Applied Geochemistry 158 (2023) 105790

Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

Applying machine learning to model radon using topsoil geochemistry

M. Banríon [a,*], M. Cobelli [b], Q.G. Crowley [a]

[a] Geology Department, School of Natural Sciences, Trinity College, Dublin 2, Ireland
[b] School of Physics, AMBER and CRANN Institute, Trinity College, Dublin 2, Ireland

ARTICLE INFO

Editorial Handling by: G. Bird

Keywords:
Geogenic radon potential
Radon
Topsoil geochemistry
Machine learning
Principal component analysis

ABSTRACT

Radon is classified as a Class 1 carcinogen, being the leading cause of lung cancer in non-smokers. Understanding the prominent sources of radon helps to mitigate against the adverse effects of radon exposure. Considering soil-gas radon is the main contributor to indoor radon, it is possible that soil geochemistry can be used as a proxy for the soil radon emanation potential or geogenic radon classes for a particular location. This paper investigates the relationship between soil geochemistry and geogenic radon. A large area of 17,983 km^2 from the West, Midlands and East of Ireland was selected to represent a range of geology types and radon categories. A rigorous assessment is presented to investigate the relationship of geogenic radon and topsoil geochemistry; using univariate processes (i.e. r^2, Pearson r and heatmaps) and multivariate techniques (i.e. principle component analysis (PCA) and machine learning (ML) algorithms including Gaussian process regression, logistic regression and random forest). Here, PCA and ML techniques were used to test the utility of soil geochemistry to predict geogenic radon classes. Gaussian Process Regression yielded the highest accuracy (74%) and f1-score (0.74) of all models. The feature importance (i.e. highest ranking elements for predicting geogenic radon class) from the ML models outputs elements including [Y, Tl, Mn, Cr, Co, Be, Sc and Rb]. The PCA biplot demonstrates that these elements cluster in conjunction with higher geogenic radon categories. Multivariate data analysis reveals that certain elements important for predicting higher geogenic radon classes, also covary together within topsoil samples; here these are termed "radon-prone elements". Spatial covariance of radon-prone elements permits soil geochemistry to be used as a tool for understanding the distribution of geogenic radon. The methodology presented in this paper provides a comprehensive geo-statistical approach to investigate the relation between topsoil geochemistry and geogenic radon. This approach could be applied as a diagnostic tool to assist radon mitigation measures, hence adding value to legacy soil geochemistry datasets.

1. Introduction

Radon is a radioactive noble gas that occurs naturally and is released from soil, water and rocks. There are three main radon isotopes, ^{222}Rn (radon), ^{220}Rn (thoron) and ^{219}Rn (actinon) which have half-lives of 3.82 days, 55.6 s and 3.96 s, respectively. Uranium (^{238}U) decays to radon (^{222}Rn) as an intermediate daughter product before completely decaying to the stable lead isotope (^{206}Pb). ^{220}Rn and ^{219}Rn are products of the ^{232}Th and ^{235}U decay chains, respectively. Importantly, if ^{222}Rn or its daughter products (^{214}Po and ^{218}Po) are inhaled they can cause significant damage by emitting alpha radiation directly into lung cells; linking radon exposure to lung-cancer (Rodríguez-Martínez et al., 2021; Zagà et al., 2021).

The majority (80%) of radiation exposure received by the general public is from natural sources (United Nations Scientific Committee on the Effects of Atomic Radiation, 2011). Approximately 53% of natural radiation is caused by the inhalation of decay products from the uranium and thorium series, of which the isotope ^{222}Rn accounts for over 90% of these decay products (United Nations Scientific Committee on the Effects of Atomic Radiation, 2011; Zohuri, 2020). Radon (^{222}Rn) exposure is the leading cause of lung-cancer cases in non-smokers, with 3–15% of global lung-cancer cases attributed to radon (World Health Organization, 2009). It is estimated that up to 350 people are affected by Rn induced lung-cancer in Ireland annually (Elío et al., 2018; Environmental Protection Agency, 2022). The survival rate after 5 years since lung cancer diagnosis is between 12% and 25% (Lin et al., 2019;

Abbreviations: IRC, Indoor radon concentration; PCA, Principal component analysis; ML, Machine learning; SGRn, Soil-gas radon; GPR, Gaussian process regression; RF, Random Forest; LR, Logistic regression; GRP, Geogenic radon potential.

* Corresponding author. Museum Building, Geology Department, School of Natural Sciences, Trinity College, Dublin 2, Ireland
E-mail addresses: mbanrion@gmail.com (M. Banríon), mcobelli@tcd.ie (M. Cobelli), crowleyq@tcd.ie (Q.G. Crowley).

https://doi.org/10.1016/j.apgeochem.2023.105790
Received 13 September 2022; Received in revised form 3 September 2023; Accepted 5 September 2023
Available online 7 September 2023

Schabath and Cote, 2019).

Reduction of radon exposure is paramount to reducing radon-related lung-cancer cases. Internationally, several countries have implemented initiatives to reduce the harmful effects of radon exposure. Some initiatives include setting reference limits for acceptable levels of indoor radon concentration (IRC), and having prevention measures (i.e. radon barrier, sump, ventilation) in place for new buildings (Radiological Protection, 2019 (Ionising Radiation) Regulations, 2019). It is important to note that there is no safe limit of radon exposure (World Health Organization, 2009). However, countries often decide the 'safe' IRC level based on average background levels and the realistic ability to lower IRCs to a 'safe' level (Degu Belete and Alemu Anteneh, 2021). In Ireland, 200 Becquerels per meter cubed (Bq m^{-3}) is the reference level for IRC in houses (Environmental Protection Agency (Ireland), 2019).

Identifying radon-prone elements in topsoil may help distinguish radon-prone areas. Soil-gas radon (SGRn) is the prime contributor to IRC (World Health Organization, 2007), understanding the association of soil geochemistry with the natural variation of geogenic radon distribution could allow topsoil geochemistry to be incorporated into mapping radon. The latter could help overcome the lack of national soil-gas radon data and add value to legacy soil geochemical mapping programs. It is possible radon-prone elements covary spatially for several reasons including association with different lithologies, weathering and mobility of elements. As such, it may be possible to determine the radon potential of a region using reliable soil geochemical data.

The variation of natural terrestrial radiation is largely dependent on geology (Banríon et al., 2022; Hamideen et al., 2020; Tzortzis et al., 2003). For example, granite and black shales contain radon precursor elements (Pereira et al., 2010; Petersell et al., 2015), whereas lithologies such as limestone generally lack radon precursor elements being incorporated into the bedrock (Fu et al., 2017; Khan et al., 2021). As such, the spatial variation of radon coincides with lateral differences in bedrock geology (Giustini et al., 2019; Petersell et al., 2015). The spatial correlation of bedrock geology and radionuclides (i.e. ^{226}Ra, ^{232}Th, ^{40}K) in topsoils is reported in the literature (Faanu et al., 2011; Ribeiro et al., 2018). However, few studies investigate the association of a wider range of topsoil elements with geogenic radon.

The application of multivariate geostatistical data analysis and machine learning for geochemical and environmental modelling is becoming more established (Bossew et al., 2020; He et al., 2022; Huntingford et al., 2019; Zuo, 2017). Principal component analysis (PCA) and machine learning have been applied to investigating the distribution of potentially toxic metals in topsoils at regional (Xu et al., 2021) and local scales (Wu et al., 2022). PCA and compositional data analysis have also been used to research the distribution of rare earth elements in topsoils (Ambrosino et al., 2022), as well as map pollution source of sea sediments (Somma et al., 2021), and geological and geochemical mapping at local, regional and larger scales (Ballabio et al., 2019; Wang et al., 2021; Zheng et al., 2021). Processing topsoil geochemistry using machine learning have also been applied to monitoring soil organic carbon (Sakhaee et al., 2022) and investigating heavy metal distribution in soils (Wang et al., 2023). Although machine learning has been used for modelling geogenic/indoor radon distribution (Elío et al., 2023; 2019; Petermann et al., 2021; Rezaie et al., 2021), the present research explicitly investigates the topsoil geochemical signature of geogenic radon in a 17,983 km^2 area in Ireland.

The main purpose of this study is to test multivariate statistical methods (PCA, ML) for analysing the relation between topsoil geochemistry and geogenic radon at a regional scale. Considering the lack of comparable scientific literature on this subject, several ML models are tested. The analysis is performed on a topsoil dataset obtained from Tellus Geological Survey Ireland (GSI) and compared with corresponding geogenic radon categories derived from Elío et al. 2020. The results demonstrate a correlation between topsoil elements and geogenic radon. This indicates that the geochemical signature of a topsoil sample can provide insight into the geogenic radon available for an area. The geostatistical approach used in this research confirms the feasibility of using topsoil geochemistry for assessing geogenic radon risk at a regional scale (>10 km^2).

2. Methodology

In total, 4279 shallow topsoil samples were obtained from the 2017–2019 GSI Tellus survey G5 area of the North Midlands, Ireland; including samples from counties Galway, Mayo, Roscommon, Longford, Westmeath, Offaly, Meath, Kildare and Dublin (Geological Survey Ireland Tellus programme, 2020). The GSI Tellus programme collected the topsoil samples with a 4 km^2 sample density. The topsoil samples cover a range of Quaternary sediments including till derived from limestones, cherts, alluvium, Namurian sandstones and shales, blanket peat, Lower Palaeozoic and Devonian sandstones, Devonian and Carboniferous sandstones, gravels derived from limestones.

