Newey and Powell 1987) and quantile regression within a unique paradigm based on a 'quantile-like' generalisation of regression defined via influence functions (M-regression). Tzavidis et al. (2016) extended the M-quantile regression by including random effects in order to consider the hierarchical structure in the data, whereas Alfó et al. (2017) proposed a finite mixture of M-quantile regressions with discrete random coefficients; the discrete distribution of the latter can be interpreted as a nonparametric estimate of an unspecific continuous distribution. Although M-quantile and quantile regressions cannot be directly compared as they target different location parameters, both approaches try to model location parameters that are related to the same part of the conditional distribution (Jones 1994). Why should the potential data analyst consider M-quantile regression when the main advantage of quantile regression is the more intuitive interpretation? M-quantile regression models are more flexible, in particular they allow for robustness in exchange for efficiency in inference by tuning a suitable constant of the influence function (see Sect. 3). The option to select different continuous influence functions in an M-quantile regression—in contrast to the absolute value function in a quantile regression—can offer additional computational stability.

As far as the statistical methodology is concerned, this paper extends the work by Tzavidis et al. (2016) to a random effect semiparametric M-quantile model which is able to account both for the characteristics of the soil and for the material and architectural structure of a building. It is well known, however, that radon dynamics are spatially structured due to a number of causes that may affect IRC on a local and on a large-scale over and above the available secondary information obtained via administered surveys or measurement campaigns. For this reason, we extend the

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ☑ CP | ☑ DISK

model to include a flexible component that is able to grasp this spatial effect. In particular, the proposed M-quantile model incorporates the spatial information (locations) of the data through a spline component in the linear predictor of the model, and therefore does not rely on any structural assumptions in the error terms. This is particularly relevant when, as in the case study presented later in the paper, geographically referenced measures have to be spatialised to produce maps.

The paper is structured as follows. Section 2 presents the data and the model we propose in Sect. 3. Section 4 shows the results obtained applying the suggested random effect semiparametric M-quantile model to indoor radon data. In particular, we explain how the suggested model permits us to map the phenomenon of interest across space and identify those areas more prone to high concentration and to estimate the impact of potential determinants on the pollutant concentrations. Concluding remarks are presented in Sect. 5.

2 Data description

The data used in the present paper come from an indoor radon gas monitoring survey implemented by the Agency for Environmental Protection (ARPA) from 2003 to 2005 in the Lombardy Region (northern Italy). With a surface area of 23,800 km^2 Lombardy is the fourth largest and most populated region in Italy, with about 10,000,000 inhabitants according to the last census in 2011, which corresponds to about 20% of the entire Italian population. A national survey conducted by the National Health Service from 1989 to 1994 already indicated that Lombardy is exposed to high values of IRC. This survey pointed out that the IRC in Lombardy was 116 Bq/m^3 on average, and is therefore higher than the national average of 70 Bq/m^3. Assessing the spatial variability of IRC and the population exposure to this gas is a prominent environmental and health-related issue in this part of the country.

In this paper, the problem of high IRC in dwellings is investigated by considering a sample of 900 measures of IRC collected throughout the regional territory for which a complete record of all relevant building characteristics were available. Figure 1 shows the locations of the measurement points (indicated by crosses) and the study region expressed in UTM projection.

Long-term measurements was obtained using CR-39 trace detectors that were positioned in dwellings for 12 months. The dosimeters were changed after approximately 6 months and the year-long average of the two semester values is considered in this paper, weighting the two one-semester measurements by the actual time of exposure of each detector. The average concentration is around 118 Bq/m^3 (sd 136 Bq/m^3) ranging from a minimum of 12.5 Bq/m^3 to a maximum 1762.5 Bq/m^3. As shown in Fig. 2, the IRC distribution is strongly asymmetric with a number of potential outliers, in line with other studies (Nero et al. 1986, amongst others).

A questionnaire were also administered to dwellers of each sampled unit to collect other information about the building and the rooms in addition to the IRC. IRC measurements and building information were then combined in a single dataset. In the following, we focussed on factors that are expected to affect IRC, such as the wall material (stone versus other materials such as lateritious and hollow brick), the presence of an air conditioning system, the type of connection with the soil (i.e. whether the building is in direct contact with the ground or a basement/crawlspace is present), the type of building (detached vs. non-detached), the year of construction or last renovation (before or after 1990) and the floor material (marble or granite versus other materials). In Table 1 some summary statistics of IRC conditioned to these building factors are shown. We observe that different house characteristics impact differently on the IRC level. For instance, the differences between the 20th percentiles and 80th percentiles of IRC measured in dwellings with a marble-granite floor and in dwellings with an other-material floor are 4.9 Bq/m^3 (20th percentile) and 14.2 Bq/m^3 (80th percentile), respectively. When comparing the 20th percentiles and 80th percentiles of IRC measured in dwellings with stone walls versus other material walls, the differences are 11.1 Bq/m^3 and 62.1 Bq/m^3, respectively, with a much more pronounced spread between higher quantiles. Estimating the quantiles of the IRC conditional distribution, given the covariates, is worth pursuing.

The composition of the soil on which a building is located is another important feature that can affect the IRC, since the concentration of uranium and radium varies depending on the rock lithology. Hence, it is expected that higher concentration levels tend to occur in particular geological areas. The geological composition of Lombardy varies extremely with regards to the lithological and soil typologies. In order to derive this information the data were linked to a geo-lithologic map on a scale of 1:250,000 (Borgoni et al. 2011) that partitions the territory of the Lombardy into 11 geological classes (see Fig. 3). Since the measurement points were geo-referenced, it was possible to assign one of the 11 types to each of them.

Figure 4 shows the boxplot of IRC for each geological class. The dashed lines in the figure connect the 20th and 80th quantiles of IRC conditioned to different geo-lithologic classes. We observe that the quantiles change considerably between the geo-lithologic classes.

Finally, high IRC can also be found in areas with low radium levels, especially when fractured rocks or intensive

Journal : Large 477 Dispatch : 26-12-2018 Pages : 20
Article No. : 1643 ☐ LE ☐ TYPESET
MS Code : SERR-D-18-00069 ☑ CP ☑ DISK

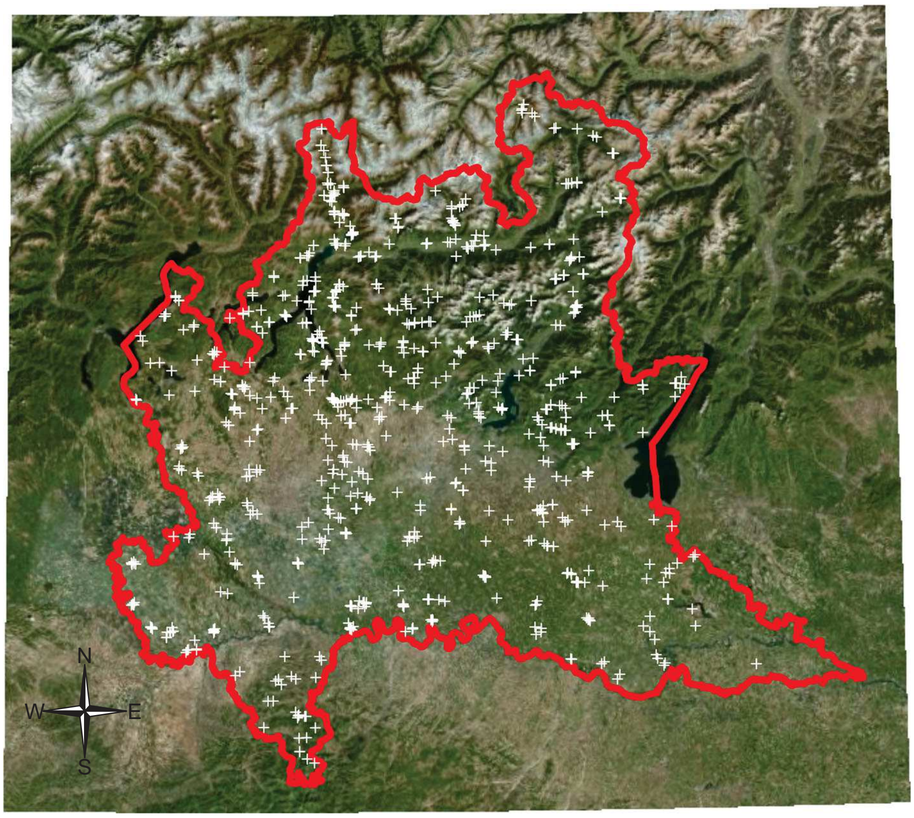


Fig. 1 Sampling locations and the study area

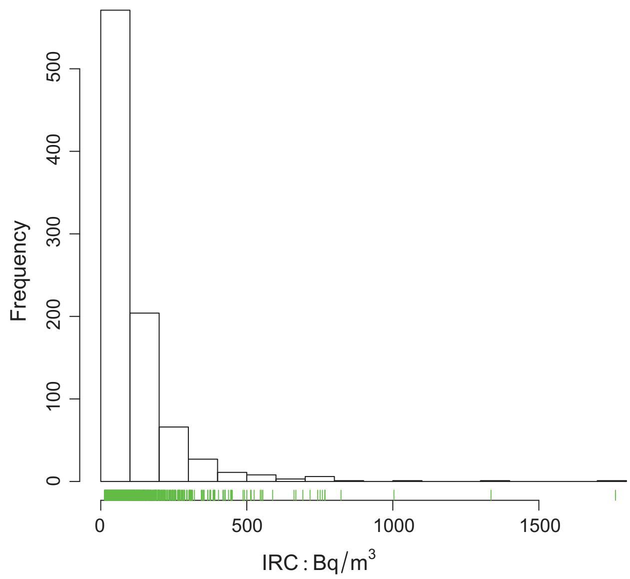


Fig. 2 Sampling distribution of IRC

tectonic frameworks are present. This may be due to the presence of faults, which may foster the gas to seep up from deeper origins and enter into homes. To assess the effect of being in the proximity of a fault, we calculated the distance of each sampling point x to a tectonic fragment A. Tectonic fragment cartography was available from a shape file where each fault is geocoded in a vector format through a set of nodes. Hence, the distance has been calculated as

$$d(x, A) = \min_{s \in A} \|x - s\| \qquad (1)$$

as suggested by Foxall and Baddeley (2002).

