Journal: Science of Total Environment

A new perspective in radon risk assessment: mapping the geological hazard as a first step to define the collective radon risk exposure

Authors and affiliations

Eleonora Benà[1], Giancarlo Ciotoli[2,3], Eric Petermann[4], Peter Bossew[4*], Livio Ruggiero[5], Luca Verdi[6], Paul Huber[7], Federico Mori[2], Claudio Mazzoli[1] and Raffaele Sassi[1]

[1]Dipartimento di Geoscienze, Università di Padova, Padova, Italy

[2]Istituto di Geologia Ambientale e Geoingegneria (IGAG), Consiglio Nazionale delle Ricerche (CNR), Rome, Italy

[3]Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy

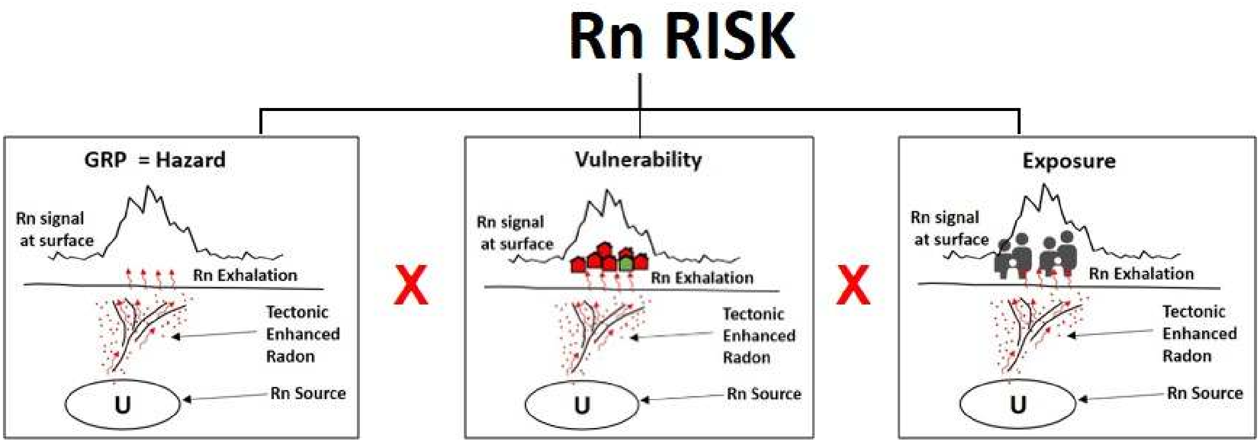
[5]Federal Office for Radiation Protection (BfS), Section Radon and NORM, Berlin, Germany (*retired)

[4]Istituto Nazionale per la Ricerca e la Protezione Ambientale (ISPRA), Roma

[6]Provincia Autonoma di Bolzano, Laboratorio analisi aria e radioprotezione, Bolzano, Italy

[7]Azienda Sanitaria dell'Alto Adige, Bressanone, Italy

Graphical abstract (mandatory)



Abstract (300 words)

Radon, a radioactive gas, is the largest source of ionizing radiation exposure for humans. As a result, its accumulation in confined environment can pose serious threats to health. In order to address the issue of radon exposure in dwellings, the European BSS directive 2013/59/EURATOM

established national reference level and guidelines to define Radon Priority Areas (RPAs). The Geogenic Radon Potential (GRP) is considered an effective hazard indicator for assessing the potential accumulation of this gas in buildings from geological sources. Several approaches, including multivariate geospatial analysis and artificial intelligence algorithms, have been applied to generate spatial continuous maps of the GRP based on soil gas point measurements and other related geo-environmental proxies.
The goal of this study is to map GRP of the central sector of the Pusteria Valley by a supervised machine learning algorithm (Random Forest), and use this map as a basis for identifying RPAs. The Pusteria Valley (North-eastern Italy) has been chosen as a pilot site due to its well-known geological, structural, and geochemical features. We then incorporate land use and population as vulnerability and exposure factors, respectively, to provide a final risk map. Results indicate that high Rn risk areas are associated with GRP values higher than 50 $kBqm^{-3}$. The use of the GRP map as a hazard component of radon risk leads to a new geological definition of RPAs. GRP is an important tool for mapping Collective Risk Areas (CRAs) since the BSS directive only addresses the identification of Individual Risk Areas (IRAs).

6

Highlights

- Mapping the Geogenic Radon Potential (GRP) using a robust supervised machine learning (i.e., random forest) as the most important spatial predictor for Indoor Radon Concentration (IRC).
- Apply the risk definition (i.e., product of hazard, vulnerability and exposure) in order to define the CRAs by intersecting the GRP map (hazard) with the type of location (vulnerability) and total population (exposure)
- Construct geological-based Collective Risk Areas (CRAs) map starting from GRP, coupled with land use (location type) and population density of the census tracts to define those areas subject to territorial planning by local authorities

Keywords

Geogenic Radon Potential, Machine learning, Pusteria Valley, Radon risk, Collective risk areas

Abbreviations: GRP = Geogenic Radon Potential; RPA = Radon Priority Area; IRC = Indoor Radon Concentration; BRS = Background Radon Source; TER = Tectonically Enhanced Radon; SRE = Surface Radon Exhalation; CRA = Collective Risk Area; IRA = Individual Risk Area;

SGRC = soil gas radon concentration; GRHI = Geogenic radon hazard index; BSS = Basic Safety Standards, TGDR = Terrestrial gamma dose rate

Introduction

Radon (^{222}Rn, hereafter Rn) is a radioactive gas considered the major source of ionizing radiation exposure for the population. Its potentially harmful effects on human health have been extensively documented (WHO, 2009). In particular, radon gas represents a serious hazard when it accumulates in indoor environments (Indoor Radon) such as in residential houses and workplaces.

Exposure to indoor radon is a serious problem that has prompted Europe to introduce legislation (Basic Safety Standards Directive 2013/59/EURATOM) which, on the one hand, establishes maximal national reference level aimed to reduce the exposure to Indoor Radon Concentration (IRC); on the other hand, promotes the public administrations to define Radon Priority Areas (RPAs). For this reason, it is fundamental to identified areas characterised by the highest Rn hazard for the population.

The concentration of radon gas in the environment can vary depending on the geological characteristics of an area. Radon produced within the Earth can migrate through permeable pathways (faults and fractures) in rocks and soil, or dissolved in groundwater, up to the shallow environment. Then, radon can diffuse into the atmosphere, or enter buildings. Geogenic Radon Potential (GRP) can be considered an optimal Rn hazard indicator. It is conceptualised as "*what Earth delivers in terms of radon*" from the geogenic sources (e.g., radionuclides content, faults and fractures) towards the atmosphere (Bossew, 2015; Bossew et al., 2020).

In particular, GRP is characterised by the interaction of three natural processes:

• the Background Radon Source (BRS), the process that produces Rn isotopes (^{222}Rn and ^{220}Rn) through the natural decay of uranium (U) and thorium (Th), which are found in varying concentrations in rocks and soil;

• the Tectonically Enhanced Radon (TER, Benà et al., 2022), the additional process allowing radon to migrate more easily to the surface through permeable pathways (e.g., faults and fractures in the crust) from deeper sources, caused by the stress increase and pressure conditions associated with tectonic activity;

• the Surface Radon Exhalation (SRE), the process by which radon gas is released from the ground into the surrounding environment. SRE includes the variables that affect radon movement in the shallow soil up to the soil/atmosphere interface (e.g., land morphology, soil permeability, humidity and temperature). This quantity of radon, which has not yet been extensively measured, represents the amount of radon that could potentially enter buildings.

However, BRS and TER represent the dominant geological radon sources.

Over the years, several approaches have been applied to estimate the GRP over an area (e.g., Neznal et al., 2004; Bossew et al., 2015; Pasztor et al., 2016; Ciotoli et al., 2017; Giustini et al., 2019; Petermann et al., 2021; Coletti et al., 2021).

