



OPEN ACCESS

EDITED BY

Antoaneta Ene,
Dunarea de Jos University, Romania

REVIEWED BY

Cucos Alexandra Laura,
Babeş-Bolyai University, Romania
Hesham M. H. Zakaly,
Ural Federal University, Russia
Giuseppe La Verde,
University of Naples Federico II, Italy

*CORRESPONDENCE

Joan F. Rey
✉ joan.rey@epfl.ch

RECEIVED 05 July 2024

ACCEPTED 02 September 2024

PUBLISHED 11 September 2024

CITATION

Rey JF, Antignani S, Baumann S,
Di Carlo C, Loret N, Gréau C, Gruber V,
Goyette Pernot J and Bochicchio F (2024)
Systematic review of statistical methods for
the identification of buildings and areas with
high radon levels.
Front. Public Health 12:1460295.
doi: 10.3389/fpubh.2024.1460295

COPYRIGHT

© 2024 Rey, Antignani, Baumann, Di Carlo,
Loret, Gréau, Gruber, Goyette Pernot and
Bochicchio. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Systematic review of statistical methods for the identification of buildings and areas with high radon levels

Joan F. Rey^{1,2*}, Sara Antignani³, Sebastian Baumann⁴,
Christian Di Carlo³, Niccolò Loret³, Claire Gréau⁵,
Valeria Gruber⁴, Joëlle Goyette Pernot¹ and
Francesco Bochicchio³

¹Western Switzerland Center for Indoor Air Quality and Radon (croqAIR), Transform Institute, School of Engineering and Architecture of Fribourg, HES-SO University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland, ²Human-Oriented Built Environment Lab, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, ³Italian National Institute of Health – National Center for Radiation Protection and Computational Physics, Rome, Italy, ⁴Austrian Agency for Health and Food Safety, Department of Radon and Radioecology, Linz, Austria, ⁵Institut de Radioprotection et de Sûreté Nucléaire, Bureau d'Etude et d'expertise du Radon, IRSN, PSE-ENV, SERPEN, BERAD, Fontenay-aux-Roses, France

Radon is a natural and radioactive noble gas, which may accumulate indoors and cause lung cancers after long term-exposure. Being a decay product of Uranium 238, it originates from the ground and is spatially variable. Many environmental (i.e., geology, tectonic, soils) and architectural factors (i.e., building age, floor) influence its presence indoors, which make it difficult to predict. However, different methods have been developed and applied to identify radon prone areas and buildings. This paper presents the results of a systematic literature review of suitable statistical methods willing to identify buildings and areas where high indoor radon concentrations might be found. The application of these methods is particularly useful to improve the knowledge of the factors most likely to be connected to high radon concentrations. These types of methods are not so commonly used, since generally statistical methods that study factors predictive of radon concentration are focused on the average concentration and aim to identify factors that influence the average radon level. In this paper, an attempt has been made to classify the methods found, to make their description clearer. Four main classes of methods have been identified: descriptive methods, regression methods, geostatistical methods, and machine learning methods. For each presented method, advantages and disadvantages are presented while some applications examples are given. The ultimate purpose of this overview is to provide researchers with a synthesis paper to optimize the selection of the method to identify radon prone areas and buildings.

KEYWORDS

radon prone areas and building, public health, statistic, geostatistics, machine learning

1 Introduction

Radon is a naturally occurring radioactive gas that might accumulate indoors and pose a health issue. It is formed from the decay of uranium found in soil, rock, and water (1, 2). Radon can enter buildings through cracks in foundations, gaps around pipes, and other openings, where it can accumulate to potentially harmful levels (2, 3). Prolonged exposure to elevated radon concentrations increases the risk of developing lung cancer, making it a leading cause of lung cancer (2, 4) along with smoking. Furthermore, high radon exposure affects individuals of all ages and backgrounds, with particularly heightened risks for smokers due to the synergistic effect of radon and smoking fumes (5, 6). For these reasons, it became a public health issue, handled by the WHO since the beginning of the 1980's (2). Radon management, regulations and their implementation vary significantly from one country to another, but the common goal is to reduce the average level of radon to which the population is exposed and thus reduce the risk. Each nation adopts its approach to address the challenges posed by radon exposure based on factors such as geographical location, geological composition, and existing infrastructure. European Basic Safety Standards, Euratom Directive 59/2013, requires that Member States identify areas with elevated levels of indoor radon concentration. According to the optimization principle in radiation protection, areas identified as radon-prone and buildings with high radon concentrations are priority targets for intervention to reduce radon-related risks.

Measuring radon levels typically involves deploying dosimeters in buildings, dwellings, schools, and workplaces to assess the concentration of radon gas over a specified period. However, conducting radon measurements in every building is barely impossible due to logistical challenges and resource constraints. The sheer number of buildings, coupled with varying access permissions and the need for prolonged monitoring periods, as all the measurement protocols require, makes comprehensive testing unfeasible. Therefore, it is crucial to explore alternative approaches to identify radon-prone areas and buildings efficiently.

Indoor radon levels can be influenced by various factors, including geological, climatic, building and occupancy characteristics (7). The presence of uranium-rich soil and rocks beneath or surrounding a building can significantly influence indoor radon levels (8–10). The lithology of the region, including the type of rocks and soil composition, plays a crucial role in determining the potential for radon generation (8). Additionally, climatic factors such as temperature, humidity, atmospheric pressure and precipitation can affect radon transport and accumulation within the soil and surrounding environment (7, 11–14). Furthermore, tectonic activity, such as faults and fractures in the Earth's crust, can create pathways for radon gas to migrate from deep geological layers to the surface and into buildings (15–17). Therefore, a comprehensive understanding of the geological, climatic, and tectonic characteristics of an area is essential for assessing and mitigating indoor radon levels effectively. Moreover, various building factors can influence the presence of radon and its concentration levels indoors. Construction materials and building design play a significant role in determining radon infiltration (18–21). For instance, the presence of cracks in the foundation or walls can provide pathways for radon gas to enter a building from the surrounding soil (3). Additionally, the type of flooring, the age and the type of building may all influence

indoor radon concentrations (22, 23). The ventilation system of a building also plays a crucial role in radon mitigation, as proper ventilation can help dilute radon concentrations and more generally promotes a better indoor air quality (18, 24, 25). Moreover, occupancy patterns, such as the number of occupants, the duration of time spent indoors and their activities, can influence radon levels and indoor air quality by affecting indoor air circulation and mixing (7, 22, 26).

A comprehensive understanding of the factors influencing indoor radon levels is essential for guiding well-defined and sustainable public health policies to mitigate the health risks associated with elevated indoor radon levels. By investigating and analyzing these different influences on indoor radon levels, and their inter-relationship, it may be possible to predict indoor radon levels, or at least, the probability of exceeding specific radon levels. These approaches include basic statistical analysis, regressive statistical analysis, geostatistical methods and machine learning (ML) methods. In today's landscape, the abundance of available methods poses a challenge for researchers in selecting the most appropriate method relative to its own context of application. This paper thus aims to delve into main methods used for investigating indoor radon concentrations, with a special focus on high levels, and to assess their respective applications. More specifically, this paper seeks to present a systematic review of statistical methods to identify radon prone areas and buildings as a priority target of intervention to reduce radon related risk, aiming to comprehensively evaluate effectiveness and suitability of the different methods across different contexts and scenarios.

2 Methodology

2.1 Document selection

The literature searches for methods aiming at (1) identifying areas with high radon levels and at (2) identifying buildings with high radon levels that have been performed separately. However, for both purposes, a systematic review of the published studies has been performed using the PRISMA methodology (27). The latter methodology is commonly used by researchers to perform systematic literature review and meta-analysis (27). Among the databases available, we used the following online databases to search for documentation: Web of Science, Scopus and PubMed.

The defined research strategy aimed to collect all the documents, regardless of the document type, the year of publication, the language, and its availability (i.e., open source). All documents have been collected up to the 10th of May 2024. The keywords used for the searches are reported below in Table 1 for the identification of radon prone areas and in Table 2 for the identification radon prone buildings. The wildcard "*" was used to encompass different words that are variations of the same term. The references in the retrieved articles were also assessed as potentially relevant.

2.2 Data extraction and evaluation of the papers

Selected documents were evaluated and information about each one was extracted using a standardized datasheet. The latter was used

TABLE 1 Research criteria to select documents for identifying radon prone areas.

Criteria	Where	Keywords
I	Title OR Abstract	"radon" "(222)Rn" "222Rn" "Rn-222"
II	Title OR Abstract	"radon prone area*" "priority area*" "high radon level*" "elevated radon level*" "high radon concentration*" "high radon exposure*" "high radon potential" "high background" "radon affected area*"
III	Title OR Abstract	"method*" "approach*" "mapping"
IV	Title OR Abstract	NOT "transform*"

Each document selected contained one keyword, at least, in each criterion keywords list. "Or" is implemented with its logical sign: "|". The wildcard "*" was used to encompass different words that are variations of the same term. The criteria I-IV were combined using the "And" Boolean operator.

TABLE 2 Research criteria to select documents for identifying radon prone buildings.

Criteria	Where	Keywords
I	Title OR Abstract	"radon" "(222)Rn" "222Rn" "Rn-222"
II	Title OR Abstract	"building*" "dwelling*" "house*" "school*" "workplace*" "indoor*"
III	Title OR Abstract	"factor*" "characteristic*" "feature*" "parameter*"
IV	Title OR Abstract	"identif*" "detect*" "predict*" "forecast*" "affect*" "influenc*" "impact*"
V	Title OR Abstract	"high*" "elevated" "exceed*" "above" "quantile*"
VI	Title OR Abstract	"approach*" "method*"
VII	Title OR Abstract	"logistic regression" "quantile regression" "random forest" "machine learning"

Each document selected contained one keyword, at least, in each criterion keywords list. "Or" is implemented with its logical sign: "|". The wildcard "*" was used to encompass different words that are variations of the same term. The criteria I-VI were combined using the "And" Boolean operator.

to collect: (1) main information of the paper evaluated; (2) main characteristics of the datasets to which the method has been applied (e.g., number and type of measurements); (3) main information of the method used (e.g., pro and cons, accuracy, conditions for applicability). Based on this information, the paper was either rejected or retained for inclusion in the review. The standardized form used in this study is presented in [Figure 1](#).