3 A random effect semiparametric M-quantile model for IRC

The M-quantile of order q, $MQ_y(q|\mathbf{X};\psi)$, for the conditional density of an outcome variable y given auxiliary variables $\mathbf{X}$, $f(y|\mathbf{X})$, is defined by Breckling and Chambers (1988) as the solution of the integral equation $\int \psi_q\{y - MQ_y(q|\mathbf{X};\psi)\} f(y|\mathbf{X}) dy = 0$. Here, ψ_q is the derivative of an asymmetric loss function ρ_q, called the (asymmetric) influence function. In particular, $(\mathbf{x}_i^T, y_i)$, $i = 1, \ldots, n$, denotes n observations of a random sample, y_i is the outcome variable and $\mathbf{x}_i^T$ are the p-vectors of the covariates $\mathbf{X}$. A linear M-quantile regression model of y_i given the auxiliary variables $\mathbf{x}_i$ is given by

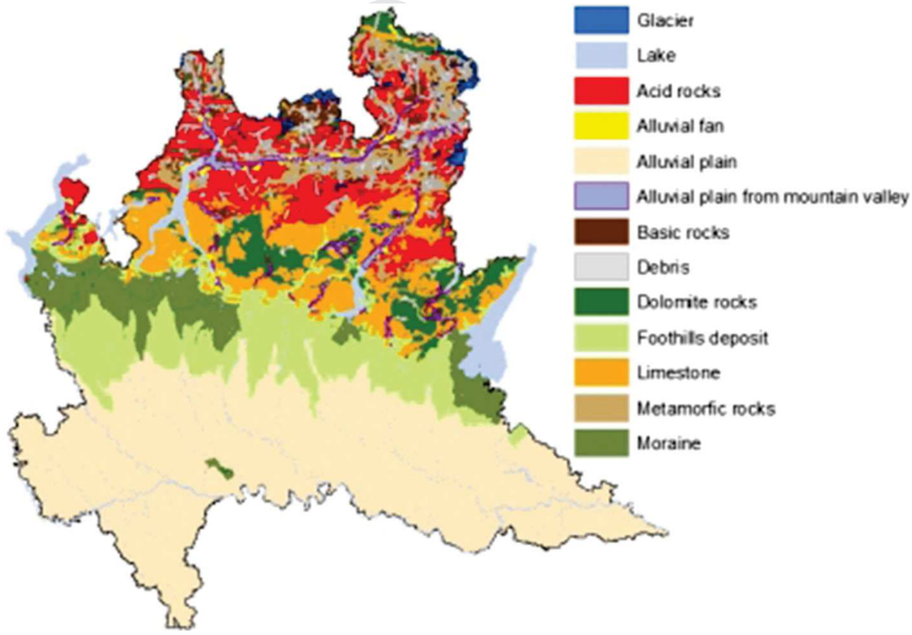
$$MQ_{y_i}(q|\mathbf{x}_i;\psi) = \mathbf{x}_i^T \boldsymbol{\beta}_q,$$

and $\boldsymbol{\beta}_q$ can be estimated minimising

Table 1 IRC summary statistics by dwelling characteristics

	N	Mean	SD	Min	Q20	Median	Q80	Max
Wall material								
Other	788	110.4	114.3	12.5	40.3	73.2	150.7	1003.9
Stone	112	171.1	233.3	16.0	51.2	98.7	212.8	1762.5
Conditioning system								
No	840	120.2	138.8	12.5	41.9	76.7	164.7	1762.5
Yes	60	85.4	84.5	19.9	40.6	65.7	95.7	555.2
Connection with the ground								
In contact	352	135.9	143.0	13.8	45.0	84.6	195.3	1003.9
Basement	348	106.4	130.5	12.5	39.1	70.9	145.5	1762.5
Type of building								
Single	323	95.4	130.3	13.8	35.4	64.7	130.0	1762.5
Not single	577	130.5	137.9	12.5	46.0	84.2	187.8	1336.2
Year construction/last renovation								
Before 1990	525	117.0	134.6	12.5	41.0	74.4	161.4	1336.2
After 1990	375	119.2	138.6	16.0	42.6	79.0	160.9	1762.5
Floor material								
Other material	853	118.9	137.3	12.5	41.9	76.5	161.2	1762.5
Marble-granite	47	100.8	100.8	19.9	37.0	63.6	147.0	742.9

Fig. 3 Geo-lithological classification of the regional territory



$$\sum_{i=1}^{n} \rho_q\{r_{iq}\}, \tag{2}$$

Here, $r_{iq} = (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_q)/\sigma$, σ is a scale parameter, the asymmetric loss function is $\rho_q\{r_{iq}\} = 2\rho\{r_{iq}\}[qI(r_{iq} > 0) + (1-q)I(r_{iq} \leq 0)$ and $I(\cdot)$ is the indicator function. The regression parameters could differ for different values of q. M-quantile, quantile and expectile regression models can be obtained as special cases by using different specifications for the asymmetric loss function ρ.

See details in Bianchi et al. (2018). Throughout the paper the Huber loss function (Huber 1981) is used to define the linear M-quantile regression model:

$$\rho_q\{r_{iq}\} = 2\begin{cases} (c|r_{iq}| - c^2/2)|q - I(r_{iq} \leq 0)| & |r_{iq}| > c \\ (r_{iq}^2/2)|q - I(r_{iq} \leq 0)| & |r_{iq}| \leq c, \end{cases} \tag{3}$$

where c is a tuning constant. Conventionally, in an M-regression, the data analyst tunes this constant to provide a trade-off between robustness and efficiency. Huber (1981)

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

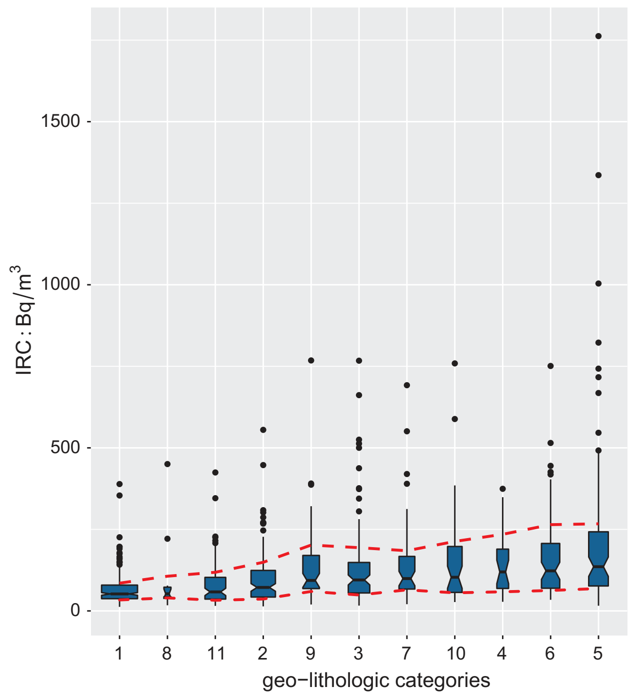


Fig. 4 Geo-lithological classification of the regional territory. Class labels: 1 Alluvial plain, 2 Foothill deposit, 3 Limestone, 4 Alluvial fan, 5 Debris, 6 Dolomite rocks, 7 Acid rocks, 8 Basic rocks, 9 Metamorphic rocks, 10 Alluvial plain, and 11 mountain valley. Dashed lines connect the 20th and 80th IRC quantiles of each geo-lithological category. Larger boxes indicate a large sample size in the corresponding geo-lithological class

proposes a value between 1 and 2. In particular, the author suggests 1.5. In the `rlm` function of the package `MASS` in `R` the default value for the tuning constant is 1.345. It corresponds to 95% of the efficiency of the estimates under normality. It means that when the errors follow the normal distribution, setting c equal to a large value, say 100, is the most appropriate choice. In this case if a smaller value, say 1.345, is used it will reduce the efficiency of the estimates because the tuning constant will offer unnecessary robustness. The tuning constant $c = 1.345$ is used throughout the paper. To overcome this ad hoc approach to selecting the tuning constant, Bianchi et al. (2018) extended the data-driven method by Wang et al. (2007) to an M-quantile regression to estimate c via likelihood equations. This approach could also be applied for the semi-parametric M-quantile models we propose in this paper, but this is beyond the scope of this work, however, and is left for future research.

Recently, M-quantile regression models that consider the two-level hierarchical structure in the data by including random effects were proposed by Tzavidis et al. (2016). The maximum likelihood method has been used by the authors for the estimation of model parameters.

The model suggested in this paper also includes a random intercept term to account for the clustering of data sharing the same geological substratum. However, as mentioned above, radon dynamics shows regularities when monitored in space both on the local or on the large-scale that are typically far from being linear. Hence, to adjust for these potential non-linear effects in the regression, we also add a flexible component to the linear predictor of the mixed effect M-quantile model. In particular, we use penalised splines. Penalised splines are effective tools for a number of reasons. Firstly, they are reasonably simple to implement, being a relatively straightforward extension of a linear M-quantile regression. Secondly, their flexibility enables the inclusion in a wide range of modelling features. More specifically, penalised splines account for spatial dependencies in the IRC data in the semiparametric M-quantile regression model adopted in the case study presented below. This component is expected to grasp not only large-scale dependencies of the radon data but also spatial local effects on the concentration field. Splines rely on a set of basis functions to handle non-linear structures in the data. In this paper, we assume that the spatial pattern of the variable of interest can be explained as a function of the location of a point that is represented by its cartographic coordinates. Thus, a bivariate smoothing spline is included in the additive specification of the model and is specified in terms of a set of bivariate basis functions. Following Ruppert et al. (2003), Pratesi et al. (2009) suggested the use of radial basis functions to derive low-rank thin plate splines.

The model we propose for a specified M-quantile q is:

$$MQ_y(q|\mathbf{X},\mathbf{Z},\mathbf{Z}_{sp};\psi) = \mathbf{X}\boldsymbol{\beta}_q + \mathbf{Z}\mathbf{u}_q + \mathbf{Z}_{sp}\boldsymbol{\gamma}_q, \tag{4}$$

where $\mathbf{X}$ is a matrix of dimension $n \times p$ of auxiliary variables, $\boldsymbol{\beta}_q$ is the $p \times 1$ vector of M-quantile regression coefficients; $\mathbf{u}_q$ is a $G \times 1$ vector of geological categories and γ_q is a $K \times 1$ vector of random effects associated with the spline matrix; $\mathbf{Z}$ is an incidence $n \times G$ matrix coding the point-geological class hierarchy; $\mathbf{Z}_{sp}$ is a $n \times K$ spline matrix and K is the number of spline knots. More specifically (Opsomer et al. 2008),

$$\mathbf{Z}_{sp} = \left[C(\boldsymbol{w}_i - \mathbf{k}_j)\right]_{1\le j\le K}^{1\le i\le n}\left[C(\mathbf{k}_j - \mathbf{k}_k)\right]_{1\le j,k\le K}^{-1/2} \tag{5}$$

$\mathbf{k}_j$ and $\mathbf{k}_i$, $j = 1,\ldots,K$, $i = 1,\ldots,K$, being two-dimensional vectors representing the cartographic coordinates of knots j and k. $\boldsymbol{w}_i$ is a two-dimensional vector representing the cartographic coordinates of the sampling location i and $C(\mathbf{s}) = \|\mathbf{s}\|_2^2 \log\|\mathbf{s}\|_2$ where $\mathbf{s} \in \mathbb{R}^2$ and $\|\mathbf{s}\|_2$ is the Euclidean norm of $\mathbf{s}$ in $\mathbb{R}^2$.

