*Year :* 2012

# Analysis and Spatial Modeling of the Indoor Radon Swiss Data

Tapia Rafael

Tapia Rafael, 2012, Analysis and Spatial Modeling of the Indoor Radon Swiss Data

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive.
http://serval.unil.ch

UNIL | Université de Lausanne

Faculté des géosciences et de l'environnement
Institut de géomatique et d'analyse du risque

# Analysis and Spatial Modeling of the Indoor Radon Swiss Data

THESE DE DOCTORAT
présentée à la Faculté des géosciences et de l'environnement
de l'Université de Lausanne par

Rafael Tapia

Ing., M.Sc.
Universidad Nacional Agraria La Molina, Lima – Peru
International Training Center (ITC), Enschede –The Netherlands

**Jury:**

| | |
|---|---|
| President: | Prof. Torsten Vennemann |
| Thesis Director: | Prof. Mikhail Kanevski |
| Internal Expert: | Prof. François Bavaud |
| External Expert: | Dr. Vasily Demyanov |

Lausanne, 2012

UNIL | Université de Lausanne
Faculté des géosciences et de l'environnement
bâtiment Amphipôle
CH-1015 Lausanne

# IMPRIMATUR

Vu le rapport présenté par le jury d'examen, composé de

| | |
|---|---|
| Président de la séance publique : | M. le Professeur Torsten Vennemann |
| Président du colloque : | M. le Professeur Torsten Vennemann |
| Directeur de thèse : | M. le Professeur Mikhail Kanevski |
| Expert interne : | M. le Professeur François Bavaud |
| Expert externe : | M. le Docteur Vasily Demyanov |

Le Doyen de la Faculté des géosciences et de l'environnement autorise l'impression de la thèse de

## Monsieur Rafael TAPIA

*Master in Geoinformation Science and Earth Observation*
*Specialisation « Geoinformatics »*
*University of Twente*

intitulée

## ANALYSIS AND SPATIAL MODELING OF THE INDOOR SWISS DATA

Lausanne, le 18 juin 2012

Pour le Doyen de la Faculté des géosciences et de l'environnement

Professeur Torsten Vennemann, Vice-Doyen

*A mis queridas hijas Astrid y Sofia*

*A mis padres Mario y Ana Maria*

# Analysis and Spatial Modeling of the Indoor Radon Swiss Data

**Rafael Tapia**

Institut de géomatique et d'analyse du risque

**Résumé**

Ce travail de recherche s'intéresse à une problématique importante pour la santé publique en Suisse ; il s'agit de la pollution induite par l'accumulation de radon à l'intérieur des logements. La modélisation spatiale du radon d'intérieur en Suisse est une tâche complexe qui représente un défi en raison des nombreux facteurs qui doivent être pris en compte. L'analyse de données du radon doit être abordée simultanément d'un point de vue non seulement statistique mais aussi d'un point de vue spatial. Étant un processus de nature multivariée, c'était important de définir l'influence de chaque facteur. En particulier, il était intéressant de définir l'influence de la géologie, puisqu'elle est souvent considérée comme étant un facteur prépondérant influençant les concentrations du radon d'intérieur. Cette importance de la géologie a pu être vérifiée à l'aide des données suisses, mais il est également apparu évident qu'il ne s'agissait pas du seul facteur à prendre en compte pour la modélisation spatiale.

L'analyse statistique de données, au niveau uni-varié et multi-varié et a été suivi par une analyse spatiale exploratoire. Plusieurs outils proposés dans la littérature ont été essayés et adaptés, y compris des méthodes basées sur le concept de fractalité, sur les algorithmes de désagrégation et sur les méthodes d'analyse par fenêtres glissantes. L'utilisation de l'Indice de Morishita par Quantiles (QMI) a été proposée comme une procédure permettant d'évaluer l'agrégation spatiale en fonction du niveau de radon. Les méthodes de désagrégation ont été appliquées dans le but d'approcher au mieux les paramètres de l'histogramme global. La phase d'analyse exploratoire multi-échelle a été réalisée en partant d'une échelle régionale jusqu'à une échelle locale. La distribution spatiale des données par secteurs a été optimisée pour faire face aux conditions de stationnarité requises par les modèles géostatistiques. Des méthodes simples de modélisation spatiale telles que les K plus proches voisins (KNN), la variographie et les réseaux de neurones de régression généralisée (GRNN) ont également été proposés comme des outils exploratoires.

Ensuite, différentes méthodes d'interpolation spatiale ont été appliquées à un sous-ensemble de la base de données utilisées. Une séquence d'analyse allant de la méthode la plus simple à la plus complexe a été adoptée et les résultats rassemblés pour trouver des définitions communes pour les paramètres de continuité et de voisinage. Dans le but de réduire les variations importantes des mesures à l'échelle locale, un filtre fondé sur la validation croisée a été proposé (le CVMF). A la fin du chapitre correspondant, une série de tests permettant d'évaluer l'homogénéité des données et la robustesse des méthodes ont été effectuées. Ceci a mené à conclure à l'importance de produire une distribution non biaisée

des données pour l'étape de validation. Des limitations propres aux fonctions linéaires et aux méthodes de régression dans leur capacité à reproduire des distributions statistiques asymétriques ont été également mises en évidence.

La dernière partie a été dédiée aux méthodes de modélisation avec une interprétation probabiliste. La transformation des données et les simulations ont permis d'utiliser des modèles de distribution multi-gaussienne et ont aidé à la prise en compte de l'incertitude relative au processus de pollution par le radon d'intérieur. La transformation par catégorisation et la classification ont été présentées comme des solutions pour faire face à la présence de valeurs extrêmes. Des scénarios de simulation ont été proposés, y compris une alternative pour la reproduction de l'histogramme global basée sur le domaine spatial d'échantillonnage. La simulation Gaussienne séquentielle (SGS) s'est révélée être la méthode donnant les informations les plus complètes, tandis que la classification a fourni une manière plus robuste d'interpoler les données. Une mesure d'erreur adaptée à la fonction de décision utilisée pour classifier dans le cas des catégories biaisées a été définie. Parmi les méthodes de classification, le réseau neuronal probabiliste (PNN) semble être mieux adapté à la modélisation de catégories définies avec des seuils élevés, de même que pour l'automatisation. Les séparateurs à vaste marge (SVM) sont avantageux lorsqu'il s'agit de classifier des catégories avec un nombre d'éléments équilibrés.

Une des conclusions générales de cette recherche est qu'on ne peut pas juger une certaine méthode d'estimation comme étant la meilleure à toutes les échelles et pour tous les voisinages. Par contre, la simulation doit être considérée comme la méthode de base, tandis que toutes les autres méthodes peuvent donner des informations complémentaires permettant d'accomplir une cartographie d'aide à la décision efficace.

# Analysis and Spatial Modeling of the Indoor Radon Swiss Data

**Rafael Tapia**

Institut de géomatique et d'analyse du risque

**Abstract**

The present research deals with an important public health threat, which is the pollution created by radon gas accumulation inside dwellings. The spatial modeling of indoor radon in Switzerland is particularly complex and challenging because of many influencing factors that should be taken into account. Indoor radon data analysis must be addressed from both a statistical and a spatial point of view. As a multivariate process, it was important at first to define the influence of each factor. In particular, it was important to define the influence of geology as being closely associated to indoor radon. This association was indeed observed for the Swiss data but not probed to be the sole determinant for the spatial modeling.

The statistical analysis of data, both at univariate and multivariate level, was followed by an exploratory spatial analysis. Many tools proposed in the literature were tested and adapted, including fractality, declustering and moving windows methods. The use of Quantile Morisita Index (QMI) as a procedure to evaluate data clustering in function of the radon level was proposed. The existing methods of declustering were revised and applied in an attempt to approach the global histogram parameters. The exploratory phase comes along with the definition of multiple scales of interest for indoor radon mapping in Switzerland. The analysis was done with a top-to-down resolution approach, from regional to local levels in order to find the appropriate scales for modeling. In this sense, data partition was optimized in order to cope with stationary conditions of geostatistical models. Common methods of spatial modeling such as K Nearest Neighbors (KNN), variography and General Regression Neural Networks (GRNN) were proposed as exploratory tools.

In the following section, different spatial interpolation methods were applied for a particular dataset. A bottom to top method complexity approach was adopted and the results were analyzed together in order to find common definitions of continuity and neighborhood parameters. Additionally, a data filter based on cross-validation was tested with the purpose of reducing noise at local scale (the CVMF). At the end of the chapter, a series of test for data consistency and methods robustness were performed. This lead to conclude about the importance of data splitting and the limitation of generalization methods for reproducing statistical distributions.

The last section was dedicated to modeling methods with probabilistic interpretations. Data transformation and simulations thus allowed the use of multigaussian models and helped take the indoor radon pollution data uncertainty into consideration. The categorization transform was presented as a solution for extreme values modeling through classification. Simulation scenarios were proposed, including an alternative proposal for the

reproduction of the global histogram based on the sampling domain. The sequential Gaussian simulation (SGS) was presented as the method giving the most complete information, while classification performed in a more robust way. An error measure was defined in relation to the decision function for data classification hardening. Within the classification methods, probabilistic neural networks (PNN) show to be better adapted for modeling of high threshold categorization and for automation. Support vector machines (SVM) on the contrary performed well under balanced category conditions.

In general, it was concluded that a particular prediction or estimation method is not better under all conditions of scale and neighborhood definitions. Simulations should be the basis, while other methods can provide complementary information to accomplish an efficient indoor radon decision mapping.

# Acknowledgements

I would first like to thank Prof. Mikhail Kanevski for the opportunity he gave me to work in this thesis, for his generosity and for sharing his knowledge with me. Without his constant support this work would not have been possible. I fill also indebted to Dr. Vasily Demyanov, to Prof. Francois Bavaud and to Prof. Torsten Vennemann, for their revision and comments to improve the thesis. I specially want to acknowledge Prof. Michel Maignan for his inspiring teachings in geostatistics.

I also want to mention the colleagues at IGAR, with whom we collaborate on the indoor radon research. In particular to Vadim Timonin, Devis Tuia, Alexei Pozdnukhov and Pierre Emmanuel Huguenot. I want to thank them and all the colleagues at the IGAR institute, to Marj Tonini, Carmen Vega Orozco, Jean Golay, Loris Foresti, Michele Volpi and Giona Matasci, above all for their friendship and for the good environment they have created. I wish to made my thanks extensive to the dynamic research group of Prof. Michel Jaboyedoff and to himself in particular, for the opportunity we had to collaborate in the remote sensing field.

I wish to thank the Swiss Federal Office of Public Health (OFSP) for providing the Swiss Radon Data used in the present research. In particular to Martha Gruson for her support. I also want to acknowledge the IBRAE research team for the use of the GSO and MLO software packages made in this thesis.

## List of Acronyms

| | |
|---|---|
| ALSOS | Alternating Least Square with Optimal Scaling |
| BLUE | Best Linear Unbiased Estimator |
| CATPCA | Categorical PCA |
| CV | Cross Validation |
| Df | fractal Dimension |
| DSSIM | Direct Sequential Simulation |
| ESDA | Exploratory Spatial Data Analysis |
| EVT | Extreme Value Theory |
| FOPH | Federal Office of Public Health (Switzerland) |
| GRNN | General Regression Neural Network |
| GSLIB | Geostatistical Software Library |
| GSO | GeoStatOffice software |
| IDW | Inverse Distance Weighting |
| IK | Indicator Kriging |
| KNNR | K Nearest Neighbors Regression |
| KCVMF | KNNR Cross Validation Mean Filter |
| MDS | MultiDimensional Scaling |
| MGK | MultiGaussian Kriging |
| MI | Morishita Index |
| MW | Moving Windows |
| NLPCA | Non linear PCA |
| NS | Normal Scores transformation |
| OK | Ordinary Kriging |
| ORaP | Ordinance on Radiological Protection |
| PCA | Principal Component Analysis |
| PNN | Probabilistic Neural Networks |
| QMI | Quantile Morisita Index |
| RV | Random Variable |
| SPSS | Statistical Package for the Social Sciences |
| SGS | Sequential Gaussian Simulation |
| SK | Simple Kriging |
| SVM | Support Vector Machine |
| SUMDA | Statistical Univariate and Multivariate Analysis |
| WHO | World Health Organization |

# Contents

# Chapter 1

# Introduction

Indoor radon accumulation is an important public health threat that requires countermeasures in order to be reduced. It has been identified as a significant factor causing lung cancer, second only to smoking habits. It has been calculated that in Switzerland indoor radon is responsible for about 40 % of the population's exposure to radiation, as shown in Figure 1.1 (56).

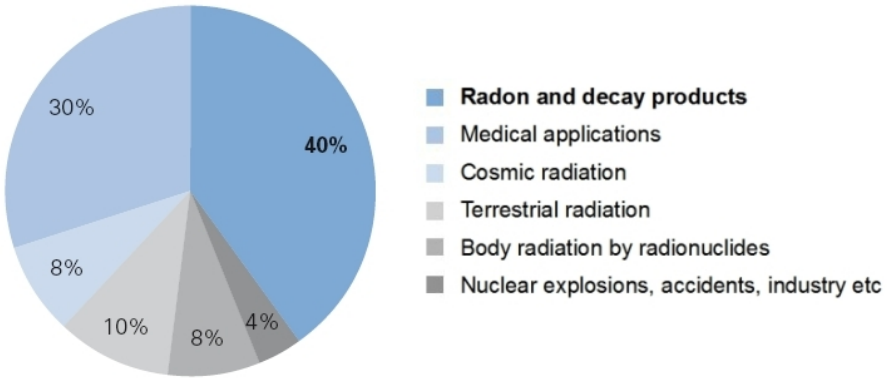



Figure 1.1: *Radiation sources in Switzerland (OFSP (56))*

A crucial measure to mitigate the effect of this pollutant is the accurate identification of dwellings with mean values that can be labeled as critical. Cartography of indoor radon spatial distribution is a tool that helps identify areas exposed to higher accumulation of the gas. Spatial modeling of indoor radon measures and cartographic representation can provide valuable information at any stage of a radioprotection program. This information can be used to plan sampling campaigns, to provide local authorities with status reports, to identify areas that require critical action and ultimately, to help identify buildings with a high probability of being above permissible limits of radon concentration. Accurate and realistic mapping of indoor radon is required for decision-making.

## 1.1 Indoor radon accumulation process

Distinct physical processes inside dwellings govern indoor radon gas accumulation. Its main source is the underlying soil body, and from there gas is diffused directly to buildings through fractures in the floor or enters in a diluted form through water. Once inside the dwelling, it is transported upward by the advection caused by hot/cold air interchange and more intensively by ventilation. Figure 1.2 attempts to sketch the process using a house model.

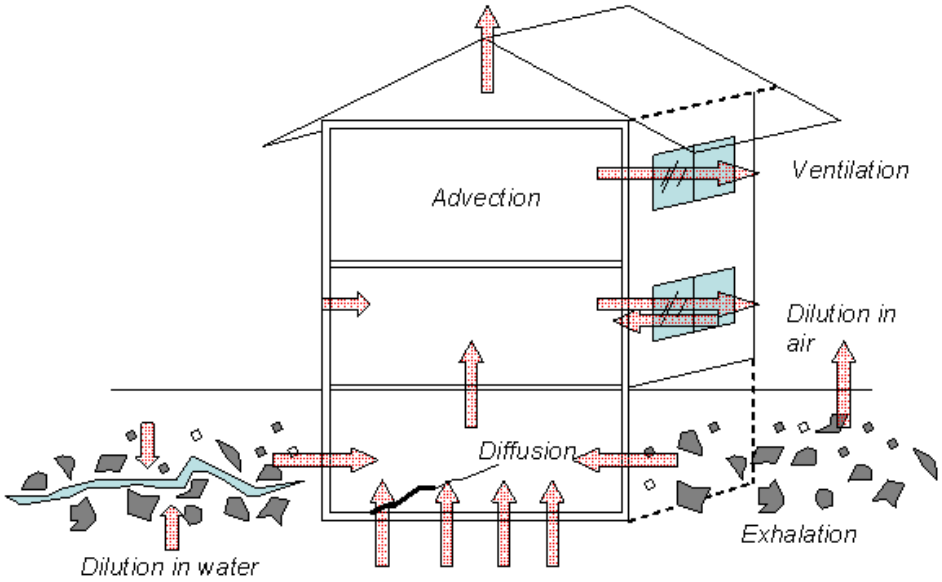



Figure 1.2: *Physical processes in indoor radon accumulation)*

Research shows that indoor radon measurements are not stable during the course of a day, a month or even a year for the same house. In fact, it depends on a variety of factors starting with the nature of the underlying lithological material, the insolation of the building, the rates of ventilation and even atmospheric conditions with annual variations. The mean life of the 222 isotope is 3.5 days; in this period it can be harmful for humans if it is inhaled at high concentrations on a daily basis. To evaluate chronic exposure, indoor radon concentrations are calculated by placing detectors in dwellings over long periods of time and then normalized to mean annual concentrations.

Indoor radon is thus, like any event in nature, a multivariate process resulting from the interaction of many factors. Some factors can be measured and some have unknown behaviors, which result in noise addition. The influence of these factors is also dependent on scale factors. There is a group of variables that can be labeled as having a local influence, while others are more global. For instance, building insolation is only pertinent to dwelling conditions and therefore is a local variable. On the contrary, geology is more global because it is present in an entire territory. Independent of the scale of influence of the radon accumulation factors, there is a coupled effect from their interaction.

On a global scale, outdoor measurements provide evidence that natural factors like geology, soil permeability and seasonality have influence on exhalation rates and could be regarded as a threat for indoor accumulation. On a local scale, these potentially hazardous concentrations could accumulate in houses as a result of human activity, including the influence of building and room types, floor level, ventilation and other factors, which are reported as indicators for higher rates of indoor radon concentrations.

Since it is a complex phenomenon, the influence of all the triggering factors results in a high local variability of measurements, with records differing even between neighboring buildings. Indoor radon accumulation is often particular to every single building. Statistical distributions of data sets are also not very homogeneous. The mean level of concentration for a country like Switzerland is about 230 $Bq/m^3$, but is possible to find values up to 20,000 $Bq/m^3$. The presence of these extreme values also poses difficulties for spatial modeling. Another obstacle is related to the spatial distribution of data. Since radon is measured only inside buildings, they often follow clustered patterns, leaving large areas of the national territory without sample coverage.

## 1.2 Soil-geology factors

The first methods of studying radon's spatial distribution were closely related to geological studies, since it is a natural radioactive gas generated in soil. Lot of interesting research was made with radon prediction using geological data. In Germany (45), the initial objective of surveys was to predict the so called radon-potential on a regional scale, which was complicated by small-scale variations of rock composition and gas permeability associated to the soil texture. The uranium content of rocks is largely variable. Therefore, it is better to measure radon concentration in soil gas directly. These measurements, together with soil permeability, were used as parameters for an empirical classification in Germany based on the radon-potential. To attain a regional radon-potential distribution, geological maps were generalized combining units under lithostrati-graphical and radon-relevant aspects. The authors, however, mention the validity of this procedure for regional mapping is disputable; although it is suitable for small-scale investigations, for example, in construction site assessment.