2.3 Classification of the results

The systematic literature review resulted in a large and dense dataset, both in terms of identified methods and their application

scope. Although the literature search has been conducted separately on statistical methods for identifying buildings and areas with elevated radon levels, we present the methods without distinguishing between their areas of application, as many methods serve both purposes. Aiming to ease the presentation of the results, we classified the different methods according to their type. The *Application* section of each type of method lists the work in which the method has been used, highlighting their application scope, which is often guided by the type of input data. The proposed classification is not a clear-cut classification, since some methods described in different classes may have features in common. Four broad categories were identified: (1) descriptive statistical methods; (2) traditional regression (quantile and logistic) methods; (3) geostatistical methods; (4) machine learning methods.

The methods selected for this review will be described in sections dedicated to the four distinct categories, with relevant literature cited as examples of their application. All the methods can generally be used for mapping purposes, that is to characterize a territory in terms of its radon potential thus identifying areas with high radon concentrations. With this aim, these methods might be applied to different types of data, such as indoor radon data or soil radon concentration measurement data. Alternatively, the same methods can be used to identify factors that predict high concentrations in buildings; in this case, the data of interest are typically indoor radon data.

2.4 Evaluation of the methods performances

Working with predictive models (or algorithms) leads to the need to evaluate the performance of a model, that is its ability to make accurate predictions on unseen data.

Several metrics exist to evaluate the performance of a model, and the choice depends on the type of model, whether it is either a regression model (continuous output) or a classification model (nominal or binary output). To test how well a predictive model generalizes, the data set is usually split into training and test data, where the test data is only used for performance evaluation and should be independent from the training set. Alternatively leave-one-out cross validation (k-folding) can be used to test the performance.

In *k-folding* the dataset is subdivided into *k* different folds (dataset portions) which are in turn extracted to test the model built on the remaining (*k-1*) dataset. Since the model is trained and validated multiple times on different subsets of the data, that "crossed" between training and validation roles, *k-fold* cross-validation helps to reduce the risk of overfitting, which occurs when a model performs well on the training data but poorly on unseen data.

A widely used tool for measuring the performance of a classification model (e.g., a logistic model with a binary outcome) is the confusion matrix ([Figure 2](#)), from which various performance metrics, such as accuracy and AUC-ROC curves. Are calculated.

Accuracy is a measure of how well the models' predictions fit real data, based on the number of correct (true) predictions out of all predictions made. The AUC (Area Under the Curve) measures the area underneath the Receiver Operating Characteristic (ROC) curve, that is the plot between the true positive rate (also known as sensitivity) and the false positive rate (also known as 1-specificity).

DOI	Authors	Year	Title	Journal	Abstract	Useful for the review	Reason for rejecting	Summary of findings

Summary of findings								
Country	Area covered by the survey	Rn concentration indexes (average, range)	Number and type of measurements	Method applied	Conditions for the application	Strengths of the method applied	Limits of the method applied	Accuracy of the method applied

FIGURE 1 Standardized form used to evaluate the selected documents.

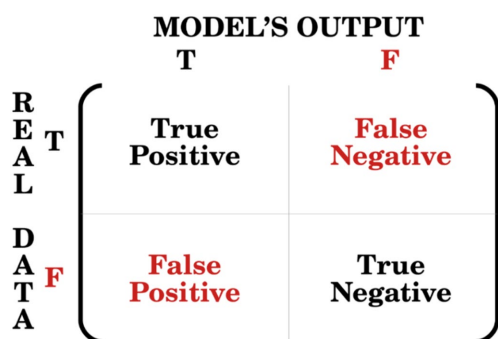


FIGURE 2 Relationship between real data and model's output within a confusion matrix.

Both the accuracy and the AUC-ROC metrics range from 0 to 1, where a higher value indicates better model performance. A value of 0.5 suggests that the model's predictions are not meaningful or useful for distinguishing between classes in the dataset, that is, the model has no discriminative power (equivalent to random guessing). Thus, the 0.5 value could be used as a benchmark for comparison.

Among the most used metrics to evaluate the performance of a regression model are MAE (Mean Absolute Error), that measures the average absolute differences between predicted and actual values, RMSE (Root Mean Squared Error), that is based on the differences between predicted and actual values, and the *R*-squared, which measures the proportion of the variance explained by the model. A better model's performance is associated with higher *R*-squared values and lower MAE or RMSE values. When the RMSE decreases or the *R*-squared increases, the model's performance improves.

3 Results

3.1 Document selection

3.1.1 Radon prone areas

The process of documents selection and exclusion is depicted in Figure 3. Overall, 575 records were found in two databases (PubMed, Scopus) and duplicate records were removed. A total of five

publications were added due to personal communication which sums up to a total of 492 selected publications. All identified papers were in English.

After screening titles and the abstracts of these papers, 445 papers were discarded for different reasons. The major reason for neglecting a publication was that the focus was brought on general radon risk, geology and mining or measurement techniques. Afterwards, another six publications were discarded due to a too broad and general overview. At the end, 47 publications remained and were further examined and analyzed for the identification of areas with high radon levels.

3.1.2 Radon prone buildings

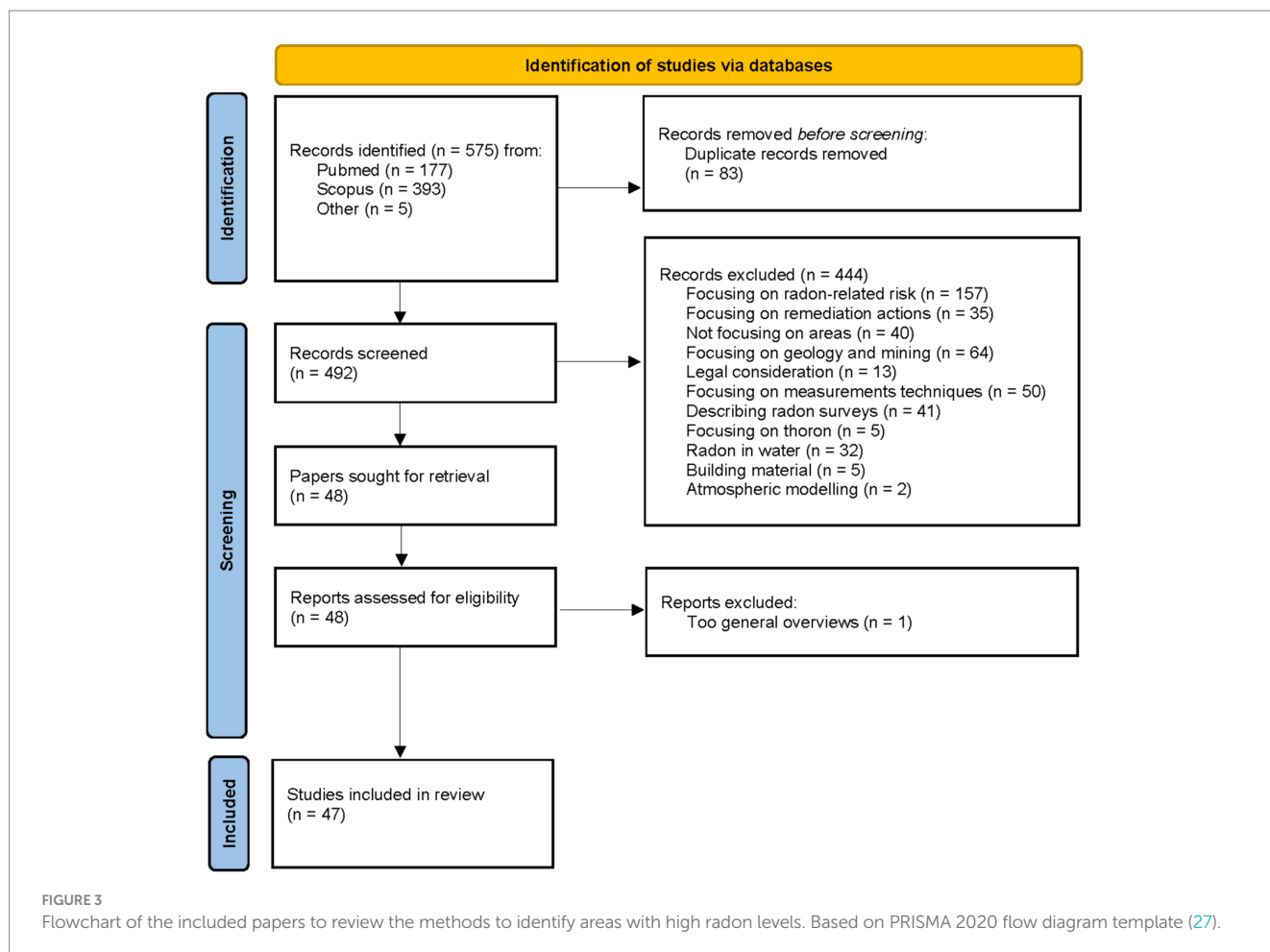
A total of 252 papers were identified after removing the duplicates from the 403 records found. All identified papers were in English. The process of study selection and exclusion is shown in Figure 4.

After the screening of the titles and the abstracts of these papers, 216 papers were discarded since they did not deal with the issue of interest for the present review (i.e., mostly papers describing methods able to find explanatory variables of the average indoor radon levels within buildings). One additional paper found among the references of the retrieved articles was also preselected. Overall, 37 full text articles were further examined. Of these, 11 papers were deemed of interest as they specifically address methods aimed at identifying buildings with high radon concentrations.

3.2 Focusing on predicting high radon levels

Different approaches to focus on the prediction of high radon levels exist. In our work we are mainly interested in methods that predict areas and buildings with high radon levels. Therefore, predictive models can be powerful tools and with a solid database and a profound training and test data split it is possible to create accurate models that generalize. In general, predictive models are not designed to predict certain values or value ranges like very small or very high values. But modifying the target value it is possible to focus on special value ranges, for example to focus on high radon values. Different techniques to increase the impact of high values exist.

One strategy to focus on high radon levels is to keep the actual radon concentration as target value without transformation or



modification. Indoor radon concentrations typically approximately follow a log-normal distribution (28). The lognormality of the target variable in a predictive model is an interesting feature, moreover when focusing on high radon levels, where the few but not often occurring high radon levels might be the most interesting ones. In a regression task when evaluating the model performance, the error of prediction and target value are calculated. The error of predictions is commonly a measure of distance, as the mean squared error or mean absolute error. Observations with higher deviations of prediction and target value contribute more to the error than observations with low deviations. Transforming the target value changes how observations with high indoor radon concentrations contribute to the prediction error. If the actual log-normal distributed indoor radon concentrations without a transformation or aggregation are used as target variable, higher indoor radon concentrations can contribute more to the prediction error than low concentrations, because they have a higher potential for deviation. By log-transforming the indoor radon concentrations this effect vanishes and the potential of contributing to the prediction error becomes balanced, and the focus of prediction is shifted from observations with high indoor radon concentration to average values. Therefore, predicting the log-transformed target value of a log-normal distribution increases model performance and because great deviations from the average values are less penalized.