Differently from the model suggested by Pratesi et al. (2009), we assume that the coefficients of the spline matrix in the linear predictor are random coefficients. A practical

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

advantage of the mixed model representation of the spline lies in fitting the model. The usual penalised spline-fitting criterion requires estimating a penalising or smoothing parameter prior to model estimation. Cross-validation is usually suggested as an appropriate way to tackle the problem. The mixed model representation avoids this step since the model can be estimated directly using routines that are appropriate for linear mixed models. Furthermore, including random coefficients for the spline basis components permits us to account for the bias due to omitted variables or unmeasured confounders. In addition, as advocated by Ruppert et al. (2003), treating the coefficients of the knots as random leads to a smoother representation of the estimated effect, compared to using fixed effects specification, and avoids data overfitting. As mentioned above, the spline component of the model is expected to catch the spatial regularity of IRC data and is used to visualise the results by smoothed maps. Hence, the random effect specification of the spline seems to be more appropriate for this end. We define the following modified estimating equations, extending the idea of asymmetric weighting of the residuals to estimate the regression coefficients and the variance components (Tzavidis et al. 2016; Borgoni et al. 2018):

$$\mathbf{X}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{1/2}\psi_q\{\mathbf{r}_q\} = \mathbf{0} \tag{6}$$

$$\begin{aligned}
&\frac{1}{2}\psi_q\{\mathbf{r}_q\}^T\mathbf{U}_q^{1/2}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{1/2}\psi_q\{\mathbf{r}_q\}\\
&\qquad -\frac{K_{2q}}{2}tr\left[\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^T\right] = 0\\
&\frac{1}{2}\psi_q\{\mathbf{r}_q\}^T\mathbf{U}_q^{1/2}\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{1/2}\psi_q\{\mathbf{r}_q\}\\
&\qquad -\frac{K_{2q}}{2}tr\left[\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\right] = 0\\
&\frac{1}{2}\psi_q\{\mathbf{r}_q\}^T\mathbf{U}_q^{1/2}\mathbf{V}_q^{-1}\mathbf{V}_q^{-1}\mathbf{U}_q^{1/2}\psi_q\{\mathbf{r}_q\} - \frac{K_{2q}}{2}tr\left[\mathbf{V}_q^{-1}\right] = 0.
\end{aligned} \tag{7}$$

Let $\mathbf{r}_q = \mathbf{U}_q^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_q)$ denote the vector of scaled residuals with components r_{ijq}, $\mathbf{U}_q$ the diagonal matrix with diagonal elements u_{ijq} equal to the diagonal elements of the covariance matrix $\mathbf{V}_q$ and $\psi_q(r)$ the derivative of a loss function ρ_q. The covariance matrix $\mathbf{V}_q$ is defined by $\mathbf{V}_q = \boldsymbol{\Sigma}_{\epsilon_q} + \mathbf{Z}\boldsymbol{\Sigma}_{u_q}\mathbf{Z}^T + \mathbf{Z}_{sp}\boldsymbol{\Sigma}_{\gamma_q}\mathbf{Z}_{sp}^T$, with $\boldsymbol{\Sigma}_{u_q} = \sigma_{u_q}^2\mathbf{I}_G$, $\boldsymbol{\Sigma}_{\gamma_q} = \sigma_{\gamma_q}^2\mathbf{I}_K$, and $\boldsymbol{\Sigma}_{\epsilon_q} = \sigma_{\epsilon_q}^2\mathbf{I}_n$, where $\sigma_{u_q}^2$, $\sigma_{\gamma_q}^2$ and $\sigma_{\varepsilon_q}^2$ are the quantile-specific variance components. $\mathbf{I}_n$ is an identity matrix of size n and $K_{2q} = E[\psi_q(\boldsymbol{\varepsilon})\psi_q(\boldsymbol{\varepsilon})^T]$ with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$. To obtain estimators of $\boldsymbol{\beta}_q$, $\sigma_{u_q}^2$, $\sigma_{\gamma_q}^2$, $\sigma_{\varepsilon_q}^2$, Eqs. (6) and (7) are solved iteratively. For Eq. (6) a Newton–Raphson algorithm is used and for (7) the fixed-point iterative method is implemented to get the estimates. The algorithm is implemented by the authors in a function in the R software (R Core Team 2017). A sandwich estimator is adopted to make inference on the model parameters. Details of the estimation algorithm and the variance estimators are reported in Tzavidis et al. (2016).

4 M-quantile modelling of geocoded radon data

In this section, the model discussed in Sect. 3 is applied to the IRC data presented in Sect. 2. The set of covariates introduced in Table 1 that may potentially have an effect on IRC are those included in the M-quantile models considered hereafter.

As mentioned in the introduction of this paper, IRC tends to vary across space showing regular possibly non-linear patterns of values due to a number of environmental, geological and anthropic factors. Geographical coordinates of the measurement locations can be considered as a surrogate variable of all these factors that may happen to be unmeasured or even unmeasurable. In this paper, we propose including this component in three ways. Firstly, the IRC of the study region tends to be quite different for the south compared to the north (Borgoni et al. 2011), hence, a trend surface model (Cade et al. 2005; Koenker and Mizera 2004) is specified semiparametrically by a bivariate thin plate spline transformation of the cartographic coordinates, as discussed in Sect. 3. Secondly, IRC values detected in buildings that are built on the same type of soil can be expected to be more similar than those detected in different geo-lithological areas. As the data show a hierarchical structure, we explicitly consider this aspect in the model specification, including a random effect to capture the variability within geological areas, ending up in the semi-parametric M-quantile random effect model presented in Sect. 3. Thirdly, since, as mentioned above, high IRC can also be found in areas where faults are present, the distance to the nearest tectonic fragment is also included in the model.

In this section, we investigate two different issues related to IRC modelling: (1) the identification of radon-prone areas, and (2) the identification of those characteristics that make a building more exposed to high IRC. The latter analysis also allows for an identification of those building profiles that are more exposed to higher concentration and for which it can be appropriate to plan remediation actions or actions that are able to prevent gas accumulation.

Some preliminary analyses reported in Appendix 1 of this paper clearly show that the data may contain outliers that prevent the Gaussian assumptions from being met, leading to biased and inefficient estimates of the model parameters. Several papers (see Huggins 1993; Huggins

Journal : Large 477 Dispatch : 26-12-2018 Pages : 20
Article No. : 1643 ☐ LE ☐ TYPESET
MS Code : SERR-D-18-00069 ☑ CP ☑ DISK

and Loesch 1998) suggested the robust estimation of mixed models to protect against departures from normality. This can be obtained using a loss function in the log-likelihood that increases with the regression residuals at a slower rate than the squared loss function. As described in the previous sections, resorting to the M-quantile approach permits us to robustly estimate the relationship between IRC and a set of covariates in a natural manner.

4.1 Radon-prone areas: identification and mapping

Health radon-related concerns have generated a growing interest in identifying those regions of the territory where high IRC are expected, the so-called 'radon-prone areas' (RPA). A number of different approaches have been suggested for this, mostly based on various cluster detection methods adopted to either delineate spatial clusters or improve the understanding of the spatial dynamic of radon using an automatic detection of those regions of space that are 'anomalous', 'unexpected' or otherwise 'interesting' (Sarra et al. 2016).

According to the World Health Organization (2009), various definitions of radon-prone areas exist. Typically, countries define RPA as regions where the estimated percentage of homes, whose radon concentrations exceed a reference value τ, oversteps a threshold q. The Italian legislation is compliant to the WHO suggestion to define RPA as regions where 'there is a high probability of finding high (indoor) radon concentrations' (art. 10-ter,comma 2, D.L.vo 241/00). The above-mentioned definitions suggest that an RPA is a region where $P(\text{IRC} > \tau)$ is *high*: *high* meaning $P(\text{IRC} > \tau) > q$ and with q denoting a fixed threshold, although different reference levels are suggested by different local authorities. Below, we exemplify the procedure using $q = 0.15$. Indicating by ξ_{1-q} the $(1-q)$-quantile of IRC, this also means that $\xi_{1-q} > \tau$. Hence, one can equivalently define an RPA as a region where a sufficiently high-order quantile of IRC is above the reference level τ.

Moving from this definition, RPA identification based on conditional quantiles of the radon distribution sounds more appropriate than basing it on cluster algorithms. Borgoni et al. (2010) also suggested a quantile-based approach adopting conventional kriging procedures coupled with Monte Carlo sequential (Gaussian) simulations to approximate the conditional distribution of radon at each point in space. Directly modelling the tail of the distribution using an M-quantile or a quantile approach (Fontanella et al. 2015), seems, however, a more direct and natural way to operate.

For the rest of this section we considered the 85th M-quantile as the reference level to identify RPA. In order to estimate such an M-quantile at different locations in space, the following semiparametric M-quantile random effect regression model is employed:

$$\begin{aligned} & MQ_{y_i}(0.85|d_{ij}, x_{1ij}, x_{2ij}, \mathbf{z}_i, \mathbf{z}_{spi}; \psi) \\ & \quad = \alpha + d_{ij}\delta + x_{1ij}\beta_1 + x_{2ij}\beta_2 + \mathbf{z}_i\mathbf{u} + \mathbf{z}_{spi}\gamma, \end{aligned} \tag{8}$$

where

- d is the fault distance;
- x_1 and x_2 are the coordinates (respectively, longitude and latitude) of the measurement points in UTM projection;
- $\mathbf{z}_i$ is a vector of geo-lithologic class indicators, i.e 0–1 variables;
- $\mathbf{z}_{spi}$ is the row of the Z_{sp} matrix defined in Eq. 5 pertinent to sampling dwelling i located in geo-lithologic class j.