In later research, sampling density was increased to intervals of 25 km, and 5 km for test areas with higher outdoor radon values (46). The permeability factor was put aside, assuming a pessimistic scenario of high soil permeability all over the country (which was found to be the more frequent case). Positive correlations with fracturing and mineralization of the older formations, such as granite and even sands in glacial deposits, helped refine classification of radon-potential based on geogenic factors; but no agreement was reached on the regional distribution of long-term indoor measurements. A similar interference from glaciations and fluvial processes was found in studies carried out in the US (26) and in Norway (67). In the Northeastern and Mid-Atlantic States that are unaffected by glaciations, bedrock geology was used to predict indoor radon, especially in basement homes, while in the Northern Appalachian Highlands and Appalachian Plateau, researchers were unable to

approximate the correlation of bedrock geology to indoor radon.

A geogenic radon potential approach was also adopted by the Czech Republic (73). In this case, researchers deemed it important to integrate geological radon surveys for predictions. When lognormal distribution was analyzed along administrative boundaries, they found that in 40% of the cases, the model criteria was not met. On the contrary, similar data distribution was observed in 5 by 5 km on square units. A spatial support related to geological units was proposed because of its finer resolution and close relation to radon pruning. Modeling based on geological information was based on the use of transfer functions per geological unit. The assumption of a lognormal distribution for the transfer functions was proposed to predict indoor radon accumulation. However, it was pointed out that the procedure relies on large enough representative data to calculate transfer factors, for both indoor and soil radon. Also, Neznal (53) et al. had worked on the development of a uniform methodology to assess the risk of radon penetrating the underlying soil or bedrock in order to determine the radon index of a site before construction.

Earlier research in the UK (52) used indoor measurements and geological boundaries to produce radon-potential maps that were more detailed than the grid square ones. However, lateral variations in geological formations contributed to higher variation in the results, making it unfeasible to extrapolate results to areas where there were insufficient homes. Later, Miles et al. (50) proposed the use of kriging on data grouped by geology because they found less radon-potential discontinuities than for administrative boundaries.

The remaining drawback of the geology-modeling approach is the spatial arrangement complexity of the geology itself. There is an inherent ontology included on geological maps that contributes to uncertainty. Additionally, indoor radon measurements are taken in urban areas where there is a degree of surface disturbance that can modify the natural properties of the soil. While outdoor soil gas records are easily linked to the geological unit, this relation is lost when it comes to indoor radon. For instance, in the English Midlands (17), surveys of complex areas ranging from one principal lithology to several had shown that soil radon values could vary in a very erratic way over distances of a few meters.

## 1.3 Other environmental factors

Following the natural succession of interactions made by radon gas on its way to dwellings, all the conditionals imposed by the environment should be mentioned. Regarding soil depth, radon detectors were placed at four different depths in intervals of 0-2 m along a 2,200 m long profile to analyze the variation of soil radon on uniform geology according to depth and season (32). An exponential function was fitted for the measurements in the months of June and August. The seasonal effect was sharp. More gas was registered in summer as shown on a deducted graphical function (Figure 1.3). In addition, a drastically lower soil radon level was observed in areas at depths close to, or under, the ground water table.

Iakovleva (30) concluded that the effects of meteorological factors (atmospheric pressure and temperature) on the radon concentration in the soil air are observed even at a depth

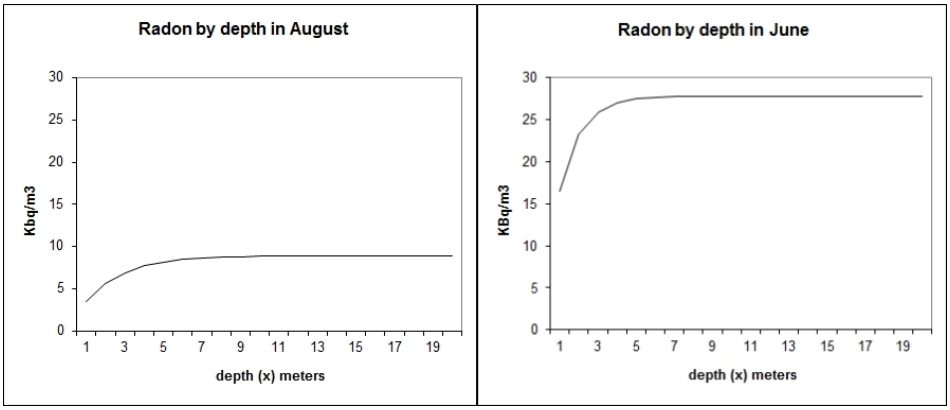



Figure 1.3: *Seasonal variation of radon in soil gas at different depth (based on data from Jonson, 2001)*

of 70 cm, and may vary nearly tenfold within the same area, depending on meteorological conditions and soil type. Thus, a single measurement of the radon concentration is not representative and may result in serious errors when predicting indoor radon levels. Iakovleva also pointed out that the average monthly radon concentration recorded in April and September closely corresponds to annual average values. In Finland (80), adjustments for the effects of outdoor temperature and wind speed were applied to the measurements over a 2 month period to make them comparable to annual values. Marley (48) investigated the radon progeny inside buildings and postulated that the variability of radon is primarily dependent on three atmospheric variables: barometric pressure, vapor pressure and wind variation, acting at the source of the radon path. Finkelstein (20) demonstrated the possibility of Rn gas anomalies being caused by changes in atmospheric temperature due to ventilation and the positive agreement between Rn gas concentrations.

## 1.4 Dwelling conditions

Although there can be a high content of naturally occurring radioactive elements in building materials, their contribution to indoor radon is often neglected. An example of high material exhalation was found in houses in Tomsk city where slag was used as the primary building material (29).

Ventilation rates and basement isolation were found to be the most influential environmental factors for indoor radon accumulation. In Italy (61), radon 222 was evaluated in commercial areas of the Savona Province. The main variables that were related to radon concentration were the age of the building, the level above ground, the season of the year, wind exposure, active windows and the type of heating system. Commercial stores equipped with individual air heating/cooling systems exhibited radon concentrations that were three times higher than those heated by centralized systems.

## 1.5 European research on indoor radon mapping

Indoor radon risk has increasingly gained the attention of public health authorities in the last decade, and methods for the analysis of data have simultaneously been developed. Among the methods of representing indoor radon concentration, radon mapping has attracted special attention. Various approaches exist in different countries, particularly in Europe, where extensive indoor measurements have been made in some cases.

Let's examine some of the documented studies. For instance, public health authorities in the UK produced maps of indoor radon distribution at 1 km and 5 km grid resolutions, representing median values and the probability of exceeding a certain threshold value assuming lognormal distribution (50). For areas with less sampling, information was supplemented using a correlation with geological units. Studies in Austria (22) used a mixture of information sources. The level of risk was defined by a radon-potential measurement for dwellings, standardized in relation to radon-related parameters and then related to geological units.

In Denmark (1) indoor radon exceeding 200 $Bq/m^3$ was presented on a municipal scale using a normal transform. Corrections were made using soil composition (sand and gravel). Similarly, in Finland (80), soil composition was taken into account for risk, since individual cases were detected exceeding the action level of 400 $Bq/m^3$ for specific types of soils. In the Czech Republic (73), transfer-functions were calculated from geological units together with indoor and outdoor measurements, which resulted in a radon potential for new building sites. Radon-prone areas in Germany (45) (46) were determined directly from outdoor measurements and geological units. Correlations of indoor radon with geological units at regional and municipal scales were found in Norway (67). In addition, in Norway (66), very high radon concentrations (up to 50,000 $Bq/m^3$) were recorded in houses located on highly permeable sediments.

In many countries radioprotection institutions have established permissible legal values of indoor radon concentration, which are represented on maps. These thresholds are used as indicators for dwelling remediations and are also used to plan resampling campaigns. Indicator values and the scale of analysis vary from country to country and with repect to the European legislation. For instance, the Commission of the European Atomic Energy Community (19) recommends taking remediation measures on existing buildings exceeding 400 $Bq/m^3$ of indoor radon and preventive measures to not exceed 200 $Bq/m^3$ in new constructions. Additionally, Euratom recommends identifying regions, sites and building characteristics that are likely to be associated with high indoor radon levels in order to serve as a guideline for surveys.

In 2010, an updated map of indoor radon created with the arithmetic mean per square cell, of 10 by 10 km of data up to 2009 was collected on a European scale (16). It was the first initiative to make a map that is compatible with the existing data. Some countries have a better covering and large datasets while others have scarce information. In this European map (Figure 1.4), Switzerland appears to have comparatively higher indoor radon levels as well as other countries having large mountainous regions or granitic lithology, such as the Czech republic, Austria, Finland or the north of Italy.

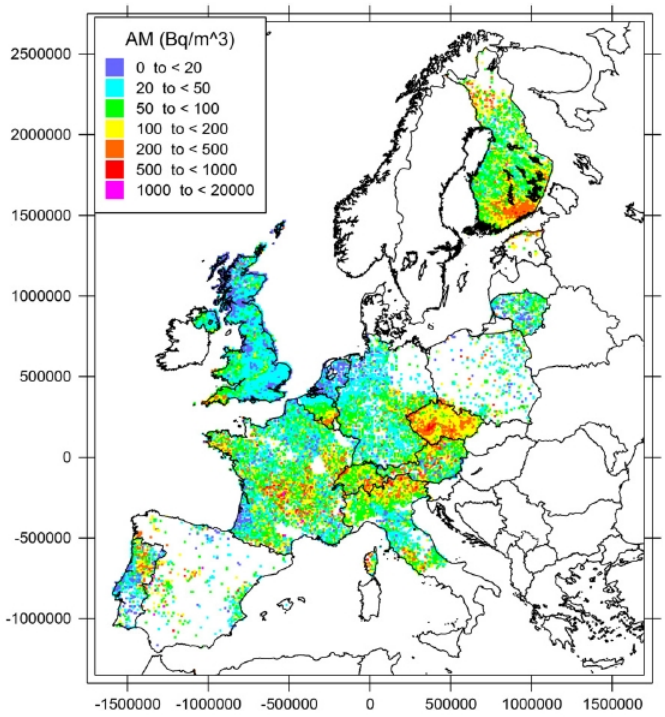



Figure 1.4: *Arithmetic means of indoor Rn concentration in ground floor rooms in 10 by 10 km cells (16)*

## 1.6 Indoor radon mapping in Switzerland

### 1.6.1 Public health mapping

Since 1981, the public health authorities in Switzerland have conducted and promoted several campaigns to measure the level of this radioactive gas inside dwellings. The indoor radon cadaster, finalized in 2004, accounted for around 60,000 measurements (25). By 2008, it had increased to nearly 140,000 measurements (Exactly 139,482). Considering that each lodgment has an average of two measurements, around 70,000 dwellings are being monitored. Though it is a very large survey, it represents a relatively low percentage of the existing dwellings and buildings. If we consider that, in Switzerland, there are around 3,800,000 dwellings (54), the indoor radon survey covers no more than 2 % of the target units.

Concerning admissible values of indoor radon, Article 110 of the Ordinance on Radiological Protection (OraP) indicates that cantonal authorities should take the necessary actions to remediate buildings where 1000 $Bq/m^3$ is surpassed and to monitor cases where more than 400 $Bq/m^3$ is detected (62).

However, these thresholds are quite elevated from the point of view of radioprotection. In (10), the investigators found a statistically significant association between radon concentration and lung cancer, even when the analysis was restricted to people in homes with in-

door radon concentrations below 200 Bq/m$^3$. The risk of lung cancer was 20% higher (95% confidence interval 3-30%) for individuals living in dwellings with indoor radon concentrations between 100-199 Bq/m$^3$ (mean: 136 Bq/m$^3$) compared to those living in dwellings with indoor radon concentrations under 100 Bq/m$^3$ (mean: 52 Bq/m$^3$).

This research was based on a very large number of individuals, and its results are coherent with other surveys conducted in the United States and China. Based on this evidence, in 2009, the World Health Organization (WHO) proposed a limit of 100 Bq/m$^3$ as the directive value (82). The Federal Office of Public Health's (FOPH) has elaborated an alternative map to show how the risk panorama will change if the directive value is lowered to 100 Bq/m$^3$ instead of the actual Swiss directive of 200 Bq/m$^3$ (Figure 1.5) (57).

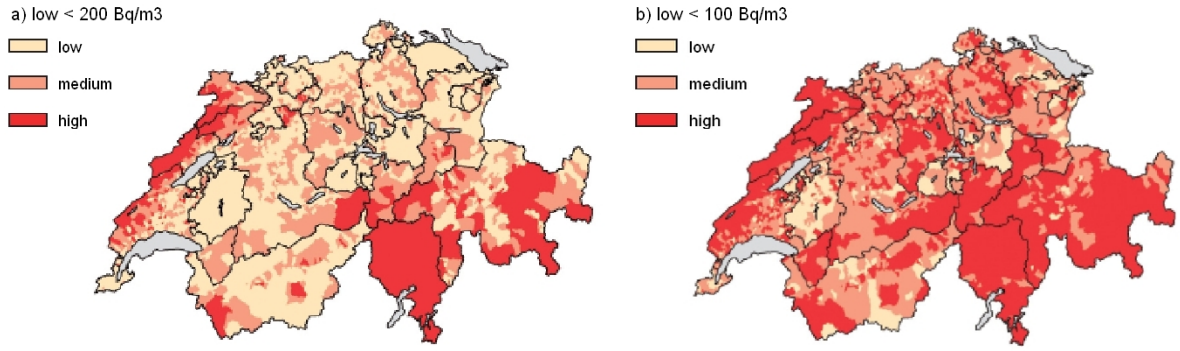



Figure 1.5: *Map of indoor radon risk considering the directive value of a) 200 Bq/m$^3$ and b) 100 Bq/m$^3$)*

Based on the measurements corresponding to inhabited dwellings, maps of mean indoor radon are produced by the FOPH on a municipality level. This information is made available via Internet by the public health authority and expressed on categorical levels of indoor radon (mean values that are below 100 Bq/m$^3$ are considered to have a low concentration of indoor radon, between 100-200 Bq/m$^3$ they are classified as medium and over the threshold of 200 Bq/m$^3$ they are considered to be areas with a high concentration. In Figure 1.6 the 2011 version of the indoor radon map for Switzerland (58) is presented.

### 1.6.2 Research on indoor radon mapping methods for Switzerland

Different geostatistical procedures and machine learning algorithms were proposed by the IDIAP research group for the analysis and mapping of indoor radon in Switzerland (Demyanov and Kanevski (12) (37) (39)). Based on their experience with radiation modeling after the Chernobyl accident (44), (43), the IBRAE (Institute of Nuclear Safety of the Russian Academy of Sciences) developed a pool of methods and software. The development of a comprehensive package for spatial data analysis called the Geostat Office Software (GSO) was of particular relevance (36).

This package includes not only current methods of geostatistics, as kriging and simulations, but also exploratory tools and machine learning algorithms. In the simplified structure

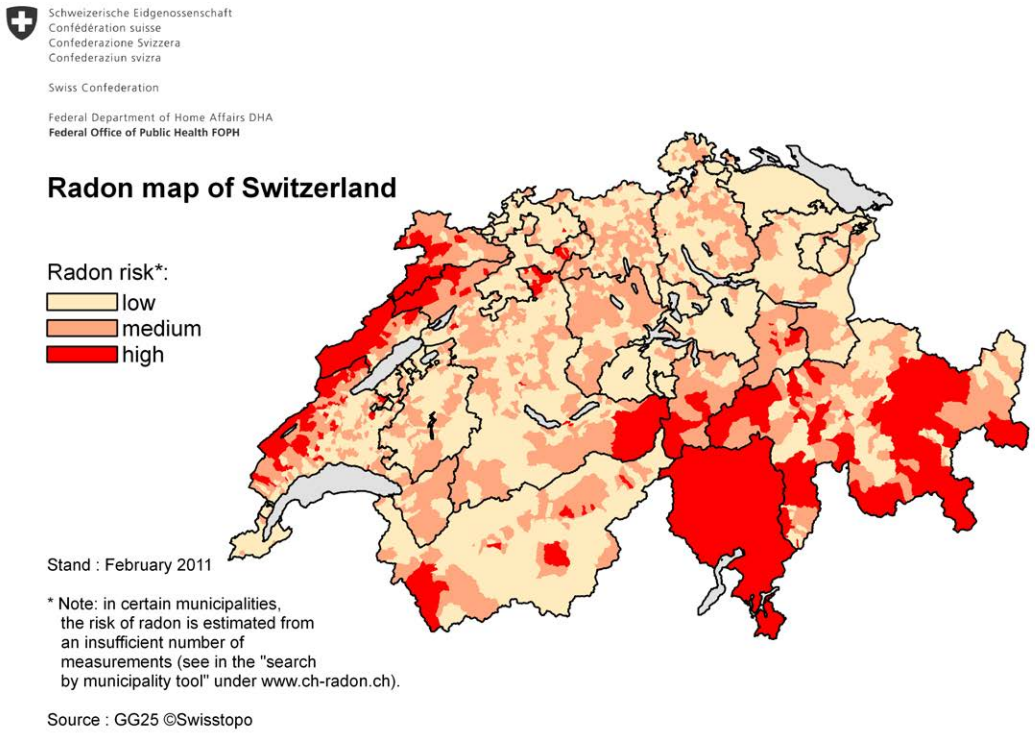



Figure 1.6: *Radon map of Switzerland, version 2011*

scheme of the software (Figure 1.7), the developed tools are presented in sections.