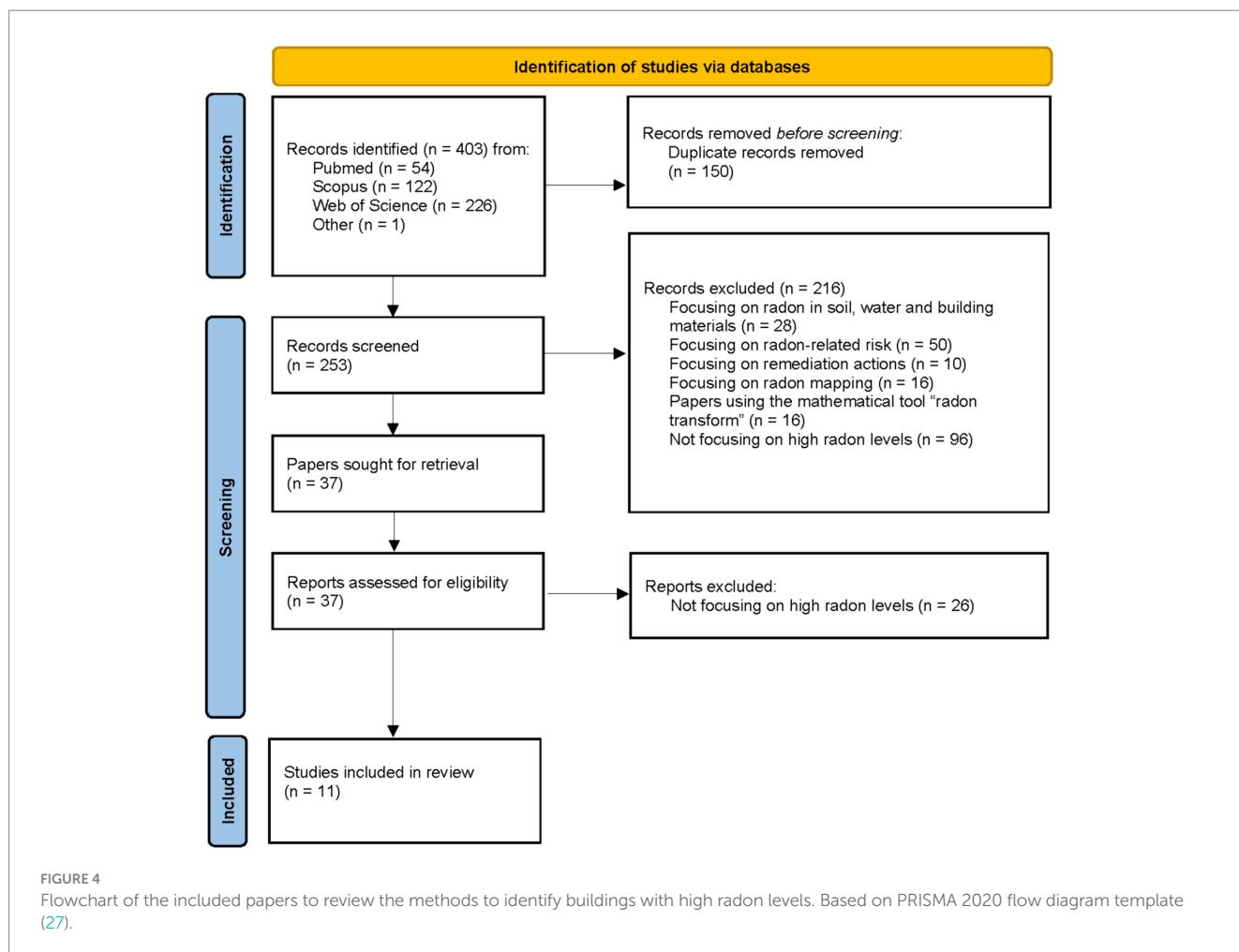
The aggregation of indoor radon concentrations by using the geometric mean or other central aggregations has a similar effect on

the predictions. By transforming or aggregating the indoor radon concentrations higher indoor radon concentrations lose impact, but model performance will increase. Using the untransformed log-normal distributed indoor radon concentrations is an example of increasing the impact of high values by penalizing them during training. Similar results can be achieved when weighing the samples according to their target value, where higher values get higher weights and have a greater impact on the performance evaluation.

Duplicating samples with high values can have a similar effect. Doing so no new information is added to the data set but during training the model more often sees these samples and therefore tries best to predict these high values. As indoor radon concentrations typically show a log-normal distribution, high values are underrepresented and by duplicating these high values the data set gets more balanced.

By transforming the regression into a classification task, the performance evaluation shifts from a distance metric yes/no decision for each class, where the exact value of the target value is not important anymore but the overall class. For example, indoor radon concentrations could be divided in two classes, lower and greater equal to 300 Bq/m^3 . This could lead to more robust predictions for each class and therefore also for high radon levels,

Another strategy is to focus on the prediction on high radon levels is to adopt the loss function, as will be discussed in the quantile regression section and machine learning chapter, where machine



learning methods can be transformed to a quantile regression task by using the tilted absolute value loss function for neural networks or by using quantile regression forests, which is an adoption of random forests.

3.3 Statistical methods to identify areas and buildings

Four broad classes of methods have been identified: (1) descriptive statistics methods; (2) regression methods; (3) geostatistical methods; (4) and machine learning methods. These classes of methods are presented below, with the aim of highlighting their respective strengths, limitations, and performance, if reported.

3.3.1 Descriptive statistics methods

Descriptive statistical methods aim to analyze, summarize, and explore data without the willingness to predict. These methods are mandatory to explore data, which includes the calculation of basic aggregates as means or ranges, or the visualization through different kind of plots, i.e., correlation analysis, ANOVA, or outlier detection. Descriptive statistical analysis is therefore the basis of any further investigation or predictive method, while it also produces its own findings. For example, these methods are used to produce maps describing a territory based on measurements of a certain variable.

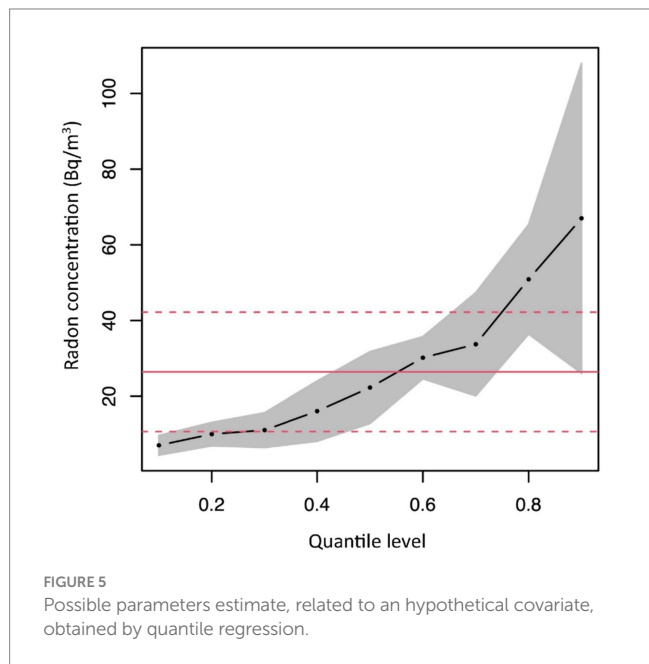
Two main descriptive statistical methods were applied to indoor radon dataset, presented in Table 3. These methods are correlation analysis, and statistics by geographic grouping.

Correlation analysis has been applied in different context to assess the statistical link (e.g., Pearson's correlation coefficient) between indoor radon and geological data (29–32), uranium content data (30), gamma ray survey (30, 33) or radon-222 exhalation rate measurements (34). Indeed, indoor radon measurements are generally used to confirm a mapping of a high radon area when the map was previously created with another dataset than indoor radon concentration (e.g., geology, field measurement) (30, 34, 35).

A common approach to delineate radon prone areas is based on the calculation of descriptive statistics (from a dataset of radon concentration measurements) typically for each of the administrative units in a territory (36–39). Miles and Appleton (40) applied some statistics by geographic grouping by combining grid square and geological mapping methods. The resulting maps were more accurate: within each geological combination with more than 100 radon measurements, the variation of radon potential was mapped using a 1 km² of the national grid. Radon potential was attributed to each grid square based, at least, on the nearest 30 house radon measurement results to that square. Grouping indoor radon measurements with geology on 1 km² of the national grid squares allowed to identify new radon affected areas that remained unidentified so far, especially in areas where radon measurements were rare. However, high indoor

TABLE 3 Selected descriptive statistics methods and their respective descriptions.

Method	Description
Correlation analysis	Correlation is a statistical measure which provides information about the relationship between two variables. The strength of the linear correlation is determined by the correlation coefficient ranging from -1 to +1. Correlation analysis does not imply causation but rather can be used to determine the strength and direction of the correlation between two variables.
Statistics by geographic grouping	When data are georeferenced and numerous, summary statistics (e.g., arithmetic mean, standard deviation, ...) of the variable of interest can be produced by geographic grouping such as grids-square or geological/administrative units.



radon levels measured within a single square with low measurement density may overestimate the potential.

3.3.2 Regression methods

Regression is a common statistical method for studying relationships between variables. In general, a response variable Y is a function of a set of n observed predictor variables X_1, \dots, X_n , also defined as covariates, such that $Y = f(X_1, \dots, X_n)$. With regression methods the only information that is obtained about the relationship between y and the covariates X is how the *mean* of the response variable Y varies as each X_i varies; in other words, the function is defined for the expected value of Y conditional on the covariates:

$$E(Y|X_1, \dots, X_n) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \quad (1)$$

With the coefficients β_i describing the impact of each covariate on the average level of the response variable. Focusing exclusively on variations in the average may under or overestimate or fail to detect the real impact some variables may have on a response variable. If, therefore, we apply a classical regression model that predicts effects of several factors (predictors) on the mean value of radon concentration, the results obtained may not provide a complete picture of the relationships between those variables, especially if we are interested in the impact that certain factors have on high radon concentration values (right tail of the radon concentration distribution).