The bedrock geology includes Devonian Granites, Ordovician Sandstones, conglomerates and Silurian sandstones and siltstones in the west, with a range of Carboniferous Limestones that predominate throughout the study area. There are occurrences of Silurian siltstones, sandstones and shales in the north-east of the area, and the northern portion of the Leinster Granites are located in the south-east of the study area. A detailed description of all the bedrock lithologies in this study area can be viewed on the Geological Survey of Ireland (GSI) online map viewer (www.gsi.ie (https://dcenr.maps.arcgis.com/apps/webappviewer/index.html?id=ebaf90ff2d554522b438ff313b0c197a&scale=0)).

The geogenic radon potential (GRP) categories developed by Elío (2020) were derived from GSI Tellus airborne radiometric surveys; where equivalent soil-gas radon and soil permeability were used to model GRP with a 1 km^2 resolution (Elío et al. 2020). The soil permeability categories were estimated from the Groundwater Subsoil Permeability (GWSP) map of Ireland and the all-Ireland Quaternary map (Elío et al. 2020).The airborne geogenic radon potential (GRP) categories; High (H), Moderate-High (M-H), Moderate-Low (M-L) and Low (L), published by Elio et al. (2020) are used as dependent variables in the machine learning models (Fig. 1B). The four groups are assigned to two classes. The Low category is assigned to class 1 and the Moderate-Low, Moderate-High and High groups are assigned to class 2. The Low GRP category has significantly less probability (6.93%) of having indoor radon concentration above the 200 Bq m^{-3} reference level compared to the remaining GRP classes (average 16.98%) (Elío et al., 2020).

The data for the independent variable(s), of shallow topsoil geochemistry, were downloaded from the Geological Survey Ireland/data and maps/geochemistry website (Geological Survey Ireland Tellus programme, 2020). The dataset investigated is '6117xxA-6174xxA_-Shallow_Topsoil_Download_v1.0.xlsx' (Geological Survey Ireland Tellus programme, 2020). Each shallow topsoil (0.05–0.20 m depth) sample is composed of five subsamples; four of which are collected from each corner of a 20 m square and the fifth is collected from the centre of that square corresponding with the GPS location for that sample (Knights et al., 2020).

The GSI Tellus protocol for preparing samples prior to analysis includes fan oven drying at 30 °C, carefully breaking clumps, dry sieving, dry sieving to obtain < 2 mm soil fraction, pulverising to obtain 63 μm fraction using an agate ball mill (Young et al., 2016). Tellus topsoil geochemistry was analysed using ICPMS for multiple elements (Al, Ba, Ca, Cr, Cu, Fe, K, Li, Mg, Mn, Na, Ni, P, S, Sr, Ti, V, Zn, Zr, Ag, As, Be, Bi, Cd, Ce, Co, Cs, Ga, Ge, Hf, Hg, In, La, Mo, Nb, Pb, Rb, Sb, Sc, Se, Sn, Te, Th, Tl, U, W, Ta, Au, Pd, Pt, Re and Y) (Geological Survey Ireland Tellus programme, 2020). Several elements (Ta, Au, Pd Pt and Re) were

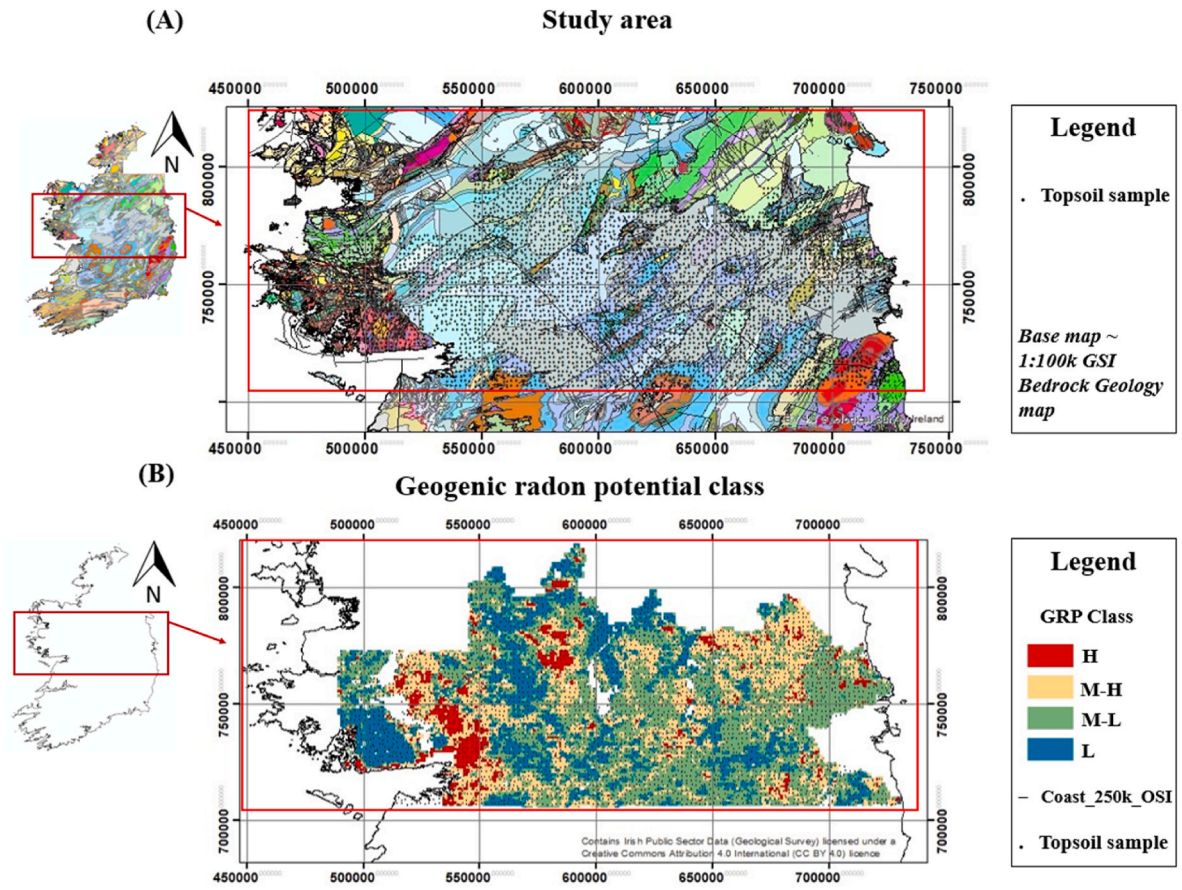

Fig. 1. 1.A Study area in relation to GSI bedrock geology map (1:100 k); The topsoil sample refers to shallow topsoil geochemical samples collected by the GSI Tellus topsoil geochemical survey (G5 north midlands study area), and 1.B Geogenic radon potential class for study area (derived from Fig. 5, Elío et al. 2020). Both maps were made using Arc map 10.8 and utilizing GSI basemaps and OSI Coast - National 250k Map of Ireland from Tailte Éireann.

omitted for further analysis due to more than 5% of values being below the detection limit. Only 4130 of 4279 topsoil samples from the Tellus north midlands dataset were used for analysis. The 149 samples not used were not within the boundaries of a radon potential classes reported by Elío et al. (2020), as such they could not be used for investigating the link with geogenic radon potential.

2.1. Data closure

Geochemical data is a subset of compositional data in which the concentrations of elements represent a proportion of the entire composition, leading to the constant sum (or data closure) issue (Aitchison, 1982). The centred log ratio transformation is used on data prior to multivariate analysis (i.e. PCA, Machine Learning models) in this study to accommodate the data closure restraint.

Clr-transformation projects the dataset into Euclidean space, allowing for the interpretation of geochemical results without the issues of spurious correlation of dependent variables (Aitchison, 1982; Grunsky and Caritat, 2019; Wang et al., 2021). The isometric log ratio (ilr) transformation is used on data prior to univariate data analysis (i.e. correlation heat maps, Pearson r and r^2). The ilr method transforms a composition from an Aitchison-simplex to D-1 (dimension minus 1) Euclidian vector space, retaining isometry (Juan José Egozcue et al., 2003). The ilr transformed dataset was back-transformed to obtain the original dimensionality of the starting dataset; this was done using 'R' software version 4.0.2 and the *'compositions'* package.

2.2. Principal component analysis

Principal Component Analysis (PCA) is an affine geometric transformation technique, which allows for increased interpretability of a multivariate dataset (Tolosana-Delgado and McKinley, 2016). PCA is a linear dimensionality reduction technique based on Euclidean methods, that transforms multiple variables into a smaller number of principal components (PC's), where principal component 1 (PC1) represents the direction of maximal variance of the dataset (Mueller et al., 2020). After PC1 the second principal component (PC2) represents the second largest variance of the data and is orthogonal to PC1 (Vermeesch, 2013). Assuming each PCA variable contributed to the variance of the dataset equally, each PC would explain (100/n)%, where n is the number of variables. Considering there are 47 elements used in the analysis, each PC would explain 2.1% (100/47) of the variance if they contributed equally.

The first several (n) PC's approximate the composition of elements that cause the highest degree of variation in the dataset. The last PCs explain the least variance, which may correspond to quasi-constant elements that could be utilized for studying immobile and mobile element migration if a geospatial/geochemical dataset is used (Tolosana--Delgado and McKinley, 2016). PCA aids in the interpretation of elemental variance within a compositional dataset and can be used in the process discovery phase of research (Grunsky, 2010; McKinley et al., 2018).