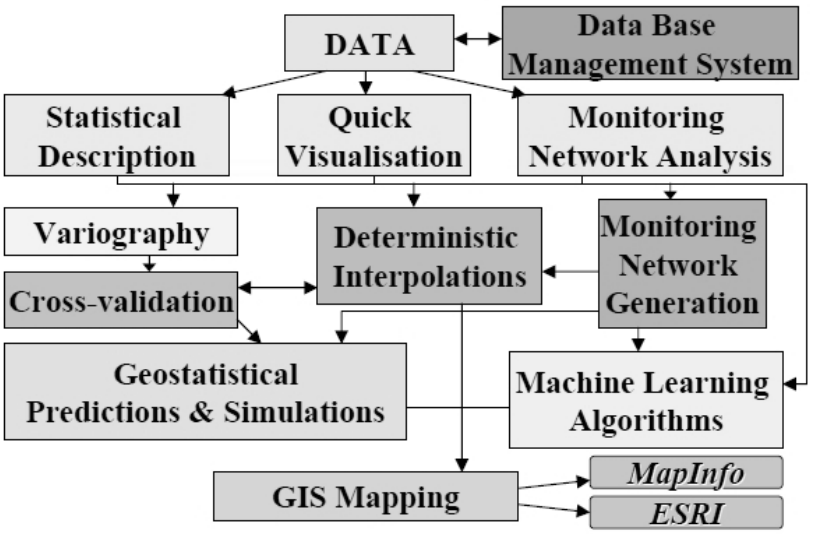



Figure 1.7: *The GSO software and tools structure scheme (34)*

Within the so-called monitoring network analysis and generation, many procedures based on fractality and clustering were included . Many other tools were implemented to help with data visualization and exploratory analysis. The implementation of machine learning programs such as the Multilayer Perceptron (MLP), General Regression Neural Networks (GRNN), Probabilistic Neural Networks (PNN) and Support Vector Machines (SVM) has been an attempt to provide different approaches to spatial data analysis and to combine such methods (41). Also of great relevance, was the proposal to analyze the properties of the spatial distribution of the 'monitoring network' or sample locations using fractality and other tools (39).

Other research on modeling indoor radon spatial distribution in Switzerland followed with an emphasis on the comparative use of methods (6) (38) (70) (28). The methods proposed are based on an 'efficient radon' concept, with which the information necessary for modeling indoor radon distribution is deemed to be contained within the data itself. The use of neural network tools as classifiers were also proposed as a way to simplify and automatize the task of indoor radon mapping (7) (72).

Among the tested methods, sequential Gaussian simulations appear to be a helpful tool for the task of indoor radon mapping (40) (68). The univariate distribution was also studied using extreme value theory (76), and the multivariate nature of the process was described (69).

## 1.7 Motivations

The review of the literature about indoor radon problem has motivated to focus the present research towards a number of interesting issues. From a statistical point of view, pollution by radon gas indoor dwellings is an interesting physical phenomenon because of the influence of many factors in the process. The accumulation of this radioactive gas inside dwellings is not only due to source exhalation factors but also due to human living conditions. Some authors have proposed to use these explanatory variables to normalize radon measurements as an expected radon concentration in a standard situation (21). Other authors have also proposed to refine spatial predictions using geological information (1) (51). Then, it will be important to perform a multivariate analysis of the Swiss indoor radon in order to determine the contribution of explanatory variables and the options to use them for spatial modeling.

Data skewness and spatial clustering express the complexity of the indoor accumulation process and poses a challenge for the spatial modeling task. In the research done for Switzerland and for other countries, there is a common view that measurements present a heavily skewed statistical distribution with the presence of extreme values (12) (15) (76). The Swiss dataset have shown to deviates from lognormality at a certain scale of analysis (76) while, for other countries, the lognormal distribution was used as a prediction model (1). In this sense, there is a need to adapt the modeling tools for the presence of extreme values.

For what concerns the spatial distribution of samples, indoor radon has usually showed a high spatial clustering. This property was successfully quantified using fractality and other

topological tools (39). The author also pointed out the significance of clustering (and declustering) on the estimation of histograms. In (70) the use of a domain constrained to the sampled area was used to avoid presenting results for unpopulated areas where the uncertainty prediction is high due to the lack of neighboring data. In (74) the authors have highlighted that, using fractality measurements, the spatial distribution of indoor radon samples can be considered as non-clustered if the predictive space were constrained to these populated regions. The alternative spaces of spatial distribution were called *validity domains*. The use of functional clustering methods as proposed in (37) can be used to characterize indoor radon for different sampling schemas. Determining the influence of spatial clustering over spatial estimations and the proper declustering solutions to obtain valid results are also main motivations for this thesis.

An additional problem for indoor radon Swiss data is that sampling density is not homogeneous at administrative levels. For instance, in Switzerland there are cantons that have a large number of samples while others have a lower amount. The same happens on a European scale, where important inter-country sampling schema differences can be found (15). The problem of finding a proper indoor radon mapping methodology considering a multiscale prediction framework is an interesting subject to be addressed in this research.

Regarding the application of spatial estimation methods, there are also many pending questions. The use of a variety of methods, going from geostatistics to machine learning, have been proposed for indoor radon (12) (39) (6). In particular, the potential of simulation methods were tested. Simulation methods have proved to provide less smoothed maps than regression methods (33) which is an advantage for extreme events like indoor radon (39) (38) (40). However, simulation and other prediction methods still require being adapted to the particular conditions of the indoor radon data. Model parameters and hyper parameters must be optimized by taking particular spatial and statistical distributions into account. In (68) the effect on neighboring parameters using constrained nets for simulations was analyzed. It was an attempt to optimize predictions neighboring parameters and to find a link with clustering of indoor radon data. Optimization should also work towards an automation of the modeling and mapping processes with the goal of handling large volumes of data. These requirements have helped to delineate the present research.

Finally, data transformation into categories can enlarge the options of spatial modeling when direct modeling of raw data is not feasible. Classification methods and neural network modeling are robust solutions to non-linearity and extreme values (7). The adaptation of these methods for indoor radon mapping also requires further research.

## 1.8 Objectives

Based on the exposed motivations a general objective and specific objectives were outlined for the present thesis. The general objective is to propose an operational modeling methodology for indoor radon mapping in Switzerland. The specific objectives are stated as follows:

- To analyze different factors in a multivariate perspective in order to measure their relation with the indoor radon concentration. To analyze in particular the correlation of

indoor radon with the geological factor.

- To test and propose exploratory tools to characterize indoor radon spatial clustering and to analyze the effects of declustering techniques.
- To make a multi-scale analysis in order to find optimal scales for indoor radon mapping.
- To propose procedures for the optimization of parameters and data transformation for modeling with regression, simulation and classification methods.
- To test and validate different prediction methodologies under different conditions of statistical distributions, spatial distributions and data transformations.
- To propose simulation scenarios to obtain a statistical and spatial distribution approaching realistic representations.
- To integrate the analysis and modeling methodologies in a workflow towards an automation of indoor radon mapping for decision-making.

## 1.9 Methodological overview

In Figure 1.8, a diagram of the conceptual methodology to be used in the thesis is presented. The indoor radon data analysis and modeling process are at the center of the graph. Modeling starts with the statistical univariate and multivariate data analysis (SUMDA) and the exploratory spatial data analysis (ESDA). These exploratory steps will help to optimize parameters and will feed the modeling process. The exploratory phase will also include simpler modeling tools that can help to define the optimal parameters for other more complex methods. The definition of the scales of analysis is also an important ESDA component.

Parameter optimization is an essential procedure to automatize modeling and estimation. An automatic processing chain can be created with the collaborative work of different modeling methodologies on a bottom-up complexity flow. As shown by the arrows, an analysis of the data can be done following alternative flows. It can go in the direction of linear predictions, towards simulation or towards classification. The linear prediction methods are pursuing precision in the results. The idea in this case, is to honor sample data without doing transformations. Nevertheless, linearity is not an assumption that always holds for data, and linear relations can be enhanced through the use of filtering procedures.

Simulations may require doing a nscore transformation in order to obtain Gaussian simulations. In this case, the prediction goal is extended to data and uncertainty modeling. Another alternative is to simplify and to transform data into categories to make use of classification methods. A further important component to assure good parameter definitions is to obtain error measures for the estimations and to evaluate the performance of methods as a function of data consistency (method robustness).

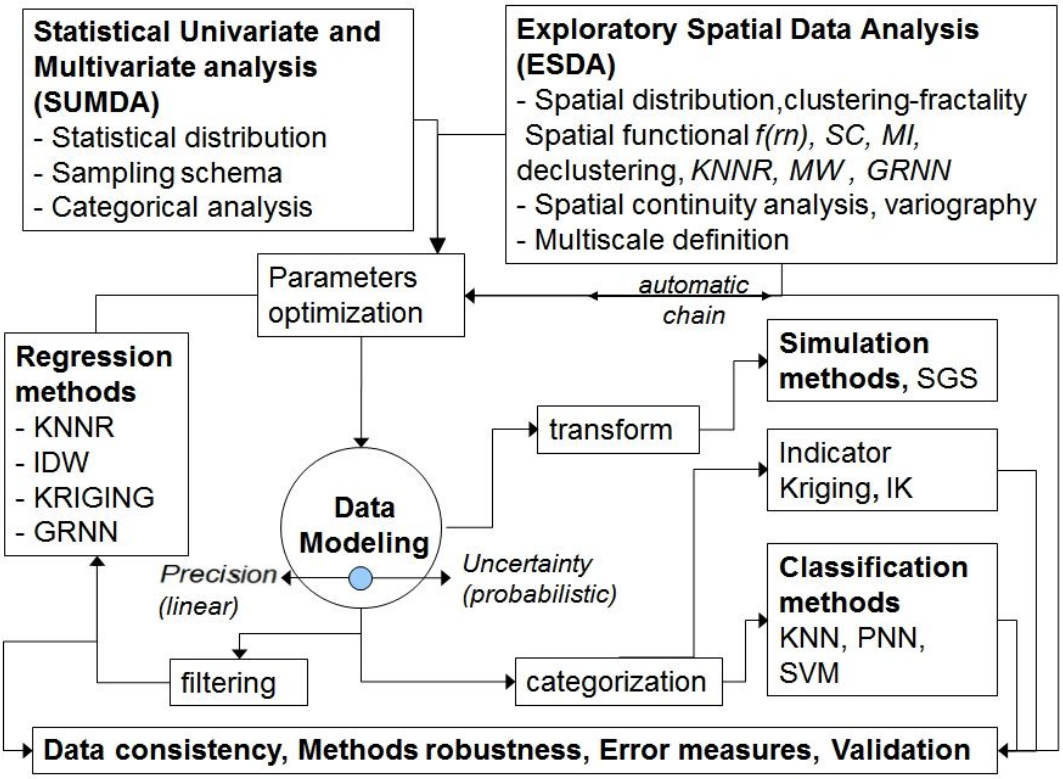



Figure 1.8: *Conceptual methodological diagram for indoor radon modeling*

# Chapter 2

# Statistical univariate and multivariate data analysis

## 2.1 Data description

The data used in the present thesis correspond to the collection for the entire Swiss territory up to 2008. It includes 139,483 records and is called the total set. It was provided by the Federal Office of Public Health (FOPH) with the restriction of not including coordinates in maps or making a detailed publication of data. It is always possible to consult the corresponding FOPH web page to obtain a detailed report about the indoor radon values per region.

For the purpose of interpolation analysis (from chapter 3 onwards) a set of data was selected considering samples taken at the ground floor of inhabited places. Data having the same coordinates, which correspond to repetitions in different rooms for the same dwelling, were removed. Then, the sampling points were superimposed with a coverage of polygons representing the populated areas in Switzerland.

### 2.1.1 Indoor radon build-up spatial domain

For the particular case of indoor radon measurements, the buildings spatial domain constrains the location of samples. As previously explained, indoor radon measurements are acquired inside buildings. Therefore, it is wise to conduct a trial on the characterization of this spatial domain.

Using the topographic maps of Switzerland at a scale of 1:25000, a buffer zone was defined around the representation of buildings to obtain an approximation of the built-up area. A buffer area with a radius of 120 meters was used. This distance gives more continuity to polygons after dissolving, and it seems coherent with a zone of urban expansion (considering that the information from charts are not completely up-to-date). After using the buffer parameter, a complex shape with many thousands of polygons resulted. Afterwards, this layer was simplified considering a vertex threshold, by dissolving the limits between adjacent polygons. The polygons were dissolved and simplified in order to have integrated

areas. The built-up domain is a good approximation of the buildings' land cover, and though it is a simplified version, it is already composed of 19,197 polygons. This domain is where indoor measurements are supposed to be contained and where predictions must be done with priority. The built-up area, thus calculated, covers 36% of the Swiss territory. A map of this area is presented in Figure 2.1.

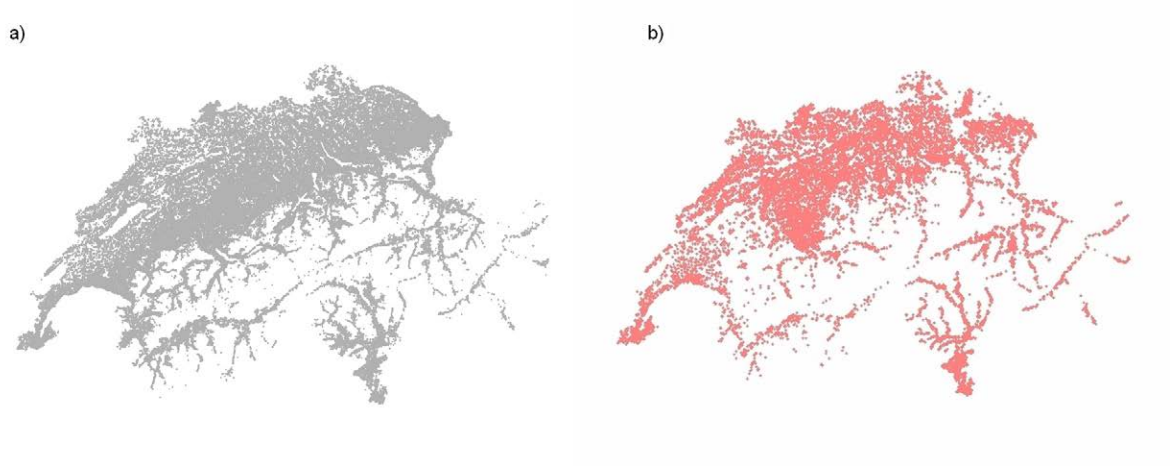



Figure 2.1: *a) Spatial domain of built-up areas for the Swiss territory, b) Indoor radon sampling coverage*

### 2.1.2 The Swiss indoor radon dataset for inhabited buildings and ground floor

The reselection of indoor radon samples contained within this built-up domain, considering the inhabited condition and ground floor, includes 41,787 samples. It is, therefore, shorter than the total dataset which included 139,483 records. This data selection is more coherent with the public health problem, since it corresponds to inhabited conditions, which also excludes measurements in cellars. Besides this, according to sampling policies, detectors should preferably be placed on the ground floor.

This dataset, with 41,787 samples, has a mean value of 163 Bq/m$^3$, a variance of 102,873, a maximum value of 15,045 Bq/m$^3$ and a skewness of 12.16. These data were subsequently used to conduct analyses at cantonal or natural regional levels.

### 2.1.3 The canton of Bern dataset

The canton of Bern has a dataset collected up to the year 2005. The canton of Bern's data is particularly interesting since it includes additional ancillary information about dwelling conditions. Besides indoor radon values, inhabited condition, floor level and the type of room, there is information about basement material, year of building, type of house, date of starting record and date of ending record. Unfortunately, this ancillary information was not recorded for all measurements, and in order to have a more detailed analysis of variables,

the database had to be cleaned with a resulting reduction in the number of measurements. The geotechnical unit, the elevation and the 1990 population variables were also added as was done for the total dataset.

This dataset has a mean indoor radon of 227 Bq/m$^3$, a standard deviation of 456 Bq/m$^3$ and a positive skewness of 9. Skewness is lower in comparison to the total dataset but data is still neither normally nor lognormally distributed.

### 2.1.4 The three data examples

Three data examples of the spatial distribution of points were used to illustrate the methods of spatial analysis. In Figure 2.2a there is a postplot for a toy dataset 1, with 400 points regularly distributed among 20 columns on 20 rows, with a 52.63 unit separation between points, giving a lattice of 1,000 by 1,000 units (19 separation distances by 52.63 units). Figure 2.2b shows a graph for toy set2, which consists of 400 points randomly distributed within a domain of 1,000 by 1,000 units. Finally, Figure 2.2c shows the distribution of a selection of 1310 indoor radon sample points (called set3).

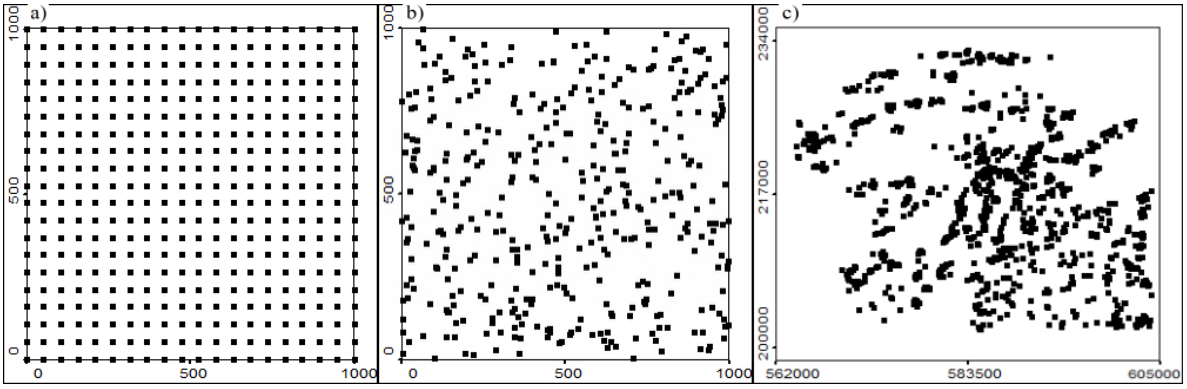



Figure 2.2: *a) points with regular distribution (toyset 1), b) points randomly distributed (toyset 2) c) subset of indoor radon sample points (toyset 3)*

### 2.1.5 The indoor radon set 3

Set3 is a square selection of indoor radon data from the canton of Bern dataset, which can be used to conduct a local spatial analysis. It is a real case study with lower values from the southern area and with higher values to the north in the Jura region. It is a heterogeneous selection, which is irregularly distributed and quite representative for the national dataset. The statistical distribution showed a high variance and a heavy skewness.

Set3 consists of 1,710 indoor radon measurements considering only inhabited dwellings and measurements at the ground floor. The spatial distribution of measurements is presented in Figure 2.3 superimposed over a topographic map of the area. It can be observed that distribution is irregular and clustered, and in general, it follows the distribution of urban

areas. The information presented is also restricted to cantonal limits. Agricultural land, forested areas, mountains and lakes are also not expected to have measurements.