Neznal et al. (2004) proposed an early method to define the GRP that has been widely used due to its simplicity up until this day. The method was based on the measure of two quantities: the Rn concentration in the soil and the soil permeability. Equation 1 reports the Neznal formula to calculate the GRP (dimensionless):

$$GRP_{Neznal} = \frac{SGRC}{-\log_{10} k - 10} \tag{1}$$

where SGRC is the soil gas radon concentration ($kBqm^{-3}$) measured at a depth of about 0.8 m, and k is the soil gas permeability in m^2.

More recently, Pasztor et al. (2016) and Ciotoli et al. (2017) applied multivariate geospatial analysis (regression kriging and geographical weighted regression, respectively) for GRP modelling by using SGRC and selected environmental proxies for the first time.

In the past three years, researchers have developed more advanced multivariate techniques, such as regression kriging (Coletti et al., 2021) and machine learning (ML) algorithms (e.g., random forest, etc.), which include several predictors associated with the geogenic Rn component (Petermann et al., 2021). However, it is important to emphasise that all these regression techniques require a response variable (i.e., SGRC or IRC).

However, many European countries lack sufficient SGRC measurements and permeability data to support the mapping of GRP. As a consequence, the calculation of a Geogenic Radon Hazard Index was proposed by Bossew et al. (2020). The concept of the GRHI arose from the need to determine a specific indicator using regionally accessible geological variables. However, it is difficult to maintain consistency between GRHI scores in neighbouring regions when using multiple predictors. In other words, GRHI values for areas with comparable geological factors but different data sources should be nearly equivalent.

In order to produce consistent maps, Cinelli et al. (2015) proposed a first method to assign weights to continuous or categorical input variables (covariates) based on their contribution to the index. The weighted "mean class" is then used to calculate the Geogenic Radon Hazard Index (GRHI), a dimensionless quantity. The weights allocated to each covariate are determined by the observed correlations with the GRP in regions where GRP data are effectively accessible. Another way to avoid the issue is the application of other more subjective techniques which does not require a response variable (i.e., Spatial Multi Criteria Decision Analysis, SMCDA, Ciotoli et al., 2020). Since the GRP represents the amount of radon that could potentially enter buildings, it is considered as the most significant spatial predictor of the IRC; therefore, it is crucial to map the GRP as accurately as possible using a robust methodology. In this regard, the BSS European Directive 59/2013, transposed into Italian law by Legislative Decree n.101/2020, emphasizes further the identification of RPAs, originally defined as those areas where the annual average IRC in a significant number of dwellings is expected to exceed the reference level of 300 Bqm^{-3}. However, the concept and interpretation of "significant number of buildings" in the European Directive remained unclear.

In a recent study, Petermann et al. (2022) emphasized that rather than the collective concept of geogenic risk, the interpretation of "significant number" of buildings is factually based on the concept of geogenic hazard, in that it is assumed to mean something like a relevant percentage of buildings within an area notwithstanding the number of houses or of people affected.

The latter accounts for the number of people affected, based on the notion that the possible detriment caused by Rn exposure in an area - the number of lung cancer fatalities - depends on hazard and presence of people who can be harmed. After all, according to the BSS, it is the detriment that should be reduced by Rn policy.

On the other hand, staying with the hazard- or individual risk-type notion of RPA, there is no uniform decision at the regional scale regarding the selection of the reference level (RL) and the threshold of probability percentage (p_0) of buildings exceeding the RL. In general, the majority of European nations (including Finland, Germany, Greece, Montenegro, and Spain) adhere to the European Directive adopting the recommended reference level of 300 Bqm^{-3} and a common probability threshold of 10% (Bossew, 2018). Particularly, Germany only considers IRC measurements in rooms on the ground floor of buildings with basements, whereas Spain only considers measurements in rooms on the ground or first level. Ireland has an RL of 200 Bqm^{-3} and a p_0 of 10%. Other countries, including Austria and Switzerland, define distinct priority levels based on RL and measurer IRC (Bossew, 2018). Italy has an RL of 300 $Bqm^{-3}$and a p_0 of 15% (D. Lgs. n. 101/2020). A map of the confusing diversity of RPA definitions across Europe has been shown in Bossew and Suhr (2023, see Fig. 2 in the cited paper).

As reported in Bossew et al., 2021, the goals of radiation protection from Rn indoor are twofold:

- to protect people from high Rn exposure in order to reduce individual risk (even if few people are involved);

- to avoid high exposure to the community, because the harm to society is proportional to the collective risk.

But how the European Community protects people from Rn risk?

In order to limit radon exposure and thereby lower the population's likelihood of contracting lung cancer, the regulatory program aims to identify RPAs and implement ad-hoc mitigation plans. European legislation aims to reduce the detriment from Rn exposure (i.e., the number of lung cancer deaths) and as a consequence, reduce the collective exposure. Collective exposure can be assessed by introducing the concept of collective risk, a complement of the individual risk concept (as interpreted by the BSS directive, "classical" RPA). Collective risk can be figured as consisting of many little individual risk zones. Accordingly, the demand to identify RPAs focused only on building remediation rather than land use planning is a controversial subject affecting both European and regional legislation.

In the light of these considerations, and in the absence of any specific approach for defining RPAs at the European level, we propose to map collective risk areas (or detriment, CRAs) as the combination of the GRP, vulnerability and exposed factors, as a complement to mapping individual risk areas (IRAs) associated with IRC (i.e., "traditional" RPA). Rn abatement policy must take care of these areas in order to minimize the damage, while also preserving areas of high individual risk. The goal of this study is to show the effectiveness of a CRAs map to define Rn risk areas in a test site, the Pusteria Valley (Bolzano, northern Italy), based on the GRP map (i.e., hazard factor) generated by using multivariate machine learning (MML) technique (e.g., Random Forest, RF). The Pusteria Valley was chosen because it is well-known from a geological, structural, and geochemical point of view with a lot of available data (Benà et al., 2022), and also because it was already the target of IRC measurements (Minach et al., 1999). The obtained GRP map (hazard) was merged with the land use type (vulnerability) and population (exposure) of census tracts available from the ISTAT (i.e., *Istituto Nazionale di Statistica*) website in order to identify risk areas that may be subject to territorial planning by local authorities. The CRAs map can be merged with the RPAs map (*sensu stricto*).

The construction of GRP maps is a fundamental tool for both Rn hazard and risk analysis and, according to a new more geological view, as a base map for the identification of Radon Priority Areas (RPAs). This is important for collective risk assessment (land-use planning and prevention), and for individual risk assessment (i.e., more strategic planning of indoor surveys, and specific remediation actions).

2. Methods

A dataset including different variables (e.g., response and predictors) was used to elaborate the GRP map of the study area by applying a ML technique (i.e., Random Forest, Breiman, 2001) and predict the radon values at the center points of a 50x50 m fishnet. The resulting GRP map will be then used as the hazard factor in the risk equation (Eq. 2, see section 2.4) and multiplied by census tract indicator of land use and population density representing the vulnerability and the exposure factors, respectively.

Figure 1 shows the flowchart of the applied procedures. Data were processed using ArcGIS Pro 3.1.2 (copyright 2023@ESRI Inc.) and Scikit learn library in Python PyCharm 2023.1.2 (Copyright © 2010–2023 JetBrains s.r.o.).