Each sample has a distinct score for each principal component, with the scores representing the samples in the transformed Euclidean space. To each principal component are associated loadings representing the

size and direction of the contribution from each of the original variables (i.e. elements) (Abdi and Williams, 2010).

The positioning and size of each loading are significant; loadings with similar directions, depict variables with similar variation patterns. In comparison, loadings with large differences between them have lower correlation (Dempster et al., 2013).

2.3. Machine learning

Many geological and environmental studies have utilized machine learning (ML) for classification processes, including random forest (RF) for determining mineral prospecting, determining aggregate provenance and radon mapping (Daviran et al., 2021; Dornan et al., 2020; Petermann et al., 2021). Gaussian process regression (GPR) has been used for its powerful interpolation techniques for natural hazard mapping, soil science and investigating heavy mineral pollution (Colkesen et al., 2016; Pham et al., 2021; Wang et al., 2022). Logistic regression (LR) has been applied to environmental pollution studies, mapping geogenic radon, and geochemical exploration (Aditya et al., 2018; Elío et al., 2017; Nathwani et al., 2022).

For the purposes of this study, four ML models are compared to investigate the suitability of using topsoil geochemistry for predicting geogenic radon potential categories. Results from RF, GPR and LR are compared with a baseline/control model; which assigns topsoil samples to geogenic radon classes based on equal probability. The ML algorithms (RF, LR, GPR) have been chosen in this study due to their robustness and ability to analyse large (> 1000 samples) multivariate datasets.

In simple terms, Random Forest (RF) can be used as a supervised classification algorithm that aims to predict which group an observation belongs to. A random forest classifier is an ensemble model that builds multiple decision trees on different subsets of the dataset and employs averaging to enhance predictive accuracy while mitigating overfitting (Breiman, 2001; Farhadi et al., 2022; Shang et al., 2019). A more elaborate explanation of the mathematical theory underpinning RF can be found in the literature (Biau and Scornet, 2016; Breiman, 2001; Schonlau and Zou, 2020). The 'RandomForestClassifier' function in the sklearn library (python version 3.6) was used to implement the random forest model.

Gaussian Process Regression (GPR), is a nonparametric classification ML algorithm based on Bayesian probability. The GPR model initially forms a prior distribution, then uses the training dataset to reallocate probabilities and form a posterior probability distribution. Thorough explanations of the theory underlying GPR can be found in the literature (Bernardo et al., 1998; Bousquet et al., 2011; Kanagawa et al., 2018). The 'GaussianProcessClassifier' function in the sklearn library (python version 3.6), with a squared exponential kernel was used to train and test the GRP model.

Logistic regression (LR) is a statistical method for modelling the relationship between a binary dependent variable and one or more independent variables by estimating probabilities using a logistic function. More details regarding the LR mathematical model are published in the literature (Hastie et al., 2009; Kirasich et al., 2018; Sperandei, 2014). Specifically, the sklearn library is used in python (version 3.6), utilizing the 'LogisticRegression' function.

Cross-validating data is important for determining the stability of model parameters. Spatial cross-validation aims to eliminate overfitting in the model due to spatial autocorrelation (Pohjankukka et al., 2017; Talebi et al., 2022). The Tellus North Midlands G5 dataset was spatially cross-validated using 5-fold k-means cluster spatial blocking. This was achieved by splitting the dataset into 5 separate spatial blocks, 4 blocks were used for training the model and the remaining block was used to test the model. To aid in balancing the model, equal amounts of samples from classes 1 and 2 were randomly extracted during the training and testing process. The 5-fold cross-validation was repeated 10 times for RF, GPR, LR and the control model.

3. Results and discussion

The primary aim of this case study is to investigate the applicability of topsoil geochemistry for classifying geogenic radon potential at a regional scale. Initially, PCA is deployed to determine the extent of variation within the dataset and identify clusters of elements. Several ML models are analysed to investigate if any of the ML models tested can fit the topsoil geochemistry to predict the geogenic radon class (GRP). Summary statistics (i.e. mean, min, max, quartiles 1–3, standard deviation, skewness and Kurtosis) as well as Pearson r, and r^2 results are included in the supplementary materials (*supp. matt.*).

3.1. Principal component analysis results

The PCA biplot (Fig. 2) shows the data-cloud colour coded to GRP class. It is important to note that the PCA biplot doesn't use information regarding the assigned GRP class when calculating PCA loadings. However, the data-cloud depicts low GRP scores concentrating in the positive PC1-PC2 biplot region (i.e. top right), and samples collected from regions with higher GRP class are represented by lower PC1-PC2 scores on the biplot (i.e. bottom left) (Fig. 2). As such, the data-cloud shows directionality, indicating GRP class decreases as PC1 and PC2 scores increase.

PC1 accounts for ~ 60% of the total variance. The positive PC1 loadings (> *0.01*) relate to the elements S, Na, Se, Sr and Ca (Fig. 2) (Table 1 supp. matt.). The lowest PC1 loadings *(0.17 to -0.18)* are associated with the elements Be, V, Cr, Al, Co, La, Ce, Ga, Li, Tl, Ni, Rb, Y, Fe, Sc and Th (Fig. 1) (*supp. matt.*).

PC2 accounts for 9.5% of topsoil geochemistry variance. The more positive PC2 loadings *(0.18 - 0.36)* are associated with Se, Hg, Sb, Sr, S, Pb, Ca, Cd, Sn, Bi and Ge (*supp. matt.)*. The most negative PC2 loadings *(−0.06 to -0.14*) are linked to Rb, Li, Cs, Th, Ce, Al, K, Ga, La and Cr. Low GRP class mainly clusters with positive PC1 and PC2 scores and higher GRP classes cluster with negative PC1 and PC2 scores (Fig. 2).

The PCA results indicate that radon-prone elements have geospatial restraint, suggesting topsoil geochemistry can be used as a tool to understand the geogenic radon potential in an area. Explicitly, if a topsoil sample contains elevated concentrations of elements with negative PC1 and PC2 loadings, then the topsoil may correlate with higher geogenic radon potential. Conversely, if topsoil samples contain elevated concentrations of elements with positive PC1– PC2 loadings may indicate the soil has lower geogenic radon potential.

The PC1 score map was superimposed onto the 1:100 k GSI bedrock map using ArcMap 10.8.1. The full legend for the different bedrock geologies can be found on GSI's online map viewer (www.gsi.ie accessed 2022). High PC1 scores occur above various limestones including the Waulsortian limestones, Croghan limestone Formation, Ballymore limestone Formation, Cong Canal Formation and the majority of the Burren Formation (Fig. 3a). However, some of the highest PC1 scores also occur above the Devonian granites in Co. Galway and Co. Dublin (Fig. 3b), where a higher concentration of uranium and radon-prone elements would be expected (i.e if soil formed from the underlying bedrock and insignificant chemical and physical weathering occurred). The lack of radon-prone elements in soils above the Devonian granites indicates that (a) the soil did not form from the underlying bedrock or, more plausibly, (b) sedimentary processes including weathering and erosion has occurred that has mobilised radon-prone minerals/elements to be deposited above different bedrock geologies (Moles and Moles, 2002). Mobility of elements in soils due to glacial activity is shown for Northern Ireland, north of the study area (Dempster et al., 2013).

Low PC1 loadings (relating to elements including Tl, Y, Mn, Cr, Ni, Co, Al, V, Sc, Be and Zr) are observed, in higher concentrations, above Namurian (undifferentiated) sandstones, siltstones and shales, as well as in soils above siltstones and sandstones in the east of the study area including the Balrickard Formation (Namurian sandstone and shale), Walshestown Formation (Namurian shale, sandstone limestone),

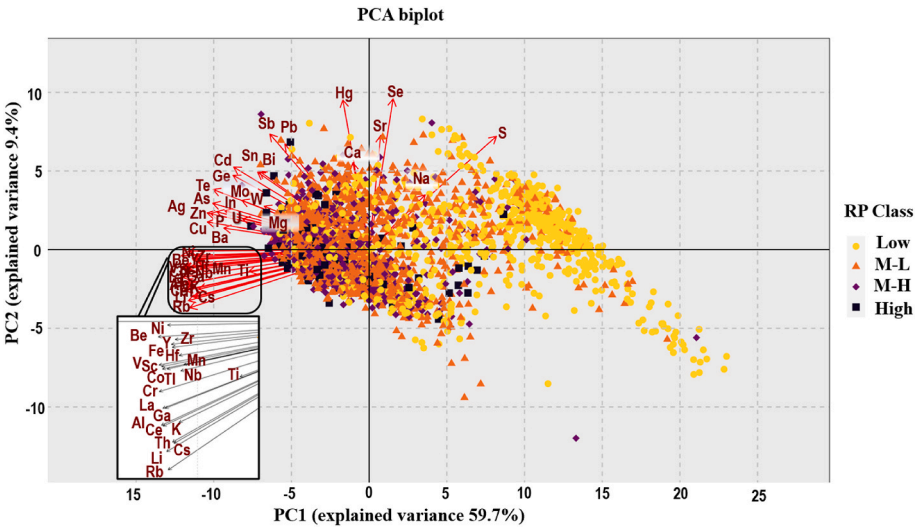

Fig. 2. PC1 vs PC2 biplot for Tellus North Midlands topsoil geochemistry data. PCA score results are grouped to Geogenic Radon Potential class (i.e. topsoil sample grouped to Low (L), Moderate-Low (ML), Moderate-High (MH), and High (H) GRP class derived from Elío et al., 2020b); 'R' software was used; specifically, the 'prcomp' library with the centre and scale argument set to T = true.