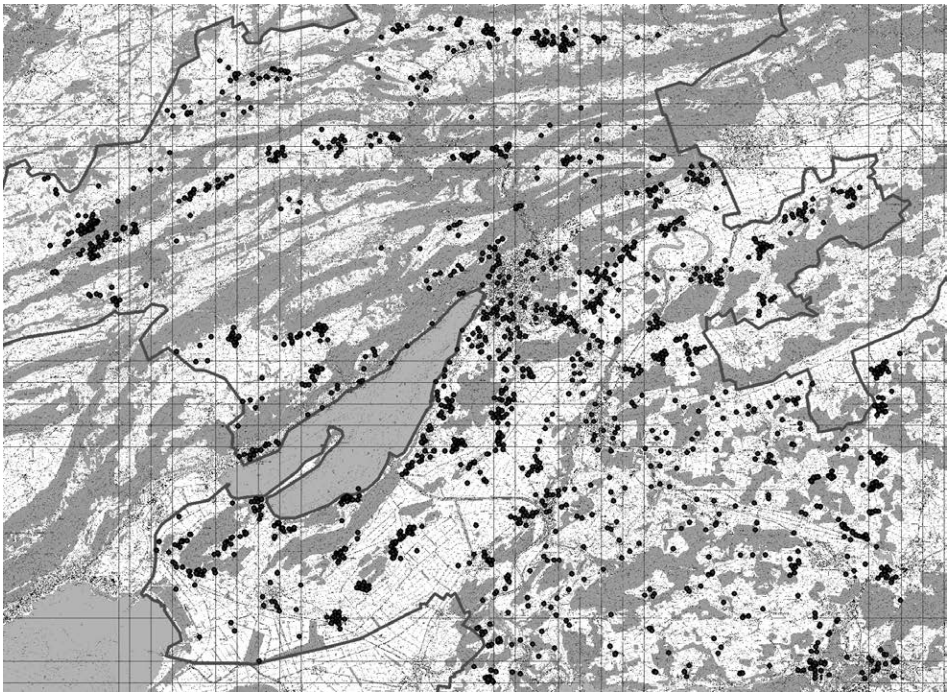


Figure 2.3: *Topographic map of the study area including location of samples for the toyset3*

The dataset was subdivided into a set of training data with 1,310 samples and a set for validation with 400 values. The set with 400 values from the same area was reserved for further independent validations. The spatial distributions of values for these two sets are presented in Figure 2.4

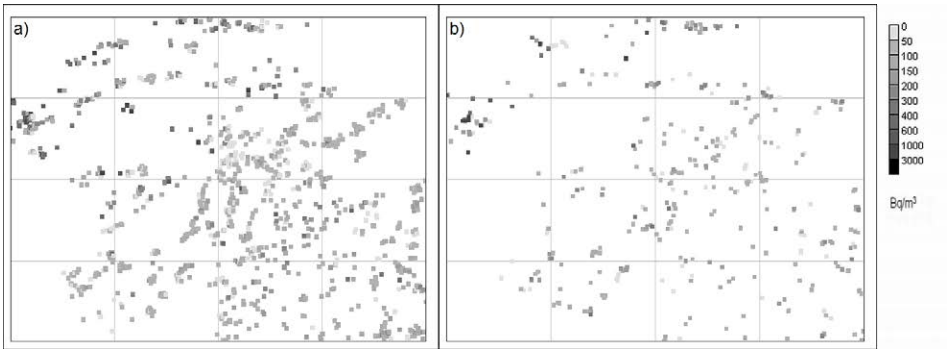



Figure 2.4: *Map of set3 sample locations for the a) training set and b) validation set*

As will be explained in chapter 3, set3 was partitioned into two subsets called set3A, and set3B. These subsets presented different statistical and spatial properties and helped to carry out a comparative analysis.

**Statistical distribution of the training set3**

An initial analysis of the datasets includes the calculation of the statistical parameters of data distribution. For indoor radon records, which contain values far from the median, a useful representation is given with boxplots (as shown in Figure 4.23). In this graph, points that exceed the one and a half interquantile range (the difference between the third and first quartile limit), counting from the third quartile may be considered as outliers if compared to a normal distribution. This group of high values contributes to the production of a high sample variance (46,929) in the dataset, while the sample mean is 142 $Bq/m^3$, the median is 92 $Bq/m^3$ and the skewness coefficient is 6.08.

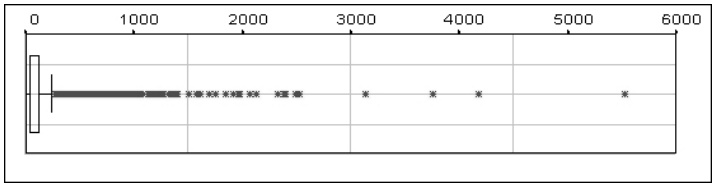



Figure 2.5: *Representation of values distribution by boxplot graph*

It is also necessary to create a probability density function (*pdf*) for the training dataset, which is used for modeling. As the distribution is heavily skewed, it will pose difficulties to carry out a graphical comparison with the histograms of results; therefore a selection of 95% of the data was used to make a graph for the core of the data as shown in Figure 2.6.

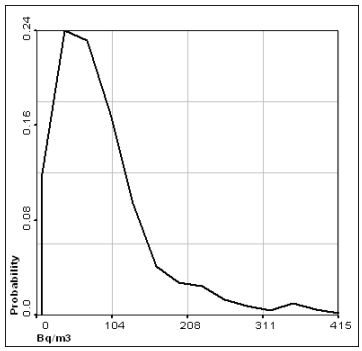



Figure 2.6: *Probability distribution of set3 training values, data is represented up to the 95th centile.*

## 2.2 Indoor radon sampling schema in Switzerland

### 2.2.1 Evolution of the sampling schema

The first step to understanding data is to analyze the conditions they were collected in, including the sampling design chosen and the number of samples. The Federal Office of Public Health (FOPH) of Switzerland is in charge of the radon program, which includes the measurement, mapping, regulations, training and communication related to indoor radon. The

FOPH has promoted the execution of these activities in all cantons in Switzerland and has centralized the results of its measurement campaigns since 1982. Fortunately, ancillary data were also collected along with the indoor radon values, and it can be used to analyze the evolution of the sampling design, not only by location, but also by inhabited condition and floor level of the building.

In the first campaign, the measurements were concentrated in the canton of Neuchatel, at the locality of La-Chaux-de-Fonds, where high values were found. The canton of Argovie also started to take measurements, which presented lower values. In Figure 2.7 a progression of the mean indoor radon levels between 1982 and 2008 is presented. It is apparent that the higher means were present in 1982, 1986 and 1991. As mentioned, in 1982, most of the measurements were taken from the Jura's mountainous region, while in the following years they were concentrated in the plain region, where it is more likely that lower values are to be found. In 1986, another important campaign was carried out in the Jura region, and in 1991 the campaign was concentrated in canton of Grisons, also a mountainous area. The elevated values of radon gas in certain regions can partially be explained by lithology. Boehm (3) found that areas with crystalline basement rocks, on average, show three times higher radon activity in water (35Bq/L) than regions in sedimentary basins (12Bq/L).

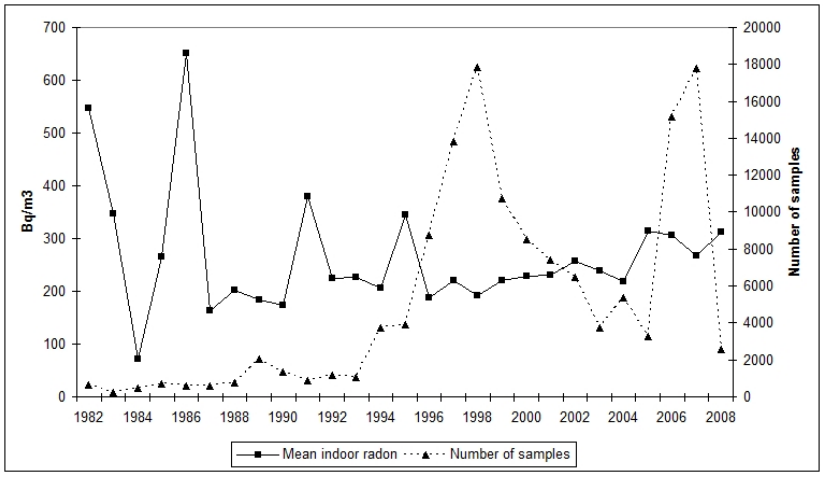



Figure 2.7: *Evolution of the mean indoor radon and the number of samples in Switzerland*

In the early years of the survey, an important percentage of samples were taken in vulnerable regions, and a high percentage of detectors were placed at floor level -1 (the cellars), as seen in Figure 2.8b. These factors can result in higher means, but the campaigns were reduced in the number of samples (as seen in Figure 2.7). Later, larger campaigns were carried out in 1998 and 2007. The 1998 campaign differs from the 2007 one in that surveys were previously more extensive and heterogenous. Communes with lower and higher risk were involved at the beginning, while lately, communes having higher mean values have increased the number of samples. A remarkable case is the canton of Ticino, where massive surveys were promoted in 2007 and 2008. Another difference between these two surveys was the inhabited condition of the building and the floor on which detectors were placed. As seen

in Figure 2.8a, the percentage of inhabited buildings increased during the 2007/2008 season, as well as the ground floor becoming the preferred level for detector placement (2.8b).

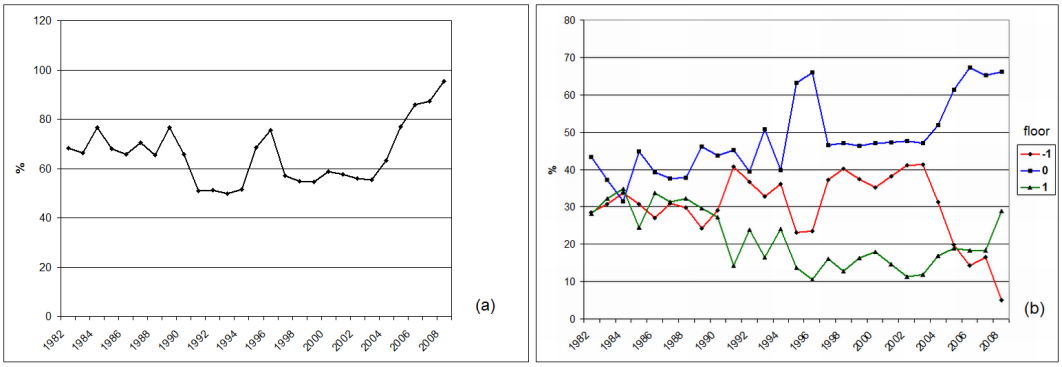



Figure 2.8: *a) Evolution of the percentage of inhabited buildings sampled and b) Evolution of the percentage of samples for three floor levels*

The indoor radon sampling strategy has evolved from a more random distribution towards targeting inhabited buildings and less cellars. This is more clearly illustrated in Figure 2.9.

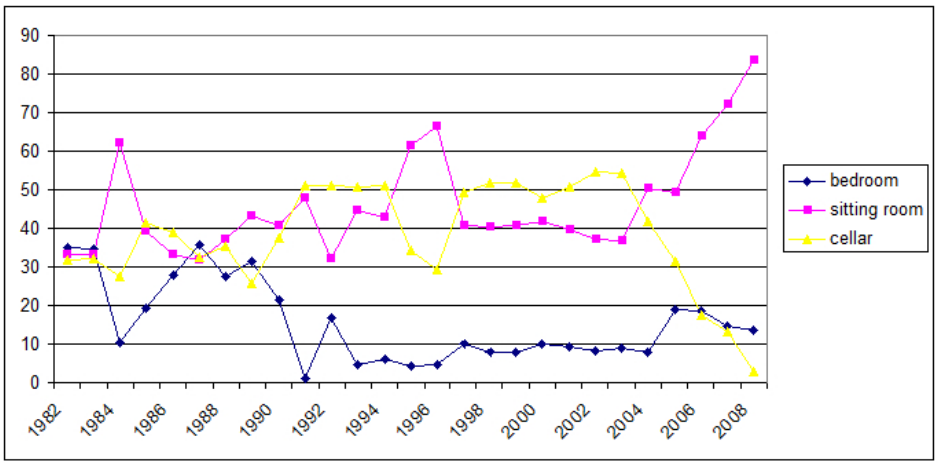



Figure 2.9: *Evolution of the percentage of samples by type of room*

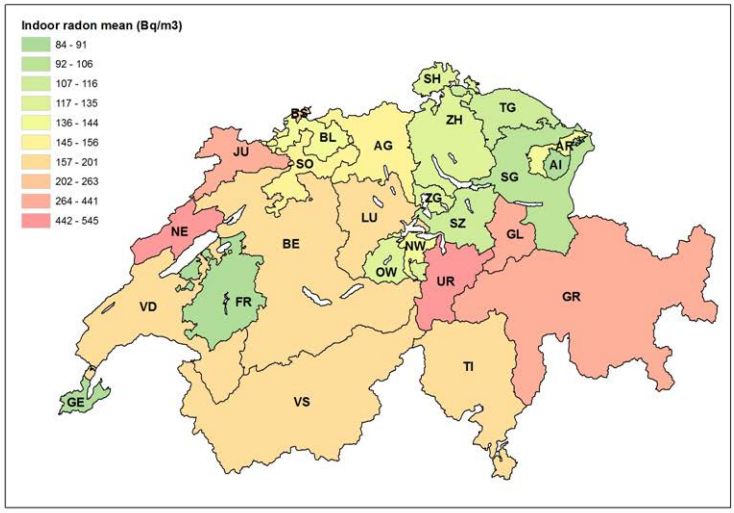
The percentage of sampled living rooms increased in comparison to cellars and bedrooms. The living room is probably the space that is least susceptible to variations during the measurement period.

### 2.2.2 Evolution of sampling between 2005 and 2008 on a cantonal level

As mentioned above, it is possible to see, from Figure 2.7, that the years 1998 and 2007 indicate the picks of two major sampling campaigns. Between these two campaigns, in 2005, there is a decline in the number of samples, which can be considered as a possible break point

in sampling strategy and design. Regarding the mean value per date of samples, early years appear to be more erratic. The first samples were taken in areas more affected by indoor accumulation and hence, the mean indoor radon for some of these early sets exceeded 500 Bq/m$^3$. In 1984 there was a sample mean under 100 Bq/m$^3$ (The Argovie campaign). In the following years, campaigns were generalized, providing a more constant mean value. More intensive campaigns were initiated in 1994, and from that year on, the mean indoor radon level detected had a tendency to increase. This is probably due to a preferential sampling on a cantonal level for more exposed cantons as previously explained.

Taking all the 139,483 available measurements, regardless of inhabited condition and floor factor, the indoor radon mean was calculated for each canton. In Figure 2.10 cantons are colored according to the indoor radon mean. In addition, a table with the ten cantons with most radon is presented. It should be made clear that some values appear to be particularly elevated because they include a significant number of samples taken in cellars. This measurement shows some variations in different regions.




| Canton | number samples | % of total samples | mean RN Bq/m3 |
|---|---|---|---|
| Ticino (TI) | 26823 | 19.2 | 201 |
| Bern (BE) | 21467 | 15.4 | 179 |
| Grisons (GR) | 17232 | 12.4 | 441 |
| Neuchâtel (NE) | 9481 | 6.8 | 545 |
| Zurich (ZH) | 8630 | 6.2 | 134 |
| Aargau (AG) | 7369 | 5.3 | 156 |
| Vaud (VD) | 6273 | 4.5 | 179 |
| Solothurn (SO) | 5276 | 3.8 | 149 |
| Jura (JU) | 5184 | 3.7 | 373 |
| Luzern (LU) | 4837 | 3.5 | 173 |

Figure 2.10: *Map of indoor radon mean per canton and the number of samples ranking table*

The Canton of Ticino is a region where the sampling strategy changed greatly towards massive sampling. The effect of this resampling campaign was a reduction in the mean of the samples from the period up to 2005 compared to the period between 2006 and 2008. For data of inhabited buildings and at ground floor, the mean of samples taken up to 2005 is 223 Bq/m$^3$ while samples from 2006 had a mean of 211 Bq/m3. What becomes evident, is that the national mean increased because of sampling done in exposed cantons like Ticino. On the contrary, within the canton of Ticino the mean had decreased, probably because more samples were taken in urban areas where dwellings appear to be less exposed. The spatial distribution of samples will be analyzed in more detail in the following chapter.

For the canton of Neuchatel, the tendency is quite different because the sample mean up to 2005 was 314 Bq/m$^3$ (only inhabited buildings and ground floor), while the mean for

the period between 2006 and 2008 increased to 462 Bq/m$^3$. The increment of the samples mean for later campaigns reflects a probable preferential sampling for regions or buildings deemed to be more exposed.

### 2.2.3 Influence of the inhabited condition and floor in sampling

If we consider the mean value of indoor radon per category of inhabited condition and floor variables, we can infer that there is also a global influence of these two factors. First, the mean indoor radon for samples taken in inhabited buildings (e.g. dwellings) is 183 Bq/m$^3$ while the uninhabited buildings (e.g. public places) have a mean of 363 Bq/m$^3$. Regarding floor level, the corresponding mean values for the categories -1 (cellar), 0 (ground floor) and 1 (first floor) are 315, 221 and 149 Bq/m$^3$ respectively. There can be, of course, an association of effects between factors. Say, for example, that detectors in uninhabited buildings were preferentially placed in cellars (which is, in fact, the case). Therefore, it is important to perform a multivariate analysis to study factors correlation. That analysis will be later presented.

From what is said above, it can be assumed that there is a possible bias in the sampling of higher indoor radon records. Cantons with higher indoor radon mean values have been given priority for massive surveys. The percentage of measurements in cellars is also important, and many inhabited buildings were surveyed. In a framework of health risk analysis, sampling should be constrained to the areas of exposure, which are areas where people are most likely to spend the majority of their time. Hence, inhabited buildings and ground floors were traditionally considered to be locations of common exposure to radon by health authorities.

### 2.2.4 Seasonal correction

In Switzerland, indoor radon evaluation is based on direct measurements inside houses, because this is considered to provide the most reliable information (55). The general recommendation is to situate the detector device away from air currents and to keep it for a period of three months during winter, since this is the season when air is most stable inside buildings, due to insolation. The period of measurement (the days between the starting date and the ending date), has a mean of 98 days for the Swiss dataset. Therefore, following the recommendations, the detectors were placed mainly during winter and fall (between October and March), mostly starting in the months of December and January (fig 2.11). Very few remained during spring and summer (April to September). This preferential sampling produces a bias that results in increased indoor radon values. Then, a correction is readily applied by the FOPH using the following formula:

$$A_0 = A_m \frac{N_{wi} + Nsu}{1.12 * N_{wi} + 0.88 * Nsu} \tag{2.1}$$

Where, $A_0$ is the standardized annual mean, $A_m$ is the indoor radon measurement, expressed in Bq/m$^3$; $N_{wi}$ is the exposition time in winter, and $N_{su}$ is the exposition time in

summer. As can be seen, the period of exposition in winter is penalized so that, in general, standardized indoor radon values are reduced. Figure 2.11 shows the monthly evolution of the number of samples and the standardized annual mean of indoor radon.