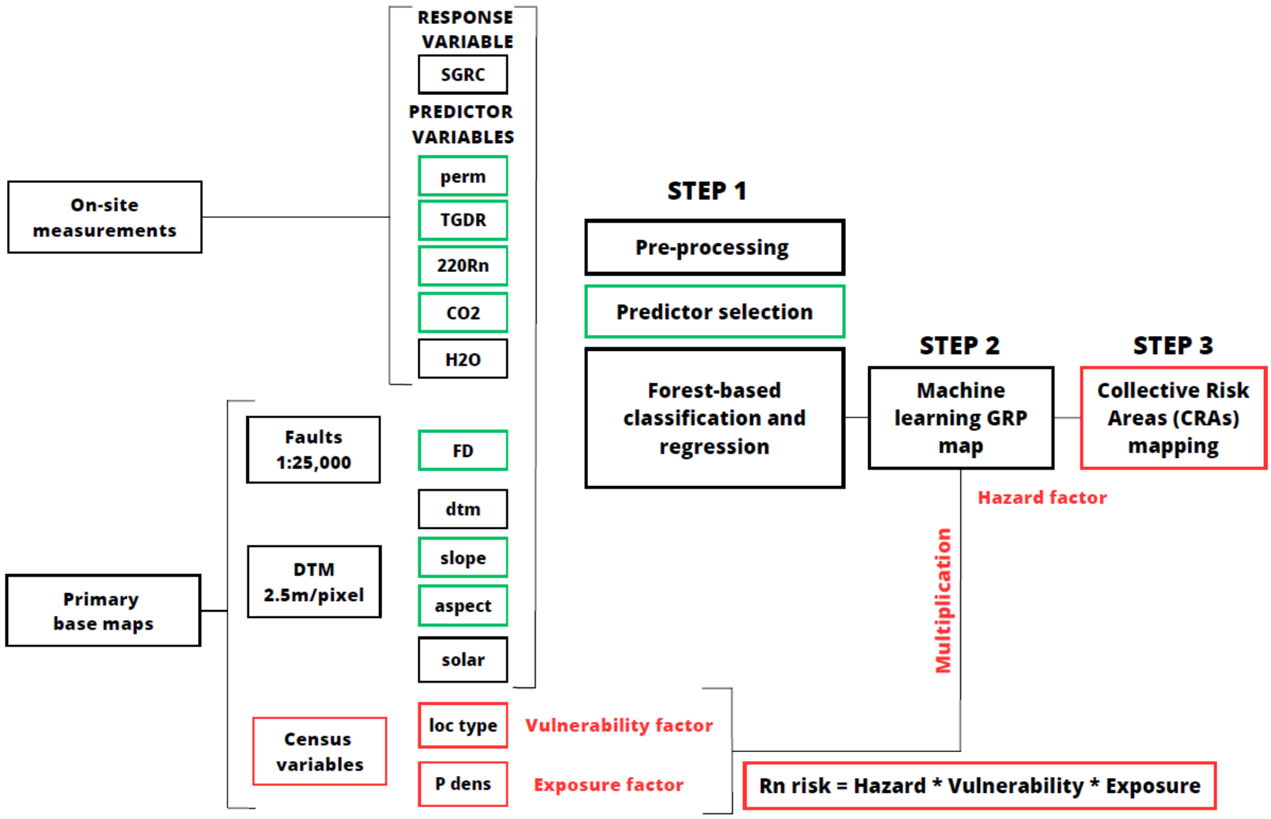


Figure 1. Flowchart of the mapping process and procedures. SGRC = soil gas radon concentration; perm = soil permeability; TGDR = terrestrial gamma dose radiation; ^{220}Rn = thoron; CO_2 = carbon dioxide concentration in soil gas; H_2O = concentration of radon dissolved in water; FD = fault density; dtm = digital terrain model; slope = slope; aspect = aspect ratio; solar = solar radiation; loc type = location type: P dens = population density; GRP map = geogenic radon potential map; RPAs = radon priority areas. SGRC, permeability, TGDR, thoron, carbon dioxide, radon dissolved in water, faults, DTM were pre-processed in order to apply the forest-based classification and regression (first step) to construct the GRP map (hazard factor). The GRP map was then multiplied by the location type (vulnerability factor) and population density (exposure factor) to construct the collective risk map.

2.1 Dataset

The dataset used consists of one response variable (SGRC) and ten independent variables that were either measured on-site or derived from primary base-maps, available online from the Bolzano province's Geo-catalogue (http://geokatalog.buergernetz.bz.it/geokatalog/#!). These ten variables were selected as potential predictors for machine learning regression models.

Soil gas surveys (^{222}Rn, ^{220}Rn, CO_2) (Benà et al., 2022), TGDR, and permeability measurements were collected on-site during two separate field campaigns in summer 2021 and 2022 under similar and stable climatic conditions. The Digital Terrain Model (DTM 2.5 m resolution) and fault distribution were obtained directly from the base maps of the Bolzano Province Geo-catalogue. The examined predictors were pre-processed using geospatial analysis to generate 50x50m raster maps (see Fig. S1 in supplementary materials). The "Extract multi-value to point" tool of ArcGIS Pro was used to assign the values of the predictor grids to each observation of the collected soil gas samples. The obtained dataset, containing the predictors and the response variable (SGRC), was used to train the Random Forest (RF) model. Once the best model was found, it was applied to a regular point 50x50m fishnet corresponding to the centroids of the predictors' raster grid pixels. The final dataset consists of 27,758 points that includes complete information for all predictors. The following sections provide a detailed description of the response variable and the predictors.

2.1.1 Response variable

SGRC ($kBqm^{-3}$) was used as a response (dependent) variable in the Forest Regression algorithm to describe the GRP map. The Rn dataset consists of 278 measurements obtained in the field using the methodology and sampling pattern described by Benà et al. (2022).

2.1.2 On-site predictor variables

Five predictors were measured in the field: thoron (^{220}Rn), carbon dioxide (CO_2), TGDR, permeability and ^{222}Rn dissolved in water. The same sampling technique and pattern adopted for the measurement of Rn concentrations in soil gas was also adopted for the measuring of thoron and carbon dioxide (CO_2) (Benà et al., 2022).

TGDR measurements

TGDR measurement have been performed at 76 sampling points using the NaI γ-ray portable scintillometer (Scintrex GRS-500) pre-set on the total count rate window corresponding to an energy interval range between 80-3000 keV. The device was held 1 m above the ground for a measuring time equal to the time needed to reach a 3% accuracy. The sensitivity factor of the

Scintrex GRS-500 is 3.40 cps/nGyh^{-1} and allows the counting rates to be converted into the IS unit of the gamma dose rate (μSv/h, Giustini et al., 2019, 2022).
Geostatistical analysis (experimental variogram calculation, modelling, kriging) was used to obtain a prediction map of the TGDR (see Fig. S2 a and b in supplementary materials). This variable is used as a proxy of the BRS contribution (i.e. radionuclides content in rocks) of the geogenic radon component.

Permeability

Soil gas permeability directly affects radon gas migration from the deep source (mainly by advection along faults), and in the shallow soil (by diffusion prevalent mechanism) (Nuhu et al., 2021; Neznal et al., 2005). High permeability allows the upward migration of radon, enabling its exhalation to the atmosphere, while the presence of a shallow soil layer with low permeability could increase the accumulation of radon in the soil with a consequent decrease of exhalation rate at the soil-atmosphere boundary (Castelluccio et al., 2015; Johner et al., 2001). The radon concentration in soil gas is directly dependent on the geological characteristics of the area (i.e., radionuclide content, presence of fractures and faults) and can be strongly influenced by soil permeability in terms of soil pore dimensions and soil water content (i.e., soil moisture) (Benavente et al., 2019; Lara et al., 2015). Additionally, some other physical characteristics of soils, such as soil texture and grain size, have a significant impact on the mechanisms of radon emanation and exhalation in the soil environment (Huynh Nguyen et al., 2018; Yang et al., 2019).
In the study area, the soil permeability was measured at 76 sampling points with a permeameter developed by the University of Roma Tre and directly connected to the soil gas sampling probe (Castelluccio et al., 2015). The soil is assumed to be homogeneous and isotropic, and standard state is considered; the air is assumed to be incompressible. The calculation of the final soil permeability (k) is based on Darcy's equation and expressed in m^2. Geostatistical analysis (i.e., experimental variogram calculation, modelling, kriging) was used to obtain a prediction map of the soil permeability (see Fig. S3 a and b in supplementary materials).

Radon dissolved in groundwater

Dissolved ^{222}Rn was measured at 22 captured water springs in the study area. Water samples from selected springs were already studied for their chemical-physical conditions by the *Agenzia provinciale per l'ambiente e la tutela del clima - Laboratorio analisi acque e cromatografia* (Bolzano province) in 2022.