Denhamstown Formation (Silurian greywacke sandstone and siltstone) and Clatterstown Formation (Silurian siltstone and sandstone). Low PC1 scores also occur above the limestones and shales located in the north east of the study area (i.e. Mornington Formation, Crufty Formation, the Lucan Formation, the Loughshinny Formation, Clontail Formation (calcareous red-mica greywacke).

The highest PC2 scores occur above the Shannapheasteen and Galway Granite and the Leinster Granite (Fig. 3b). Higher PC2 scores also occur scattered in the middle and east of the study area above various limestones. The most negative PC2 loadings (<−0.06) are linked to Rb, Li, Cs, Th, Ce, Al, K, Ga, La and Cr (*supp. matt.)*. Low PC2 scores are found above the undifferentiated Visean limestones, dark fine-grained limestone and shale, as well as in proximity to black mudstone, siltstone and greywacke (Fig. 3b)

The inverse trend of GRP class decreasing as PC1-PC2 values increase can also be observed by comparing the GRP class map (Fig. 1b) with the PC1-2 score maps (Fig. 3). The lowest PC1 scores (Fig. 3a) correspond with areas that show moderate-high geogenic radon potential (GRP) in the North Midlands (Fig. 1b). Areas exhibiting a low GRP class (Fig. 1b) correspond to areas with high PC1-PC2 scores (Fig. 3).

A canonical heatmap is shown for the shallow topsoil elements measured in the Tellus North Midlands G5 study area, using ilr-transformed data (Fig. 4). The covariance of elements that cluster together on the PC1-PC2 biplot (Fig. 2) is reinforced by showing relative positive correlation on the canonical heat map (Fig. 3). Low PC1-PC2 loadings including La, Y, Zr, Hf, Tl, Sc, Be and Mn, as well as Cr, Al, Ga, V, Ni, Co, Li, and Rb show relatively positive correlation on the heatmap (Fig. 4). In comparison, high PC1-PC2 loadings including S, Na, Se, Sr, Hg and Ca are positively correlated with each other; and mostly exhibit negative correlation with the low PC1-PC2 loadings (Figs. 2 and 4).

3.2. Machine learning model performance

The accuracy, precision, recall, f1-score and ROC-AUC results for RF, LR, GPR and the control machine learning model are reported in Table 1. The GPR model is the most accurate (74% (f1-score 0.74)), followed by LR (74% (f1-score 0.73)) and RF (73% (f1-score 0.73)), although the accuracies of these models are within uncertainty of one another (Table 1). The GPR, LR and RF models are significantly more accurate compared to the control model accuracy (50.2%).

The performances of the ML models prove a link between topsoil geochemistry and geogenic radon class, as they are capable of identifying patterns in the training set to predict the radon class on unseen data with an average f-1 score of ~0.73.

The average true positive, true negative, false positive and false negative results from each model were used to compute a simplified confusion matrix (Table 2). Type I errors (false positives) represent samples that were mistakenly classified in a high GRP class. Type II errors (false negatives) occur when a model misclassifies a high-risk GRP sample as being in the low-risk GRP class.

The type I error is lowest in the LR model (22%), followed by GPR (23%) and RF (32%). The control model has the highest rate of type I and type II errors (49.9%). The RF model has the lowest rate (23%) of type II errors followed by GPR (29%) and LR (29%). Overall, the percentage of correctly classified samples is consistently higher in the GPR model (73.9%), LR (73.7%) and RF (72.8%) algorithms compared to the control model (50.1%) (Table 2). Models chosen for low type I errors would minimize the amount of resources spent on targeting low geogenic radon areas which are mistakenly classified as high. Whereas a model with the lowest type II error, would be the most effective for targeting areas associated with a higher geogenic radon potential probability.

3.3. Feature importance

The ML models feature importance indicates which elements are important for distinguishing low and higher GRP classes. The feature importance is calculated for each of the four models using sklearn libraries in python (Fig. 5). The LR, RF and GPR models show a pattern of elements that rank among the most important features in at least two

Table 1
Accuracy, Precision, Recall, f1-score and ROC-AUC results for Random Forest, Logistic Regression, Gaussian Process Regression and the control machine learning model. The type I and type II errors are reported after ± in the relevant cells.

Model type	Accuracy	Precision	Recall	F1 -score	ROC-AUC
Random Forest	0.73 ± 0.04	0.75 ± 0.04	0.73 ± 0.04	0.73 ± 0.04	0.81 ± 0.05
Logistic Regression	0.74 ± 0.04	0.76 ± 0.04	0.74 ± 0.04	0.73 ± 0.04	0.80 ± 0.04
Gaussian Process Regression	0.74 ± 0.04	0.76 ± 0.04	0.74 ± 0.04	0.74 ± 0.04	0.81 ± 0.04
Control	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.03	0.50 ± 0.00

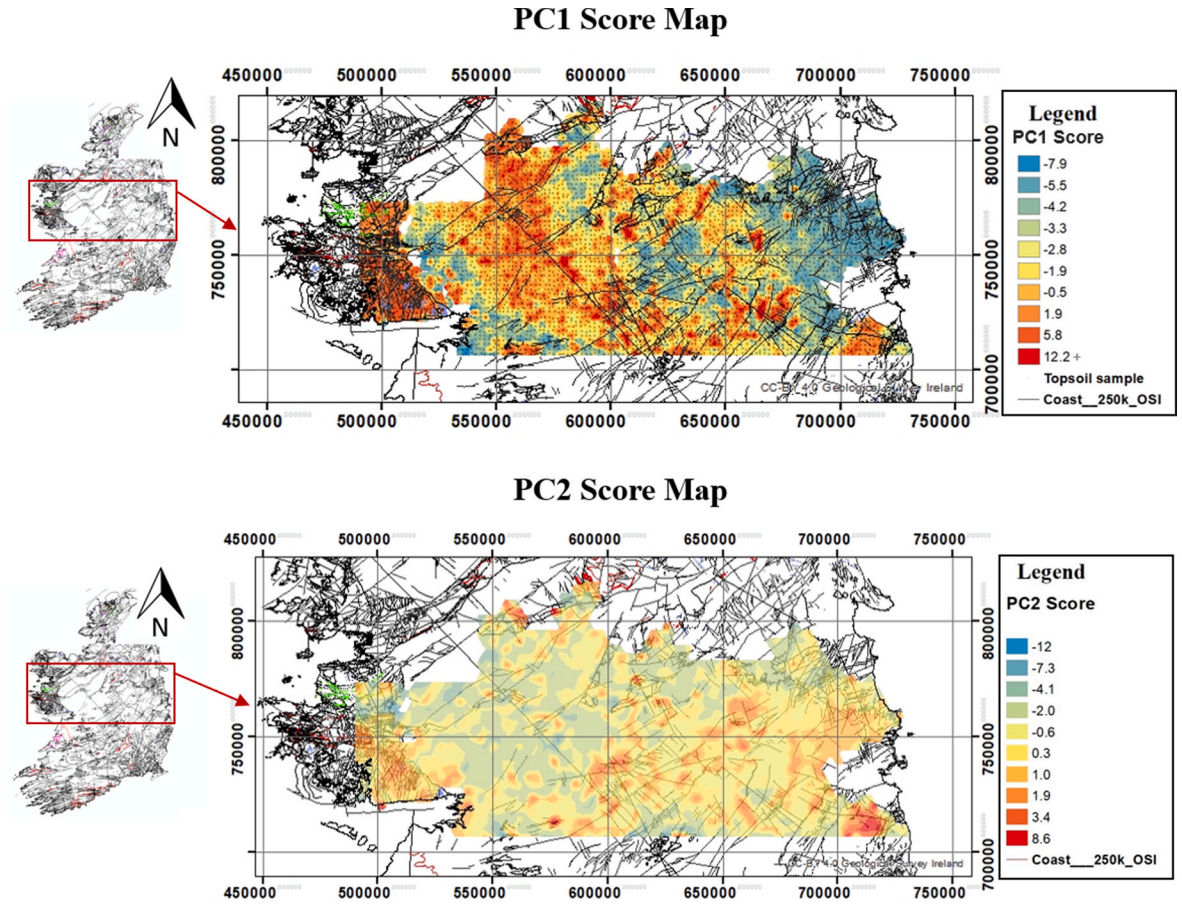

Fig. 3. (a - top) PC1 score map and (b - bottom) PC2 score map of topsoil geochemical data, superimposed onto OSI 250 k coast of Ireland and the bedrock geology linework (1:100k) GSI (www.gsi.ie). The PC score maps were made in ARCGIS using the Inverse-distance weighted geostatistical function (grid output size = 50 m, major and minor semiaxis = 2000, smoothing factor = 0.2). Irish Transverse Mercator coordinate system used.

models (Fig. 5); these elements include Y, Tl, Mn, Sc, Ba, Sb, Cr, U, Li and Rb, some elements are ranked as important features in individual models. For instance, random forest also ranks Pb, U, Zn, Sr, P and Zr among the important features, gaussian process regression additionally ranks Co, Cu, Be highly, whereas logistic regression includes Al, Rb, Li among its most important features (Fig. 5).