3.3.2.1 Quantile regression

It is possible to fit regression curves to other parts of the distribution of the response variable: the quantile regression, as introduced by Koenker and Bassett (41), is a method for estimating functional relations between variables for all portions of a probability distribution, thus implying the possibility that there is not a single rate of change describing the relationship between a response variable and predictor variables. Quantile regression extend the classical regression methods aiming to estimate the relationship between quantiles of the conditional distribution of the response variable Y and a set of observed covariates X_1, \dots, X_n (Equation 2):

$$Q_Y(\tau | X_1, \dots, X_n) = \beta_{\tau 0} + \beta_{\tau 1} X_1 + \dots + \beta_{\tau n} X_n \quad (2)$$

With τ representing the quantile level of interest, reminding that a quantile of level τ (with $0 < \tau < 1$) represents the value of the Y distribution such that τ % of the data falls below it. This approach offers a regression model for each of the quantiles of interest separately. In this way, it is possible to study the influence of explanatory variables on the shape of entire Y distribution. In fact, the parameters $\beta_{\tau i}$ represent the impact of the covariate X_i on the specific quantile of level τ , allowing these parameters to be different for different quantile levels (e.g., Figure 5).

The feature of different slope coefficients at different points in the distribution is particularly useful if the underlying data exhibits heteroscedasticity, that is, the quantile regression model is particularly well suited to detect and describe the heteroscedasticity and how it may act as an “effect modifier” of a predictor at different levels of the response variable (42).

The quantile regression can be applied without assuming any parametric distribution and without specifying the variance and covariance structure of the error for the response variable. The parameter estimation approach, based on the minimization of a check loss function, makes the estimates more robust than those obtained through the classical regression described by the conditional mean model (Equation 1). Quantile regression model allows to evaluate the impact of several factors simultaneously on the quantile $Q_Y(\tau)$, however generally this method is not used to handle a very large number of covariates.

Starting from this original idea as presented in Koenker and Bassett (41), the quantile regression approach was further developed to make it suitable for a wide variety of data analysis settings. In particular, one extension of the quantile regression approach stems from the need to handle spatial data, such as radon concentration measurements, the difficult of which is due to the possibly highly complex spatial dependence among the various measurement sites.

Moreover, the quantile regression approach was also extended to the Bayesian framework (43–45). A comprehensive review of the quantile regression class of methods, their applications, and relevant literature, is reported in the “Handbook of quantile regression” (46).

3.3.2.1.1 Applications

Quantile regression can be particularly useful in the radon context, because it allows to focus on the right tail (i.e., the high percentiles) of the radon distribution, which is associated with a higher lung cancer risk.

Although a vast literature is available for quantile regression methods and many of its applications, few papers were found focusing on radon concentration (47–50), and all used methods that develop from the original quantile regression (and also incorporating spatial dependence) to evaluate the impact that predictors potentially have on high radon concentration values. In these papers the analyses are applied to indoor radon concentration datasets, in dwellings or schools. Factors included in the model are related to building characteristics (building destination of use, contact with ground, building materials, building age, dwelling floor), and in some cases also radiometric data, geological data and altitude are included.

Using a quantile regression approach, these papers generally showed that the effect of explanatory variables may change quite significantly depending on the level of indoor radon concentration (e.g., single building, direct contact with ground). In few cases no clear pattern of influence across quantiles is observed as for instances building age (50) and geological factors (49).

3.3.2.2 Logistic regression

Logistic regression is a very simple and powerful tool to develop a prediction model for a binary outcome (the dependent variable). In the present context, the binary outcome is the indoor radon concentration (IRC) above (event “1”) or below (event “0”) a certain threshold. Logistic regression is a statistical model predicting the probability p of an event taking place (e.g., IRC above a threshold) depending on the linear combination of one or more independent variables. Usually, when one wants to understand the relationship between one or more predictor (or explanatory) variables and a continuous response variable, it is possible to use a linear regression model. However, in our case, the response variable is categorical (high = 1/low = 0 IRC), and a classical linear regression as described in [1] is not suitable to predict the probability of the event being 1, as it is not limited in a predefined interval, while probability is (by definition, from 0 to 1). Therefore, the relationship between some predictor variables and the probability p to find a radon concentration higher than a fixed threshold T may be described through a function whose values span in the $[0,1]$ interval (Equation 3), such as the S-shape logistic curve (Figure 6):

$$p = P(Y > T | X_1, \dots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}} \quad (3)$$

Applying a proper function (i.e., logit function) to the response variable (the probability p), a model linearization is obtained, and it is possible to easily estimate the values of the coefficients, although their interpretation is not trivial.

Logistic regression is straightforward to use to discriminate between the many factors influencing the probability to measure high

indoor radon levels, but the method requires a threshold to be set to identify what are considered high values. The threshold should be a reasonable value that considers both the data available (“high” radon concentration measurements might be a small number in a dataset) and the particular interest to use a specific threshold (because it has a regulatory or recommendatory value) (Table 4).

3.3.2.2.1 Applications

In the selected papers, the logistic model is applied using different threshold values. In Dai et al. (51), the threshold considered is 148 Bq/m³ (action level for remediation in the United States), while in Vukotic et al. (52) the threshold value was fixed to 200 Bq/m³, in line with the UE guidelines for indoor radon levels. Often different analyses are implemented using several thresholds, as done in Borgoni et al. (53), who used four different threshold values (100, 200, 300, and 400 Bq/m³), and in Stanley et al. (54), who used the two thresholds 100 and 500 Bq/m³.

All the previously discussed papers applied logistic regression with the main objective of identifying the major features influencing IRC. However logistic regression was also used to identify radon-prone areas, feeding the model not only with covariates describing building characteristics, but also with several geogenic and/or geophysical factors, including geological data and airborne geophysical parameters (29, 55, 56).

Therefore, even if logistic regression might not be entirely classified within the geostatistical methods (see next section), such an approach can also be useful to enhance predictive power of geogenic radon maps with a high level of accuracy [76.5% in (55), and 77.9% in (29)].

3.3.3 Geostatistical methods

In the context of identifying buildings and areas with high radon levels, geostatistical methods have been largely explored so far. Indeed, geostatistical tools allow us to predict radon related variables based on the distribution of known observations, in areas where it has not been measured. This is possible because closer observations show higher autocorrelation than the most distant observations, which is known as the first law of geography (57): “Everything is related to everything else, but closer things are more related than distant things.” Naturally, geostatistical methods depend on georeferenced data, which is now generally used in radon surveys. They can be applied with only the

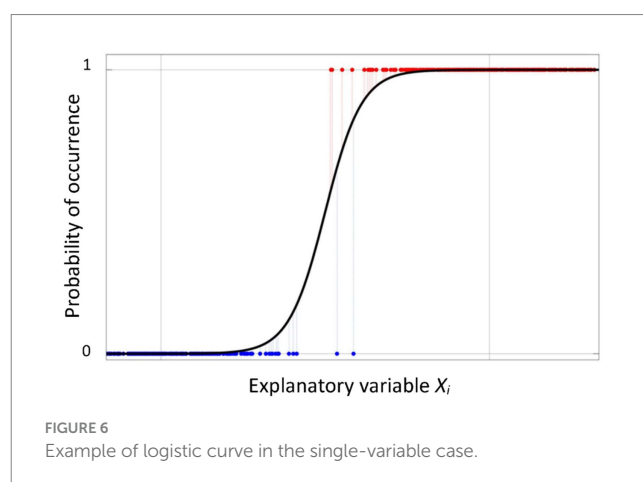


TABLE 4 Selected regression methods and their respective descriptions.

Method	Description
Quantile regression	Quantile regression is a statistical method for estimating functional relations between variables for all portions of a probability distribution, thus implying the possibility that there is not a single rate of change describing the relationship between a response variable and predictor variables.
Logistic regression	Logistic regression is a statistical model for a binary outcome (the dependent variable). Logistic model predicts the probability of an event taking place (in our context, indoor radon concentration above or below a certain threshold) depending on the linear combination of one or more independent variables.

TABLE 5 Selected geostatistical methods and their respective descriptions.

Method	Description
Inverse Distance Weight (IDW)	Interpolation and deterministic method which relies on existing data measurements. Radius (max. Distance of influencing measurements) and power (influence pondered to distance) determine the result of interpolation.
Kriging	Interpolation and deterministic method based on existing data measurements and their statistic relationship. According to different parameters, the resulting interpolation is associated with an uncertainty map issued from the statistical modeling.
Geographically weighted regression	Multiparameter predictive tool which uses different explanatory variables, including non-stationary variables (climate, geographical coordinates), to predict an indicator. Each explanatory variable is given a coefficient, such as in a linear regression.
Local polynomial interpolation	Two-dimensional interpolation based on a polynomial function applied as a filter on a part of the total surface. Polynomial function will vary according to its location.
Global polynomial interpolation	Two-dimensional interpolation based on a polynomial function applied to the entire surface investigated.

data of interest or use additional georeferenced variables. Methods using many different input parameters might overlap with regression methods and machine learning methods, where for instance

coordinates are only one predictor in bunch of input parameters. In this chapter, application of classical geostatistical methods such as inverse distance weight, kriging and geographically weighted regression are presented. Table 5 synthesizes some of the most used geostatistical methods in the radon context.

3.3.3.1 Applications

Inverse distance weight (IDW) has already and, above all, largely been applied through different studies in different contexts. Overall, authors decided to use IDW because the method has the advantage of complying well with noisy data and it might interpolate short-range variation (58–61). Radon measurements carried out in buildings might present a high spatial variability, which results in very few spatial correlations between the measurements. IDW is therefore the ideal tool to answer this kind of dataset (60). Moreover, this method presents another advantage in its ease of implementation, which makes it cost-and time-efficient. However, authors noticed some counter-performances of IDW. Since IDW allows to identify small scale variations, this method tends to miss the global trends in the overall dataset (58). Finally, the results provided by IDW are considered as valuable by the different authors (62) although it is necessary to implement IDW carefully knowing all its strengths and weaknesses.

Among all the geostatistical methods introduced in this chapter, kriging and its variations are the most implemented method to assess high radon areas (58, 59, 61, 63–69). Most of authors note the difficulty to include all related elements and carry out the different steps of the methodology, although some software allows to simplify its application, such as ArcGIS (65, 66). Nevertheless, this method presented really good results: Sabbarese et al. (68) concluded that kriging allowed them to identify radon prone areas even the ones where very few data was recorded. Similarly, a geogenic radon potential map, highlighting radon-prone areas was created using kriging and co-kriging (67). Moreover, it has been highlighted that kriging is the best linear unbiased spatial predictor (58, 61). It has been noticed that an application of kriging algorithm allows to avoid smoothing effects (63). Beyond the difficulty to implement this method, it has been shown that the calculated weights not only depend on the distance but also on the direction and orientation of the closest data in scarce data area. Cafaro et al. (65) underlined a strong need to deeply understand radon's underlying layers, such as geology, and to have a homogeneous spatial repartition of radon data among the studied spatial extent. Indeed, unreliable measurements might be produced by non-stationary effects and the lack of correlation between geology and indoor radon levels, especially in karstic areas (65). Akkala et al. (58) also underlined non-stationarity issues in real-world datasets. Kriging may be declined in different models, such as simple, universal, conditional or disjunctive kriging (70). The latter model revealed to be very effective to identify radon prone areas using only indoor radon measurements (69). Finally, Bachirou et al. (59) compared the performances obtained with ordinary kriging and IDW. No significant differences were observed in the prediction errors of the two techniques adopted, applied in a similar context and with an identical dataset (59).