We use 50 knots for the spline, obtained by applying the partitioning clustering algorithm CLARA (Kaufman and Rousseeuw 1990) to the sampling locations. The estimated parameters are reported in Table 2. The linear effect of the cartographic coordinates has been found to be not significant. However, the estimate of the variance of γ is about three times its estimated standard error pointing out a strong effect of the random component of the spline. This demonstrates that the spatial variability due to location is definitely relevant in the IRC dynamic. We notice that the two linear terms are retained in the model in the following analysis to correctly specify the spline component.

Since the aim of this analysis is to classify the areas of the region according to their proneness to radon without considering any particular building typology, the structural and architectonic characteristics of the building are not included in the model. The issue of assessing the impact of building factors will be addressed in the following section.

The estimates of the 85th M-quantile are depicted in Fig. 5 where the surface has been discretised via a grid of

Table 2 Estimates of the RPA model

	Estimate	Std. error	p value
Intercept (α)	178.75	275.62	0.52
Fault distance (δ)	− 76.14	44.01	0.08
Longitude (β_1)	237.45	385.68	0.54
Latitude (β_2)	− 154.31	387.98	0.69
σ^2_ϵ (individual)	7758.4	220.07	
σ^2_u (geo-lithology)	302.4	402.06	
σ^2_γ (spline)	75250.2	28397.95	

Journal : Large 477 Dispatch : 26-12-2018 Pages : 20
Article No. : 1643 LE TYPESET
MS Code : SERR-D-18-00069 CP DISK

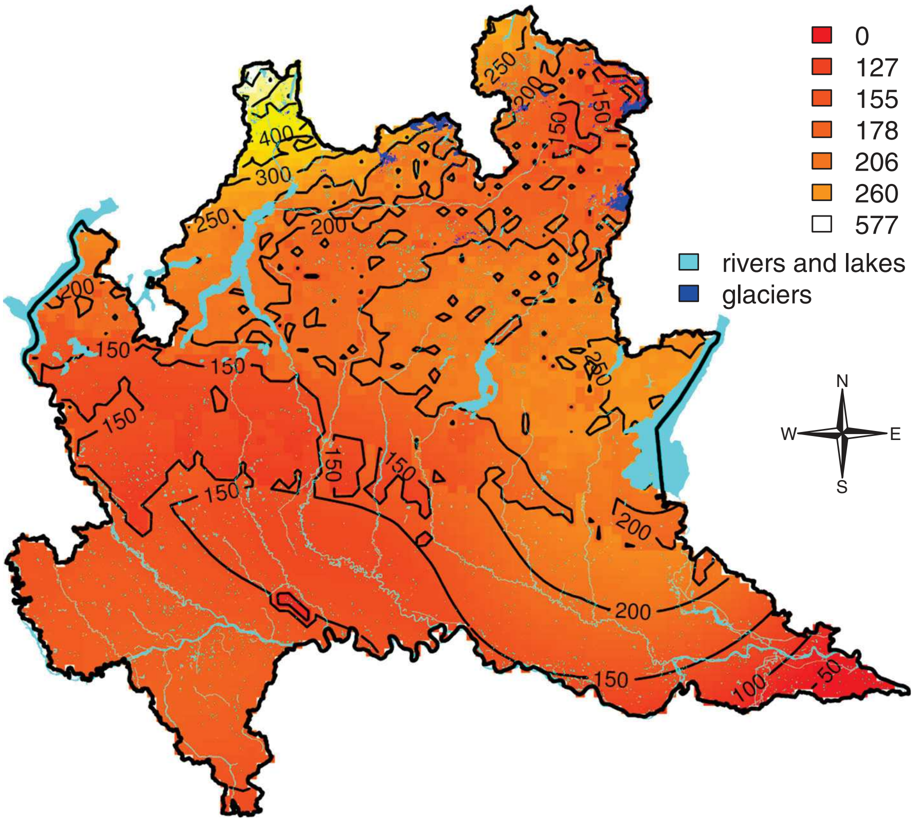


Fig. 5 Fitted surface of the 85th M-quantile. The values in the legend represent the minimum, the maximum and five equispaced percentiles of the 85th M-quantiles estimated by the model in Eq. (8) at the grid points

4651 points internal to the administrative boundary of the study region. For each point of the grid, the geo-lithologic class has been retrieved by overlaying the point on the map in Fig. 3 and the distance to its nearest tectonic lineament has been calculated by applying Eq. (1). Analogously, the spline base $\mathbf{z}_{sp}$ has been calculated for each location of the grid by applying Eq. (5). As far as the fixed effects are concerned, the value of the covariates are multiplied by the estimated coefficients shown in Table 2, whereas the random-effects vector $\mathbf{u}$ and γ in Eq. (8) have been predicted using a modified Fellner equation (Fellner 1986) as proposed by Borgoni et al. (2018) for the three-level M-quantile random effect models. The issue of predicting random effects is also discussed by Geraci and Bottai (2014) and Tzavidis et al. (2016). The estimated M-quantiles are calculated by summing the different components according to Eq. (8) and a raster of 4,651 pixels ~~are~~ eventually obtained and ~~are~~ displayed in Fig. 5.

In order to identify radon-prone areas, Fig. 5 has been transformed into a binary map by colouring those pixels in red, where the estimated M-quantile is above the reference level. Figure 6a, b show the results using a reference level τ corresponding to 200 Bq/m^3 and 300 Bq/m^3, respectively. The latter reference level has been suggested by the recent 2013/59/EURATOM European recommendation as a suitable reference value for the annual average indoor concentration of radon, whereas the former was suggested by the 90/143/Euratom recommendation and it was widely used in the past.

4.2 Assessing the role of influential factors on IRC

As noticed in Sect. 2 there are a number of factors in addition to space and geological dimensions that can potentially affect the concentration of radon in an indoor environment, such as building-specific characteristics. The exploratory analysis also suggests that the impact of a given characteristic can be different at different concentration levels. Hereafter, the conditional distribution of IRC is modelled using the approach introduced in Sect. 3 as a function of these building factors in order to quantify their potential effects and how they differ at different levels of IRC. Hence, the model in Eq. 8 is expanded to include these covariates as fixed effects. The baseline house is located in a building in direct contact with the ground, equipped with an air conditioning system and constructed or refurbished in 1990 or before with walls made by materials other than stone and floors made by materials other then marble or granite.

Table 3 shows the estimated parameters for three different M-quantiles, 0.25, 0.5 and 0.75. As in the section

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ☑ CP | ☑ DISK

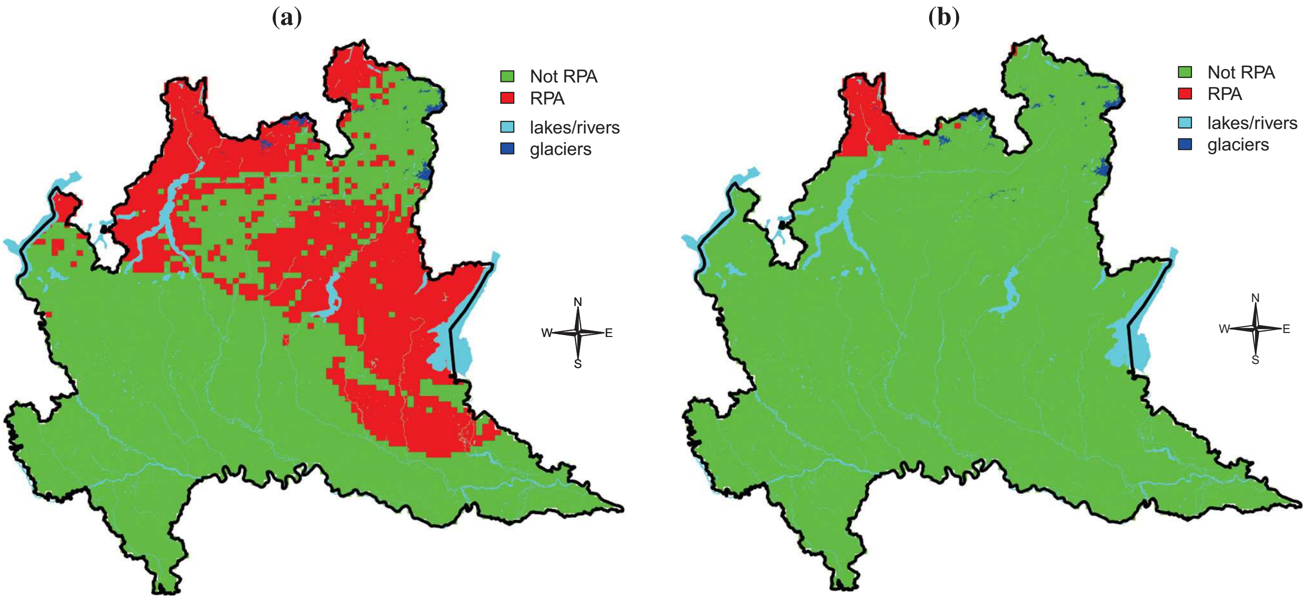


Fig. 6 Radon-prone areas (RPA) according to a reference value of **a** 200 Bq/m^3 and **b** 300 Bq/m^3

Table 3 Results—Semiparametric M-quantile random effect model: point estimates with standard errors in parentheses

	$q = 0.25$		$q = 0.50$		$q = 0.75$	
	Estimate	p value	Estimate	p value	Estimate	p value
Intercept	27.24	0.044	45.59	0.069	84.93	0.256
	(13.53)		(25.12)		(74.86)	
Distance to nearest fault	− 3.21	0.697	− 10.25	0.446	− 35.19	0.21
	(8.27)		(13.47)		(28.08)	
Floor: marble or granite	− 5.41	0.328	− 8.88	0.316	− 22.43	0.195
	(5.54)		(8.87)		(17.32)	
Wall: stone	3.77	0.338	6.53	0.301	16.86	0.173
	(3.94)		(6.31)		(12.38)	
Years of construction/last renovation: after 1990 single buildings	7.99	0.001	11.25	0.005	14.11	0.077
	(2.55)		(4.09)		(8.003)	
	6.35	0.018	10.36	0.016	24.98	0.003
	(2.69)		(4.32)		(8.49)	
Not in contact with the ground no air conditioning	− 5.73	0.032	− 9.74	0.022	− 18.24	0.029
	(2.67)		(4.28)		(8.35)	
	1.09	0.832	− 1.48	0.858	− 7.59	0.64
	(5.16)		(8.29)		(16.24)	
Longitude	33.4	0.03	42.43	0.16	87.73	0.369
	(15.44)		(30.23)		(97.85)	
Latitude	36.62	0.05	48.81	0.165	38.58	0.716
	(18.77)		(35.19)		(106.14)	
σ^2_ϵ (individual)	819.88		3172.99		6283.1	
	(48.91)		(241.93)		(421.41)	
σ^2_u (geo-lithology)	42.21		149.7		220.13	
	(26.41)		(85.48)		(92.08)	
σ^2_γ (spline)	75.96		491.78		3140.52	

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ✔ CP | ✔ DISK

above, 50 knots have been used for the radial spline, obtained by applying the CLARA algorithm to the sampling locations. Appendix 2 provides a short sensitivity analysis where the estimated parameters of quantile and M-quantile regression models are compared. Although the two approaches cannot be directly compared, as these models target different location parameters, the results show that the coefficients of the M-quantile regression model are in the same direction as the ones based on quantile regressions.