Logistic Regression ranks elements Y, Al, Rb, Mn, Sc, Li, Cr and K, with high feature importance (Fig. 5) which group together on the PCA biplot with the negative PC1-PC2 loadings (Fig. 2). The majority of the next highest-ranking features also occur in the negative PC1-PC2 biplot cluster (i.e. Ga, Cs, Nb, Ce and Tl). Gaussian Process Regression also highly ranks elements with negative PC1-PC2 loadings i.e. Y, Tl, Mn, Sc, Co, Be and Cr. Other highly rankly features include elements Cu, Ba, and U which have positive PC2 and negative PC1 loadings. Random Forest highly ranks elements from various PC1-PC2 biplot clusters, namely negative PC1-PC2 loadings (Tl, Y, Mn and Sc) and positive PC2 – negative PC1 cluster (Ba, P, U, Zn, Sb and Pb). Elements with low PC1-PC2 loadings are consistently ranked highly on the feature importance for several ML models. However, there are occurrences of more positive PC1 and PC2 loadings having higher feature importance in the ML models, e.g. Pb, Sb, Ca, Sr, Hg and Se. The latter results indicate that the models use the geochemical signature from areas classified as low GRP when performing a prediction.

Low PC1 and PC2 loadings are more frequently ranked among high feature importance, reinforcing the results shown by the PCA directionality of low PC1 - 2 scores being associated with higher GRP class (Fig. 2).

3.4. Geological meaning of affinities

The affinities between shallow topsoil geochemistry and geogenic radon potential across the study area are influenced by different geological processes. At any given location, the geochemistry of each topsoil sample is an integration of previous processes involved in rock formation, weathering processes, glacial processes and soil formation processes, and possibly anthropogenic processes (Dosseto et al., 2011; Heimsath et al., 1997; Johnson and Watson-Stegner, 1987; Shepherd, 1989). It is important to highlight that affinities between elements can change if chemical and physical parameters change. Below is an elaboration of some possible suggestions of topsoil affinities.

The affinity between Co with Ni (0.85), Cr (0.68) and V (0.62) in high geogenic regions (Low PC1 areas – Fig. 4) could be related to sulphide, ferromagnesian and oxide minerals. For example, Ni and Co can covary together in siegenite ($(Ni,Co)_3S_4$), pyrrhotite (Fe(1-x)S) or pentilandite ($(Fe,Ni)_9S_8$) as stoichiometric lattice substitutions in hydrothermal sulphide deposits; or Ni, Co, along with Cr and V can be adsorbed onto fine-grained inorganic particles (Jansson and Liu, 2020; Kovalev et al., 2014; Loring, 1976). As such, it is possible for covariances between Ni, Co, Cr and V formed as a result of different geological processes. However, considering the geology underlying Co, Ni, V and Cr covariances is dominated by sedimentary lithologies, it is likely the [Co, Ni, V, Cr] affinity relates to those sedimentary processes, namely adsorption onto clay-rich or shale layers within limestones. However, the Navan Zn–Pb deposit in Co. Meath, that formed as a result of hydrothermal sulphides interacting with carbonate host lithologies, could contribute to the Co, Ni and Fe affinity (Ashton et al., 1980; Johnston et al., 2013; Marks et al., 2017; Yesares et al., 2019). The latter highlights that heterogenous

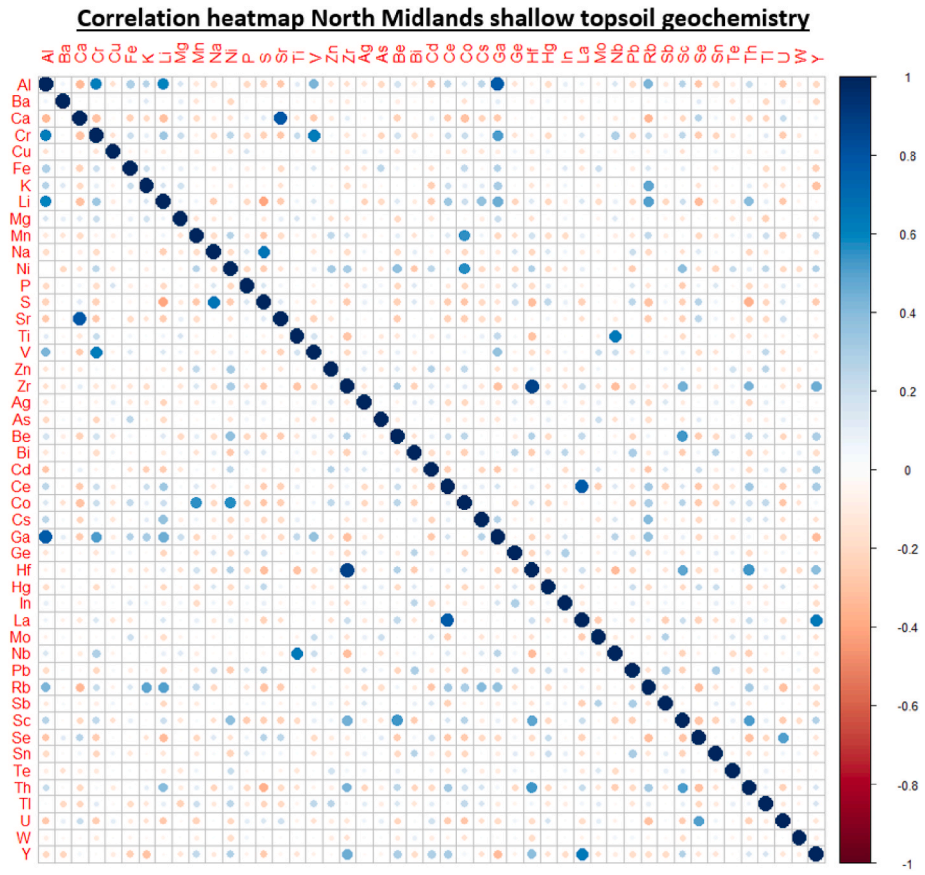

Fig. 4. Shallow topsoil geochemistry ilr-transformed canonical heatmap. Blue and positive numbers up to +1 depict positive correlations and red and negative values down to −1 depict negative correlations between variables.

Table 2
Confusion matrix results for Random Forest, Gaussian Process Regression, Logistic Regression and a Control model applied to classifying topsoil samples into Geogenic Radon Potential classes. The Asterix '*Type I' refers to the percentage rate of false positives, while '**Type II' refers to the percentage rate of false negatives.

Model	Confusion matrix	Predicted GRP class 1	Predicted GRP class 2
Random Forest	GRP class 1	13330 (68%)	6170 (32%) *Type I
	GRP class 2	4424 (23%) **Type II	15076 (77%)
	Percentage of samples correctly classified: 72.8%		
Gaussian Process Regression	GRP class 1	13198 (77%)	3886 (23%) *
	GRP class 2	6302 (29%) **	15614 (71%)
	Percentage of samples correctly classified: 73.9%		
Logistic Regression	GRP class 1	12930 (78%)	3682 (22%) *
	GRP class 2	6570 (29%) **	15818 (71%)
	Percentage of samples correctly classified: 73.7%		
Control model	GRP class 1	13742 (50.1%)	13670 (49.9%) *
	GRP class 2	13498 (49.9%) **	13570 (50.1%)
	Percentage of samples correctly classified: 50.1%		

geological processes across the study area could be contributing to the affinity between elements.

It is possible that physical weathering and leaching of the granites could be the potential source of Li, Cs and Rb covariation observed in the high geogenic radon regions (i.e. above the bedrock Formations in the east of the study area and in tills derived from limestones and tills derived from Silurian and Namurian sandstones and shales). It's possible these [Li, Rb, Cs] element affinities within soils correlating to higher geogenic radon potential areas are due to clay minerals illite and chlorite, which are reported to incorporate traces of radon precursor elements (i.e. uranium and radium) by adsorption in phosphorous-rich environments (Benedicto et al., 2014; Kim et al., 2017; Liao et al., 2020; Mei et al., 2022). Confirming the geological provenance of element affinities within the topsoil would require specific mineralogical analysis, petrographic microscopy and/or provenance studies on bedrock and soil samples to validate geological processes and controls responsible for the geochemical affinities.

3.5. Limitations

There are some discrepancies between the feature importance from various models, likely connected with the different model architectures. To evaluate the feature importance on the same footing for all models, we applied the same method, specifically permutation importance, as implemented in the sklearn library (python version 3.6). Considering topsoils are partially derived from bedrock, the topsoil geochemical signatures that relate to radon-prone areas are associated with the geologies within the study region. It is important to note the limitation of applying the results of our compositional and multivariate statistical analysis to areas of distinctly different geologies.