Then, geographically weighted regression (GWR) has been used several times to identify radon-prone areas. This method has been applied several times in Italy (71–73). GWR presented the advantage of localizing small-scale variations, such as IDW, but also produces a

map of the local coefficient (73). Ciotoli et al. (71) underlined the accurateness of the model, which is commonly used, and the fact that only influential explanatory variables are needed to run the model. Finally, De Novellis et al. (72) used GWR because it allowed to estimate the indoor radon concentration by using some local environmental properties, e.g., the geogenic radon potential of the underlying soil. Although GWR presents good results, revealed by the authors, this method is very sensitive to the representativeness of the sampled data (73).

Other geostatistical methods were investigated and compared in two articles (58, 61). Local and global polynomial interpolations were applied, and their results are compared to other methods, such as kriging, IDW and machine learning methods. LPI presented the advantages of being easy to implement, to comply well with noisy data, and to interpolate local variations, while the estimation errors increased exponentially with increasing complexity in the training dataset. GPI requires less data to interpolate and was computationally less intensive than LPI. However, GPI was not suitable for extrapolation. Compared to IDW and kriging, these two methods presented lower performance scores (normalized mean square error) than kriging ones, while the latter seemed to be similar as IDW accuracy (58, 61).

Finally, some authors used GIS (Geographic Information System) software to implement different and less common methods, such as Rasters Factor Rating Method (74), PCA and correlation analysis (75), Bayesian estimation of the percentage of houses exceeding a given action level (76).

3.3.4 Machine learning methods

Machine learning is a collective term for applications using statistical models to analyze, make predictions and draw inferences from data, without following full instructions. Machine learning methods can be divided into three main categories: supervised learning, unsupervised learning and reinforcement learning. In the field of radon usually supervised and, in a few cases, unsupervised machine learning methods are used. Supervised learning uses labelled input data to generate models to predict unseen or future data. Machine learning methods have been shown to be very successful in many applications and are already used in predicting radon levels, as summarized and analyzed also in recent studies (77, 78).

Supervised learning with a numerical target value is a regression task which is already described in the quantile regression section (section 3.3.2.1). Unsupervised learning analyses data and tries to find underlying patterns and structure in the data. As examples Table 6 presents different selected methods applied in the framework of radon prediction.

Supervised learning predictive models are built using labelled training data. Model performance is evaluated on test data using metrics (e.g., loss functions) to calculate differences between the prediction and actual value (e.g., accuracy, confusion matrix, mean squared error). The prediction can be a categorical or numerical value and, based on the target value and the aim of the prediction, a suitable metric is used. A great variety of different supervised learning methods for various applications exist.

In the field of radon, various machine learning methods have been used already, especially for the prediction of indoor radon concentrations. The reason for using data-driven machine learning models is that a valid physical transport model from the soil to actual

indoor radon concentrations has not been found until now. The process depends on different factors, that also might be independent from each other such as the geogenic radon potential, the type and the technical standard of the building and the usage of the building. Also, in recent years many indoor radon concentration surveys have been carried out, which is a solid data basis for machine learning techniques.

3.3.4.1 Applications

Overall, 18 publications were selected and in the following their ability to predict high radon levels is summarized.

The selected papers showed a great variety of machine learning models. Some of the publications test different models or use several ensembled models. The most used supervised machine learning methods are artificial neural networks (58, 61, 78–81), followed by random forest (79, 82–84). Other models are only used in one publication: support vector machine (84), convolutional neural network (85), LSTM (86), k-nearest neighbor (87), mixed effect regression model (88), extreme learning machine (86), random vector functional (86), multivariate adaptive regression splines (84), boosted generalized additive and linear model (80), XGBoost (78, 79), automatic linear modeling (79), and group method of data handling (89). In four publications, unsupervised machine learning methods or semi-supervised machine learning methods, Bayesian cluster detection (50), k-medoids clustering (83), k-mean clustering (90) and Bayesian profile regression were used. None of these methods are specifically designed to focus the prediction on high radon levels. In the following the overall workflow of the selected publication is analyzed, focusing on modifications of the target variable and the used train/test split strategy.

Indoor radon concentrations, their aggregates or transformations, were the target value in 15 publications. In two publications, the geogenic radon potential was used as target variable. In three publications, the actual IRC value was used as target value which keeps the focus on high radon levels, because single high values have a greater

TABLE 6 Selected machine learning methods and their respective descriptions.

Method	Description
Random forest	Assemble of decision trees, applicable for regression and classification task. Can handle categorical and/or numerical predictors without further modifications.
Support vector machines	Linear boundaries as lines, planes or hyperplanes are used either to separate binary classes or used as prediction in a regression task. It can be extended to non-linear models using the kernel trick, where predictors are transformed to higher dimension, where non-linear relations become linear separable.
Feed forward neural network	An input is passed through connected neurons to an output. The error at the output layer, where a loss function measures the performance of the model predictions, is used to backpropagated to adjust the weights and biases of the neurons.

potential for prediction errors, as also discussed in the result section. In one publication the indoor radon concentrations were transformed in a classification task with three classes (low, high medium). In two other publications the probability exceeding a reference value is used as target value. Doing so, the former regression task is also transformed to a classification task. Both are valid strategies to predict high radon levels. In the remaining publications, log-transformed indoor radon concentrations or aggregates like the geometric mean are used for prediction. Although these models are also capable of predicting areas with high radon levels in general, they are not specially designed to predict high radon levels, but rather focus on predicting average or mean values. Still a well-designed machine learning workflow can be a very powerful tool for predicting radon levels and high radon levels. Overall, the selected papers usually do not include detailed sections of the modification of the target value (e.g., the log-transformation, aggregation) or the reason why a certain modification was used, with some exceptions discussed in more detail in the following.

Another sensitive topic when applying predictive models is the model evaluation on training and test data. In 14 publications the data splitting in training and test data is documented. In 12 of these, the data split was performed randomly without further consideration. In two publications the data split was performed on spatial criteria to avoid spatial autocorrelation of the target variable. When comparing the model performance, it is essential to consider if it was evaluated on the training, validation, or test data or if these account for the spatial autocorrelation of the target value. Models without a clear data split workflow might show high model performance, but will fail to generalize and might perform poorly on new or unseen data. Keeping these general remarks in mind, we will now give short descriptions of selected publications with different workflows, that can serve as profound examples and baseline of a machine learning workflow.

In Petermann et al. (84) three different machine learning methods (random forest – RF, support vector machines – SVM, multivariate adaptive regression splines – MARS) are tested to predict the geogenic radon potential in Germany with 36 possible predictors. The autocorrelation of the target value is described and analyzed, and a solution outlined how to deal with this autocorrelation by splitting training and test data in geographical blocks. Leave-one-out cross validation was used to explore which predictors are useful and to tune the hyperparameters of the three models. Feature importance for the predictors was calculated and an in-depth analysis of their impact on the results was carried out. Multiple metrics for performance evaluation are given and the distribution of the prediction is discussed, stating that the predictions tend to overestimate small values and underestimate high values. Performance metrics for MARS, RF and SVM were, respectively, 55.2, 52.8 and 54.4 for RMSE, 0.16, 0.22, 0.22 for R^2 , and 25.6, 23.7 and 22.2 for MAE, thus denoting a slightly higher predicting power of Random Forest (RF). The publication is the only one of the selected publications that does not predict indoor radon concentrations but covers so many aspects of a profound machine learning workflow that it can serve as a very good starting point for predicting a spatial autocorrelated target value. In the publication of Rezaie et al. (86), three different machine learning methods (long short-term memory – LSTM, extreme learning machine – ELM, random vector functional link – RVFL) were tested to predict the non-transformed indoor radon concentration. The training and test split were done randomly which does not take autocorrelation into account. The models are complex models in the field of neural

networks. The introduction and the discussion of the results are in-depth analysis also using different performance metrics. LSTM, ELM, and RVFL models performance were similar, as depicted by their respective AUC-ROC score of 0.81, 0.83, and 0.82.

In Wu et al. (78) two machine learning methods (Neural-network and XGBoost) are used to predict radon levels. The usually applied regression task is transformed to a classification task with three radon classes based on the Swedish radon legislative: 0 to 200 Bq/m³, 200 to 400 Bq/m³, and greater 400 Bq/m³. Among the selected publications this is a unique approach and relevant for also predicting high radon levels, As also stated in the result section machine learning models predicting classes, might be more robust compared to regression workflows, and therefore might also predict high values more accurately. The train and test split and the performance metric is clearly described and documented, XGboost achieved better results than the neural network: macro-F1 score were, respectively, ranging between 0.93–0.96, and 0.64–0.74 for XGboost and neural network. Interestingly the accuracy of the developed model was highest for the low and high classes. The middle class showed the highest errors among the three classes. The study shows that the transformation from a regression to a classification task could lead to more robust predictions for high radon values. The downside is that by using classes as a target variable, information about the actual numerical radon level gets lost. A comparison with a regression model using the same data and workflow would be interesting but was not applied.

In two selected publications unsupervised machine learning techniques are also applied. In Kropat et al. (83) lithological units based on their IRC distribution are clustered in classes. Six classes are later used for the prediction of IRC among other predictors to predict indoor radon concentrations with regression trees. When using such a lithological classification based on the target value, information flows from the target value to the predictor space. This could lead to models that might not generalize well and overestimate the actual importance of the lithological classes as predictors. But still, the clustering of lithological units alone can be a very interesting method, when searching for high-concentration areas or by producing an actual radon map. In Sarra et al. (50) supervised and unsupervised learning techniques were used together. In this publication, a quantile regression model was built on building characteristics to predict indoor radon concentrations in standardized homes. These standardized indoor radon concentrations best reflect the geogenic radon potential. In a second step the distributions of the standardized radon concentrations are used to cluster lithological units into areas of radon hazard.

3.4 Assessment of performance metrics

With regards to the various applications identified within the literature review conducted using the PRISMA method, we provided a synthesis table of performance metrics used to assess performances of the different methods. Table 7 lists and highlights the connections between the various methods, if any.

4 Discussion

As demonstrated by the results, various methods exist for identifying radon-prone areas and buildings, ranging in complexity

TABLE 7 Synthesis of performance metrics with their relative description and relationships.