The water was sampled directly from the captured springs using glass bottles. Rn concentrations were measured using RAD7 in the sniff mode connected to Big Bottle RAD H_2O and drystick (drierite desiccant) accessories. Prior to the measurements, the system was purged to guarantee that the moisture (water content) inside the system was reduced to less than 10% humidity. The sampled bottle was then connected in a closed air-loop mode to the RAD7 (Durridge Company Inc.). During system operation, continuous circulation gradually enriches the air contained in the closed loop with the Rn dissolved in the water sample. Each measurement was performed with a 5-minute integration period and was repeated until the difference between the last two readings is less than 5-10%. The final result was calculated by averaging the previous two integrations. Thiessen polygons was constructed to create a map of areas of influence around the water springs. Water springs represents the centroid of the Thiessen polygons in which the measured dissolved radon value (i.e., the centroid) is assumed to be representative of the area underlying the entire polygon. The resultant map was transformed in a 50x50m raster grid and used as predictor in the RF model.

2.1.3 Derived predictor variables

Fault density

Faults and fractures represent the main pathway for radon, and other gases (CO_2 and CH_4) migration in the subsoil from deep sources (see Ciotoli et al., 2007, 2014, 2017, 2020; Giustini et al., 2019). Therefore, the network of the fractured zone characterising the study area has been used as a proxy of the secondary permeability. The distribution of the main faults in the study area (Keim et al., 2013) was converted into a fault density (FD) map using the quadratic kernel density function (Silverman, 1986), as described in Benà et al. (2022).

Digital terrain model

The Digital Terrain Model (DTM) of the study area (i.e., elevation) was used as a proxy of the meteorological conditions which may strongly affect radon migration and exhalation mechanisms. The mobility of radon can be impacted by the presence of slopes, hills, and depressions, which can alter air flow and soil pressure (Gundersen et al., 1992). Radon may not build up as much in areas with rough terrain because air circulation and groundwater drainage may be improved. On the other hand, low-lying areas and depressions may act as radon traps, resulting in higher levels of the gas (Sukanya et al., 2021). Furthermore, Griffiths et al. (2014) highlighted how crucial it is to take topographic interactions into account when estimating radon concentrations across different geographical areas. The DTM (2.5 m/pixel) of the Bolzano province is available on the Geo-catalogue of the Bolzano province (*Rete Civica dell'Alto Adige, https://geoportale.retecivica.bz.it/default.asp*).

The "Surface Parameters" tool of Spatial Analyst" in ArcGIS Pro was applied to the DTM to create maps of further potential proxies: slope, solar radiation (e.g., Areal Solar Radiation) and aspect ratio. The slope can be used as a proxy of soil moisture and shallow soil meteorological conditions; the solar radiation is used as a proxy of the microclimate/temperature. Aspect (i.e., slope exposure) refers to the compass direction of the downhill slope faces in relation to the sun. Into details, slope conditions such as the angle, aspect, and elevation of a land surface can strongly influence local weather patterns and microclimates acting as a proxy of meteorological conditions in different ways (e.g., sun exposure, rainfall distribution, wind patterns, temperature gradients), all of which may impact radon generation and movement (Zalloni et al., 2018).

2.2 Predictor selection

Predictor selection was conducted using Least Absolute Shrinkage and Selection Operator (Lasso) regression. Least Absolute Shrinkage and Selection Operator (Lasso) regression is an extension of ordinary least squares (OLS) regression used in statistical modelling and machine learning (ML) to estimate the relationships between variables and make predictions (Tibshirani, 1996, 2011; Durrant et al., 2021). This technique aims to find an equilibrium between model simplicity and accuracy by introducing a penalty term into the traditional linear regression model, which enables sparse solutions in which some coefficients are forced to be exactly zero. LASSO is especially useful for variable selection because it can automatically identify only the most significant and discard irrelevant or redundant variables, especially if we assume that many of the features do not contribute significantly to the target variable (Durrant et al., 2021; Handorf et al., 2020). It also helps to prevent overfitting by removing variables with low predictive value, potentially making the model more robust across datasets. Furthermore, because it can choose between correlated explanatory variables, it can aid in the optimization of models with high multicollinearity. In simple words, the Lasso regression adds a penalty term to the MSE used in linear regressions. This penalty term is proportional to the sum of the absolute values of the variable coefficients. The Lasso regression seeks the coefficient values that minimize the sum of the MSE and the penalty.

The Lasso regression cost function is defined as follows (Eq. 3):

$$J(\beta) = \left(\frac{1}{n}\right) * \sum (y_i - \hat{y}_i)^2 + l * \sum |\beta_j| \tag{3}$$

where

- J(β) is the cost function
- n is the number of data or physical samples (statistically, the sample size)
- y_i is the actual output for the i-th sample
- $\hat{y}_i$ is the predicted output for the i-th sample
- βj represents the coefficients (weights) associated with each feature
- l is the regularization parameter that controls the amount of regularization applied to the model. Higher values of λ led to more regularization, resulting in a more pronounced feature shrinkage and potentially some coefficients becoming exactly zero.

In this work, Lasso regression was applied in Python code using the scikit-learn module (sklearn.linear_model.Lasso).

2.3 Machine learning and GRP mapping

Machine learning (ML) algorithms allow to solve very complex problems. First, generating a model based on processing the dataset and then, predicting the values of a new input data point by executing the created model (supervised machine learning) (Rebala et al., 2019).
In the literature, recent works have applied ML techniques for spatial prediction in a number of studies that deal with environmental science (e.g., landslide applications, Micheletti et al., 2014, Tehrani et al., 2022; soil mapping, Hengl et al., 2017, GRP mapping, Petermann et al., 2021; time series analysis, Janik et al., 2018). ML can handle complex multi-dimensional non-linear relationships and mostly makes no or weak assumptions of the underlying distribution of the data (Fouedijo and Klump, 2019). Furthermore, ML based approaches have been proven to outperform classical geostatistical models for several prediction tasks dealing with highly complex systems (e.g. Nussbaum et al., 2018; Hengl and MacMillan., 2019; Li et al., 2019). ML models display a high performance due to their ability to reflect the influence and interplay of a multitude of factors.
Random Forest (RF) is an ensemble classifier algorithm developed by Breiman (2001) typically used in classification and regression problems providing an output based on a Decision Trees structures. Decision Tree is a regression model built using a series of decisions based on variable values. Splitting values are determined to best separate subsets of data to take one path or the other. Random Forest is a method of averaging many Decision Trees created from a bootstrap sample of the full training set using a subset of predictors (=mtry) at each split in order to reduce overfitting by a single Decision Tree. It uses bagging (i.e., bootstrap aggregation) to create numerous Decision Trees by sampling a subset of training data with replacement and constructing the model based on the sampled training set (Rebala et al., 2019).

In this study, we have used Scikit learn code in python to apply a supervised machine learning method (i.e., Random Forest) to model the relationships between the SGRC (response variable) and the nine predictors described in the section 2.1.2 (^{220}Rn, CO_2, TGDR, permeability, fault density, digital terrain model DTM, slope, aspect ratio and solar radiation).

2.4 Radon risk mapping

2.4.1 Risk concept

The development of GRP maps is a valuable tool for hazard analysis; this map, coupled with vulnerability and exposure factors, it is critical to assess the collective risk, i.e., the risk to which the general public is exposed by geological causes.

Furthermore, the map of the collective risk can be combined with the indoor measurements (thus including the knowledge of the geological base processes) to better delineate Radon Priority Areas, and manage the individual risk in terms of remediation activities.

As above mentioned, we can define the risk as the product of hazard, vulnerability and exposure (Eq. 2).

$$Risk = Hazard * Vulnerability * Exposure \tag{2}$$

The application of the risk definition in order to mapping the CRA represents a first and easy method to assess the collective Rn exposure in the study area.