4. Conclusion

Exposure to indoor radon gas is a major contributor to lung cancer globally (Gaskin et al., 2018; World Health Organization, 2009), and it is in interest of human wellbeing to provide rigorous research aimed at understanding natural radon distribution. Soil-gas is the main contributor to indoor radon, and considering there is no national soil-gas radon dataset available, our study sets out to investigate if topsoil geochemistry can be used to predict geogenic radon class. Multiple methods of

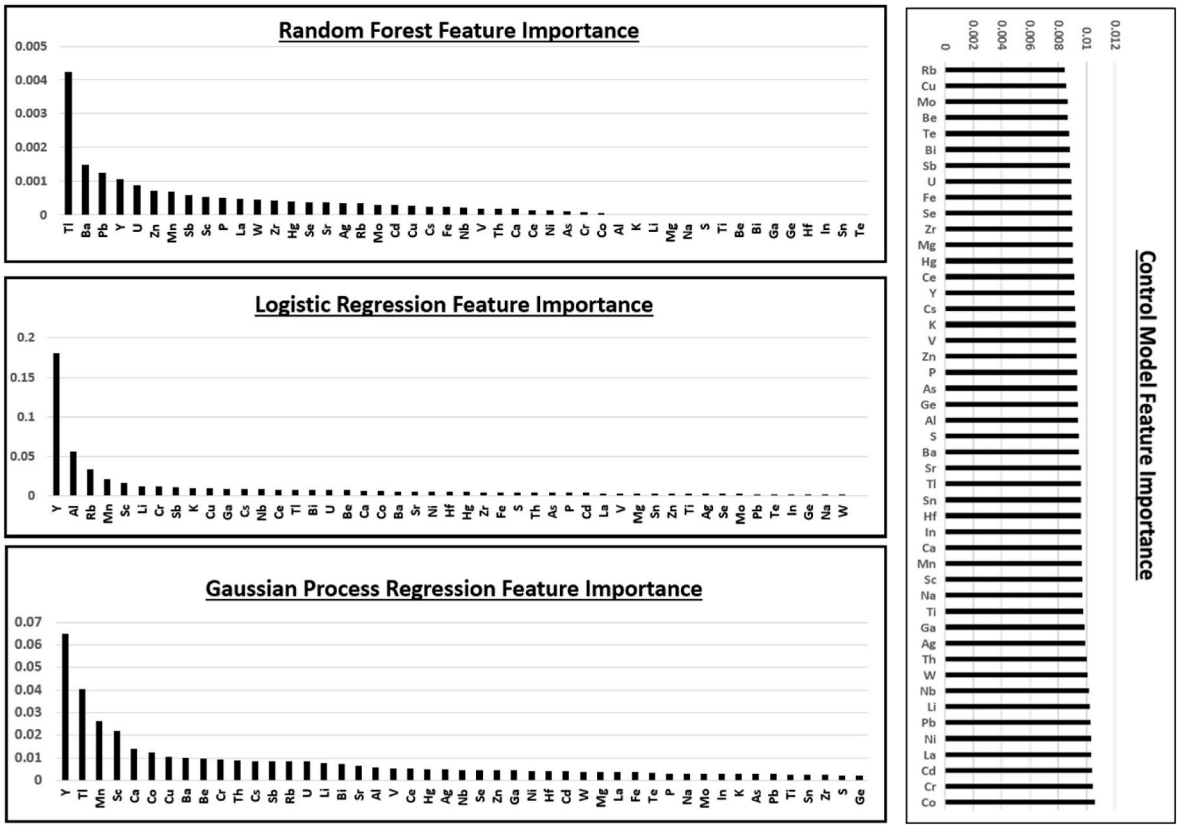

Fig. 5. Feature importance (i.e. frequency in which a variable ranks highly for classification purposes) for each of the four machine learning models, random forest, gaussian process regression, the control model and logistic regression.

analysing data are used to investigate the extent topsoil geochemistry can be used to understand geogenic radon distribution. The ML models tested can use topsoil geochemistry to predict radon class with ~74% accuracy (f-1 score ~ 0.73). Elements that are relatively more correlated with geogenic radon consistently group together when comparing multivariate and univariate results for regional scale data (Tellus 'North Midlands' Ireland). The feature importance from several machine learning algorithms designed to test the relation of topsoil geochemistry to geogenic radon, reinforce the hypotheses that topsoil geochemistry can be indicative of geogenic radon. In particular, Y, Tl, Mn, Cr, Co, Al and Sc are elements commonly associated with elevated geogenic radon. Additionally, Pb, Sb, Ca and Se are among elements frequently negatively correlated with geogenic radon. The results demonstrate the value that topsoil geochemical surveys can have in determining radon prone areas. The analytical approach used in this paper contributes a thorough and novel method for extracting useful information from topsoil geochemical data. Machine learning results demonstrate a rigorous method for testing and interpreting large datasets applied to geogenic radon studies, although the methodologies are also applicable to other areas of research such as resource exploration (e.g. topsoil geochemistry applied to predicting underlying lithologies opposed to geogenic radon). Overall, we contribute a better understanding of the correlation between topsoil geochemistry and geogenic radon, and provide a robust methodological approach for interpreting compositional data. As such the methods presented here could be applied as a diagnostic tool to assist radon mitigation measures, adding value to legacy soil geochemistry datasets.

Role of funding source

The partial SUSI grant provided to Méabh Banríon's Ph.D. has no involvement or role in the study design, the collection, analysis and interpretation of data; in writing of the report; and in the decision to submit the article for publication. The financial scholarship provided to Matteo Cobelli from IRC has no involvement or role in any aspect of this research project, including research design, sampling, analysis, report writing or decision to submit the article for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research contains Irish Public Sector Data (Geological Survey Ireland) licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Explicitly, the Tellus midlands topsoil geochemistry dataset was used (file name: 58xxxxA-65xxxxA_Shallow_Topsoil_Download_v1.1, https://www.gsi.ie/en-ie/data-and-maps/Pages/Geochemistry.aspx).

Méabh H. Banríon would like to thank Student Universal Support Ireland (SUSI) for granting partial financial assistance. Matteo Cobelli thanks the Irish Research Council for financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeochem.2023.105790.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comput. Stat. 2, 433–459. https://doi.org/10.1002/wics.101.

Aditya, C., Chandana, R.D., Nayana, D., Gandhi, P., Vidyav, astu, 2018. Detection and prediction of air pollution using machine learning models. Int. J. Eng. Trends Technol. 59, 204–207. https://doi.org/10.14445/22315381/IJETT-V59P238.

Aitchison, J., 1982. The statistical analysis of compositional data. J. Roy. Stat. Soc. B 44, 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Ambrosino, M., Albanese, S., De Vivo, B., Guagliardi, I., Guarino, A., Lima, A., Cicchella, D., 2022. Identification of Rare Earth Elements (REEs) distribution patterns in the soils of Campania region (Italy) using compositional and multivariate data analysis. J. Geochem. Explor. 243, 107112 https://doi.org/10.1016/j.gexplo.2022.107112.

Ashton, J., Black, A., Geraghty, J., Holdstock, M., Hyland, E., Bowden, A., Earls, G., O'Connor, P., Pyne, J., 1980. The geological setting and metal distribution patterns of Zn-Pb-Fe mineralization in the Navan Boulder Conglomerate. The Irish minerals industry 1990, 171–210.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 355, 113912. https://doi.org/10.1016/j.geoderma.2019.113912.

Banríon, M.H., Elío, J., Crowley, Q.G., 2022. Using geogenic radon potential to assess radon priority area designation, a case study around Castleisland, Co. Kerry, Ireland. J. Environ. Radioact. 251–252, 106956 https://doi.org/10.1016/j.jenvrad.2022.106956.

Benedicto, A., Missana, T., Fernández, A.M., 2014. Interlayer collapse affects on cesium adsorption onto illite. Environ. Sci. Technol. 48, 4909–4915. https://doi.org/10.1021/es5003346.

Bernardo, J., Berger, J., Dawid, A., Smith, A., 1998. Regression and classification using Gaussian process priors. Bayesian stat. 6, 475.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Bossew, P., Cinelli, G., Ciotoli, G., Crowley, Q., Cort, M., Elío, J., Gruber, V., Petermann, E., Tollefsen, T., 2020. Development of a geogenic radon hazard index—concept, history, experiences. Int. J. Environ. Res. Publ. Health 17, 4134. https://doi.org/10.3390/ijerph17114134.

Bousquet, O., von Luxburg, U., Rätsch, G., 2011. Advanced lectures on machine learning: ML summer schools 2003, canberra, Australia. In: Revised Lectures. Springer Berlin Heidelberg. February 2-14, 2003, Tübingen, Germany, August 4-16, 2003.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Colkesen, I., Sahin, E.K., Kavzoglu, T., 2016. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. J. Afr. Earth Sci. 118, 53–64. https://doi.org/10.1016/j.jafrearsci.2016.02.019.

Daviran, M., Maghsoudi, A., Ghezelbash, R., Pradhan, B., 2021. A new strategy for spatial predictive mapping of mineral prospectivity: automated hyperparameter tuning of random forest approach. Comput. Geosci. 148, 104688 https://doi.org/10.1016/j.cageo.2021.104688.

Degu Belete, G., Alemu Anteneh, Y., 2021. General overview of radon studies in health hazard perspectives. J. Oncol. 2021, 6659795 https://doi.org/10.1155/2021/6659795.

Dempster, M., Dunlop, P., Scheib, A., Cooper, M., 2013. Principal component analysis of the geochemistry of soil developed on till in Northern Ireland. J. Maps 9, 373–389. https://doi.org/10.1080/17445647.2013.789414.

Dornan, T., O'Sullivan, G., O'Riain, N., Stueeken, E., Goodhue, R., 2020. The application of machine learning methods to aggregate geochemistry predicts quarry source location: an example from Ireland. Comput. Geosci. 140, 104495 https://doi.org/10.1016/j.cageo.2020.104495.