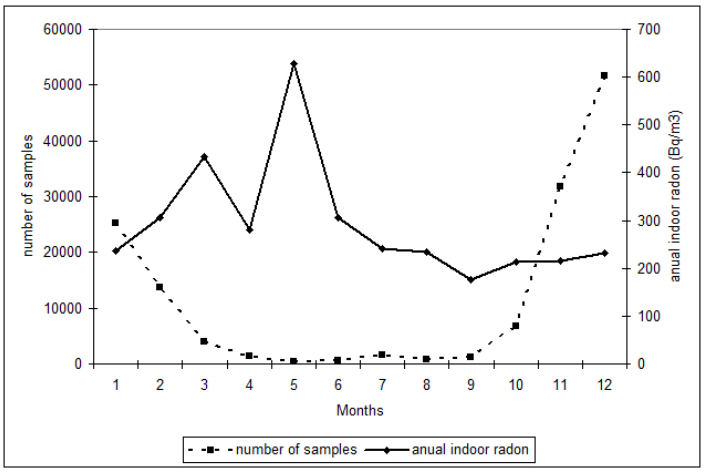



Figure 2.11: *Number of samples and mean indoor radon per month of starting date measurement*

As seen, indoor radon concentration in winter appears to be lower than in summer. The idea is to assume equal conditions inside a building (exposition, ventilation etc.). During the year, if conditions are normal, a higher presence of radon is expected in summer when exhalation from the soil is higher. Thus, the variable used in the present study is the corrected annual indoor radon level, however, from here on, it will be referred to, as the indoor radon value for simplicity.

## 2.3 Univariate and categorical analysis of the indoor radon dataset

### 2.3.1 Univariate analysis of the indoor radon dataset

The dataset used for the present research corresponds collection for the whole Swiss territory to the up to 2008, which includes 139,483 records. The mean for the total dataset is 243 Bq/m$^3$ and the median is 95 Bq/m$^3$ with a standard deviation of 659 Bq/m$^3$, a variance of 434,761, a distribution skewness of 28 and a kurtosis of 2,724. The dataset is highly positively skewed and the distribution is far from normal and also not lognormal, as seen in the Q-Q plots in Figure 2.12 .

Figure 2.12a shows that the positive skewness is due to outliers with very high values, with a maximum record of 90571 Bq/m$^3$. In Figure 2.12b deviations from lognormal distribution are also reflected in the lower and upper tail. The high kurtosis indicates a heavy clustering of centralized values; this is why the lower tail deviates from normal distribution. Nevertheless, in the case of the logarithmic transform, deviations are smaller. Does that mean that distribution will approximate a lognormal? Is it an extreme type of distribution? These questions were well addressed in the work of Tuia (76). Based on the analysis

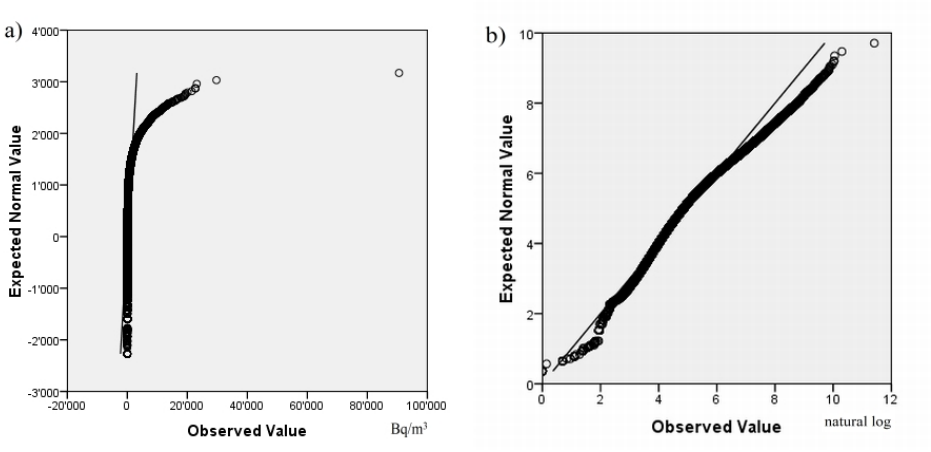



Figure 2.12: *a) Q-Q plot for the indoor radon distribution and b) Q-Q plot with natural logarithm transform*

of 29,000 records from ground floors in inhabited buildings they observed that the total data was better approximated by a lognormal than by a Gumbel EVT (Extreme Value Theorie) distribution function, due to this heavy lower tail. Nevertheless, the right heavy tail (values over 200 Bq/m$^3$) was better approximated with an EVT and an exponential distribution function. They have also observed that when dividing data by regions, only regions with high radon concentration and only right tails can be better modeled with EVT. In such cases, it was the lognormal distribution that had the best fitting.

To verify if some of the statements of the referred study hold also for the new 2008 database, a number of analyses were done. A set of around 61,000 records for ground floors and inhabited buildings was used. The mean value was 189 Bq/m$^3$ and the median 90 Bq/m$^3$. The distribution skewness was 10 and the kurtosis was 200 (slightly lower than for the total dataset). The limit value, for what is considered the outlier value, is 365 Bq/m$^3$. This limit is calculated adding 1.5 times the interquartile range to the third quartile limit. In turn, the interquartile range is the difference between the third quartile and the first quartile limit. The values over 365 Bq/m$^3$ (approx. 10% of the data) were taken out to see their distribution without these so called 'outliers'. This group of data was still positively skewed with a value of 1.3 and a kurtosis of 1.2. In Figure 2.13a, a histogram for data lower than 365 Bq/m$^3$ with a normal curve on top is presented. There is a clustering of values around 50 Bq/m$^3$, which is reflected by a light kurtosis. In fact, a value of 3 is often considered to decide whether or not a distribution is Leptokurtik. The positive skewness illustrates the disagreement with normal distribution.

A Kolmogorov-Smirnov test of normality was run for log transformed indoor radon data to see if they were lognormally distributed. The result was that statistically it cannot be considered normal, but it was closer than raw data. This approximation can be seen graphically in Figure 2.13b; here the lower and higher heavy tails are less conspicuous but can be also identified. The same test of normality was also run for the dataset below 365 Bq/m$^3$, with the same results. Data distribution by moving windows and at regional levels

are also important points to examine, and they will be addressed in the chapter dedicated to spatial analysis.

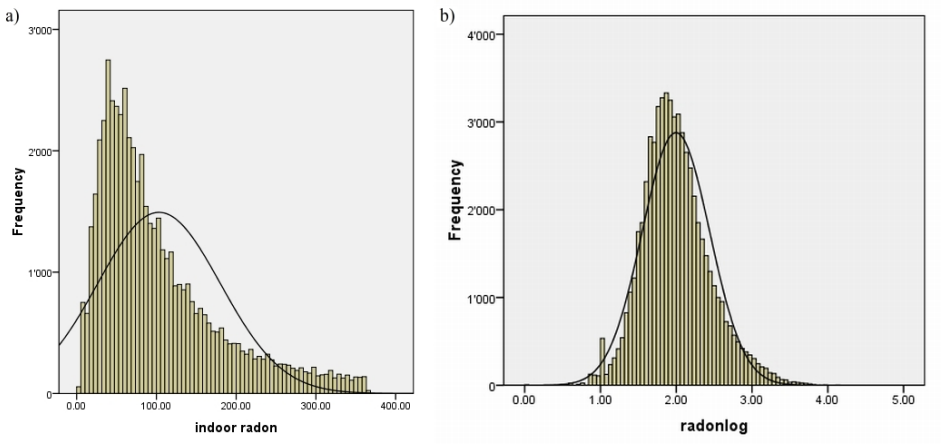



Figure 2.13: *a) histogram for values below 365 Bq/m$^3$ with normal curve b) logtransfomed histogram with normal curve*

### 2.3.2 Analysis of the total indoor radon data per category

For the main portion of samples in the total dataset, additional and interesting information about floor level, inhabited building condition and type of room where the detectors were placed have been recorded. In addition to these variables, it was possible to incorporate information about the corresponding geotechnical unit and the elevation for each sample. For these two variables it was possible to find maps with a good definition of classes, in order to obtain more detailed information. These variables also correspond to environmental factors, which are interesting to analyze and compare to dwelling conditions.

First, the mean indoor radon levels were calculated per category to see the variable's behavior at a glance (Table 2.1). This table includes the first and the second most frequent categories (the one having more cases) and then the category with the maximum mean indoor radon level and the corresponding mean values.

Table 2.1: *Mean indoor radon per variable and categories*

| Variable | most frequent categories and mean radon (Bq/m3) | | | | category with the highest mean radon | | |
|---|---|---|---|---|---|---|---|
| | category | mean | category | mean | category | mean | % samples |
| **habited** | inhabited | 183 | uninhabited | 363 | uninhabited | 363 | 33.1 |
| **floor level** | ground floor | 221 | cellar | 316 | floor -2 | 602 | 0.3 |
| **type of room** | living room | 160 | store room | 358 | store room | 358 | 29.9 |
| **altitude (meters)** | 190-500 | 172 | 501-1000 | 226 | 1001-1500 | 528 | 12.0 |
| **geotechnical** | GT3 | 185 | GT6 | 256 | GT23 | 704 | 0.1 |
| **canton** | TI | 186 | BE | 155 | NE | 390 | 6.6 |

It is possible to say that, typically, samples were taken from inhabited buildings on the

ground floor, in the living room, from an altitude between 190 and 500 m.a.s.l, over the geotechnical unit 3, with sand and silt, and in the canton of Ticino (TI). For the maximum values, the categories are more or less in accordance with literature. In general, greater values were recorded in uninhabited buildings, at floor level -2, in store rooms, at altitudes between 1001 and 1500 m.a.s.l., and over the geotechnical unit 23, which corresponds to granites, quartz and diorites. It should be noted, that some categories with maximum radon values are scarcely represented in the samples. For instance, just 0.3 % of the samples are taken at floor level -2 and only 0.14 % of the samples are over the geotechnical unit 23. Regarding the cantons, Neuchatel has the highest indoor radon concentrations. As will be analyzed further, this is mainly due to geological factors.

Next, each variable was analyzed individually to observe differences in indoor radon concentrations per category. Some categories were binned together when they had a low number of samples. The floor level variable was previously simplified to consider ony the most sampled categories, that is levels -1, 0 and 1. The floor level exhibits significant differences in mean indoor radon values as is shown in the report (2.2) and in the Analysis of Variance (ANOVA) tables with the F-test (table 2.3).

Table 2.2: *Report table of indoor radon per floor level category*

| floor level | N | Mean | Median | Std. Deviation | Skewness |
|---|---|---|---|---|---|
| -1 | 40787 | 3.25E+02 | 116 | 889.84 | 31.69 |
| 0 | 71000 | 2.21E+02 | 94 | 555.42 | 14.01 |
| 1 | 24631 | 1.50E+02 | 73 | 316.67 | 12.44 |
| Total | 136418 | 2.39E+02 | 95 | 647.47 | 29.63 |

Table 2.3: *ANOVA table for indoor radon per floor category*

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| **indoor radon** | **Between Groups** | **5.21E+08** | **2** | **2.61E+08** | **627.183** | 0 |
| *** floor level** | Within Groups | 5.67E+10 | 136415.00 | 415404.19 | | |
| | Total | 5.72E+10 | 136417.00 | | | |

The F-test indicates that the null-hypothesis can be rejected with very high probability since it is far below the F-value. The null hypothesis states that the floor level groups have similar indoor radon means, therefore it is possible to say that the concentration of radon differs from floor to floor. The mean test per category was also conducted for the variables type of room, geotechnical units and binned elevations, also showing significant statistical differences between categories.

For the inhabited condition, the inhabited category has a mean value of 183 $Bq/m^3$ while the mean value for the uninhabited category is 363 $Bq/m^3$. What comes to mind, is that perhaps this high significant difference can partially be explained by the fact that most samples from uninhabited buildings were taken in cellars (floor = -1). As seen in the floor level analysis, cellars have an indoor radon mean value of 325 $Bq/m^3$, while the mean values for the levels 0 and 1 are 221 and 150 $Bq/m^3$ respectively (table 2.2). Facing this preferential sampling for uninhabited cellars, the next logical question that arises concern the influence

of both factors over indoor radon concentration. The inhabited and uninhabited categories were analyzed independently considering floor levels -1, 0 and 1. The F-test indicated significant radon concentration differences between floor levels for both inhabited conditions. In table 2.4, a report of values per floor category for the inhabited group is shown. In this group, there is, as expected, an increase in indoor radon for higher floors. What is surprising, is that the mean and median values for floor 0 are higher than those for floor -1 in the case of uninhabited buildings, which can be seen in the following table (2.5).

Table 2.4: *report table for indoor radon per floor category for inhabited buildings*

| floor level (binned) | Mean | Median | N | Std. Deviation | Skewness |
|---|---|---|---|---|---|
| **-1** | 277 | 115 | 5489 | 586 | 8.4 |
| **0** | 189 | 90 | 61198 | 385 | 10.3 |
| **1** | 146 | 73 | 23996 | 302 | 13.2 |
| **Total** | 183 | 86 | 90683 | 382 | 10.7 |

Table 2.5: *report table for indoor radon per floor category for uninhabited buildings*

| floor level (binned) | Mean | Median | N | Std. Deviation | Skewness |
|---|---|---|---|---|---|
| **-1** | 333 | 116 | 35024 | 931 | 31.9 |
| **0** | 441 | 134 | 9106 | 1153 | 9.0 |
| **1** | 316 | 105 | 537 | 682 | 5.2 |
| **Total** | 355 | 119 | 44667 | 979 | 24.6 |

This particular feature of samples from uninhabited buildings has a probable explanation coming from the analysis of the third ancillary variable. When looking at the type of room, it is evident that most samples for the uninhabited category from floor 1 correspond to those of store rooms (77%) as well as to corridors and other categories. It is possible that these places are more exposed to gas accumulation and that they are often less ventilated than other rooms. We can speculate that perhaps store rooms on floor 0 are less ventilated than those on floor -1. In fact, many cellars in Swiss buildings have small windows working as ventilation outlets. This influence is not certain, but what is evident, from literature, is the influence of the ventilation factor.

### 2.3.3 Indoor radon in inhabited dwellings and on the ground floor

In the scope of public health protection, indoor radon distribution analysis must be restricted to inhabited conditions, in other words, to places where people are most exposed. Thus, samples from cellars should not be taken into consideration. Only 2% of the samples were taken in the inhabited floor level -1 that corresponds also to bedrooms. Concerning floor levels over the ground floor (level 0), the F-test have shown significative mean indoor radon value differences between levels 0 and 1. On the contrary, floors 1 and 2 are very similar (as seen in table 2.6) and therefore can be analyzed together. Samples from levels 1 and 2 are important since they account for a quarter of the inhabited samples, and they can eventually be used after normalization or as new variables.

Table 2.6: *ANOVA table for indoor radon per floor levels 1 and 2 at inhabited buildings*

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| **indoor radon** | **Between Groups** | 2942.164 | 1 | 2942.164 | 0.032 | 0.858 |
| *** floor level** | **Within Groups** | 2.13E+09 | 23238 | 91562.696 | | |
| | **Total** | 2.13E+09 | 23239 | | | |

Then, for the purpose of interpolation analysis, a set of data was selected considering samples from inhabited places and taken on the ground floor. Meanwhile, it was considered to be important to use the full dataset for the multivariate analysis.

### 2.3.4 Analysis of the Bern indoor radon data per category

On a first analysis, the dataset was split int subsets according to certain categories that are influential according to literature and which were recorded for all of the 15,456 measurements: basement material, floor and inhabited condition. This was just a trial to observe the behavior of the statistical distribution of indoor radon values after a re-selection for specific categories. In this way, the original set of 15,456 was reduced to 2,106 when re-selecting measurements in buildings with concrete basements. The corresponding frequency distribution graph is shown in Figure 2.14b. The set was then resized to 809 with a re-selection for three categories: dwellings with concrete basements, on the ground floor and inhabited. In Figure 2.14c, the frequency distribution is presented. The categories chosen, relate to lower indoor concentrations as seen in the analysis of the total dataset, therefore a reduction of the skewness was expected.

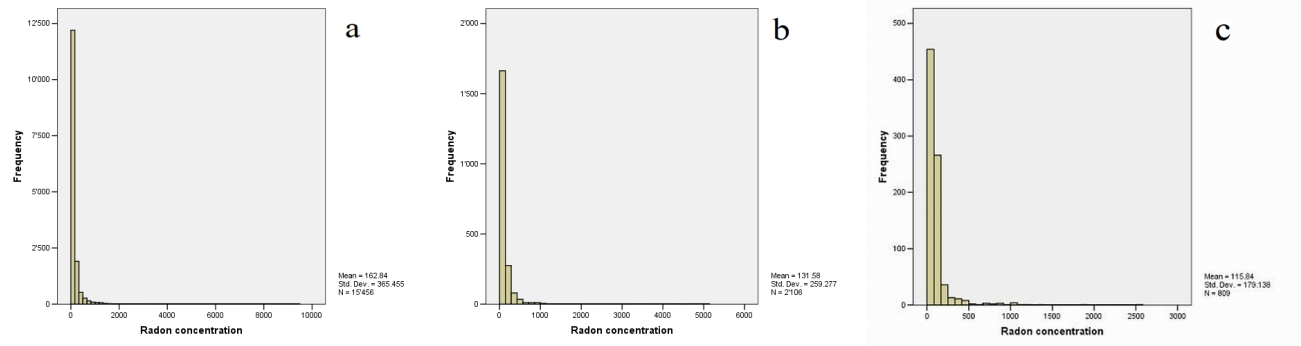



Figure 2.14: *(a) indoor radon measurements for canton of Bern , (b) concrete basement category and (c) concrete-ground floor-inhabited data distribution*

According to this simple analysis, indoor radon data distribution appears to be a very robust population of values. Even after splitting data into subsets, the distribution remains similar. Despite the mean and the standard deviation being reduced, data population retains a high positive skewness. An analysis of case summaries were then performed for indoor radon values, against independent variables, to see their influence at a glance (Table 2.7. This table includes the most frequent category (the one having the most cases), the categories hav-

ing the minimum and the maximum mean of indoor radon and the corresponding values.

Table 2.7: *Mean indoor radon per variable and categories*

| Variable | most frequent categories and mean radon (Bq/m3) | | | | category with the highest mean radon | | |
|---|---|---|---|---|---|---|---|
| | category | mean | category | mean | category | mean | % sam |
| **Habited condition** | uninhabited | 268 | inhabited | 168 | uninhabited | 268 | 59 |
| **Floor level** | -1 | 266 | 0 | 205 | -3 | 2390 | 0 |
| **Type of room** | cellar | 267 | living room | 164 | children's | 298 | 0.5 |
| **Altitude cutoffs** | 600-800 | 193 | 800-1000 | 216 | 1200-1400 | 482 | 5.6 |
| **Geotechnical unit** | GT3 | 169 | GT6 | 220 | GT28 | 763 | 0.3 |
| **Type of building** | farm | 201 | individual | 267 | school | 354 | 2.8 |
| **Year of construction** | 1801-1899 | 217 | 1900-1909 | 209 | 1950-1959 | 318 | 5.6 |
| **Type of basement** | natural soil | 221 | concrete pav | 185 | sanitary fill op | 561 | 0.5 |
| **Type of dossimeter** | E | 221 | G | 706 | G | 706 | 1.2 |
| **Population90** | 50 | 222 | 100 | 218 | 350 | 258 | 0 |

What can be seen from this first analysis is that most samples were taken from uninhabited buildings, at floor level -1, in cellars, at an altitude between 600 and 800 m.a.sl., and with a subsoil composed of gravel and sand. Concerning the building, the typical sample corresponds to farms, that were constructed between 1801 and 1899, and have natural soil floors in the basement. The most common type of dossimeter is the E type and the most frequent population density is 50 inhabitants per $km^2$.