Figure 7 shows the estimated effect by M-quantile of each covariate that we have included in the model. Confidence bands across the M-quantiles are also reported in the graphs to display the sampling variation. For each considered M-quantile, the band is obtained by calculating the point-wise 95% confidence interval of the regression coefficients and it is displayed in the graph by a grey-shaded area around the line. It can be seen that the variation between M-quantiles is diverse, sometimes even extensively so, and it tends to increase at the edges of the M-quantile order, where such an increase can be quite large. This is quite typical for quantile modelling, since estimates too far from the centre of the distribution usually cannot be determined with high precision. This problem can possibly be exacerbated by the hierarchical structure of the data. The clustering of the measurement points (about 900 in all) in 11 classes implies that the tail of the distributions cannot be well frequented by the data, as is also shown by the conditional boxplot in Fig. 4, contributing to reducing the information.

Concerning the random components of the model at the lower quantile, a large part of the variation is due to individual variability whereas both the variability due to geo-lithology and space are definitely minor, as one could expect. Moving towards higher quantiles, the spatial component tends to become more and more relevant. Figure 8 shows the estimated spline effects at the three considered M-quantiles. The estimated effects are obviously larger at higher quantile orders. The maps also show that a substantial homogeneity in space exists at the lower quartile (Fig. 8a) apart from some picks in the mountains in the far north and south-east of the region. At high orders, the spatial dynamics tend to vary more due to the large-scale tendency over the investigated region and due to local effects caught by the semiparametric component of the model.

We finally observe that an important part of spatial prediction refers to the measurement of uncertainty. The predicted spline effects at M-quantiles have an uncertainty that could be estimated following Ruppert et al. (2003) and Opsomer et al. (2008). As Ruppert et al. (2003) noticed, the mixed model formulation of penalised splines is a convenient artefact for estimating the smoothing parameters while the ML or REML variance component estimation provides estimates of the smoothing parameter that generally behaves quite well. The standard errors derived according to the approach suggested by these authors are expected to account for both the error components (variance and squared bias) and can be somewhat wider than those obtained without using a mixed model representation. A similar approach can also be adopted to calculate the standard error for the predicted bivariate spline effects at M-quantiles. However, this is beyond the scope of this work and it is left for further research.

Quite surprisingly, the geo-lithological component remains negligible even at the highest quantile. This can be due to the spatial resolution of the geological and lithological information. The maps available for this analysis are scaled 1:250,000, hence, different geological structures can be mixed up in different classes because of a low resolution inducing the inhomogeneity of the geological units, and the geo-lithological effect can be watered down.

5 Discussion and conclusions

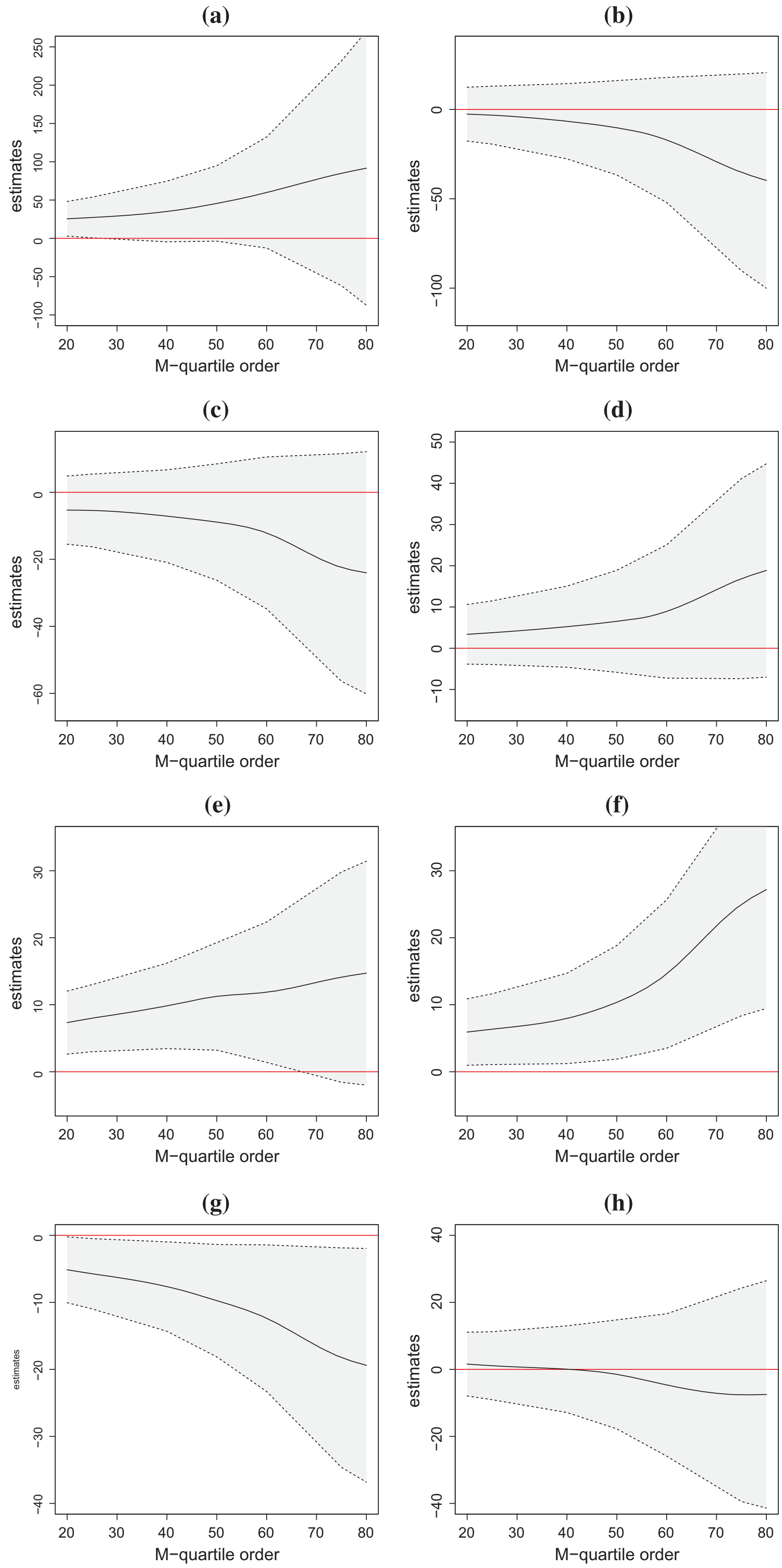
In this paper, a semiparametric random-effect M-quantile model is introduced in order to investigate radon concentration within buildings. It has been shown how the model can be estimated within the framework of robust maximum likelihood by using a numerical optimisation method based on the Newton-Raphson and fixed-point algorithms that apply the data of an indoor radon gas monitoring survey carried out by the Agency of Environmental Protection of Lombardy (Italy) in 2003.

The proposed model allows for an investigation of how a set of covariates acts at different levels (M-quantiles) of the IRC distribution while accounting for the hierarchical nature of the data. In particular, a set of building characteristics are included in the model as well as information concerning the geological structure of the soil.

One of the building characteristics that was found to be statistically significant at M-quantiles was whether the building is in direct contact with the ground and the building type. Buildings in contact with the ground and detached buildings are found to have a higher indoor concentration than other buildings and the impact of those variables becomes larger as the M-quantile order increases (i.e. for those situations more seriously affected by large concentrations). This is not an unexpected result. Unlike many other indoor air pollutants that are correlated to outdoor air pollution, radon gas concentrations in homes are related primarily to the ingression of radon from ground sources. Hence, being in contact with the ground fosters gas accumulation. Condominiums are often constructed out of concrete. The radium content of the concrete is typically

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

Fig. 7 Estimated coefficients of M-quantile regressions: **a** intercept, **b** fault distance, **c** floor material, **d** wall material, **e** year from construction/last renovation, **f** single building, **g** not in contact with the ground, **h** air conditioning system. Shaded areas represent the 95% confidence intervals



Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

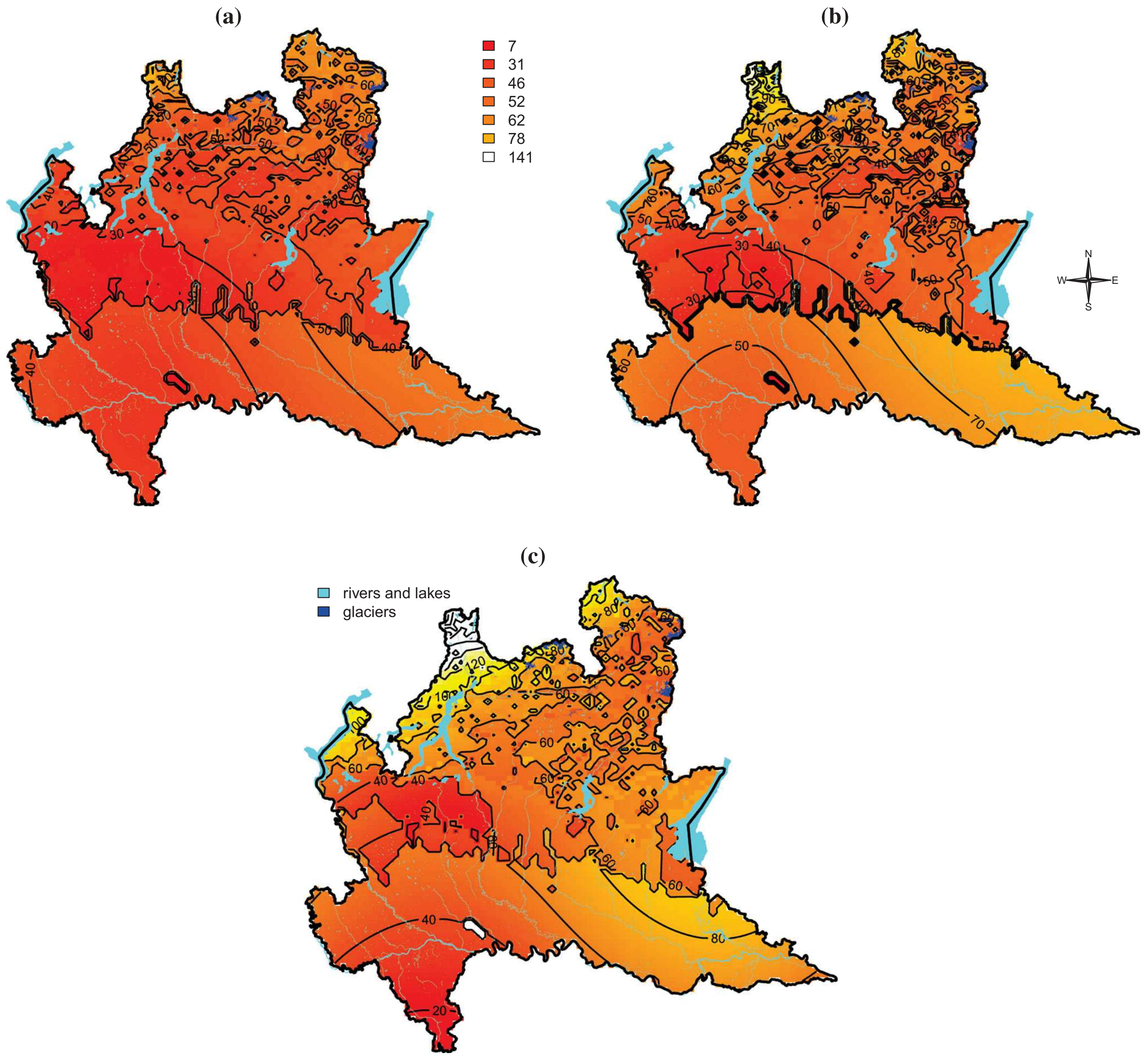


Fig. 8 Spline effects at the three considered M-quantiles: **a** 25th, **b** median, and **c** 75th

not high and this may explain why high-density housing is characterised by lower radon concentration. The statistical significance of the regression coefficient associated with whether the building was built or refurbished after 1990 is mild and not completely clear at all M-quantile orders. There is widespread belief that increased weatherproofing and the energy efficiency of homes significantly contribute to the increase in residential radon concentrations. Nonetheless, uncertainty remains about their actual impact, in particular, whether energy efficiency guidelines include the consideration of air exchange rates and ventilation. Finally, although not all the variables considered in the model have been found to be statistically significant, it can be observed that the estimates tend to be larger in modulus moving towards higher M-quantiles, suggesting that building characteristics can be expected to be effective potential levers for moderating critical situations.

Our findings provide useful indications in this direction, helping to identify those factors that mainly foster high concentration levels of the pollutant on a large and inhomogeneous territory with several different house typologies. Using the estimated regression coefficients, it is possible to classify the different typologies of buildings based on a selected M-quantile of the IRC distribution and to provide a ranking of the dwellings according to their proneness to IRC. To this end, we considered the

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ☑ CP | ☑ DISK

M-quantile regression of order $q = 0.75$ that models how dwelling characteristics impact at a high level of the pollutant distribution. More specifically, fixating the spatial location of a building to a prefixed value, we combined the variables found to be statistically significant at the M-quantile $q = 0.75$ (p values smaller than 0.05), namely whether the building is in contact with the ground and whether it is a single building, with the geo-lithological classes obtaining a set of $2 \times 2 \times 11 = 44$ different building profiles for which the considered M-quantile has been estimated. For instance, for a single building that is not in contact with the ground located in a debris class, replacing the unknown parameters in Eq. (4) with their estimates (see Table 3) and ignoring the spatial component of the model, gives:

$$\begin{aligned}\widehat{MQ}_y(0.75) = {} & 84.93 + 24.98 \times I(single\ building = 1) + \\ & (-18.24) \times I(not\ in\ contact\ with\ the\ ground = 1) + \\ & 20.89 \times I(geo - lithological\ class = debbris) = 112.55\end{aligned}$$

where 20.89 is the estimated residual for the debris class. These profiles are ranked according to their estimated value of the 0.75 M-quantile from the highest to the lowest, obtaining a measure of the building's proneness to a high IRC level. The top-five building profiles most prone to IRC are listed in Table 4.

Selecting an arbitrary location to provide the different scenarios can be done without loss of generality. Although the actual location does affect the estimated IRC M-quantile, the ranking of profiles provided by the M-quantile model based on the building characteristics does not change, considering other locations in space given the additive nature of the model. Looking at the five profiles most prone to high IRC listed in Table 4 we found that all of them are single buildings and four out of five are in contact with the ground and located on porous soils or soils characterised by weathered carbonate rocks (such as debris and dolomite) where radon emanation is known to be high despite the low concentrations of uranium. Hence, our results can help local authorities involved in environmental protection both to identify some guidelines for new buildings and to identify those dwelling typologies already present in the territory that should be monitored in order to mitigate concentration levels.

Given the serious health-related problems induced by the exposure of humans to radon gas, the usage of monitoring surveys for the identification of areas more prone to high IRC, named radon-prone areas above, has been promoted in many countries worldwide. We demonstrated how the semiparametric M-quantile model proposed in this paper provides a natural way to identify such areas flexibly and effectively, taking into account (1) the spatial dynamic of IRC via flexible bivariate spline transformations, and (2) information related to the geological and geophysical information included in fixed and random components of the model. It is worth noticing that the information concerning the geology of the soil is conveyed via digital maps that can be linked to the dataset by GIS operations. Hence, the spatial resolution of this information may have an impact on the precision of the estimates, and having high-resolution maps for those dimensions may sensibly improve the outcome of the analysis.

We also show how the outcome of the model can be visualised by using thematic maps. Such maps inform people and local authorities of where higher concentrations can be expected. Local authorities can use the maps to differentiate construction requirements according to different locations.

Finally, we recognise that when considering a complex phenomenon such as radon gas accumulation, the set of relevant factors may be larger than the one considered in the present paper, and adding these further control variables might lead to improved results. In particular, as has been demonstrated, for instance by Kemski et al. (2009), soil radon measurements are often considered to be a potential predictor of indoor concentration since soil gas containing radon leaks into houses through cracks or holes in the foundations because of the lower air pressure observed indoors compared to outside. Radiometric data has sometimes also been used to account for the radioactivity of the soil.

Numerous weather-related factors influence the ingress of radon into buildings, including wind, barometric pressure, rainfall, and indoor and outdoor difference of temperature variations (Rowe et al. 2002). Increased wind can exert small pressure differences between the lower levels of a dwelling and the outdoors and an increased precipitation can act to impede radon emanation. Climate

Table 4 The top-five building profiles most prone to IRC

Ranking	Contact with the ground	Building type	Geo-lithological class
1	In contact	Single	Debris
2	In contact	Single	Dolomite rocks
3	In contact	Single	Alluvial fan
4	Not in contact	Single	Debris
5	In contact	Single	Limestone

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ☑ CP | ☑ DISK

parameters affect the ventilation of indoor environment as well, in turn influencing the indoor concentration of the pollutant. Unfortunately, we did not have this information at hand for this study, and hence their effect has not been investigated, although it can easily be added to the model proposed in this paper, if available.

Acknowledgements ~~The work of Nicola Salvati has been developed under the support of the Progetto di Ricerca di Ateneo 'From survey-based to register-based statistics: a paradigm shift using latent variable models' (Grant PRA2018-9).~~ The authors were further supported by the MIUR-DAAD Joint Mobility Program (57265468).

Appendix A: Preliminary data analysis

Hereafter some preliminary data analyses is reported that motivates the need for a robust approach when modelling IRC data. To this aim an ordinary random effect model for the mean IRC that reflects the hierarchical structure of the data with buildings nested in the geological classes has been fitted using the function `lmer` of the `R` package `lme4`. Figure 9a shows the normal qq-plot of the individual residuals (i.e. residuals pertinent to the building level) whereas Fig. 9b displays the normal qq-plot of the residuals estimated from the model at the geological class level. These plots show that the normality assumptions of the ordinary mixed model are violated, which is also confirmed by the Shapiro-Wilk test (p values=0.0000078 for the geological class residuals and p value= 2.2e−16 for the building residuals). Figure 10a shows the histogram of the standardised building residuals obtained by the random effect regression model, whereas Fig. 10b displays the distribution of the standardised residuals by geological classes. The histogram appears very skewed and some classes have many large positive residuals (larger than 2). Thus, influential observations seem to be present in the data. This is also confirmed by Fig. 11 that displays the Cook's Distance for the two sets of residuals.

It is clear that the data may contain outliers and influential points that invalidate the Gaussian assumptions. In these circumstances, estimates of the model parameters are biased and inefficient and the robust approach suggested in this paper sounds more appropriate.

Appendix B: Additional results for modelling geocoded radon data

Appendix 1 provides a short comparison of the estimated parameters obtained from quantile and M-quantile regression models. The two approaches cannot be directly compared since they target different location parameters. However, both approaches try to model location parameters that are related to the same part of the conditional distribution of IRC. Table 5 reports the estimated regression coefficients for $q = 0.5$ for two approaches: (1) the proposed semiparametric M-quantile random effect regression model (semiMQRE), and (2) a semiparametric quantile regression model (semiQR). semiQR is based on an additive quantile regression model (Koenker et al. 1994) where the spatial structure is captured by bivariate splines but without accounting for the hierarchical structure in the data by a random component. The results indicate that the coefficients based on M-quantile regression models are in the same direction as the ones based on quantile regression. However, with quantile regression convergence problems