Dosseto, A., Buss, H., Suresh, P.O., 2011. The delicate balance between soil production and erosion, and its role on landscape evolution. Appl. Geochem. 26, S24–S27. https://doi.org/10.1016/j.apgeochem.2011.03.020.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300. https://doi.org/10.1023/A:1023818214614.

Elío, J., Cinelli, G., Bossew, P., Gutiérrez-Villanueva, J.L., Tollefsen, T., De Cort, M., Nogarotto, A., Braga, R., 2019. The first version of the pan-European indoor radon map. Nat. Hazards Earth Syst. Sci. 19, 2451–2464. https://doi.org/10.5194/nhess-19-2451-2019.

Elío, J., Crowley, Q., Scanlon, R., Hodgson, J., Long, S., 2017. Logistic regression model for detecting radon prone areas in Ireland. Sci. Total Environ. 599–600, 1317–1329. https://doi.org/10.1016/j.scitotenv.2017.05.071.

Elío, J., Crowley, Q., Scanlon, R., Hodgson, J., Long, S., Cooper, M., Gallagher, V., 2020. Application of airborne radiometric surveys for large-scale geogenic radon potential classification. radon 1. https://doi.org/10.35815/radon.v1.4358.

Elío, J., Crowley, Q., Scanlon, R., Hodgson, J., Zgaga, L., 2018. Estimation of residential radon exposure and definition of Radon Priority Areas based on expected lung cancer incidence. Environ. Int. 114, 69–76. https://doi.org/10.1016/j.envint.2018.02.025.

Elío, J., Petermann, E., Bossew, P., Janik, M., 2023. Machine learning in environmental radon science. Appl. Radiat. Isot. 194, 110684 https://doi.org/10.1016/j.apradiso.2023.110684.

Environmental Protection Agency, 2022. New EPA (Ireland) Radon maps show more homes and workplaces at risk from cancer-causing gas. https://www.epa.ie/news-releases.

Environmental Protection Agency, 2019. EPA (Ireland) Protocol for the Measurement of Radon in Homes and Workplaces.

Faanu, A., Darko, E., Ephraim, J., 2011. Determination of natural radioactivity and hazard in soil and rock samples in a mining area in Ghana. West African J. Appl. Ecol. 19.

Farhadi, S., Afzal, P., Boveiri Konari, M., Daneshvar Saein, L., Sadeghi, B., 2022. Combination of machine learning algorithms with concentration-area fractal method for soil geochemical anomaly detection in sediment-hosted irankuh Pb-Zn deposit, Central Iran. Minerals 12. https://doi.org/10.3390/min12060689.

Fu, C.-C., Yang, T.F., Chen, C.-H., Lee, L.-C., Wu, Y.-M., Liu, T.-K., Walia, V., Kumar, A., Lai, T.-H., 2017. Spatial and temporal anomalies of soil gas in northern Taiwan and its tectonic and seismic implications. J. Asian Earth Sci. 149, 64–77. https://doi.org/10.1016/j.jseaes.2017.02.032.

Gaskin, J., Coyle, D., Whyte, J., Krewksi, D., 2018. Global estimate of lung cancer mortality attributable to residential radon. Environ. Health Perspect. 126, 057009 https://doi.org/10.1289/EHP2503.

Geological Survey Ireland Tellus programme, 2020. Tellus Shallow topsoils geochemistry survey data, release survey area G5. https://www.gsi.ie/documents/Tellus_Geochemistry_Survey_Block_G5_Shallow_Topsoil_Data_Release_Notes_Oct20.pdf.

Giustini, F., Ciotoli, G., Rinaldini, A., Ruggiero, L., Voltaggio, M., 2019. Mapping the geogenic radon potential and radon risk by using Empirical Bayesian Kriging regression: a case study from a volcanic area of central Italy. Sci. Total Environ. 661, 449–464. https://doi.org/10.1016/j.scitotenv.2019.01.146.

Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochem. Explor. Environ. Anal. 10, 27–74. https://doi.org/10.1144/1467-7873/09-210.

Grunsky, E.C., Caritat, P. de, 2019. State-of-the-art analysis of geochemical data for mineral exploration. Geochem. Explor. Environ. Anal. 20, 217–232. https://doi.org/10.1144/geochem2019-031.

Hamideen, M.S., Bdair, O.M., Chandrasekaran, A., Saleh, H., Elimat, Z.M., 2020. Multivariate statistical investigations of natural radioactivity and radiological hazards in building materials mainly used in Amman Province, Jordan. Int. J. Environ. Anal. Chem. 100, 189–203. https://doi.org/10.1080/03067319.2019.1635123.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Ma, X., Zhu, Y., 2022. A review of machine learning in geochemistry and cosmochemistry: method improvements and applications. Appl. Geochem. 140, 105273 https://doi.org/10.1016/j.apgeochem.2022.105273.

Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1997. The soil production function and landscape equilibrium. Nature 388, 358–361. https://doi.org/10.1038/41056.

Huntingford, C., Jeffers, E.S., Bonsall, M.B., Christensen, H.M., Lees, T., Yang, H., 2019. Machine learning and artificial intelligence to aid climate change research and preparedness. Environ. Res. Lett. 14, 124007.

Jansson, N.F., Liu, W., 2020. Controls on cobalt and nickel distribution in hydrothermal sulphide deposits in Bergslagen, Sweden - constraints from solubility modelling. GFF 142, 87–95. https://doi.org/10.1080/11035897.2020.1751270.

Johnson, D.L., Watson-Stegner, D., 1987. Evolution model of pedogenesis. Soil Sci. 143, 349–366.

Johnston, J., Raub, T., Ashton, J., 2013. Mineral Magnetism Identifies the Presence of Pyrrhotite in the Navan Zn-Pb Deposit, Ireland: Implications for Low Temperature Pyrite to Pyrrhotite Reduction, Timing of Mineralization and Future Exploration Strategies. Mineral deposit research for a hi-tech world. https://doi.org/10.13140/2.1.2447.3601. Uppsala 323–325.

Kanagawa, M., Hennig, P., Sejdinovic, D., Sriperumbudur, B.K., 2018. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences arXiv preprint arXiv:1807.02582.

Khan, F., Khattak, S.A., Wazir, Z., Waqas, M., 2021. Spatial distribution of radon concentrations in balakot-bagh (B–B) fault line and adjoining areas, lesser himalayas, north Pakistan. Environ. Earth Sci. 80, 291. https://doi.org/10.1007/s12665-021-09569-8.

Kim, E., Ahn, H., Jo, H.Y., Ryu, J.-H., Koh, Y.-K., 2017. Chlorite alteration in aqueous solutions and uranium removal by altered chlorite. J. Hazard Mater. 327, 161–170. https://doi.org/10.1016/j.jhazmat.2016.12.051.

Kirasich, K., Smith, T., Sadler, B., 2018. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Sci. Rev. 1, 9.

Knights, K.V., Szpak, M., Mather, J., Collins, L., 2020. Geological Survey Ireland Tellus Geochemical Survey: Shallow Topsoil Data from the Border and West of Ireland (No. Version 1.0). Geological Survey Ireland.

Kovalev, S.G., Puchkov, V.N., Kovalev, S.S., 2014. First findings of siegenite (CoNi2S4) in picritic and picrodoleritic complexes of the Southern Urals. Dokl. Earth Sci. 457, 796–802. https://doi.org/10.1134/S1028334X1407023X.

Liao, R., Shi, Z., Chen, Y., Zhang, J., Wang, X., Hou, Y., Zhang, K., 2020. Characteristics of uranium sorption on illite in a ternary system: effect of phosphate on adsorption. J. Radioanal. Nucl. Chem. 323, 159–168. https://doi.org/10.1007/s10967-019-06878-y.

Lin, H.-T., Liu, F.-C., Wu, C.-Y., Kuo, C.-F., Lan, W.-C., Yu, H.-P., 2019. Epidemiology and survival outcomes of lung cancer: a population-based study. BioMed Res. Int. 2019, 8148156 https://doi.org/10.1155/2019/8148156.

Loring, D.H., 1976. Distribution and partition of cobalt, nickel, chromium, and vanadium in the sediments of the Saguenay fjord. Can. J. Earth Sci. 13, 1706–1718. https://doi.org/10.1139/e76-180.

Marks, F.R., Menuge, J.F., Boyce, A.J., Blakeman, R.J., 2017. Controls on the formation of a large Zn-Pb Irish-type deposit: evidence from the Navan halo. Programme and 61.