2.4.2 Construction of CRA map

According to the risk equation, in order to construct the CRA map we identified the GRP as the hazard term, the location types and the total population of the census tracts of the study area (available on the ISTAT web site, www.istat.it/it/archivio/104317#accordions) as vulnerability and exposure factors, respectively.

The location type in the ISTAT dataset is marked by a number identifying the specific type of building areas from 1 (residential areas) to 4 (sparse houses). These numbers were reclassified in order to assigned the highest weight (4) to the area with the highest expected mean population density, as follow: (i) location type 4 = residential areas; (ii) location type 3 = housing unit; (iii) location type 2 = industrial areas; (iv) location type 1 = sparse houses.

Then, the total population and the location type have been used to calculate the population density as the ratio between the total population living in a specific location type and the total area (in km^2) of the census tract. The maps of the location type and the population density were converted in 50m

x 50m raster grid and normalised to the maximum value before constructing the final Rn risk map; the GRP map was also normalized to the maximum value.
Furthermore, these three factors (GRP, Location Type and Population Density) were multiplied using the Raster Calculator tool in ArcGIS Pro according to Eq. 2. The resulting risk map has been further standardized and the Zonal Statistic tool of Spatial Analyst in ArcGIS Pro was applied to assign a risk value to each polygon of the census tract. We considered the maximum risk value assigned to the polygon in order to visualize the risk map and to create the risk classes. The final risk map is divided into three risk classes expressed in percentage of risk (i.e., low, medium and high).

3. Results

3.1 Selected predictors, RF modelling and predictors importance

Results of LASSO regression identified 7 predictors out of the 10 candidates: TGDR, CO_2, FD, ^{220}Rn, slope, aspect and soil permeability (see table S1 in supplementary materials). DTM, solar radiation, and Rn in groundwater all show coefficients of 0, so they are excluded from the model because they are considered as being misrepresentative. Furthermore, though slope and aspect show not significant coefficients, they were however included in the RF model. The Variance Inflation Factor (VIF) was also calculated for the 7 selected predictors to evaluate multicollinearity, in order to be sure that among the 7 predictors there is no redundancy. All the selected predictors show VIF < 7 (see table S2 in supplementary materials). The selected predictors include: one geophysical parameter (TGDR), geochemical parameters (^{220}Rn and CO_2), geological parameters (Fault and Permeability), and geomorphological parameters (slope and aspect). All of these parameters are representative of the overall process at the core of Rn production (source), migration, and behaviour in shallow soil, as well as at the soil-atmosphere interface.
Before to execute the RF model, the number of trees was set at 1000. The analysis of the model performance shows R^2 of 0.93 and 0.47 for training and test data, respectively, and a RMSE of 0.30 and 0.83 for training and test data, respectively (see Fig. S4 in supplementary materials displaying the predicted vs observed values for training and test data).
The importance of the individual predictors in the RF model is considered as the relative influence of an individual predictor on the model performance (Fig. 2). The variable percentage importance shows that TGDR, CO_2, fault density, ^{220}Rn, slope, aspect and permeability have the main influence in the model performance, respectively. In particular, TGDR, a proxy of the Rn source in rocks and soil, and the CO_2 (the main carrier gas in the study area, Benà et al., 2022) represent the most influencing predictors with an importance higher than 30%. The fault density (FD) (i.e., proxy of

secondary permeability) highligths an important decrease in the percentage range of 10-15%. ^{220}Rn and slope show an importance lower than 10% followed by the aspect ratio and soil permeability lower than 5%, respectively.

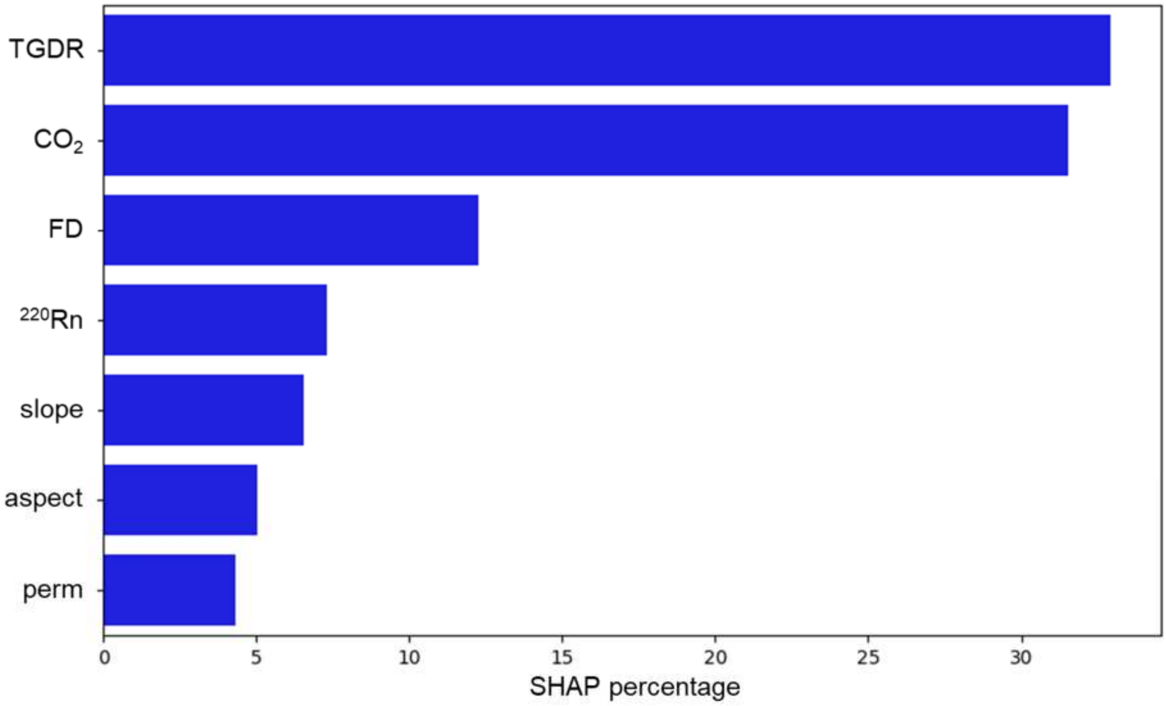


Figure 2. Feature importance based on SHAP value percentage in the RF model. The predictors are ordered by decreasing importance.; X-axis: SHAP percentage; Y-axis = selected predictors. TGDR = terrestrial gamma dose rate; CO_2 = carbon dioxide; FD = fault density; ^{220}Rn = Thoron; perm = soil permeability.

Furthermore, we constructed the SHAP diagram by using the “shap” library in Python to highlight the impact of each selected predictors on the model prediction (Fig. 3). The Y-axis of the SHAP diagram reports the 7 selected predictors in descending order of importance in the RF model from TGDR (the most influent) to the soil permeability (the less influent). The X-axis of the SHAP diagram represents the SHAP values quantifying the impact of a single feature on the model’s output: positive SHAP values indicate that the feature positively contributes to the output, while negative values suggest a negative contribution. Red and blue dots represent the contribution of individual features to the prediction compared to a reference value. Red dots represent positive contributions and indicate that the feature is increasing the predicted output. Blue dots represent negative contributions and indicate that the feature is decreasing the predicted output. In particular, the SHAP diagram pointed out that positive values of TGDR, CO_2, FD, ^{220}Rn, slope and permeability exert the main influence in the model output; while, aspect is the only variable that has influence in the model output for negative values.

Figure 3. SHAP diagram. Y-axis: reports the 7 selected predictors in descending order of importance in the RF model; X-axis: the SHAP values quantifying the impact of a single feature on the model's output: positive SHAP values indicate that the feature positively contributes to the output, negative values suggest a negative contribution. Red and blue dots represent the contribution of individual features to the prediction compared to a reference value. Red dots represent positive contributions indicating that the feature is increasing the predicted output. Blue dots represent negative contributions indicating that the feature is decreasing the predicted output. TGDR = terrestrial gamma dose rate; CO_2 = carbon dioxide; FD = fault density; ^{220}Rn = Thoron; perm = soil permeability.