Category	Performance metric	Description	Relationships with other metrics
Classification	ROC-AUC	Measures the ratio between the true positive rate and false positive rate, represented by the area under the ROC curve.	AUC provides a single metric to compare different models' ability to identify different classes.
	Confusion Matrix	Table listing true positives, true negatives, false positives, and false negatives, which help at describing classification performance.	Basis for calculating precision, recall, F1-score, and accuracy.
Regression	Mean Absolute Error (MAE)	Average of the absolute differences between predicted and observed values.	MAE is less sensitive to outliers compared to MSE and RMSE. Often compared with these metrics to assess error distribution.
	Mean Squared Error (MSE)	Average of the squared differences between predicted and observed values, more sensitive to outliers.	Squaring amplifies the impact of larger errors (outliers), leading to more conservative models compared to MAE.
	Root Mean Squared Error (RMSE)	Square root of MSE, having the same units as the original data.	Directly related to MSE. RMSE is easy to interpret due to the same unit as the original dataset.
	R-squared (R^2)	Part of variance in the dependent variable that is predictable from the independent variables. R^2 ranges from 0 to 1 and can be expressed as a percentage	Often compared with Adjusted R-squared to assess the impact of adding more predictors.
	Adjusted R-squared (Adj. R^2)	R-squared adjusted to the number of predictors included in the model. Score decreases with the addition of non-significant predictors.	Adjusts R-squared to prevent overfitting by adding too many predictors.
Model selection	Akaike information criterion (AIC)	Measure of the relative quality of statistical models, by balancing fit and complexity of the model.	Often compared with BIC. Both metrics penalize model complexity, but AIC is less strict.
	Bayesian Information Criterion (BIC)	Similar to AIC but with a stronger penalty for models with more parameters.	Stricter than AIC, and often preferred when overfitting is a concern.

and implementation. These methods encompass a spectrum of approaches, from basic statistical methods to AI methods. Based on our findings, the present study led to the production of [Table 8](#), delineating the strengths and weaknesses inherent in each method's category. This evaluation sheds light on the efficacy and limitations of each approach, aiding in the discernment of optimal methodologies for high radon levels prediction. Moreover, specific circumstances under which each method proves to be the most suitable are highlighted, thus facilitating an informed decision-making regarding the selection and deployment of radon evaluation strategies in diverse environmental and geographical contexts. [Table 8](#) provides guidance for stakeholders involved in radon risk assessment and management efforts.

Overall, the methods investigated in our research consistently yielded results aligning with researchers' expectations. Basic statistical methods demonstrated robust performance alongside ease of application, guaranteeing a comprehensible analysis. Quantile and logistic regression methods emerged as effective tools for forecasting elevated indoor radon levels, offering both simplicity in implementation and clarity in interpretation. Meanwhile, geostatistical methods, with their spatial component, emerged as efficient in predicting indoor radon concentrations in unmeasured areas, thereby facilitating the identification of radon-prone regions. However, researchers must take care in methodological application and parameterization to ensure accurate and reliable results. Finally, machine learning methods examined in the analyzed papers are

generally not tailored specifically for predicting high radon values. Still, the methods applied can be used to specially target high radon levels and even more with the modifications discussed in section 3.2., as adoption of the loss function or keeping the target value on linear scale. An impact assessment of the inclusion of different modifications within ML methods to specifically predict high radon levels must be more explored.

These findings underscore the significance of methodological choice and careful consideration in the pursuit of effective high radon risk assessment. More generally, the analyzed publications used diverse data sources, manipulations, models, and performance metrics, making it challenging to compare them and draw definitive conclusions about the most effective workflows for predicting high radon areas and buildings. Consequently, we propose further research to determine the best-practice among the methods investigated tailored for predicting elevated radon levels and to determine the optimal strategy aimed at identifying as many dwellings as possible with high radon concentrations.

While our literature review synthesizes key insights regarding the identification of radon-prone areas and buildings, our study still presents some limitations. The application of methods across varied datasets, transformed and homogenized with various prior manipulations, models, and contextual settings, introduces a significant challenge to direct comparison. Moreover, the absence of standardized metrics for method performance evaluation complicates the assessment process, hindering the ability to benchmark the

TABLE 8 Strengths and weaknesses of the different classes of method identified.

Category	Strengths	Weaknesses	Recommendations
Descriptive statistics	Provide good results while being easy to implement.	Some analyses are subject to interpretation.	Strong knowledge of statistics and the data is required. Useful to guide on more specific and appropriate analyses.
Regression	Fairly well-known and widespread method; produces results that are fairly easy to interpret and that are fairly familiar to many people	It works better when the number of covariates is not very large. Certain prerequisites/assumptions must be verified for applicability. To perform a logistic regression, a specific reference level should be fixed.	Strong knowledge of the dataset and of the relationships between variables is required. Certain. Assumptions on which the model is based should be checked.
Geostatistical tools	Powerful and adapted methods to predict radon levels where it has not been measured.	It is sometimes time-consuming and difficult to apply. Many parameters must be included.	Spatial behavior of the predicted variables and predictors must be known.
Machine learning tools	Various methods and strategies to deal with different research aims and scopes, the most powerful predictive models available.	The implementation and learning can be difficult and computationally expensive. Usually, black boxes and additional analysis for interpretation is needed.	Clear workflow from data handling to performance evaluation, as modifications of the target variable or the train/test split strategy.

Recommendations of use are provided by the authors, based on the literature review.

different approaches. Additionally, the interpretation of strengths and weaknesses within literature often remains subjective, influenced by authors' individual perspectives and experience. These limitations underscore the need for further refinement and standardization in methodologies to enhance the robustness and comparability of research findings in this critical field of study.

5 Conclusion

This paper aims to provide researchers with a systematic literature review of the various methods employed to identify radon-prone areas and buildings. Through our review, based on the PRISMA methods, we identified different methodologies that can be categorized into four main classes: descriptive statistics, regression methods, geostatistics, and ML methods. While these categories were established by the authors, they are not always distinct, as certain methods may fall into multiple classes. However, this classification has enabled to highlight diverse application contexts alongside their corresponding results. By examining the available techniques within each category, this paper offers insights into the effectiveness and applicability of different approaches for addressing the challenge of the identification of radon prone areas and building, i.e., areas and buildings with high radon levels.

The investigated methods primarily focus on identifying the predictors of elevated concentrations or delineating the characteristics of clusters within buildings and areas with heightened levels of radon. The methods tackled in this study serve dual purposes: identifying radon prone buildings and radon prone areas. Nevertheless, machine learning and geostatistical approaches are predominantly deployed to identify areas with elevated radon levels, often due to the availability of georeferenced data. Conversely, quantile and logistic regression methods are more frequently utilized to identify predictors of elevated radon concentrations within buildings. However, in some recent studies, these regression methods' original iterations have been

expanded to accommodate the spatial correlation inherent in radon concentration measurements.

After evaluating the diverse papers and their application of various methods, several questions emerged for consideration: at firstly, there is a query regarding whether it is advantageous to employ specific methodologies that focus on high radon concentrations, as opposed to more conventional methods targeting average (geometric or arithmetic) levels. Notably, one method discussed in this review, quantile regression, demonstrated the potential for certain explanatory variables to exhibit varying impacts depending on the indoor radon concentration targeted. Secondly, a concern arises regarding the extent to which the outcomes of the applied methods rely on the characteristics of the dataset under analysis. Factors such as the age of the datasets or the geographic area from which the data are gathered may significantly influence the results, particularly concerning the identification of factors impacting high indoor radon levels.

For instance, when applying logistic regression to study different datasets, containing very different sets of predictors, the results as expected are variegated while generally in agreement; however, in some cases the use of different dataset may lead to apparent contradictions. For instance, in Stanley et al. (54) old buildings were found to have lower probabilities of being associated with high IRC while in Vukotic et al. (52) the findings were the opposite. Clearly the concept of "old" has two very different meanings in Europe and in North America and the types of buildings considered were not easily comparable.

Our analysis revealed the challenge of comparing the different methods due to the multitude of datasets, data manipulations, models, and performance metrics involved. Indeed, it is quite difficult to compare different analyses on different datasets, applying different investigation strategies and different statistical approaches, since (1) the performance of a method strongly depends on the dataset under examination and even on the outcome variable (e.g., radon concentration on a continuous scale, on a binary scale, on a log-scale), (2) the performance metrics/indicators used are different depending

on the model applied. Results exhibit a high degree of dependency on local factors and the characteristics of the dataset being analyzed. As a result, it becomes evident that there is no universally optimal method applicable across all scenarios. Rather, the selection of the optimal method is contingent upon the specific attributes of the available dataset and the practical feasibility of applying a particular methodology. This underscores the importance of carefully considering the nuances of each situation when choosing an appropriate approach for identifying and addressing effectively high indoor radon concentrations.

This paper represents the initial step to review and evaluate the diversity of methods used in identifying radon-prone areas and buildings. Serving as a foundation for future work, this paper lays the groundwork for applying methodologies, as outlined herein, to available datasets. The challenge is to come to a deeper comprehension of how different methods perform when applied to the same dataset, thereby enhancing our ability to effectively address a robust comparison. Through this paper, we hope to offer valuable insights as a basis to guide future efforts in the identification of radon prone areas and buildings, and ultimately, reduce the population exposure.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JR: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization. SA: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization. SB: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation, Methodology. CD:

References

1. Porstendörfer J. Properties and behaviour of radon and Thoron and their decay products in the air. *J Aerosol Sci.* (1994) 25:219–63. doi: 10.1016/0021-8502(94)90077-9
2. World Health Organization. WHO handbook on indoor radon: A public health perspective. Geneva, Switzerland: World Health Organization (2009).
3. Bruno RC. Sources of indoor radon in houses: a review. *J Air Pollut Control Assoc.* (1983) 33:105–9. doi: 10.1080/00022470.1983.10465550
4. World Health Organization. (2023). Radon. Available at: <https://www.who.int/news-room/fact-sheets/detail/radon-and-health> (Accessed July 12, 2023).
5. Abbasi A, Tulucu F, Issa SAM, Algethami M, Mirekhtyari F, Zakaly HMH. Modelling and calculation of risk due to radon concentration in residential houses, Nicosia, Cyprus. *J Radioanal Nucl Chem.* (2022) 331:1335–41. doi: 10.1007/s10967-022-08203-6
6. Forouzanfar MH, Lily Alexander H, Anderson R, Bachman VF, Biryukov S, Brauer M, et al. Global, regional, and National Comparative Risk Assessment of 79 Behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet.* (2015) 386:2287–323. doi: 10.1016/S0140-6736(15)00128-2
7. Rey JE, Goyette S, Gandolla M, Palacios M, Barazza F, Pernot JG. Long-term impacts of weather conditions on indoor radon concentration measurements in Switzerland. *Atmos.* (2022) 13:92. doi: 10.3390/atmos13010092
8. Ielsch G, Cuney M, Buscail F, Rossi F, Leon A, Cushing ME. Estimation and mapping of uranium and radium content of geological units in France. *J Environ Radioact.* (2017) 166:210–9. doi: 10.1016/j.jenvrad.2016.05.022
9. Nunes LJR, Curado A, Lopes SI. The relationship between radon and geology: sources, transport and indoor accumulation. *Appl Sci.* (2023) 13:7460. doi: 10.3390/app13137460
10. Singh S, Kumar A, Singh B. Radon level in dwellings and its correlation with uranium and radium content in some areas of Himachal Pradesh, India. *Environ Int.* (2002) 28:97–101. doi: 10.1016/S0160-4120(02)00012-0
11. Groves-Kirkby Chris J, Crockett Robin GM, Denman Anthony R, Phillips Paul S. (2012). Climatic and seasonal influences on radon time series in an environment of low anthropogenic activity. Available at: <http://nectar.northampton.ac.uk/4721/>
12. Mose DG, Mushrush GW, Chrosniak CE, Morgan WF. Seasonal indoor radon variations related to precipitation. *Environ Mol Mutagen.* (1991) 17:223–30. doi: 10.1002/em.2850170402
13. Rey J, Goyette S, Palacios M, Barazza F, Gandolla M, Goyette J, et al. Influence of some specific meteorological events on indoor radon dynamic in Western Switzerland. *J Phys Conf Ser.* (2021) 2042:012138. doi: 10.1088/1742-6596/2042/1/012138
14. Rey JE, Goyette S, Pernot JG. Weather impacts on indoor radon short-term measurements in Switzerland. *Atmos.* (2023) 14:1163. doi: 10.3390/atmos14071163

Investigation, Resources, Writing – review & editing. NL: Writing – original draft, Writing – review & editing, Formal Analysis, Investigation. CG: Writing – review & editing. VG: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. JG: Supervision, Writing – review & editing. FB: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The entire paper and research were funded by Euratom research, RadoNorm, under grant agreement No 900009.

Acknowledgments

The authors would like to extend their sincere gratitude to the financial partner for their invaluable support in funding this research. We also acknowledge the technical and administrative assistance provided by our affiliated institutions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

15. King C-Y, Bi-Shia King WC, Zhang W. Spatial radon anomalies on active faults in California. *Appl Geochem*. (1996) 11:497–510. doi: 10.1016/0883-2927(96)00003-0
16. Moreno V, Bach J, Font L, Baixeras C, Zarroca M, Linares R, et al. Soil radon dynamics in the Amer fault zone: an example of very high seasonal variations. *J Environ Radioact*. (2016) 151:293–303. doi: 10.1016/j.jenvrad.2015.10.018
17. Zafar WA, Ahmed J, Barkat A, Nabi A, Mahmood R, Manzoor S, et al. Spatial mapping of radon: implication for fault delineation. *Geochem J*. (2018) 52:359–71. doi: 10.2343/geochemj.2.0526
18. Collignan B, Le Ponner E, Mandin C. Relationships between indoor radon concentrations, thermal retrofit and dwelling characteristics. *J Environ Radioact*. (2016) 165:124–30. doi: 10.1016/j.jenvrad.2016.09.013
19. Demoury C, Ielsch G, Hemon D, Laurent O, Laurier D, Clavel J, et al. A statistical evaluation of the influence of housing characteristics and Geogenic radon potential on indoor radon concentrations in France. *J Environ Radioact*. (2013) 126:216–25. doi: 10.1016/j.jenvrad.2013.08.006
20. Ivanova K, Stojanovska Z, Tsenova M, Kunovska B. Building-specific factors affecting indoor radon concentration variations in different regions in Bulgaria. *Air Qual Atmos Health*. (2017) 10:1151–61. doi: 10.1007/s11869-017-0501-0
21. Sextro RG. Understanding the origin of radon indoors—building a predictive capability. *Atmospheric Environ*. (1987) 21:431–8. doi: 10.1016/0004-6981(87)90023-0
22. Borgoni R, De Francesco D, De Bartolo D, Tzavidis N. Hierarchical modeling of indoor radon concentration: how much do geology and building factors matter? *J Environ Radioact*. (2014) 138:227–37. doi: 10.1016/j.jenvrad.2014.08.022
23. Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Major influencing factors of indoor radon concentrations in Switzerland. *J Environ Radioact*. (2014) 129:7–22. doi: 10.1016/j.jenvrad.2013.11.010
24. Collignan B, Powaga E. Impact of ventilation systems and energy Savings in a Building on the mechanisms governing the indoor radon activity concentration. *J Environ Radioact*. (2019) 196:268–73. doi: 10.1016/j.jenvrad.2017.11.023
25. Yang J, Busen H, Scherb H, Hürkamp K, Guo Q, Tschiersch J. Modeling of radon exhalation from soil influenced by environmental parameters. *Sci Total Environ*. (2019) 656:1304–11. doi: 10.1016/j.scitotenv.2018.11.464
26. Lv M, Liu S, Cao Q, Zhang T, Liu J. Influence of ventilation on indoor air quality. In: Y Zhang, PK Hopke and C Mandin, editors. Handbook of indoor air quality. Singapore: Springer Singapore (2022). 1–38. doi: 10.1007/978-981-10-5155-5_54-1
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg*. (2021) 88:105906. doi: 10.1016/j.ijsu.2021.105906
28. Bossew P. Radon: exploring the log-Normal mystery. *J Environ Radioact*. (2010) 101:826–34. doi: 10.1016/j.jenvrad.2010.05.005
29. Dardac M, Elio J, Aghdam MM, Banrion M, Crowley Q. Application of airborne geophysical survey data in a logistic regression model to improve the predictive power of Geogenic radon maps. A case study in Castleisland, county Kerry, Ireland. *Sci Total Environ*. (2023) 894:164965. doi: 10.1016/j.scitotenv.2023.164965
30. Drolet J-P, Martel R, Poulin P, Dessau J-C, Lavoie D, Parent M, et al. An approach to define potential radon emission level maps using indoor radon concentration measurements and Radiogeochimical data positive proportion relationships. *J Environ Radioact*. (2013) 124:57–67. doi: 10.1016/j.jenvrad.2013.04.006
31. Kemski J, Siehl A, Stegemann R, Valdivia-Manchego M. Mapping the Geogenic radon potential in Germany. *Sci Total Environ*. (2001) 272:217–30. doi: 10.1016/S0048-9697(01)00696-9
32. Watson RJ, Smethurst MA, Ganerod GV, Finne I, Rudjord AL. The use of mapped geology as a predictor of radon potential in Norway. *J Environ Radioact*. (2017) 166:341–54. doi: 10.1016/j.jenvrad.2016.05.031
33. García-Talavera M, García-Pérez A, Rey C, Ramos L. Mapping radon-prone areas using γ -radiation dose rate and geological information. *J Radiol Prot*. (2013) 33:605–20. doi: 10.1088/0952-4746/33/3/605
34. Ielsch G, Ferry C, Tymen G, Robé MC. Study of a predictive methodology for quantification and mapping of the Radon-222 exhalation rate. *J Environ Radioact*. (2002) 63:15–33. doi: 10.1016/S0265-931X(01)00133-3
35. Bulko M, Holý K, Brandýšová A, Müllerová M, Masarik J. Study of the possibility of using radon potential maps for identification of areas with high indoor radon concentration. *J Radioanal Nucl Chem*. (2021) 328:651–7. doi: 10.1007/s10967-021-07673-4
36. Ambrosino F, La Verde G, Sabbarese C, Roca V, D'Onofrio A, Pugliese M. The first indoor radon mapping in the Campania region, Italy. *Isot Environ Health Stud*. (2023) 59:192–201. doi: 10.1080/10256016.2023.2189249
37. Fernández A, Sainz C, Celaya S, Quindós L, Rábago D, Fuente I. A new methodology for defining radon priority areas in Spain. *Int J Environ Res Public Health*. (2021) 18:1352. doi: 10.3390/ijerph18031352
38. Liu H, Wang N, Chu X, Li T, Zheng L, Yan S, et al. Mapping radon Hazard areas using 238U measurements and geological units: a study in a high background Radiation City of China. *J Radioanal Nucl Chem*. (2016) 309:1209–15. doi: 10.1007/s10967-016-4717-5
39. Miles J, Ball K. Mapping radon-prone areas using house radon data and geological boundaries. *Environ Int*. (1996) 22:779–82. doi: 10.1016/S0160-4120(96)00183-3
40. Miles JCH, Appleton JD. Mapping variation in radon potential both between and within geological units. *J Radiol Prot*. (2005) 25:257–76. doi: 10.1088/0952-4746/25/3/003
41. Koenker R, Bassett G. Regression quantiles. *Econometrica*. (1978) 46:33. doi: 10.2307/1913643
42. Koenker R, Bassett G. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*. (1982) 50:43. doi: 10.2307/1912528
43. Lum K, Gelfand AE. Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Anal*. (2012) 7. doi: 10.1214/12-BA708
44. Reich BJ, Fuentes M, Dunson DB. Bayesian spatial quantile regression. *J Am Stat Assoc*. (2011) 106:6–20. doi: 10.1198/jasa.2010.ap09237
45. Yu K, Moyeed RA. Bayesian quantile regression. *Statist Probabil Lett*. (2001) 54:437–47. doi: 10.1016/S0167-7152(01)00124-9
46. Koenker R, Chernozhukov V, He X In: L Peng, editor. Handbook of quantile regression. 1st ed. New York: Chapman and Hall/CRC (2017)
47. Borgoni R, Carcagni A, Salvati N, Schmid T. Analysing radon accumulation in the home by flexible M-quantile mixed effect regression. *Stoch Env Res Risk A*. (2019) 33:375–94. doi: 10.1007/s00477-018-01643-1
48. Borgoni R. A quantile regression approach to evaluate factors influencing residential indoor radon concentration. *Environ Model Assess*. (2011) 16:239–50. doi: 10.1007/s10666-011-9249-3
49. Fontanella L, Ippoliti L, Sarra A, Valentini P, Palermi S. Hierarchical generalised latent spatial quantile regression models with applications to indoor radon concentration. *Stoch Env Res Risk A*. (2015) 29:357–67. doi: 10.1007/s00477-014-0917-0
50. Sarra A, Fontanella L, Valentini P, Palermi S. Quantile regression and Bayesian cluster detection to identify radon prone areas. *J Environ Radioact*. (2016) 164:354–64. doi: 10.1016/j.jenvrad.2016.06.014
51. Dai D, Neal FB, Diem J, Deocampo DM, Stauber C, Dignam T. Confluent impact of housing and geology on indoor radon concentrations in Atlanta, Georgia, United States. *Sci Total Environ*. (2019) 668:500–11. doi: 10.1016/j.scitotenv.2019.02.257
52. Vukotic P, Stojanovska Z, Antovic N. Developing a method for predicting radon concentrations above a reference level in new Montenegrin buildings. *J Environ Radioact*. (2021) 227:106500. doi: 10.1016/j.jenvrad.2020.106500
53. Borgoni R, Tritto V, de Bartolo D. Identifying radon-prone building typologies by marginal modelling. *J Appl Stat*. (2013) 40:2069–86. doi: 10.1080/02664763.2013.804906
54. Stanley FKT, Irvine JL, Jacques WR, Salgia SR, Innes DG, Winquist BD, et al. Radon exposure is rising steadily within the modern north American residential environment, and is increasingly uniform across seasons. *Sci Rep*. (2019) 9:18472. doi: 10.1038/s41598-019-54891-8
55. Elio J, Crowley Q, Scanlon R, Hodgson J, Long S. Logistic regression model for detecting radon prone areas in Ireland. *Sci Total Environ*. (2017) 599-600:1317–29. doi: 10.1016/j.scitotenv.2017.05.071
56. Petermann E, Bossew P. Mapping indoor radon Hazard in Germany: the Geogenic component. *Sci Total Environ*. (2021) 780:146601. doi: 10.1016/j.scitotenv.2021.146601
57. Tobler WR. A computer movie simulating urban growth in the Detroit region. *Econ Geogr*. (1970) 46:234–40. doi: 10.2307/143141
58. Akkala A, Bhatt D, Devabhaktuni V, Kumar A. Knowledge-based neural network approaches for modeling and estimating radon concentrations. *Environ Prog Sustain Energy*. (2013) 32:355–64. doi: 10.1002/ep.11617
59. Bachirou S, Saïdou CK, Ii JENN, Bongue D, Abba HY, Hosoda M, et al. Mapping in a radon-prone area in Adamawa region, Cameroon, by measurement of radon activity concentration in soil. *Radiat Environ Biophys*. (2023) 62:427–39. doi: 10.1007/s00411-023-01042-3
60. Elio J, Crowley Q, Scanlon R, Hodgson J, Long S. Rapid radon potential classification using soil-gas radon measurements in the Cooley peninsula, county Louth, Ireland. *Environ Earth Sci*. (2019) 78:359. doi: 10.1007/s12665-019-8339-4
61. Yerrabolu P, Mareddy L, Bhatt D, Aggarwal P, Kumar A, Devabhaktuni V. Correction model-based ANN modeling approach for the estimation of radon concentrations in Ohio. *Environ Prog Sustain Energy*. (2013) 32:1223–33. doi: 10.1002/ep.11731
62. Banrion MH, Elio J, Crowley QG. Using Geogenic radon potential to assess radon priority area designation, a case study around Castleisland, co. Kerry, Ireland. *J Environ Radioact*. (2022) 251–252:106956. doi: 10.1016/j.jenvrad.2022.106956
63. Bertolo A, Bigliotto C, Giovani C, Garavaglia M, Spinella M, Verdi L, et al. Spatial distribution of indoor radon in Triveneto (northern Italy): a geostatistical approach. *Radiat Prot Dosim*. (2009) 137:318–23. doi: 10.1093/rpd/ncp252
64. Borgoni R, Quatto P, Somà G, De Bartolo D. A geostatistical approach to define guidelines for radon prone area identification. *JISS*. (2010) 19:255–76. doi: 10.1007/s10260-009-0128-x
65. Cafaro C, Bossew P, Giovani C, Garavaglia M. Definition of radon prone areas in Friuli Venezia Giulia region, Italy, using geostatistical tools. *J Environ Radioact*. (2014) 138:208–19. doi: 10.1016/j.jenvrad.2014.09.003
66. Cafaro C, Giovani C, Garavaglia M. Geostatistical simulations for radon indoor with a nested model including the housing factor. *J Environ Radioact*. (2016) 151:264–74. doi: 10.1016/j.jenvrad.2015.10.002