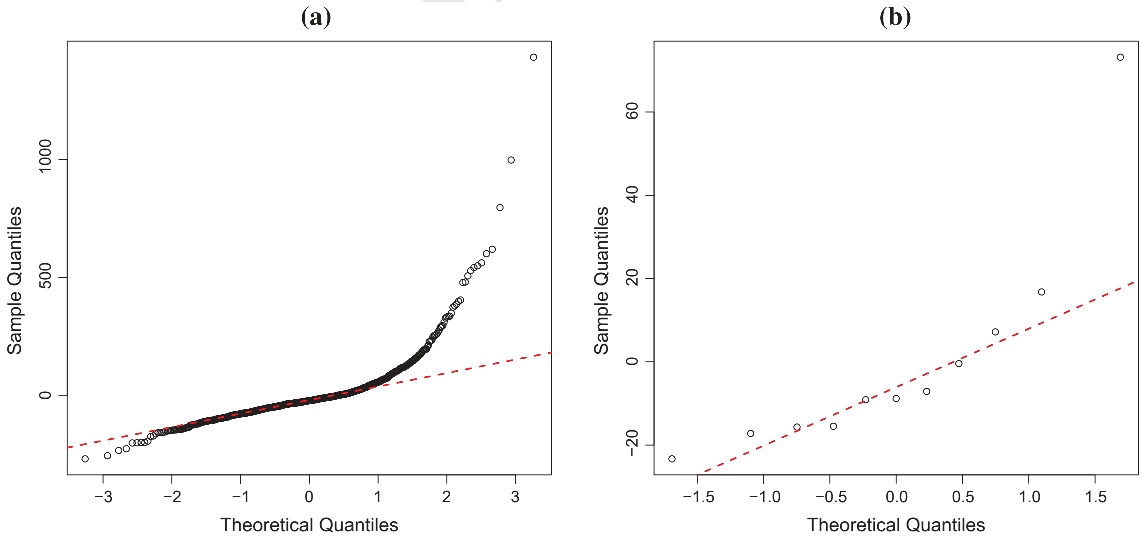


Fig. 9 QQ-plot of building residuals (**a**) and of geological class residuals (**b**) estimated by the two-level random effect regression model for the mean IRC

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

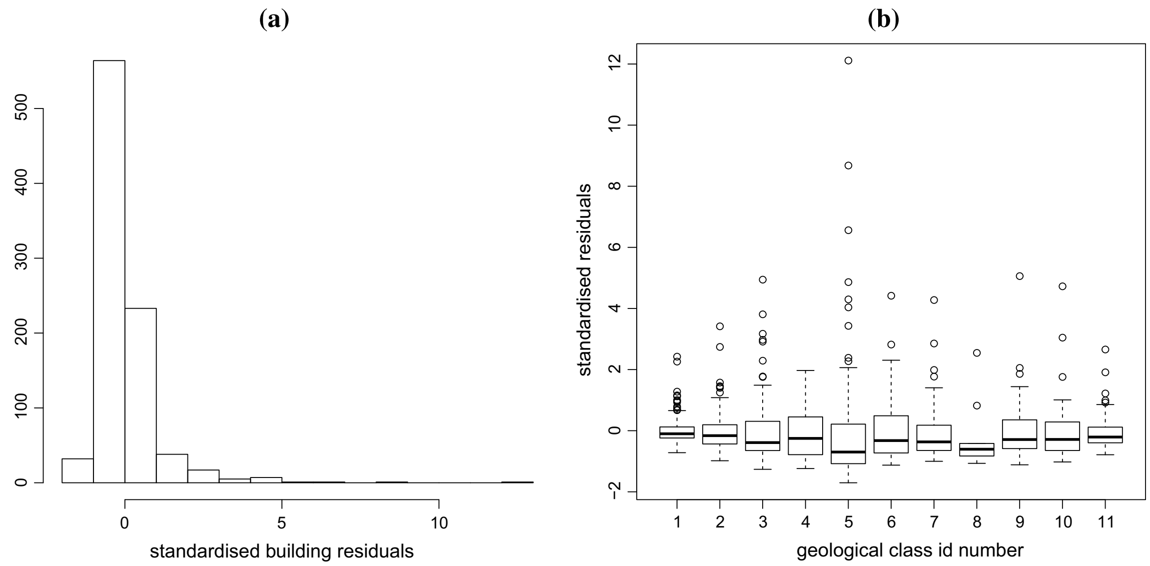


Fig. 10 Histogram of standardised building residuals of the two-level random effect regression model for the mean IRC (**a**); boxplots of standardised building residuals by geological classes (**b**)

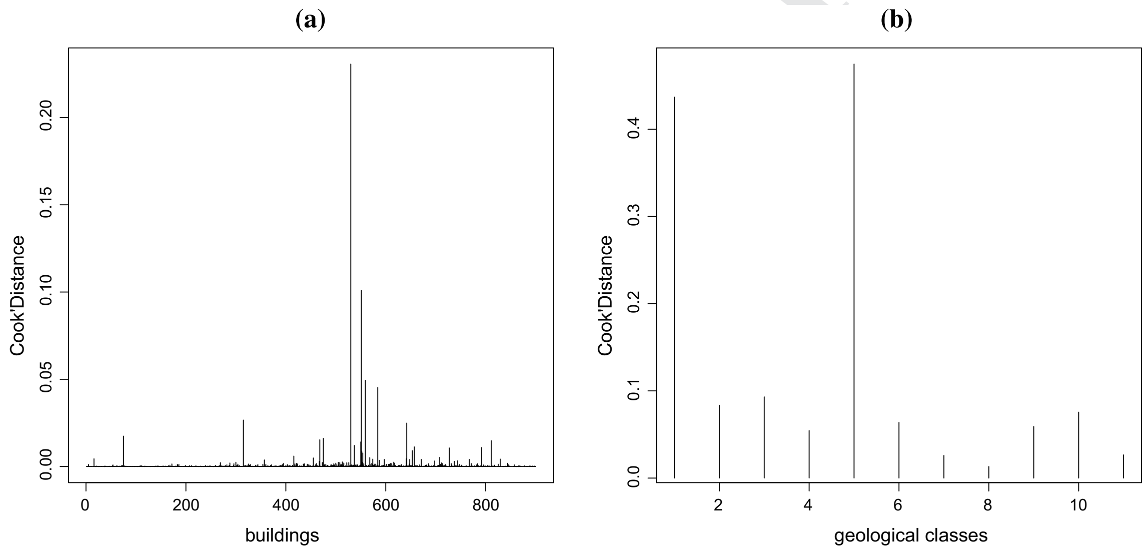


Fig. 11 Cook Distance of building residuals (**a**) and of geological class residuals (**b**) estimated by the two-level random effect regression model for the mean IRC

Table 5 Results—Semiparametric M-quantile and quantile regression models for $q = 0.5$: Point estimates with standard errors in parentheses

	semiMQRE		semiQR	
	Estimate	*p* value	Estimate	*p* value
Intercept	45.59	0.069	44.56	0.000
	(25.12)		(11.02)	
Distance to nearest fault	− 10.25	0.446	− 8.14	0.553
	(13.47)		(13.71)	
Floor: marble or granite	− 8.88	0.316	− 7.41	0.406
	(8.87)		(8.90)	

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

Table 5 (continued)

	semiMQRE		semiQR	
	Estimate	p value	Estimate	p value
Wall: stone	6.53	0.301	− 1.98	0.827
	(6.31)		(9.07)	
Years of construction/last renovation: after 1990 single buildings	11.25	0.005	7.61	0.112
	(4.09)		(4.78)	
	10.36	0.016	9.62	0.043
	(4.32)		(4.74)	
Not in contact with the ground no air conditioning	− 9.74	0.022	− 9.63	0.048
	(4.28)		(4.87)	
	− 1.48	0.858	− 1.10	0.883
	(8.29)		(7.48)	
Longitude	42.43	0.16	–	–
	(30.23)		–	
Latitude	48.81	0.165	–	–
	(35.19)		–	
σ_ϵ^2 (individual)	3172.99		–	
	(241.93)		–	
σ_u^2 (geo-lithology)	149.7		–	
	(85.48)		–	
σ_γ^2 (spline)	491.78		–	

of the algorithm sometimes occurred. On the other hand, estimation with the M-quantile approach was smoother but the interpretation of the estimated parameters is more difficult.

Finally, Fig. 12 presents the estimated effects obtained from M-quantile and quantile-mixed regression models by quantile for each explanatory variable that is considered in the model. In particular, the solid line represents the proposed semiparametric M-quantile random effect regression model and the dashed line stands for an additive quantile regression model (Geraci 2018) which includes a bivariate spline to capture the spatial structure as well as random effects to account for the hierarchy of the data (fitted by the R package *aqmm*). Note that we only plot the point estimates (without the point-wise 95% confidence intervals) in order to avoid an overload of Fig. 12. The results confirm that the results based on both models are in the same direction.

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | □ LE | □ TYPESET
MS Code : SERR-D-18-00069 | ✔ CP | ✔ DISK