The next step in model interpretation is understanding the effect of an individual predictor on the model output. Partial dependent plots (PDPs) were constructed by using "pdpbox" library in Python to analyse the relationship between a target feature and the model's predicted outcome while considering all other features as fixed (see Fig. 5S a-g and the related explanation in supplementary materials). It helps to visualize the relationship between a target feature and the model's predicted outcome. The PDP of each predictor is calculated by accounting for the average effect of the other predictors in the model (Petermann et al., 2021).

3.2 GRP map

Random forest algorithm has been applied to construct the GRP map of the study area by using SGRC as response variable and the 7 selected predictors (i.e., TGDR, CO_2, fault density, ^{220}Rn, slope, aspect, permeability). The final predicted GRP map ranges between a minimum value of 7.21 kBq·m^{-3} and a maximum value of 182 kBq·m^{-3} (Fig. 4). According to the results reported in Benà et al. (2022), we consider high GRP values those exceeding 50 $kBqm^{-3}$, i.e., the local background. Higher GRP values extend along the E-W direction from Falzes to Chienes (central part), to Terento municipalities, accordingly to the direction of the wide fracture zone belonging to the Pusteria fault system. High GRP values are linked to the Tectonically Enhanced Radon (TER) quantity (Benà et al., 2022).

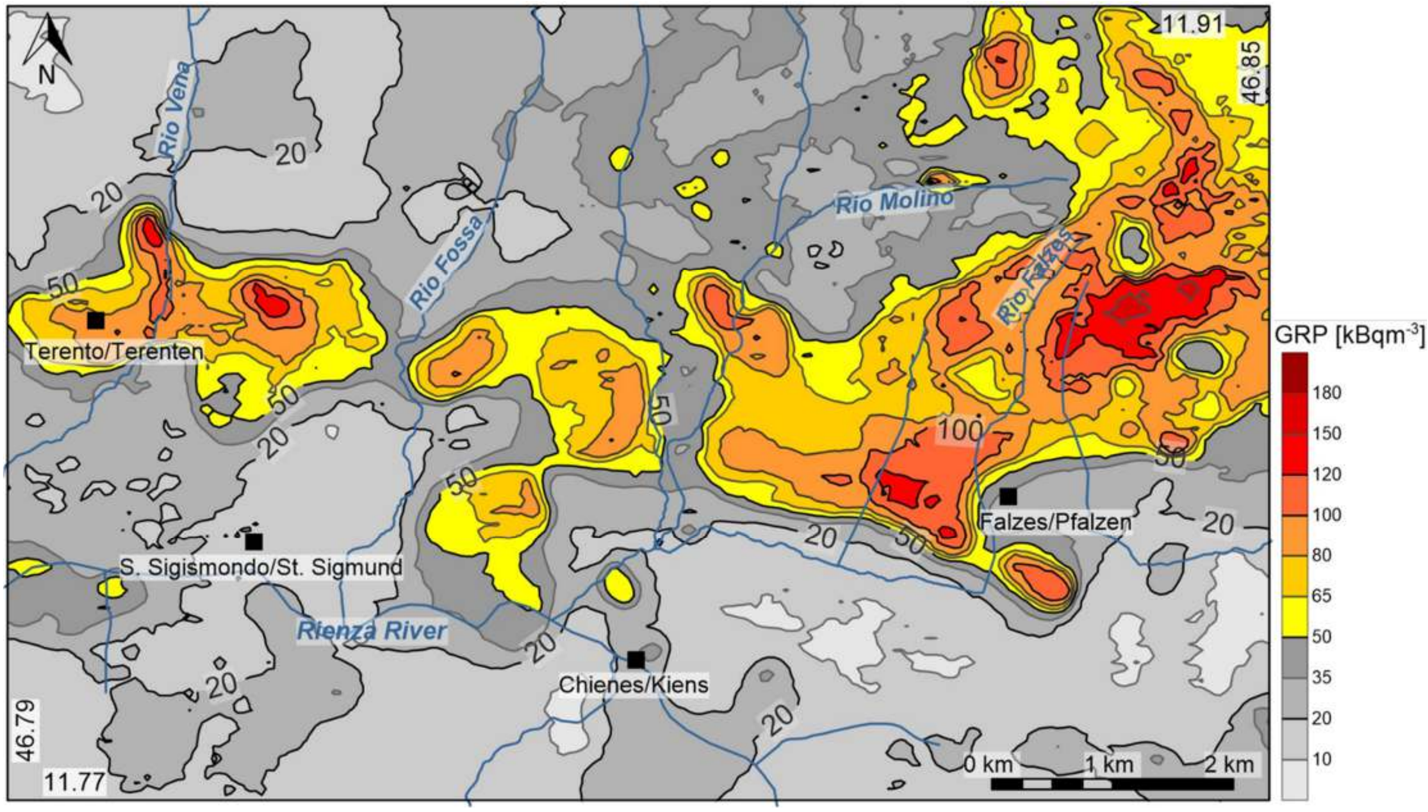


Figure 4. Geogenic Radon Potential ($kBqm^{-3}$) map of the study area.

3.3 The CRA map

Figure 5 shows the CRA map of the study area, representing the density of collective risk and obtained by multiplying the GRP map, the location type (vulnerability) and the population density (exposure factor). The map was divided into three risk classes using the natural breaks method as follow: i) risk < 5%, low risk (in white); ii) 5%<risk<50%, medium risk (in orange); iii) risk >50% high risk (in red). The CRA map is linked to table 1 which summarizes some parameters characterizing the three defined risk classes: (i) the average GRP value in $kBqm^{-3}$; (ii) the average population density expressed in number of people per km^2; (iii) the location type (i.e., 4, 3, 2, 1); (iv) the total area covered by the considered risk class.