67. Nuhu H, Hashim S, Saleh MA, Sanusi MSM, Alomari AH, Jamal MH, et al. Soil gas radon and soil permeability assessment: mapping radon risk areas in Perak state, Malaysia. *PLoS ONE*. (2021) 16:e0254099. doi: 10.1371/journal.pone.0254099
68. Sabbarese C, Ambrosino F, D'Onofrio A, Pugliese M, La Verde G, D'Avino V, et al. The first radon potential map of the Campania region (southern Italy). *Appl Geochem*. (2021) 126:104890. doi: 10.1016/j.apgeochem.2021.104890
69. Salvi F, Raspa G, Torri G. Parametrization identification and characterization of the radon priority areas for indoor radon risk management. *J Environ Radioact*. (2023) 261:107120. doi: 10.1016/j.jenvrad.2023.107120
70. ESRI. (2024). "What are the different kriging models?". Available at: <https://desktop.arcgis.com/fr/arcmap/latest/extensions/geostatistical-analyst/what-are-the-different-kriging-models-.htm> (Accessed June 16, 2024).
71. Ciotoli G, Voltaggio M, Tuccimei P, Soligo M, Pasculli A, Beaubien SE, et al. 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. *J Environ Radioact*. (2017) 166:355–75. doi: 10.1016/j.jenvrad.2016.05.010
72. De Novellis S., Pasculli A., Palermi S. (2014). "Innovative modeling methodology for mapping of radon potential based on local relationships between indoor radon measurements and environmental geology factors", 109–119. New Forest: UK
73. Pasculli A, Palermi S, Sarra A, Piacentini T, Miccadei E. A modelling methodology for the analysis of radon potential based on environmental geology and geographically weighted regression. *Environ Model Softw*. (2014) 54:165–81. doi: 10.1016/j.envsoft.2014.01.006
74. Mancini S, Guida M, Cuomo A, Guida D. A Geogenic approach for the radon monitoring and the exposure assessment at a regional scale: the results of the Rad_Campania project. *Adv Geosci*. (2020) 52:87–96. doi: 10.5194/adgeo-52-87-2020
75. Martins LMO, Pereira AJSC, Sousa Oliveira A, Sanches Fernandes LF, Pacheco FAL. A new radon prediction approach for an assessment of radiological potential in drinking water. *Sci Total Environ*. (2020) 712:136427. doi: 10.1016/j.scitotenv.2019.136427
76. Sarra A, Nissi E, Palermi S. Residential radon concentration in the Abruzzo region (Italy): a different perspective for identifying radon prone areas. *Environ Ecol Stat*. (2012) 19:219–47. doi: 10.1007/s10651-011-0183-y
77. Elío J, Petermann E, Bossew P, Janik M. Machine learning in environmental radon science. *Appl Radiat Isot*. (2023) 194:110684. doi: 10.1016/j.apradiso.2023.110684
78. Wu P-Y, Johansson T, Sandels C, Mangold M, Mjörnell K. Indoor radon interval prediction in the Swedish building stock using machine learning. *Build Environ*. (2023) 245:110879. doi: 10.1016/j.buildenv.2023.110879
79. Dicu T, Cucuș A, Botoș M, Burghel B, Florică Ș, Baciuc C, et al. Exploring statistical and machine learning techniques to identify factors influencing indoor radon concentration. *Sci Total Environ*. (2023) 905:167024. doi: 10.1016/j.scitotenv.2023.167024
80. Li L, Blomberg AJ, Stern RA, Kang C-M, Papatheodorou S, Wei Y, et al. Predicting monthly community-level domestic radon concentrations in the greater Boston area with an ensemble learning model. *Environ Sci Technol*. (2021) 55:7157–66. doi: 10.1021/acs.est.0c08792
81. Timkova J, Fojtikova I, Pacherova P. Bagged neural network model for prediction of the mean indoor radon concentration in the municipalities in Czech Republic. *J Environ Radioact*. (2017) 166:398–402. doi: 10.1016/j.jenvrad.2016.07.008
82. Benà E, Ciotoli G, Petermann E, Bossew P, Ruggiero L, Verdi L, et al. A new perspective in radon risk assessment: mapping the geological Hazard as a first step to define the collective radon risk exposure. *Sci Total Environ*. (2024) 912:169569. doi: 10.1016/j.scitotenv.2023.169569
83. Kropat G, Bochud F, Jaboyedoff M, Laedermann J-P, Murith C, Palacios M, et al. Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: an application to Switzerland. *Sci Total Environ*. (2015) 505:137–48. doi: 10.1016/j.scitotenv.2014.09.064
84. Petermann E, Meyer H, Nussbaum M, Bossew P. Mapping the Geogenic radon potential for Germany by machine learning. *Sci Total Environ*. (2021) 754:142291. doi: 10.1016/j.scitotenv.2020.142291
85. Rezaie F, Panahi M, Lee J, Lee J, Kim S, Yoo J, et al. Radon potential mapping in Jangsu-gun, South Korea using probabilistic and deep learning algorithms. *Environ Pollut*. (2022) 292:118385. doi: 10.1016/j.envpol.2021.118385
86. Rezaie F, Kim SW, Alizadeh M, Panahi M, Kim H, Kim S, et al. Application of machine learning algorithms for Geogenic radon potential mapping in Danyang-gun, South Korea. *Front Environ Sci*. (2021) 9:753028. doi: 10.3389/fenvs.2021.753028
87. Pegoretti S, Verdi L. Machine learning for the analysis of indoor radon distribution, compared with ordinary kriging. *Radiat Prot Dosim*. (2009) 137:–328. doi: 10.1093/rpd/ncp254
88. Apte MG, Price PN, Nero AV, Revzan KL. Predicting New Hampshire indoor radon concentrations from geologic information and other covariates. *Environ Geol*. (1999) 37:181–94. doi: 10.1007/s002540050376
89. Rezaie F, Panahi M, Bateni SM, Kim S, Lee J, Lee J, et al. Spatial modeling of Geogenic indoor radon distribution in Chungcheongnam-Do, South Korea using enhanced machine learning algorithms. *Environ Int*. (2023) 171:107724. doi: 10.1016/j.envint.2022.107724
90. Popit A, Vaupotič J. Indoor radon concentrations in relation to geology in Slovenia. *Environ Geol*. (2002) 42:330–7. doi: 10.1007/s00254-002-0526-y