This profile of sampling shows a certain tendency to target old uninhabited buildings where large levels of cumulated gas are often found. This sampling schema differs from the total dataset, where just 20% of samples corresponds to uninhabited places. We can presume, that it was easier to collect ancillary data in uninhabited buildings, since it is likely that people were more willing to provide information about such places. Therefore, it should be kept in mind that data is not representative of the total set, but it can help to understand the physical process of cumulation using vast information.

Not so surprising is that the highest mean natural radiation was found at the -3 floor level, in cases where built on top of open sanitary fills, when there is presence of gneiss and schist in subsoil or simply when the dossimeter used was of type G (not frequently used). According to table 2.7, some categories of variables appear to be more associated with higher radon values. The next question to arise is, which variables are best associated with indoor radon values?

## 2.4 Multivariate analysis of indoor radon

### 2.4.1 Multidimensional scaling (MDS)

The research done by Shepard and Kruskal (63) (47) initially proposed the use of monotone regression to achieve MDS. They introduced the concept of dissimilarity between a centroid and a vector representation of categories in order to analyze data labeled in categories. The

central idea was to obtain a model with a monotone relationship, either ascending or descending, between the distances in the data configuration and a dissimilarity model where a configuration is the representation of $n$ objects $x_1, \ldots, x_n$ by $n$ points in a $t$ multidimensional space.

The problem of multidimensional scaling, broadly stated, was to find $n$ points whose interpoint distances matched, in some sense, the experimental dissimilarities/similarities of $nf$ objects. Shepard showed that simply by requiring a satisfactory similarity monoticity, without making use of variability and distributional assumptions in any way (as is the case for PCA), one obtains very tightly constrained solutions. In other words, he showed that the rank order of the dissimilarities is itself enough to determine the solution for MDS.

In the methodology proposed by Kruskal (47), the distances measured between two objets or samples $i$ and $j$ (symbolized by $d_{i,j}$) must fit to a model with monotone ranked dissimilarities $\delta_{i,j}$. In order to see how well the distances match the experimental dissimilarities, he presented a scatter diagram of the $M$ possible combinations of $n$ objets (Figure 2.15a) where $M$ is also the number of dissimilarities/distances between $n$ objets so that $M = n(n-1)/2$. As shown in the scatter diagram, each star corresponds to a pairs of points. Star $(i, j)$ has abscissa $d_{i,j}$ and ordinate $\delta_{i,j}$.

The author emphasized from the beginning that instead of being dissimilarities, the experimental measurements may be similarities, confusion probabilities, interaction rates between groups, correlation coefficients or other measures of proximity or dissociation of the most diverse kind. It should be also noted that similarities can always be replaced by dissimilarities (for example, replace $\delta_{i,j}$, by $k - \delta_{i,j}$,). Since the procedure uses only the rank ordering of the measurements, such a replacement does harm to the data.

Now, it is possible to see that there is no exact linear relation between distance $d_{i,j}$ and dissimilarities $\delta_{i,j}$. The solution proposed by Kruskal was to chose a monotone sequence of numbers $\hat{d}_{i,j}$ as "nearly equal" to the $d_{i,j}$, as possible. In terms of the scatter diagram, this means that as we trace out the stars one by one from bottom to top, we always move to the right, never to the left.

By 'nearly equal' it was implied that the goal of MDS is to concentrate on reducing the total differences between a monotone sequence of numbers $\hat{d}_{ij}$ chosen as nearly equal to the distances $d_{ij}$. The cost of adjusting to monotonicity was called the raw Stress $S^*$ and was calculated as follows:

$$S^* = \sum_{i<j} (d_{ij} - \hat{d}_{ij}) \tag{2.2}$$

This stress must also be normalized in order to have a comparative measure without the effect of data scaling. Thus, the definition of the normalized stress is:

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})}{\sum_{i<j} d_{ij}^2}} \tag{2.3}$$

Figure 2.15b shows the scatter diagram of monotone ranked points $\hat{d}_{ij}$ approaching $d_{ij}$. The normalized stress measure was minimized, in this example, to 10 %.

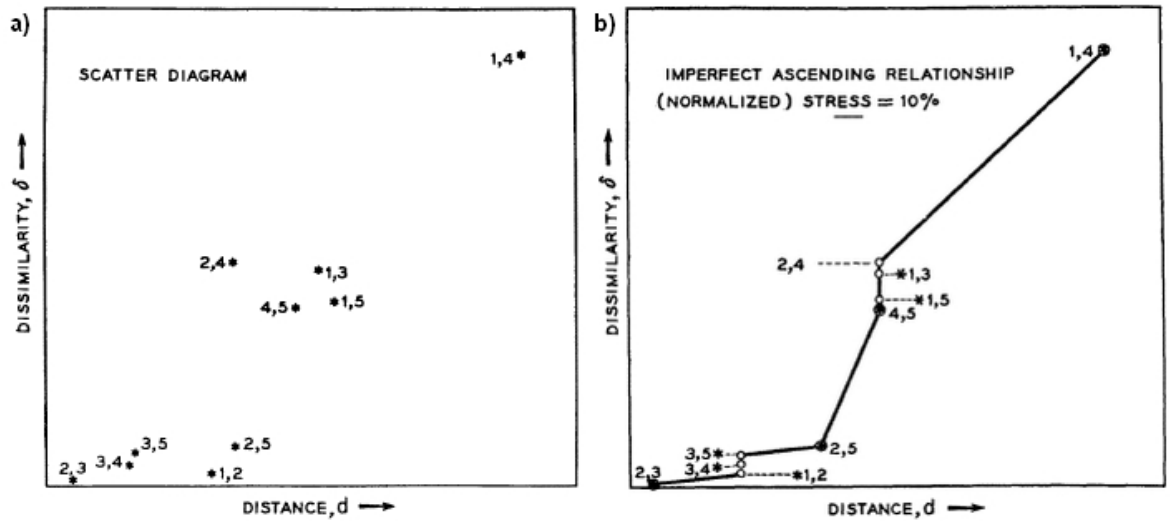



Figure 2.15: *a) scatter diagram of distances against dissimilarities b) monotone ranked points $\hat{d}_{ij}$ approaching $d_{ij}$. Figures taken from (47)*

### 2.4.2 Principal Component Analysis (PCA) and Non-linear PCA (NLPCA) methodologies

The PCA optimization problem was originally defined by Pearson in terms of finding low-dimensional subspaces (lines and planes) of best (least squares) fit to a cloud of points, and connecting the solution to the principal axes of the correlation ellipsoid. This concept was presented in modern notation by DeLeeuw (11) as minimizing the squared sums of differences between the transformations to defined subspaces $Y$ and the product of object scores $X$ by the component loadings $B$:

$$\text{SSQ} = (Y - XB') \tag{2.4}$$

In standard PCA object scores $X$ and transformations are fixed while component loadings are optimized to fit objects to a transformation based on maximum correlations. In NLPCA the loss function $\sigma(X, Y_j)$ used in the Gifi algorithms (49) it is also based on object scores $X$ and transforms $Y$, but optimal transformation is done for each variable $j$ within a common loss function. The general criterium for the transformation is to obtain categorical quantifications with a certain monoticity (as stated by Shepard and Kruskal). If the objective function minimization is expressed as:

$$\sigma(X, Y_j) = J^{-1}\text{SSQ}(X - G_j Y_j) \tag{2.5}$$

The strategy is to minimize the loss function $SSQ$ simultaneously over object scores $X$ and variable quantifications $Y_j$. At the first step the loss function is minimized with respect to $Y_j$ for fixed $X$. Essentially fitting the multivariate linear model $X = G_j Y_j +$ error for each $j \in J$. Where $G_j$ are indicator or 'dummy' matrices with entries $G_j(i, t) = 1$, if object $i$

belongs to category $t$ and $G_j(i,t) = 0$ if it belongs to some other category. In the second step of the algorithm, the loss function is minimized with respect to $X$ for fixed $Y_j$s:

$Y_j = G_j X +$ error. This strategy was called alternating least square with optimal scaling (ALSOS) and is used in the CATPCA procedure coded in the SPSS program.

### 2.4.3 Categorical Principal Components Analysis (CATPCA)

The CATPCA method gives a measure of association between variables of different type. The CATPCA method is implemented in SPSS statistical package and described in the user's documentation (65). This procedure simultaneously quantifies categorical variables while reducing the dimensionality of the data. The goal of principal components analysis (PCA) is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables. The technique is most useful when a large number of variables prohibits effective interpretation of the relationships between objects (subjects and units). By reducing the dimensionality, few components are interpreted rather than a large number of variables.

#### CATPCA objective function

CATPCA has an equivalent goal of PCA, namely to reduce the dataset to a smaller number of uncorrelated summary variables (principal components). This goal is achieved by finding linear combinations that explain, as much as possible, the variance in the data. Analogously to the situation in multiple regression, CATPCA focuses on the relationships between sets; any particular variable contributes to the results only inasmuch as it provides information that is independent of the other variables in the same set.

When all variables in a dataset are numeric and relationships are considered to be linear, PCA and CATPCA will produce exactly the same result. The difference between the methods is that CATPCA can reveal not only linear but nonlinear relations by quantifying categorical or non linearly related variables in a way that is optimal for the PCA goal. This quantification is done through an optimal scaling of variables at different levels as described below. The CATPCA procedure quantifies categorical variables using optimal scaling, resulting in optimal principal components for the transformed variables.

Initially, the variables in each set are linearly combined such that the linear combinations have a maximal correlation. Given these combinations, subsequent linear combinations are determined that are uncorrelated with the previous combinations and that have the largest correlation possible. There are many methods to calculate uncorrelated variables, but basically, the first component axis is a line with the least sum of squares from objects.

The optimal scaling approach expands the standard analysis in three crucial ways. First, it allows more than two sets of variables. Second, variables can be scaled as either nominal, ordinal, or numerical. As a result, nonlinear relationships between variables can be analyzed. Finally, instead of maximizing correlations between the variable sets, the sets are compared to an unknown compromise set that is defined by the object scores.

The objective function is to find object scores with the minimum membership differences to the compromise set under normalization restrictions. Optimal scaling is done by alternating least squares. In CATPCA, dimensions correspond to components (that is, an analysis with two dimensions results in two components), and object scores correspond to component scores. A categorical principal components analysis could be used to graphically display the relationship between variables based on the first two principal components.

Component loads reflect the correlation between the quantified variables and the principal components. Since the categorical values were included in the analysis after a numeric transformation, in CATPCA the component loads are considered, in fact, a measure of association and not a strictly linear correlation.

#### Optimal scaling, grouping and discretization

The CATPCA analysis is based on positive integer data. String variable values are always converted into positive integers by ascending alphanumeric order. User-defined missing values, system-missing values, and values less than 1 are considered missing. It is important to recode or add a constant to variables with values less than 1 to make them non-missing. The data must contain at least three valid cases (not empty, no-missing). Discretization is a method of recoding variables and is applied, by default, to fractional numeric and string variables. Fractional-valued variables are grouped into seven categories (or into the number of distinct values of the variable if this number is less than seven) with an approximately normal distribution, unless specified otherwise. String variables are always converted into positive integers by assigning category indicators according to ascending alphanumeric order. Discretization for string variables applies to these integers. Other variables are left alone by default. The discretized variables are then used in the analysis. Other options of discretization consist in ascending ranking or rounding after standardization.

The scaling levels for variables are optimized according to the data type. By default, variables are scaled as second-degree monotonic splines (ordinal). However, it is possible to select the scaling level to be used to quantify each variable. In this way, when scaling is set to be numeric, categories treated as ordered an equally spaced (interval level). For ordinal scales, the order of categories is preserved; and for nominal scales, only grouping of objets in categories is preserved.

### 2.4.4 CATPCA for the total set

In order to detect if an association between factor variables and indoor radon exists, the different quantitative and qualitative variables were analyzed together. A method that allows such a combination is the Categorical Principal Components (CATPCA). Categorical, ordinal and numeric variables exist among the ancillary data. The first step is to recode variables with negative values into positives, such as floor level. Then, for each variable an optimal scaling and, if needed, discretization should be indicated. The association between variables is then analyzed through the components of each variable for each of the two principal dimensions, as shown in Figure 2.16a.

A second CATPCA analysis was run for a selected dataset corresponding to inhabited dwellings at ground floor. The corresponding graph of components loadings is presented in Figure 2.16b.One can observe, that the dimensions 1 an 2 in the graphs correspond to the two principal components generated by analyzing a certain group of variables. Thus, dimension 1 and dimension 2 are not the same components in Figure 2.16a and Figure 2.16b.

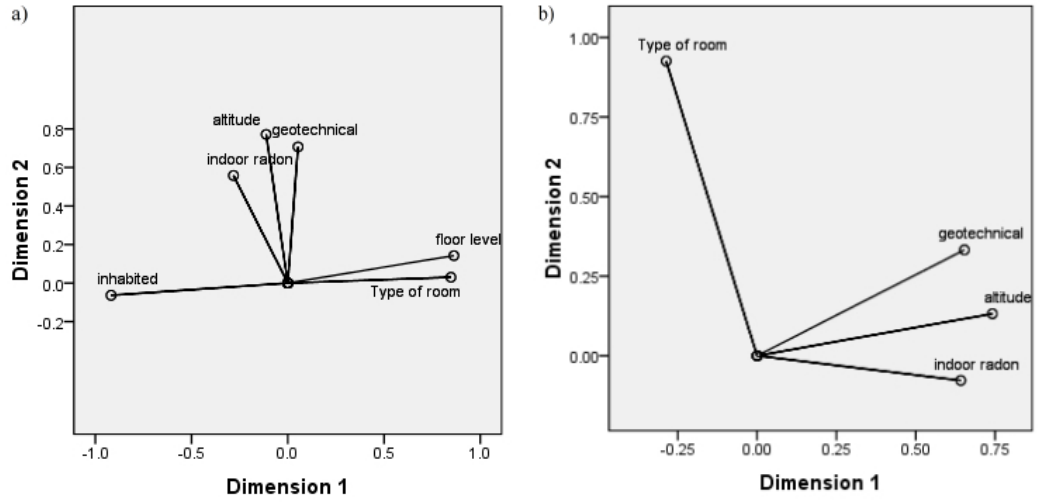



Figure 2.16: *a) CATPCA components loadings for 6 variables for the total dataset b) CATPCA for measurements on ground floor and in inhabited dwellings*

As seen in Figure 2.16a, three affine subgroups are defined by proximity of components. First the inhabited condition appears to be less associated with other variables. Another subgroup is formed by floor level and type of room. At last, indoor radon appears to be best associated with geotechnical units and altitude. The same level of close association with these two variables was found for the selection of data from inhabited dwellings taken on the ground floor.

What is important in Figure 2.16a, is that globally the influence of the inhabited condition and the floor level seems to be less related to the indoor radon variable than other variables. In Table 2.4 and Table 2.5 the combined effect of these two variables have shown to have a low correlation with indoor radon. Table 2.5 shows that for the uninhabited category, the floor level 0 had a higher indoor radon mean than floor -1. Meanwhile, samples for the inhabited category presented a higher mean value for lower floor levels. Since the behavior of the samples from inhabited dwellings seems to be more in agreement with what is expected from radon cumulation, another CATPCA run was launched for this dataset. Figure 2.17 presents the component loadings for data only from the inhabited condition.

In fact, after excluding the inhabited/uninhabited variable the floor level variable, does not show a better association with indoor radon. Some degree of a masking effect of the floor variable was expected from the inhabited condition, but further analysis indicates the contrary. Figure 2.17 indicates that geotechnical units and altitude are comparatively more associated.

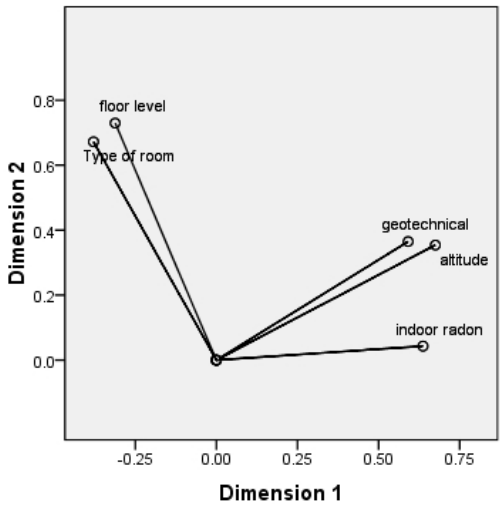



Figure 2.17: *CATPCA components loadings for 5 variables for the inhabited total dataset*

### 2.4.5 Other measures of association

Aside from the usual correlation coefficient measure of the association between variables, there are other measures that can be considered. The correlation is only meant to be used for quantitative variables, and it measures the linear association between variables.

The Eta-squared describes the ratio of variance explained in the dependent variable by a predictor while controlling for other predictors. It is a measure of effect sizes, that conveys the average difference between two groups without any discussion of the variability within the groups.

Eta was invented specifically for the situation in which there is a nominal explanatory (independent) variable and a numeric response (dependent) variable. Eta has the same kind of interpretation as Pearson's $r$, but Eta does not assume that the association is linear. Unlike r, Eta is always between 0 and 1. Eta requires the counts to be 'large enough' for its interpretation for the strength of an association to be reliable. It also performs better when the number of categories of the nominal variable is 'large'. Eta close to 0 means no association; Eta close to 1 means any association there may be is strong. Cohen (9) suggests some values to interpret the effect for Eta squared ($\eta^2$) where 0.0099 constitutes a small effect, 0.0588 a medium effect and 0.1379 a large effect.

In Figure 2.18, the bivariate correlations between indoor radon, geotechnical units, altitude and floor level is presented.

The highest correlations with indoor radon values were found for altitude, followed by floor level and geotechnical units. It is important to note, that altitude, being a numerical value, can result in the corresponding correlation being higher. Meanwhile geotechnical units were coded numerically following a criterium of lithological hardness as previously explained. Despite tests indicating that correlation is significant between all variables, the values themselves are very low.