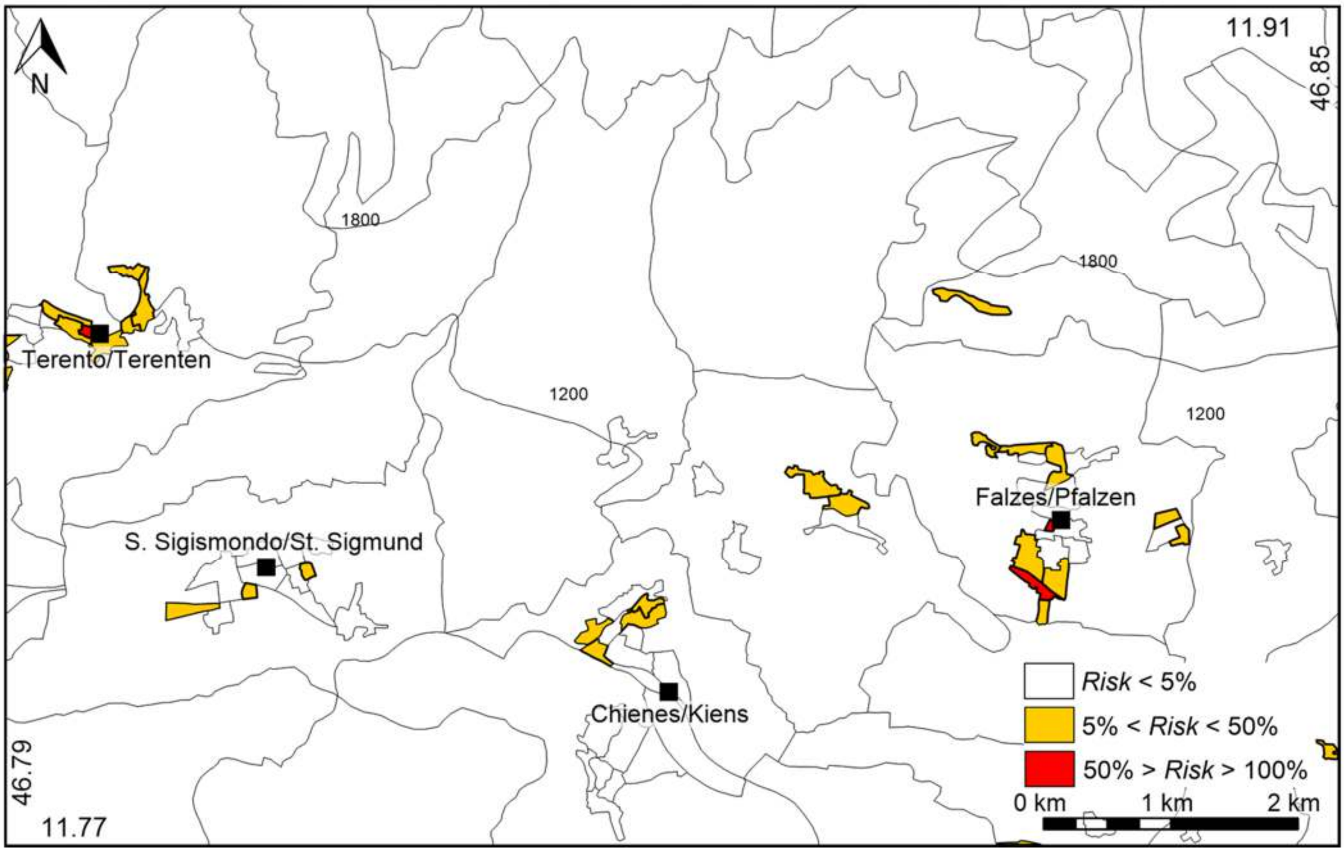


Figure 5. Map of the Collective Risk Areas.

Collective risk class	Risk level (%)	GRP mean ($kBqm^{-3}$)	Population density (people km^{-2})	Population (people)	Location type	Area (km^2)
Low	< 5	63.50	546	5927	4, 3, 2, 1	68.51
Medium	5 – 50	65.11	6116	3072	4, 3, 2, 1	0.75
High	> 50	75.88	17549	622	4	0.05

Table 1. The table reports the risk class and the correspondent percentage of risk, the mean GRP value, the population density, the location type and the extension of the area covered by the considered risk class.

Most of the study area (68.51 km^2) falls within low risk areas; this agrees with the mountainous morphology of the territory where most of the population is concentrated in the residential areas of the main municipalities (Terento, Chienes and Falzes). In general, the mean GRP values (hazard) exceed the local background value of 50 $kBqm^{-3}$ in all the three risk classes and slightly increases from low risk (63.50 $kBqm^{-3}$) to high risk (75.88 $kBqm^{-3}$). The progressively increase of the mean population density (e.g., exposure) from low to high risk areas are strictly related to the location type (e.g., vulnerability): (i) in the low risk areas most of the census tracts (33) are described as residential areas (location type = 4) and sparse houses (location type = 1, 43 census tracts); (ii) in the medium risk areas most of the census tracts are considered as residential areas (location type 4, 22 census tracts); (iii) all census tracts falling in the high risk areas are described as residential areas (location type = 4) with the highest population density. In fact, the population density increases accordingly from low to high risk areas.

4. Discussion

4.1 Interpretation of predictors in the RF model

The RF model demonstrates that all of the selected predictors influence Rn concentrations and movement in the subsoil. This result is consistent with the dependence of Rn from the geochemical and structural characteristics of the study area mainly linked to the generation and transport of Rn in the geological environment (i.e., from deep source toward the subsoil) (Benà et al., 2022). In fact, it is not surprising that the variable's importance shows clearly that GRP is primarily affected by TGDR (35%, Fig. 2) which represents the BRS contribution (e.g., the radionuclide content ^{238}U and ^{232}Th) of the main outcropping rocks (i.e., gneiss, granite, phyllite) (Tchorz-Trzeciakiewicz et al., 2021; Giustini et al., 2019, 2022). Because the survey of ambient gamma dose rate was conducted at the ground level, the correlation of TGDR with soil gas radon concentrations is likely to be stronger than with atmospheric concentrations. In the literature, Bossew et al., 2017; Cinelli et al., 2019; Melintescu et al., 2018; Sainz Fernández et al., 2017 reported a positive correlation between TGDR and GRP. The BRS contribution to the Rn amount in soil gas generates a relatively high spatial variability of Rn concentration in the soil gas, reflecting the homogeneous characteristics of the soil/rock environment at local scale (BRS). However, Rn spatial variability can increase (also at local scale) near fault zones (TER), especially in seismic areas characterised by active faults. In these areas, Rn migration from deeper sources can be increased by intense fracturing and the presence of carrier gases (mainly CO_2) that may play a dominant role for advective transport and redistribution of trace gases at surface (Wilkening, 1980; Ciotoli et al., 2007, 2014; Prasetio et al., 2023, and reference therein). This is observable in the study area along the Pusteria fault system, where radon concentrations in soil gas have a positive correlation with CO_2 concentrations (importance of about 30%, Fig. 2), suggesting a possible advective up flow caused by pressure gradients. In this faulted area, radon anomalies at surface could also be associated with elevated concentrations of radionuclide concentrations (i.e., Ra and U) in small soil particles transported by CO_2 gas molecules (Etiope & Lombardi, 1995). Furthermore, the presence of dissolved CO_2 in groundwater may promote radium dissolution and thus transport in solution (Giraults et al., 2014).

The high importance (about 15%, Fig. 2) of the fault density (interpreted as fault secondary permeability) confirms the effect of the Pusteria fault system on the Rn migration (as well as of other gases); this predictor is strictly related to the TER component (Benà et al., 2022). Indeed, damage zones related to high fracturing zones (fault areas) often exhibit a high permeability compared to the surrounding rocks and may facilitate the fluids advective transport for SGRC, thus potentially increasing radon release towards the surface and, as a consequence, Rn availability to enter buildings (IRC) (Ciotoli et al., 2007, 2014, 2016; Seminsky et al., 2014; Chen et al., 2018; Banrion et al., 2022; Zhou et al., 2023).

Similar importance of the other predictors (i.e., Tn, Slope, Aspect and Permeability) ranging from 4 to 8% can be explained by shallower processes affecting Rn movement in the soil layer, and at the soil-atmosphere interface (SRE) (Fig. 2). In the shallow environment the influence of meteorological conditions can be a complex issue, and the literature results are controversial. Air temperature and pressure on soil radon concentrations is small in comparing with total seasonal variability of this gas, and in any case the influence of these two variables is further lowered by conducting soil gas measurement campaigns during periods of stable and good weather conditions (Ciotoli et al., 2014; Beaubien et al., 2013, 2008).

The principal drivers governing diurnal and seasonal changes of radon concentration in the soil are the water-saturation and moisture-retention in the soil pore (i.e., rainfall) (King and Minissale, 1994). These two parameters directly decrease soil permeability thus preventing radon gas diffusion in the shallow soil layers (Nazaroff, 1992; Alonso et al., 2019; Beltran-Torres, 2023). High soil permeability allows ^{220}Rn to be detected at surface despite its short decay time (56 seconds).

In addition, the slope can be used as a proxy of soil moisture and meteorological conditions in absence of any other meteorological variables. High slopes also constitute zones characterized by increased soil permeability because they do not promote the retention of water and moisture in the soil pores. On the contrary, flat zones are characterized by low soil permeability because they favour the accumulation of water and moisture in the soil pore. At this regard, the SHAP diagram shows that high values of Tn, slope and permeability are positively correlated with high GRP (Fig. 3). The soil permeability may be linked to the ability of radon to migrate and escape towards the Earth surface. In fact, where permeability is high radon escapes more easily. Permeability is also linked to the fault density representing the secondary permeability.

All these predictors, except for the aspect, have an impact on the GRP values prediction for positive values and show an increasing trend up to the expected average radon value (see PDPs, in Fig. S5 in supplementary materials). On the contrary, low values of the GRP are correlated with high values of the aspect ratio (i.e., inverse correlation). The aspect identifies the compass direction that the downhill slope faces for each location; therefore, radon accumulation is easier in flat areas.