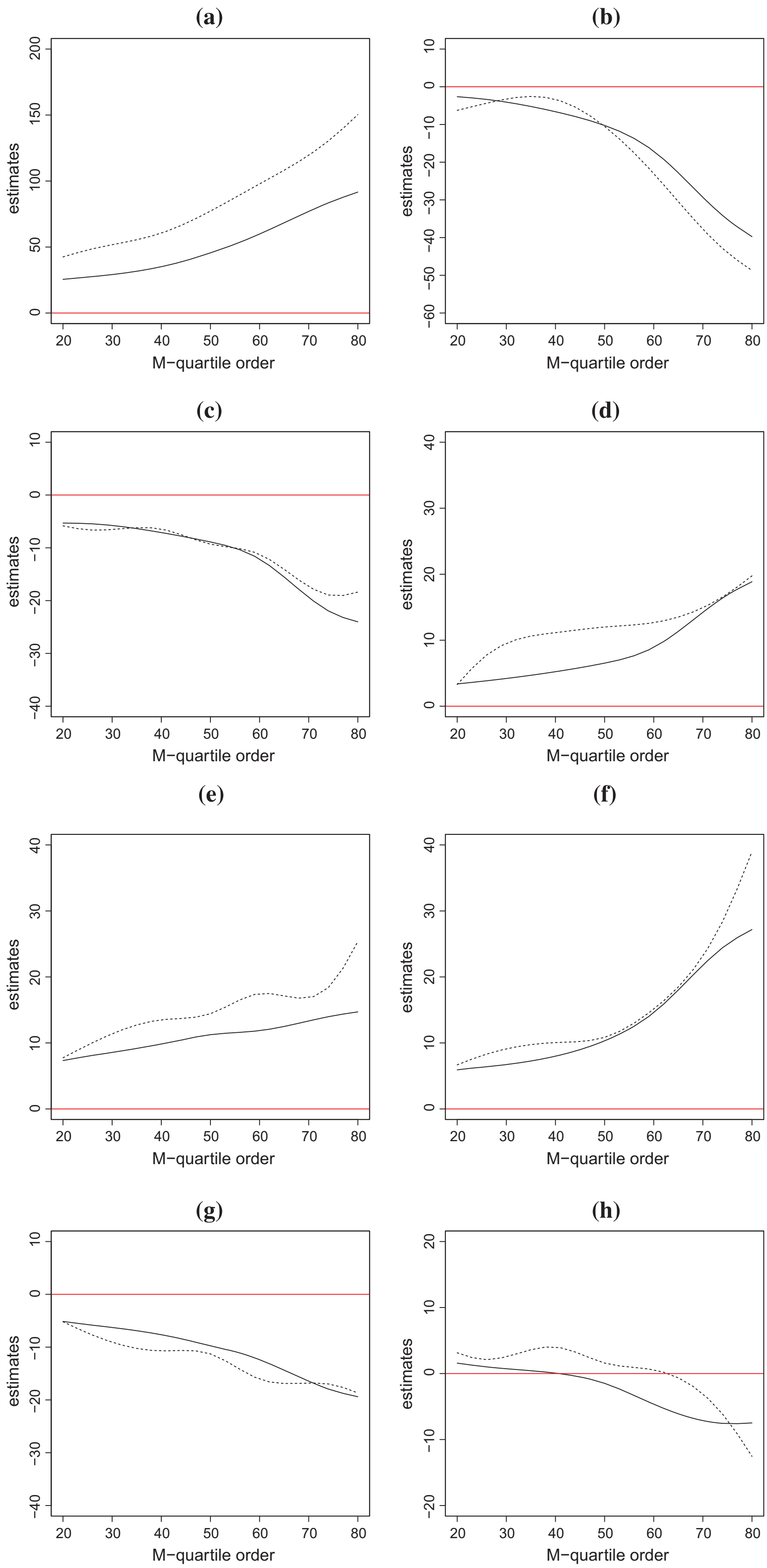


Fig. 12 Estimated coefficients of quantile regressions (dashed line) and M-quantile regressions (solid line): **a** intercept, **b** fault distance, **c** floor material, **d** wall material, **e** year from construction/last renovation, **f** single building, **g** not in contact with the ground, **h** air conditioning system

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | LE | TYPESET
MS Code : SERR-D-18-00069 | CP | DISK

References

Alfó M, Ranalli M, Salvati N (2017) Finite mixtures of quantiles and m-quantile models. Stat Comput 27:547–570

Apte M, Price P, Nero A, Revzan K (1999) Predicting new hampshire indoor radon concentrations from geologic information and other covariates. Environ Geol 37:181–194

AQ3 Bianchi A, Fabrizi E, Salvati N, Tzavidis N (2018) Estimation and testing in M-quantile regression with applications to small area estimation. Int Stat Rev 1–30

Borgoni R (2011) A quantile regression approach to evaluate factors influencing residential indoor radon concentration. Environ Model Assess 16:239–250

Borgoni R, Bianco PD, Salvati N, Schmid T, Tzavidis N (2018) Modelling the distribution of health-related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using m-quantile random effects regression. Stat Methods Med Res 27:549–563

Borgoni R, Quatto P, Soma G, de Bartolo D (2010) A geostatistical approach to define guidelines for radon prone area identification. Stat Methods Appl 19:255–276

Borgoni R, Tritto V, Bigliotto C, de Bartolo D (2011) A geostatistical approach to assess the spatial association between indoor radon concentration, geological features and building characteristics: the Lombardy case, Northern Italy. Int J Environ Res Public Health 8:1420–1440

Bosch RJ, Ye Y, Woodworth GG (1995) A convergent algorithm for quantile regression with smoothing splines. Comput Stat Data Anal 19(6):613–630

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75(4):761–771

Cade B, Noon BR, Flather CH (2005) Quantile regression reveals hidden bias and uncertainty in habitat models. Ecology 86:786–800

Chaudhuri P (1991) Global nonparametric estimation of conditional quantile functions and their derivatives. J Multivar Anal 39(2):246–269

Cinelli G, Tondeur F, Dehandschutter B (2011) Development of an indoor radon risk map of the Walloon region of Belgium, integrating geological information. Environ Earth Sci 62:809–819

Darby S, Hill D, Auvinen A, Barros-Dios J, Baysson J, Bochicchio F, Deo H, Falk R, Forastiere F, Hakama M, Heid I, Kreienbrock L, Kreuzer M, Lagarde F, MSkelSinen I, Muirhead C, Oberaigner W, Pershagen G, Ruano-Ravina A, Ruosteenoja E, Rosario AS, Tirmarche T, Tomsek L, Whitley E, Wichmann H, Doll R (2005) Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case–control studies. Br Med J 330(6485):223–226

Fellner WH (1986) Robust estimation of variance components. Technometrics 28(1):51–60

Fontanella L, Ippoliti L, Sarra A, Valentini P, Palermi S (2015) Hierarchical generalised latent spatial quantile regression models with applications to indoor radon concentration. Stoch Environ Res Risk Assess 29:357–367

Foxall R, Baddeley A (2002) Nonparametric measures of association between a spatial point process and a random set, with geological applications. J R Stat Soc Ser C 51(2):165–182

Gates A, Gundersen L (1992) Geologic controls on radon. Geological Society of America, Washington, DC (Special Paper 271)

AQ4 Geraci M (2018) Additive quantile regression for clustered data with an application to children's physical activity. ArXiv e-prints

Geraci M, Bottai M (2014) Linear quantile mixed models. Stat Comput 24(3):461–479

Green B, Miles J, Bradley E, Rees D (2002) Radon atlas of England and Wales. Report nrpb-w26, Chilton NRPB

Gunby J, Darby S, Miles J, Green B, Cox D (1993) Indoor radon concentrations in the United Kingdom. Health Phys 64:2–12

Huber P (1981) Robust statistics. Wiley, New York

Huggins RM (1993) A robust approach to the analysis of repeated measures. Biometrics 49(3):715–720

Huggins RM, Loesch DZ (1998) On the analysis of mixed longitudinal growth data. Biometrics 54(2):583–595

Hunter N, Muirhead C, Miles J, Appleton JD (2009) Uncertainties in radon related to house-specific factors and proximity to geological boundaries in England. Radiat Prot Dosim 136:17–22

Jacobi W (1993) The history of the radon problem in mines and homes. Ann ICRP 23(2):39–45

Jones M (1994) Expectiles and m-quantiles are quantiles. Stat Probab Lett 20:149–153

Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kemski J, Klingel R, Siehl A, Valdivia-Manchego M (2009) From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ Geol 56:1269–1279

Koenker R (2005) Quantile regression. Cambridge University Press, New York

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Koenker R, Mizera I (2004) Penalized triograms: total variation regularization for bivariate smoothing. J R Stat Soc Ser B 66(1):145–163

Koenker R, Ng P, Portnoy S (1994) Quantile smoothing splines. Biometrika 81(4):673–680

Kreienbrock L, Kreuzer M, Gerken M, Dingerkus M, Wellmann J, Keller G, Wichmann H (2001) Case-control study on lungcancer and residential radon in western Germany. Am J Epidemiol 89(4):339–348

Krewski D, Lubin MAJH, Zielinski JM, Catalan V, Field R, Klotz J, Letourneau E, Lynch C, Lyon J, Sandler D, Schoenberg D, Steck J, Stolwijk C, Weinberg C, Wilcox H (2005) Residential radon and risk of lung cancer: a combined analysis of seven North American case-control studies. Epidemiology 16(4):137–145

Levesque B, Gauvin D, McGregor R, Martel R, Gingras S, Dontigny A, Walker W, Lajoie P, Levesque E (1997) Radon in residences: influences of geological and housing characteristics. Health Phys 72:907–914

Lubin J, Boice J (1997) Lung cancer risk from residential radon: a meta-analysis of eight epidemiological studies. J Natl Cancer Inst 89(1):49–57

Nero A, Schwehr M, Nazaroff W, Revzan K (1986) Distribution of airborne radon-222 concentrations in US homes. Science 234:992–997

Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica 55(4):819–847

Opsomer J, Claeskens G, Ranalli M, Kauermann G, Breidt F (2008) Nonparametric small area estimation using penalized spline regression. J R Stat Soc Ser B 70(1):265–283

Organization WH (2009) WHO handbook on indoor radon: a public health perspective. WHO Library Cataloguing-in-Publication Data

Pratesi M, Ranalli M, Salvati N (2009) Nonparametric m-quantile regression using penalized splines. J Nonparametr Stat 21:287–304

Price P, Nero A, Gelman A (1996) Bayesian prediction of mean indoor radon concentrations for Minnesota counties. Health Phys 71:922–936

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Journal : Large 477 | Dispatch : 26-12-2018 | Pages : 20
Article No. : 1643 | ☐ LE | ☐ TYPESET
MS Code : SERR-D-18-00069 | ☑ CP | ☑ DISK

Rowe J, Kelly M, Price L (2002) Weather system scale variation in radon-222 concentration of indoor air. Sci Total Environ 284:157–166

Ruppert D, Wand M, Carroll R (2003) Semiparametric regression. Cambridge University Press, Cambridge

Sarra A, Fontanella L, Ippoliti L, Valentini P, Palermi S (2016) Quantile regression and Bayesian cluster detection to identify radon prone areas. J Environ Radioact 164:354–364

Shi X, Hoftiezer D, Duell E, Onega T (2006) Spatial association between residential radon concentration and bedrock types in New Hampshire. Environ Geol 51:65–71

Smith B, Field R (2007) Effect of housing factor and surficial uranium on the spatial prediction of residential radon in Iowa. Environmetrics 18:481–497

Smith B, Zhang L, Field R (2007) Iowa radon leukemia study: a hierarchical population risk model. Stat Med 10:4619–4642

Sundal A, Henriksen H, Soldal O, Strand T (2004) The influence of geological factors on indoor radon concentrations in Norway. Sci Total Environ 328:41–53

Tiefelsdorf M (2007) Controlling for migration effects in ecological disease mapping of prostate cancer. Stoch Environ Res Risk Assess 21:615–624

Tzavidis N, Salvati N, Schmid T, Flouri E, Midouhas E (2016) Longitudinal analysis of the strengths and difficulties questionnaire scores of the millennium cohort study children in England using m-quantile random effects regression. J R Stat Soc Ser A 179(2):427–452

USEPA (1992) National residential radon survey: summary report. Technical Report EPA/402/R-92/011, United States Environmental Protection Agency, Washington, DC

Wang Y, Lin X, Zhu M, Bai Z (2007) Robust estimation using the Huber funtion with a data dependent tuning constant. J Comput Graph Stat 16(2):468–481

Yu K, Lu Z, Stander J (2003) Quantile regression: applications and current research areas. Statistician 52(3):331–350

Journal : Large 477 Dispatch : 26-12-2018 Pages : 20
Article No. : 1643 ☐ LE ☐ TYPESET
MS Code : SERR-D-18-00069 ☑ CP ☑ DISK