**Correlations**

| | | indoor radon | geotechnical | altitude | floor level |
|---|---|---|---|---|---|
| indoor radon | Pearson Correlation | 1 | .086** | .143** | -.088** |
| | Sig. (2-tailed) | | .000 | .000 | .000 |
| | N | 139483 | 137870 | 139114 | 136418 |
| geotechnical | Pearson Correlation | .086** | 1 | .245** | .071** |
| | Sig. (2-tailed) | .000 | | .000 | .000 |
| | N | 137870 | 137870 | 137818 | 134827 |
| altitude | Pearson Correlation | .143** | .245** | 1 | .020** |
| | Sig. (2-tailed) | .000 | .000 | | .000 |
| | N | 139114 | 137818 | 139114 | 136050 |
| floor level | Pearson Correlation | -.088** | .071** | .020** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 136418 | 134827 | 136050 | 136418 |

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 2.18: *Crosscorrelations table between indoor radon, geotechnical units, altitude and floor level*

The levels of association with indoor radon considering $\eta^2$ were medium for floor level ($\eta^2 = 0.103$) and altitude ($\eta^2 = 0.089$); while a small association was found with geotechnical units ($\eta^2 = 0.018$). It must be pointed out that $\eta^2$ is not an adequate measure between two numerical variables like indoor radon and altitude. It can be said a priori that the non-linearity of geotechnical units and floor level is shown by its low correlation. The higher $\eta^2$ of the floor level variable can be a result of the relatively low number of categories with respect to geotechnical units. In such a case the differences between groups are higher.

In respect to category size and association measures, CATPCA is a better indicator because it normalizes categories making them comparable, and it presents the results for multiple variables together.

### 2.4.6 CATPCA for the Bern dataset

As indicated, the Bern dataset contains more ancillary information than the total dataset, and besides a certain preferential sampling, it is worth a multivariate analysis. In a first run, the relation between all the variables listed in table 2.7 was studied. The starting and ending date of measurement were not considered because it was assumed that the effect of the seasonal factor was subtracted after normalization (using the procedure explained in section 2.2.4).

Results may vary depending on the scaling chosen and on the level of discretization. It is convenient to group small categories and to correctly interpret the data type. For example, the type of dosimeter variable, with only 3 categories and a high bias in the number of samples, appear to be particularly sensitive to these parameters.

In Figure 2.19a, it is possible to visualize graphically the association of variables for the first run. In this figure, the first and the second component values are closer. For our analysis, we are interested in selecting those variables closer to the indoor radon measurement. Since the number of variables is quite large, the strategy to successively discharge the influence of the more distant variables was considered adequate. In this sense, the variables floor, type of room and inhabited condition were discharged after the first run. In the second

run, presented in Figure 2.19b, the variables year of construction, basement, population and type of house, were also discharged. Finally, the bunch of variables that appear to be most associated with indoor radon are presented in Figure 2.20. In the final run of CATPCA for the Bern dataset, indoor radon appears isolated in relation to the geotechnical units, altitude and type of dosimeter variables.

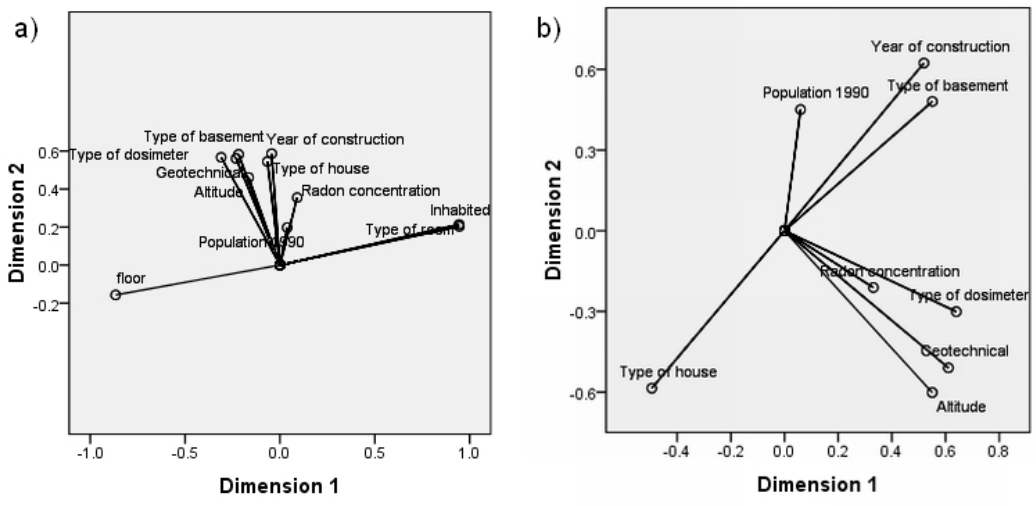



Figure 2.19: *CATPCA components loadings for the Bern dataset including a) 11 variables and b) 8 variables*

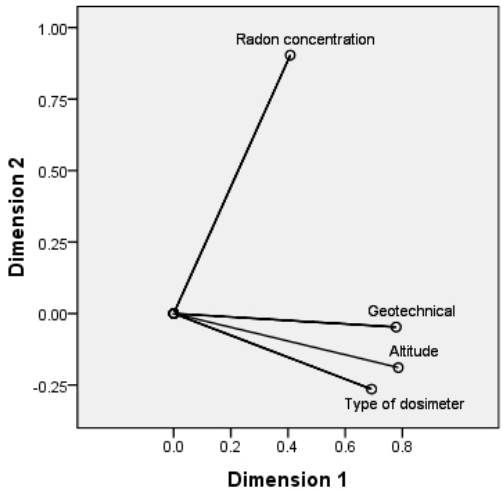



Figure 2.20: *CATPCA components loadings for the Bern dataset including 4 variables*

As mentioned, in Table 2.7, the main portion of measurements were done with an E type dosimeter, and a certain bias towards higher indoor radon values is seen with the G type dosimeter. This can explain the influence of this factor on indoor radon concentrations. Altitude is also a factor that is reported in literature as influencing radon exhalation. The

geological factor (here presented as geotechnical units) is, as expected from the literature, the factor that seems to be closer to indoor radon concentrations in comparison to other variables, despite its lower level of association. In fact, none of the variables are strongly associated with indoor radon but it can be said that geology and altitude are the closest ones. Therefore, the geotechnical factor was further analyzed.

## 2.5 Geological influence

### 2.5.1 Geological influence on a national scale

The simplified geotechnical chart of Switzerland is divided into 30 units. Composition is a basic concept in Geotechnical coding (GT_ID), and it goes from mobile materials to harder rocks. Code 1 corresponds to lakes and code 2 to glaciers. In codes 3 to 7 there are mobile formations including gravel, sand, silt and clay as materials. In codes 8 to 22 there are mainly formations with sedimentary rocks like marls, limestones, sandstones and conglomerates, and medium-grade metamorphic like schists. In codes 23 to 30 there are typically high-grade metamorphics such as quartzite, recrystalized minerals such as gneiss and formations with igneous rocks such as granite.

In Figure 2.21a, a map of geotechnical units colored with green, yellow and red is presented for the groups of units from codes 3 to 7, 8 to 22 and 23 to 30 respectively. These three simplified groups had a lithology that goes, in general terms, from high to low weathering and from low to high consolidation. According to the literature, it can be roughly said that following this criteria the units with higher codes contain more radioactive materials. Formations with igneous rocks are located to the south in the cantons of Valais, Ticino, Uri, Grissons and part of Bern. Therefore, higher radon values are expected in this region. Figure 2.21b presents a map of the actual mean indoor radon per canton colored with green, yellow and red for limits of 0 to 200, 201 to 400 and larger than 400 $Bq/m^3$ respectively. It is observed that the cantons of Ticino and Valais don't have the maximum means. On the contrary, the canton of Neuchatel to the north-west, has a large indoor radon mean value despite being made up predominately by limestones. This is due to a complex combination of factors, which are explained below. The dataset used to prepare these maps is the selection of measurements in inhabited dwellings on floor level 0.

The fact that non-igneous rock formations are directly related to high indoor radon concentrations, is particular to a geology where the content of radioactive materials is not as determinant as other factors, like the structure of the subsoil. This statement is particularly true in the case of limestones in the Jura region, where higher values of indoor radon can be found. As stated in the study of Boehm (4), a good permeability of the underground exists in karstic regions like the Jura region. The radon-containing air can stream towards buildings connected to an open karstic system in practically unrestricted quantities. The amount of radon-activity in the ground is not decisive in this case. That is to say, low radioactive materials like limestone can cumulate radon gas under these conditions.

Other factors mentioned by the same authors, which influence the presence of radon in

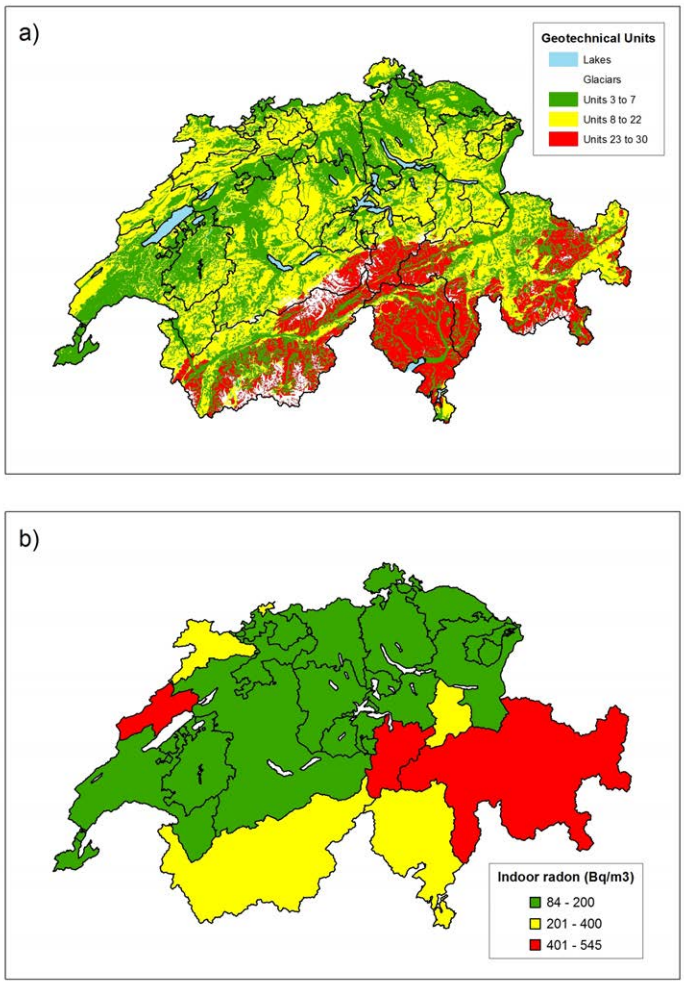



Figure 2.21: *a) Radioactive materials per geotechnical unit (approximation) , b) Mean indoor radon per canton*

buildings, besides permeability, are the existence or the absence of a deck-layer, the terrain morphology and the season. A fine-grained deck-layer is often more moist than the deeper underground; the pores are partially filled with water so that the air-exchange between atmosphere and underground is slowed down by the deck-layer. Regarding the terrain, the author observed that radon exhalation was higher downhill than uphill in summer and vice versa during winter.

In table 2.8 a summary of the number of indoor radon samples, mean and median indoor radon per geotechnical unit, is presented. Table 2.8 is arranged by the number of samples per geotechnical unit to see which units are most relevant for indoor radon analysis.

It is evident, in Switzerland, that most indoor samples - and certainly most buildings - are located on moving formations. 60% of the samples (for the units 3, 6 and 5 together) correspond to gravel, sand, silt and clay formations, followed by limestone and marl (units

Table 2.8: *Number of samples and mean indoor radon per Geotechnical unit*

| N samples | mean Rn | median Rn | GT_ID | Description |
|---|---|---|---|---|
| 4 | 131 | 52 | 2 | glacial |
| 27 | 207 | 79 | 30 | serpentines |
| 37 | 178 | 99 | 10 | red sandstone, clay schist |
| 52 | 64 | 47 | 11 | ferreous clay, often with vitrifiable sand |
| 76 | 361 | 140 | 25 | massif other layered quartzite |
| 84 | 143 | 78 | 29 | green schist and basic rocks |
| 140 | 223 | 119 | 16 | clay schist, phyllite, sandstone-breccia-conglomerates |
| 171 | 397 | 141 | 15 | conglomerates-breccia highly consolidated, with sandstone and phyllite |
| 198 | 704 | 305 | 23 | granites, quartz diorites |
| 239 | 151 | 78 | 18 | lime phyllits, lime schists with inclusions of marls, dolomites and quartz |
| 366 | 302 | 144 | 24 | quartz porphyr, tuffs porphyr, in massif and layered |
| 381 | 174 | 102 | 1 | lakes |
| 415 | 557 | 251 | 27 | conglomerates and schist breccia, rich in sericite |
| 764 | 369 | 148 | 22 | dolomites and cornieules |
| 963 | 128 | 66 | 9 | marl, dolomites, sandstone |
| 1271 | 119 | 78 | 12 | marl, sandstone and conglomerates |
| 1717 | 292 | 104 | 21 | marl schist deposits und marl |
| 1826 | 181 | 78 | 17 | marl schist, limestone phyllite with sandstone |
| 1948 | 206 | 94 | 20 | sandy limestone, marl, schists, limephyllites, silice and dolomites |
| 1998 | 151 | 82 | 14 | conglomerates middle consolidated, sandstone and marl deposits |
| 2389 | 117 | 70 | 13 | conglomerates low consolidated, sandstone and marl deposits |
| 3159 | 157 | 77 | 4 | clay limes |
| 3646 | 309 | 124 | 28 | sericite gneiss, chlorit and schits |
| 5266 | 328 | 125 | 26 | gneiss mica or biotite, with feldspat and amphibolite |
| 6970 | 304 | 114 | 7 | angular gravel bigger fragments |
| 8742 | 172 | 81 | 8 | marl (CaCO3) with sandstone |
| 12251 | 465 | 149 | 19 | massif limestone with marl deposits, intercalated with lime gravel and sandstone |
| 16822 | 207 | 92 | 5 | gravel and sand with little cementation |
| 29267 | 256 | 103 | 6 | gravel and sand with clay and silt inclusions, water carried deposits |
| 36681 | 185 | 82 | 3 | sand and silt |
| 137870 | 243 | 95 | | Total |

19 y 8), which make up 15% of the samples. Units 26 and 28 together should be mentioned as the most relevant units with igneous lithology (mainly gneiss), making up just 6.5% of the samples. The remaining percentage of samples are of clay limes (unit 4) 2.3%, different types of conglomerates 3.2% (units 13 and 14) and a group consisting mainly of marl, limestone

and sandstone (units 20, 17, 21, 12 and 9) 5.6%. The remaining units are weakly represented.

The unit with the highest indoor radon mean value is the $23^{rd}$ with 704 Bq/m$^3$, which corresponds to granite and quartz diorite formations, but it only has 198 samples. The units that are relevant in the sense of high radon and number of samples are unit 19, with 465 Bq/m$^3$, unit 26 and 28, with 328 and 309 Bq/m$^3$, unit 7 (bigger fragments of angular gravel) with a mean indoor of 304 Bq/m$^3$ and 5 % of samples. It can be said that, for Switzerland, formations of limestones, marls, gneiss and even angular gravels are relevant indicators of elevated indoor radon. All the mentioned units have a mean radon over the global mean, which is 243 Bq/m$^3$.

Based on the mean indoor values per geotechnical unit, a map was built to visualize the a priori distribution of indoor radon (Figure 2.22a). The information is colored with green, yellow and red for the 3 groups formed by thresholds of 208 and 400 Bq/m$^3$. Globally, this is a coarse representation of indoor radon distribution which fits with the map on a municipality level (fig 1.6). The main regional trends of high values to the south and north-west can be visualized. The question is, however, which level of detail can be used? At which scale does it remain accurate?

It should be noticed that the information is very coarse in comparison to the distribution of indoor radon samples. Typically, samples are clustered within the limits of urban areas and ideally, predictions should be made at this scale. Figure 2.22b shows a zoom from Figure 2.22a, where inconsistencies at local levels can be observed. In the zoom map, the categorized radon samples are superimposed over a region with three dissimilar geotechnical units. The unit in green represents gravel and sand, mainly pure other silty, sometimes with cementation (glacier cobblestone), having an indoor radon mean of 207 Bq/m$^3$. The unit in yellow has a radon mean of 256 Bq/m$^3$ (over the global mean) and represents gravel and sand, mainly pure, sometimes with a deck-layer or inclusion of clay-silt, as well as extensive carried deposits (actual stream deposits). The unit in red, having a mean of 465 Bq/m$^3$ is composed of limestone massifs, frequently with partial intercalated marl deposits, lime gravel or green sandstone. Samples were also colored according to their radon concentrations; values increase gradually from green to red.

Figure 2.22b thus shows that on a more detailed scale the variance within units appears to be more evident. It is possible to find very high indoor radon values in units with low mean values and vice versa. Higher and lower values can also appear close together, as is the case in the gravel and sand unit from the example. This erratic radon exhalation from surface units, like fluvio-glacial, was reported by (27).

As mentioned, most of the samples were taken over superficial geotechnical units, and this certainly increases the uncertainty in the correlation between geology and indoor radon in Switzerland. In addition, it was mentioned that local factors like permeability, moisture and terrain morphology can play an important roll in radon exhalation. Furthermore, the effect of all the building-related factors and weather conditions should be added. All these factors draw a complex scenario for the modeling and prediction of indoor radon. Moreover, it is certainly unfeasible to accomplish this modeling based solely on geological information. Variations can be on a very local scale. This statement was depicted with an example that

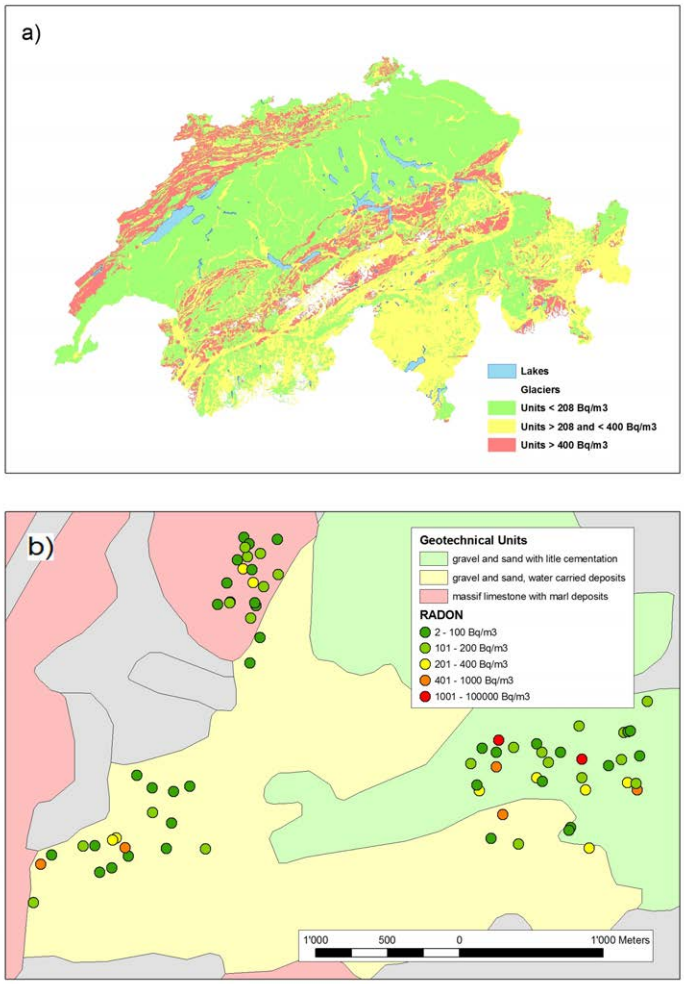



Figure 2.22: *a) Map of mean indoor radon per geotechnical unit, b) Zoom superposed with indoor radon samples*

represents the possible disagreement between sample values and geotechnical units. This high local spatial variance will be analyzed in the following chapter.