McKinley, J.M., Grunsky, E., Mueller, U., 2018. Environmental monitoring and peat assessment using multivariate analysis of regional-scale geochemical data. Math. Geosci. 50, 235–246. https://doi.org/10.1007/s11004-017-9686-x.
Mei, H., Aoyagi, N., Saito, T., Kozai, N., Sugiura, Y., Tachi, Y., 2022. Uranium (VI) sorption on illite under varying carbonate concentrations: batch experiments, modeling, and cryogenic time-resolved laser fluorescence spectroscopy study. Appl. Geochem. 136, 105178 https://doi.org/10.1016/j.apgeochem.2021.105178.
Moles, N.R., Moles, R.T., 2002. Influence of geology, glacial processes and land use on soil composition and Quaternary landscape evolution in the Burren National Park, Ireland. Catena 47, 291–321. https://doi.org/10.1016/S0341-8162(01)00190-4.
Mueller, U., Tolosana Delgado, R., Grunsky, E.C., McKinley, J.M., 2020. Biplots for compositional data derived from generalized joint diagonalization methods. Appl. Comput. Geosci. 8, 100044 https://doi.org/10.1016/j.acags.2020.100044.
Nathwani, C.L., Wilkinson, J.J., Fry, G., Armstrong, R.N., Smith, D.J., Ihlenfeld, C., 2022. Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation. Miner. Deposita. https://doi.org/10.1007/s00126-021-01086-9.
Pereira, A.J.S.C., Godinho, M.M., Neves, L.J.P.F., 2010. On the influence of faulting on small-scale soil-gas radon variability: a case study in the Iberian Uranium Province. J. Environ. Radioact. 101, 875–882. https://doi.org/10.1016/j.jenvrad.2010.05.014.
Petermann, E., Meyer, H., Nussbaum, M., Bossew, P., 2021. Mapping the geogenic radon potential for Germany by machine learning. Sci. Total Environ. 754, 142291 https://doi.org/10.1016/j.scitotenv.2020.142291.
Petersell, V., Jüriado, K., Raukas, A., Shtokalenko, M., Täht-Kok, K., 2015. Quaternary deposits and weathered bedrock material as a source of dangerous radon emissions in Estonia. Geologos 21, 139–147. https://doi.org/10.1515/logos-2015-0006.
Pham, B.T., Ly, H.-B., Al-Ansari, N., Ho, L.S., 2021. A comparison of Gaussian process and M5P for prediction of soil permeability coefficient. Sci. Program. 2021, 3625289 https://doi.org/10.1155/2021/3625289.
Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inf. Sci. 31, 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.
Radiological Protection, 2019. Act 1991 (ionising radiation) Regulations. In: Statutory Instruments Ireland, S.I. No. 30 of 2019.
Rezaie, F., Kim, S.W., Alizadeh, M., Panahi, M., Kim, H., Kim, S., Lee, Jongchun, Lee, Jungsub, Yoo, J., Lee, S., 2021. Application of machine learning algorithms for geogenic radon potential mapping in Danyang-Gun, South Korea. Front. Environ. Sci. 9 https://doi.org/10.3389/fenvs.2021.753028.
Ribeiro, F.C.A., Silva, J.I.R., Lima, E.S.A., do Amaral Sobrinho, N.M.B., Perez, D.V., Lauria, D.C., 2018. Natural radioactivity in soils of the state of Rio de Janeiro (Brazil): radiological characterization and relationships to geological formation, soil types and soil properties. J. Environ. Radioact. 182, 34–43. https://doi.org/10.1016/j.jenvrad.2017.11.017.
Rodríguez-Martínez, Á., Ruano-Ravina, A., Torres-Durán, M., Provencio, M., Parente-Lamelas, I., Vidal-García, I., Martínez, C., Hernández-Hernández, J., Abdulkader-Nallib, I., Castro-Añón, O., Varela-Lema, L., Piñeiro-Lamas, M., Fidalgo, P.S., Fernández-Villar, A., Barros-Dios, J., Pérez-Ríos, M., 2021. Residential radon and small cell lung cancer. In: Final Results of the Small Cell Study. Archivos de Bronconeumología. https://doi.org/10.1016/j.arbres.2021.01.027.
Sakhaee, A., Gebauer, A., Ließ, M., Don, A., 2022. Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms. Soil 8, 587–604. https://doi.org/10.5194/soil-8-587-2022.
Schabath, M.B., Cote, M.L., 2019. Cancer progress and priorities: lung cancer. Cancer Epidemiol. Biomarkers Prev. 28, 1563–1579. https://doi.org/10.1158/1055-9965.EPI-19-0221.
Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. STATA J. 20, 3–29. https://doi.org/10.1177/1536867X20909688.
Shang, Q., Tan, D., Gao, S., Feng, L., 2019. A hybrid method for traffic incident Duration prediction using BOA-optimized random forest combined with neighborhood components analysis. J. Adv. Transport. 2019, 4202735 https://doi.org/10.1155/2019/4202735.
Shepherd, Russell G., 1989. Correlations of permeability and grain size. Ground Water 27, 633–638. https://doi.org/10.1111/j.1745-6584.1989.tb00476.x.
Somma, R., Ebrahimi, P., Troise, C., De Natale, G., Guarino, A., Cicchella, D., Albanese, S., 2021. The first application of compositional data analysis (CoDA) in a multivariate perspective for detection of pollution source in sea sediments: the Pozzuoli Bay (Italy) case study. Chemosphere 274, 129955. https://doi.org/10.1016/j.chemosphere.2021.129955.
Sperandei, S., 2014. Understanding logistic regression analysis. Biochem. Med. 24, 12–18. https://doi.org/10.11613/BM.2014.003.
Talebi, H., Peeters, L.J.M., Otto, A., Tolosana-Delgado, R., 2022. A truly spatial random forests algorithm for geoscience data analysis and modelling. Math. Geosci. 54, 1–22. https://doi.org/10.1007/s11004-021-09946-w.
Tolosana-Delgado, R., McKinley, J., 2016. Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). Appl. Geochem. 75, 263–276. https://doi.org/10.1016/j.apgeochem.2016.05.004.
Tzortzis, M., Tsertos, H., Christofides, S., Christodoulides, G., 2003. Gamma-ray measurements of naturally occurring radioactive samples from Cyprus characteristic geological rocks. Radiat. Meas. 37, 221–229. https://doi.org/10.1016/S1350-4487(03)00028-3.
United Nations Scientific Committee on the Effects of Atomic Radiation, 2011. Report of the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2011.
Vermeesch, P., 2013. Multi-sample comparison of detrital age distributions. Chem. Geol. 341, 140–146. https://doi.org/10.1016/j.chemgeo.2013.01.010.
Wang, F., Huo, L., Li, Y., Wu, L., Zhang, Y., Shi, G., An, Y., 2023. A hybrid framework for delineating the migration route of soil heavy metal pollution by heavy metal similarity calculation and machine learning method. Sci. Total Environ. 858, 160065 https://doi.org/10.1016/j.scitotenv.2022.160065.
Wang, L., Liu, B., McKinley, J.M., Cooper, M.R., Li, C., Kong, Y., Shan, M., 2021. Compositional data analysis of regional geochemical data in the Lhasa area of Tibet, China. Appl. Geochem. 135, 105108 https://doi.org/10.1016/j.apgeochem.2021.105108.
Wang, Y., Zhao, Y., Xu, S., 2022. Application of VNIR and machine learning technologies to predict heavy metals in soil and pollution indices in mining areas. J. Soils Sediments. https://doi.org/10.1007/s11368-022-03263-3.
World Health Organization (Ed.), 2009. WHO Handbook on Indoor Radon: a Public Health Perspective. World Health Organization.
World Health Organization, 2007. International Radon Project: Survey on Radon Guidelines, Programmes and Activities (No. World Health Organization. WHO/HSE/RAD/07.01).
Wu, Y., Liu, Q., Ma, J., Zhao, W., Chen, H., Qu, Y., 2022. Antimony, beryllium, cobalt, and vanadium in urban park soils in Beijing: Machine learning-based source identification and health risk-based soil environmental criteria. Environ. Pollut. 293, 118554 https://doi.org/10.1016/j.envpol.2021.118554.
Xu, H., Croot, P., Zhang, C., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environ. Int. 151, 106456 https://doi.org/10.1016/j.envint.2021.106456.
Yesares, L., Drummond, D.A., Hollis, S.P., Doran, A.L., Menuge, J.F., Boyce, A.J., Blakeman, R.J., Ashton, J.H., 2019. Coupling mineralogy, textures, stable and radiogenic isotopes in identifying ore-forming processes in Irish-type carbonate-hosted Zn–Pb deposits. Minerals 9. https://doi.org/10.3390/min9060335.
Young, M.E., Knights, K.V., Smyth, D., Glennon, M.M., Scanlon, R.P., Gallagher, V., 2016. The Tellus geochemical surveys, results and applications. In: Young, M.E. (Ed.), Unearthed: impacts of the Tellus surveys of the north of Ireland. Dublin. Royal Irish Academy. https://doi.org/10.3318/978-1-908996-88-6.ch3
Zagà, V., Cattaruzza, M.S., Martucci, P., Pacifici, R., Trisolini, R., Bartolomei, P., Giacobbe, R., Patelli, M., Paioli, D., Esposito, M., Fabbri, V., Gallus, S., Gorini, G., 2021. The “polonium in vivo” study: polonium-210 in bronchial lavages of patients with suspected lung cancer. Biomedicines 9. https://doi.org/10.3390/biomedicines9010004.
Zheng, C., Liu, P., Luo, X., Wen, M., Huang, W., Liu, G., Wu, X., Chen, Z., Albanese, S., 2021. Application of compositional data analysis in geochemical exploration for concealed deposits: a case study of Ashele copper-zinc deposit, Xinjiang, China. Appl. Geochem. 130, 104997 https://doi.org/10.1016/j.apgeochem.2021.104997.
Zohuri, B., 2020. 2 - nuclear fuel cycle and decommissioning. In: Khan, S.U.-D., Nakhabov, A. (Eds.), Nuclear Reactor Technology Development and Utilization. Woodhead Publishing, pp. 61–120. https://doi.org/10.1016/B978-0-12-818483-7.00002-0.
Zuo, R., 2017. Machine learning of mineralization-related geochemical anomalies: a review of potential methods. Nat. Resour. Res. 26, 457–464. https://doi.org/10.1007/s11053-017-9345-4.