The model confirmed the correlations between geology and GRP and also provided insight into the utility and significance of other predictors that reflect the physical, chemical, and hydraulic properties of soil, as well as climatic predictors. On the basis of these results, further work should also consider meteorological parameters, such as soil temperature and humidity, rainfall, etc. This is especially fundamental to capture seasonal variability in models that uses IRC as response variable.

4.2 Map of the Collective Risk Areas (CRAs)

The GRP map obtained by RF regression represents radon hazard due to geological features of a specific region. It is strictly related to Rn gas directly measured in the soil and to all geological predictors (e.g., TGDR, CO_2, fault density, etc.) that significantly influence its concentration in the shallow environment, and potentially affect its movement towards homes. GRP maps, representing the most significant spatial predictor of IRC, are useful tools to evaluate the Rn risk (Bossew, 2015; Bossew et al., 2020).

As already mentioned, the European regulations aims to identify RPAs and implement mitigation plans in order to limiting radon exposure and thus reducing the risk of lung cancer to population. In an unbuilt and inhabited area, the presence of high Rn values represents only a high hazard (i.e., GRP), but not a risk. This concept is highly known and applied in the case of other natural phenomena such as in seismic microzonation studies. European legislation aims to reduce the detriment from Rn exposure (i.e., the number of lung cancer deaths) and as a consequence, reduce the collective exposure. In Figure 6, we show how GRP is a key factor in recognising of collective risk areas (CRAs).

In this paper, for the first time, we introduce the concept and define the Collective Risk Areas (CRAs) by applying the risk definition (section 2.4) consisting of three basic factors: i) the hazard, e.g., the Geogenic Radon Potential (GRP), ii) the vulnerability, e.g., the type of location, and iii) the exposure, e.g., the population.

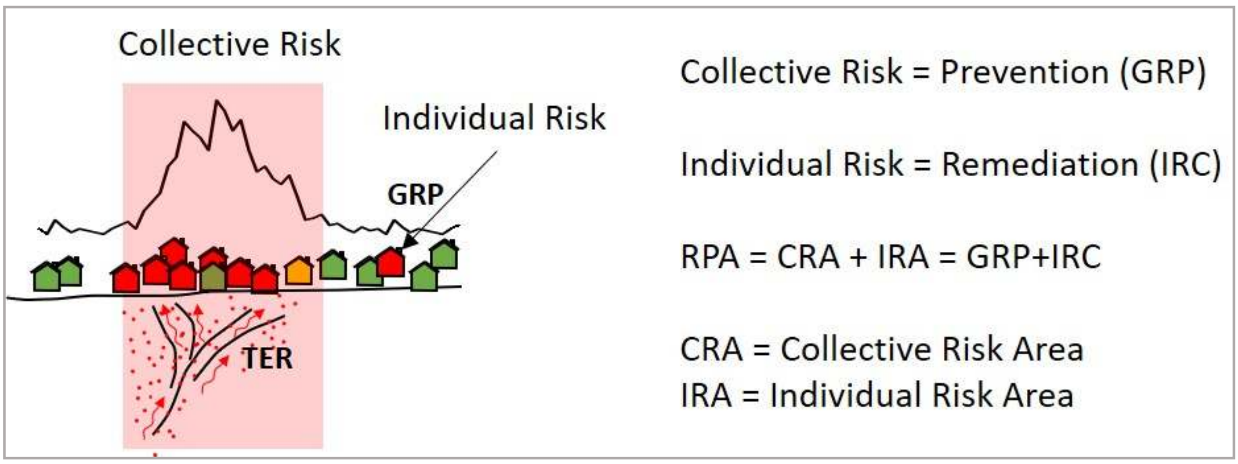


Figure 6. Summary sketch of the Collective Risk Areas concept.

Mapping the GRP is clearly the first fundamental step in defining the Rn hazard, a characteristic which cannot be mitigated. For this reason, it is important to map it as accurately as possible (i.e., by consider multiple geological variables and applying robust mapping techniques).
As reported in Benà et al., 2022, Rn values exceeding the lithological background (50 kBq m^{-3}) are considered anomalous and linked to the wide fracturing zone of the Pusteria fault system that represents Rn enhanced by tectonics (TER). However, in Benà et al.,2022, this quantity is not discussed in terms of GRP and thus it does not include the other important geological factors, such as gas permeability and deep circulation indicators (e.g., Rn in groundwater), as well as the shallow effects governed by the morphological parameters (e.g., DTM).
The identification of a threshold value of GRP is not significant to delineate CRAs, since the indoor radon risk exists even for "very low" concentrations of radon in the soil and, consequently, for very low GRP values. In fact, radon measured in the soil (GRP) is generally three order of magnitude higher than indoor radon. It is clear that every area can be affected by a potential indoor risk and all the dwellings are considered vulnerable.
However, GRP plays a key role in defining the CRAs that mainly occur along the Pusteria fault system where Rn degassing is enhanced by the intense fracturing and the GRP values are high. This is consistent with the fact that all the GRP values contribute to the risk. Therefore, the CRAs map highlights those areas with low, medium and high collective risk and, as a consequence, here the IRC values may be high for the residential areas.

5 Conclusions

The mapping and analysis of GRP (e.g., Rn hazard), obtained by using ML approach, is a fundamental tool for the delineation of CRAs according to a new, more geological, interpretation of the RPAs with respect of that reported in the BSS directive (2013/59/EURATOM).
We used the risk formula to combine the GRP map with the location type characteristic of the census tracts (e.g., the vulnerability factor) and the population density (e.g., the exposure factor). According to a geological-based interpretation of the RPAs, we can recognise hazard-based RPAs (CRAs) and detriment-based RPAs (IRAs) as complementary concepts of territorial planning and remediation actions, respectively, and not in alternative.
In particular, the obtained results highlight the following conclusions:

- Machine learning model by using the random forest technique demonstrates as a robust and high-performance method to obtain a GRP map of the study area. In particular, the obtained GRP map uses seven predictors reflecting geology (BRS and TER), soil characteristics (groundwater circulation, permeability), and meteorological conditions (DTM derivatives).

The variable importance highlights the dominant impact of Rn source but still significant contributions of the other predictors.

- As GRP is considered the most important spatial predictor of IRC, it is clear that mapping this hazard factor well represents the total amount of radon that can potentially enter buildings.
- Since GRP (e.g., soil gas concentration) values are three order of magnitude higher than the IRC, there is no reason to define GRP threshold, as the indoor radon risk can exists even for "very low" concentrations of radon in the soil and, consequently, for very low GRP values. GRP qualitative classes can serve only as delineation of zones (in the same way used in the seismic micro zonation studies) in which different land use planning strategies and/or construction types, and remediation actions should be adopted
- The absence of an unambiguous guidelines to define Radon Priority Areas (RPAs) led to the geological-based conceptualization of a complementary approach of mapping both the CRAs (in terms of prevention), as well as IRAs (in terms of building remediation actions).

This study may help policy makers to implement constructive preventive measures in those areas where new buildings are planned, and to act in terms of remediation in the RPAs sensu stricto. Future studies may aim to define the effective individual risk by constructing statistical models that also consider IRC measurements and anthropogenic factors.

Authors contribution

Eleonora Benà: Conceptualisation, data curation, methodology, writing - original draft preparation, reviewing and editing. **Giancarlo Ciotoli:** conceptualisation, methodology, writing - reviewing and editing. **Eric Petermann:** methodology, software, reviewing and editing. **Peter Bossew:** conceptualisation, reviewing and editing. **Livio Ruggiero:** reviewing and editing. **Luca Verdi:** reviewing and editing. **Paul Huber:** water springs sampling. **Federico Mori:** software, Python code. **Claudio Mazzoli:** conceptualisation, reviewing and editing. **Raffaele Sassi:** conceptualisation, project administration, supervision, funding acquisition, reviewing and editing.

Acknowledges

This research is supported by the Project SID 2021 of the University of Padova (R.S.).