For the Bern data the number of samples is very small for certain units, making it difficult to perform statistical tests of significance and correlations. Even generalizing the classification, the geotechnical surface units (such as clay, lime and sand) contain most of the samples. Sediment units has a medium number of samples, while hard rocks (as granite) are scarcely represented. Nevertheless, some differences between units can be seen: surface units have mean and median indoor radon values that are under the corresponding global mean and median, while harder rocks have values over the global parameters.

## 2.6 Conclusions about the statistical analysis

Using the criterium that indoor radon samples are strictly contained within the built-up area, a constrained domain was defined and later used for spatial analysis. The number of samples falling into this domain, after selection from ground floor and inhabited dwellings was 41,787 on a national scale.

The influence of the many factors causing indoor radon accumulation were measured and evaluated together using statistical methods. Seasonal correction has already been applied by public health authorities since winter and summer conditions have a significant influence on indoor radon. The Categorical Principal Components Analysis (CATPCA) method was used to detect the variables having a closest association to indoor radon. Altitude and the geotechnical factor appear to be most associated with indoor radon values. A more exhaustive analysis was conducted for the geotechnical factor because of its importance in describing radon gas sources. It was found that surface units with fluvioglacial materials are the most common units where samples are located. The association of indoor radon with geotechnical units was verified on large scales. On short scales of mapping, indoor radon values within units appear to be more erratic. Very low and very high values can occur close together within the same unit. In general, the geotechnical information appears to be of limited usefullness for indoor radon prediction within short distances. Statistical analysis has also shown that although geotechnical units have a significant degree of association with indoor radon the portion of explained correlation is low.

# Chapter 3

# Exploratory Spatial Data Analysis (ESDA)

This chapter is dedicated to presenting a comprehensive exploratory analysis of the spatial distribution of data points to help in the process of spatial indoor radon characterization and modeling.

One of the first motivations of this thesis was to analyze the influence of the spatial clustering of samples on indoor radon values. In this regard, many tools were tested and adapted to the particular case of indoor radon in Switzerland. The fractal dimension (the idea that objects are self-similar in space) give a baseline from which to measure clustering. In addition, the Morisita index can effectively portray the spatial clustering of points.

The functional spatial distribution (the relation between spatial distribution and indoor radon values) can also be described and quantified by means of clustering methods. Moving windows (MW) statistics can be used as a functional spatial method to relate sample values to the scale of analysis. The purpose of having methodologies to describe clustering is to compare different spatial configurations of data and spatial domains. The purpose of functional clustering methodologies is to evaluate the effect of declustering.

An idea in this part of the research, is to use interpolation methods in the modeling phase as exploratory tools of spatial properties. In this sense, important spatial exploratory tools are K Nearest Neighbor Regression (KNNR) and variography. The first describes continuity by neighborhood while variography analysis indicates the condition of spatial continuity. The hypothesis here stated is that the spatial exploratory tools should also help to define an optimal scale of analysis. In this sense exploration must be done at different scales and density of the spatial distribution.

## 3.1 Spatial characterization tools

### 3.1.1 Average distance calculation

The first spatial characterization of the data set should include a simple overview of the actual spatial density of samples. The average distance between sampling points can give an idea of how far or close they are to each other. This distance value can be used later when spatial modeling requires deciding on parameters as an interpolation grid, a lag distance or a neighborhood searching. A simple way to calculate the average distance between points is proposed in (31) and is calculated as follows:

$$\text{Average spacing} = \sqrt{\frac{\text{area covered by samples}}{\text{number of samples}}} \tag{3.1}$$

There are situations when data are highly clustered and the covered area area has an elongated shape. Then, the previous equation can give a high average distance. In these cases, it is more appropriate to calculate a minimum average distance based on the shorter side of the bounding box. This can be done with the following equation (35):

$$\text{Ravr} = \frac{\min(X_{\text{max}} - X_0, Y_{\text{max}} - Y_0)}{\sqrt{\text{number of samples}}} \tag{3.2}$$

### 3.1.2 Spatial distribution by Voronoi polygons

A simple and visual characterization of the spatial distribution of points can be achieved by building Voronoi polygons using point's locations. The areas of polygons relate to empty spaces and are good measures of clustering (as proposed in (37)). Taking all the polygon areas, it is possible to generate a histogram to analyze the level of data clustering (39). A narrow histogram distribution will indicate regularity while positive skewness will be an indicator of clustering.

### 3.1.3 Fractal dimension by sandbox counting method

The fractal dimension for the sample locations is a common clustering measure. The well-known sandbox counting method consists of measuring the departure from a homogeneous situation, for which the fractal dimension is equal to the value 2. The sandbox method can be interpreted as a measure of density of samples at different scales (75). It must be pointed out that the resulting fractal dimension $Df$ is a power relation measure based on local densities but aiming to express the existing clustering for an entire region.

A practical way to calculate $Df$ with the sandbox method for two-dimensional spaces is to set a relation between the number of samples within a circle and its corresponding surface. This relation is expressed as:

$$N \approx \text{area of the circle} \tag{3.3}$$

*where N: Number of Points*

If expressed in logarithms, this relation becomes a linear equation of this type:

$$\log N \approx \log(\pi r^2)$$
$$\log N \approx \log \pi + 2 \log(r)$$

This relation is satisfied when for an increment of the logarithm of the radius $r$, corresponds to twice the increment of the logarithm of the number of samples $N$. In other words, the slope of the linear equation fitted to the function 3.3 approaches 2 when points are regularly distributed. The value of the slope in this linear function is interpreted as the fractality dimension ($Df$) for a 2D space. In theory, one must compute the number of points within a circle starting from an infinite number of positions within the covered area and then calculate the mean of points, so that the entire fractal is analyzed for an starting radius $r$. The process should be repeated with an increment in the radius as is illustrated in Figure 3.1a. An intuition, though, is that for smaller radii no points are detected, while for radius exceeding the limits of the spatial domain, the number of points will remain constant.

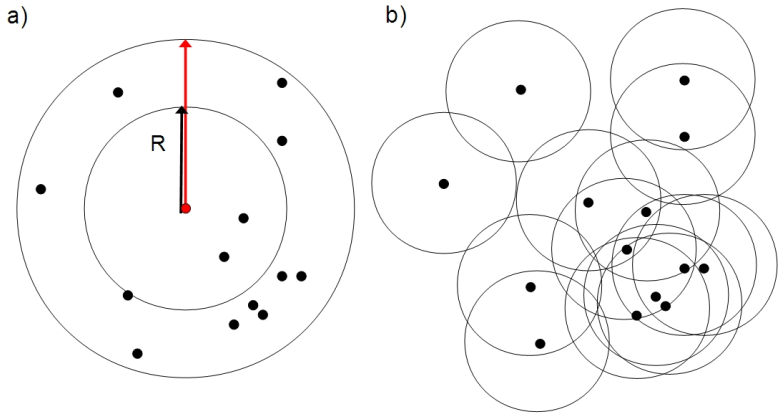



Figure 3.1: *a) The circle is enlarged to compute a sandbox index for every dimension b) the mean of points within a circle is computed*

An operational difficulty with the sandbox method is that for smaller radius the number of points are proportionally lower and, for bigger radius, the function becomes asymptotic. This edge effect must be avoided in order to approach a linear function. Then, the radius of search must not exceed the radius of the fractal. Besides, when few location points exists or they are highly anisotropic, the mean value can vary in an erratic way so that no linear function can be fitted.

This edge effect can be corrected for example for the K-functions by introducing explicit formulas (24). Nevertheless, these procedures requires additional computation efforts. A rule of thumb proposed in the present research and which proved to work for a variety

of spatial arrangements is to set the minimum cell size as the average distance of points calculated with formula 3.1. The maximum cell size can be set as half the shortest side of the bounding box in order to avoid both, the border effect and excessive anisotropy from small datasets.

Whenever the fractal dimension $Df$ is less than the Euclidean dimension $DE$, sparsely distributed phenomena with a dimension lower than $DE - Df$ cannot be detected, and indicates the difficulties arising from interpolating low-dimensional sparse data onto two dimensional regular grids (41). Using a sandbox diagram, it is also possible to analyse changes of fractality at different dimensions.

### 3.1.4 Spatial clustering by Morisita index

The measurement of the dispersion of points in space based on the Morisita index is an easy way to characterize spatial distribution. The Morisita index (MI) is a statistical index of spatial clustering (37); the area under study is divided into $C$ cells of equal size $\Delta$ and the Morisita index is then computed in the following way:

$$MI_{\Delta} = C\frac{\sum_{i=1}^{C} n_i(n_i - 1)}{N(N-1)} \tag{3.4}$$

where $i$ is the number of cells, $n_i$ is the number of points falling into the cell $i$ and $N$ is total number of points. The index is calculated for each cell size $\Delta$. The index can have values below 1 for regular spaced data, approximately 1 for random distributions and above 1 for clustered or fractal distribution. Thus, Morisita index can differentiate between regular, random and clustered spatial arrangements.

This index of dispersion is effective to describe if a set of points are distributed in space in a clustered disposition, randomly or regularly. The index approaches 1 when points are at random in space; its value is over 1 when points are clustered and is below 1 when a regular distribution is present. An advantage regarding fractal dimension is that it is possible to evaluate independently each scale or cell size, making computations more efficient for large datasets.

## 3.2 Spatial functional characterization tools

There are several methods available to describe the distribution of points in relation to a quantitative variable. They constitute the first step towards spatial prediction, and they help to detect major modeling drawbacks. For instance, the functional clustering characterization can indicate how spatially consistent data are, before using them for interpolation.

### 3.2.1 Quantile maps

The quantile map is a simple visualization tool that is used to simplify and generalize variables values in space. Data are categorized into equal-sized groups using quantile thresh-

olds. It is particularly important to give more contrast to quantiles with lower and higher values in order to detect distinctive zones. Quantile maps are especially useful for representing skewed variables, such as indoor radon, by considering equal-sized categories.

### 3.2.2 Functional Box-counting

The box-counting method relies on a different concept than the sandbox method. Instead of fixing a referential space (the cell) and computing the density or the number of points within that space, boxes containing a certain number of points are counted. This is another method for computing the dimension of a fractal. A common usage of the method is to count the number of cells containing at least one point for different cell sizes. Then, the logarithm of the cell size is plotted against the number of boxes to draw a diagram. On top of the diagram, a linear function is fitted, as is done with the sand-box method, in order to calculate the fractal dimension $Df$. Results are similar to those obtained with the sandbox method; however, the slope is negative, and the border effect is more evident for small scales (cell size). The number of cells on borders tends to be constant from a given size even if they are shortened. In other words, a box containing just one point is more likely to continue to have one point inside even when it is enlarged, simply because it is located on the border.

Another difference with the sandbox method is that local clustering cannot be perceived for the cases of spots (high aggregation of points)(75). For the particular case of indoor radon, there is a large presence of spots, and the behavior of box-counting in such cases must be evaluated.

An advantage of box-counting is that it is faster than point counting because the result is of the boolean type: contains/not contains. This is convenient for calculating $Df$ for several large subsets.

The idea behind functional box-counting (37) is to measure the fractal dimension of point's subsets according to a function variable. Then, the cumulate function $f(x) > T$ is analyzed for different thresholds $T$. The number of points will diminish for higher thresholds, independently of the data distribution, and therefore, the spectated box-counting fractal dimension will be lower.

Functional box-counting also provides insight into the level of clustering for a whole subset above a threshold. Nevertheless, it should be taken into consideration that $Df$ is dependent on the number of points and on their distribution on small scales. Small sets of points, like the locations of extreme radon values, are often interpreted as highly clustered because of empty spaces on small scales.

Therefore, if one wants to compare the level of clustering for different sets of data, one must ensure that sets are equal-sized and not cumulative. To solve this, the use of quantile subsets, such as the ones used in quantile mapping, is proposed. These were called quantile-clustering methods.

### 3.2.3 Quantile Morisita Index (QMI)

As previously discussed, MI is a useful index to indicate if clustering exists and at which scale. The advantage of MI against $Df$ is that it provides a quantification of clustering for each scale $\Delta$ independently, thus avoiding the border effect. A limitation of MI is that besides typifying clustering (below zero for regular data, one for random and above one for clustered), values above one are not an indicator of more or less clustering. The Morisita index value depends on the number of points and the size of the cell.

To avoid this class-size effect, one may use equal-sized datasets, as with the fractal dimension calculations. It is interesting to use MI to analyze the spatial distribution of equal-sized subsets defined according to quantile thresholds for the indoor radon function. This procedure was called quantile MI (QMI).

#### QMI profiles at average distance

Another way to compare the results from QMI for different sets and for a single scale is to produce profiles for a $\Delta_d$ scale of interest at distance $d$. Then, the quantile MI can be summarized using a profile at this scale of interest using the function:

$$MI(Q)_{\Delta_d} = C\frac{\sum_{i=1}^{C} n_i(n_i-1)}{N_Q(N_Q-1)} \tag{3.5}$$

where $Q$ is the quantile and $N$ is the number of points per quantile.

## 3.3 Declustering methods

Geostatistical literature proposes two interpretations of declustering. The first interpretation corresponds to declustering as an action to revert the existing spatial clustering. In this case, the sample positions will be affected. A second interpretation of declustering, is the modification of sample values only and not of the positions. This is done in order to approach the unknown global mean by assuming that a preferential sampling has been committed.

### 3.3.1 Declustering with a spatial distribution modification

In this case, declustering implies the modification of the spatial distribution of points, which in general is not desired because of the risk of losing the spatial correlation properties sought out for modeling.

#### Random declustering

One method affecting spatial distribution modification is random declustering, which simply proposes eliminating data from places where there are clustered samples. With this approach, the distribution of data tends to be more regular. This is only an option for very large

datasets, when an underlying regular distribution exists (31). For small datasets, the loss of information caused by this procedure can work against the objective of good prediction and results can be erratic because of the random selection.

#### MW averaging

Another way to modify spatial clustering without changing the sample mean is to perform averaging on a regular grid of cells. The center of each cell in the grid will be assigned the mean value of the points falling in that cell. This averaging will produce a different configuration of samples, but it will preserve the sample mean. The averaging will also cause loss of information and a reduction of the original variance.

#### Declustering by gridding

In order to eliminate spatial clustering, the option proposed here is to adjust samples to a rectangular grid ('gridding'). With this procedure, one gets rid of clustering while preserving neighborhood properties in great amounts. It can be used as a visualization tool to help reveal anisotropy of data or to relate neighborhood parameters to spatial distances. For instance, the optimal number of neighbors for prediction can be better calculated.

### 3.3.2 Declustering to approach the global mean

This option simulates values for the sample locations based on the assumption of an existing preferential sampling for certain regions. It is also assumed that the distribution of samples will provide a sample mean that deviates from the unknown global mean. The goal is then to approximate the global mean and not to modify the actual values or sample locations themselves.

When observing a dataset with spatial clustering, it can be argued that the samples at hand are not representative of the 'real population' distribution because of preferential sampling. Samples can be oddly distributed, and areas with low or high values can be over/under sampled.

It is not a trivial task to find what looks like the real population but is the central point of prediction. As will be seen, a good approximation to the global histogram is essential for its correct reproduction in sequential Gaussian procedures. Thus, declustering analysis must emphasize the global mean approximation.

Several methods have been proposed to tackle the problem of approaching real global distribution (31). Using cell and polygonal declustering, a global mean can be estimated with a linear weighted combination of data; where weights are proportional to a local mean or an area of influence.

**Cell declustering method**

The cell declustering procedure proposes to assign weights to sample values according to the number of points within a cell. A mean for each cell size is calculated from weighted values. The optimal scale of weighting depends on a predefined objective function of either reducing or increasing the estimated global mean.

The sampling space is partitioned into rectangular cells, and the mean is computed for every cell. The samples that fall within the same cell receive a weighting that is inversely proportional to the number of samples for that cell. To calculate the individual weight $w_i$ for a sample $i$ within a cell $j$ the following formula is used:

$$w_{ij} = \frac{N}{c}/n_j \tag{3.6}$$

where $N$ is the total number of points, $c$ is the total number of cells and $n_j$ is the number of points falling into the cell $j$. If many points fall within the same cell they will receive a lower weight. Therefore, clustered samples receive a lower weight.

In order to decide which cell size will be used for the weighting, a diagram of the mean of means per cell is computed for several cell sizes. The decision regarding the optimal cell size for declustering can be based on several criteria. One criterion is to match the area covered by a certain cell size with a sampling domain, like an urban area. If only cells containing at least one point are considered, the area covered by the cells will approach the sampling domain. The range of cell declustering sizes producing the best approximation to the sampling domain can provide a hint for deciding upon the cell parameter.

The important criterion to decide upon the optimal cell size is whether the declustering should reduce or increase the sample mean. As mentioned, this decision will be easy if one knows in advance that a preferential sampling was focused on either low or high values. This kind of information is normally not available. In the present research, the use of some space filling procedures is proposed in order to determine tendencies towards the global mean. The spatial domain and the tendency criteria will be combined to decide upon the optimal parameter of declustering cell size.

**Polygonal declustering method**

The polygonal declustering method (31), like the cell declustering method, applies the same idea of weighting data values according to a spatial support. In this case, the area of influence of each point is used. The idea is simply to calculate the area of a Voronoi polygon constructed around a sample (VorArea) and to calculate the weight $w_i$ for each sample $i$ as the proportion over the sum of all areas (TotalArea). To estimate the global mean after declustering, all weighted values area summed as indicated in equation 3.7:

$$\text{Estimated Global Mean} = \frac{1}{N}\sum_{\imath=1}^{N} w_i \cdot v_i \tag{3.7}$$

where $w_i = \frac{VorArea_i}{TotalArea}$; $N$ is the number of samples and $v_i$ is the value of the $i^{th}$ sample.

Voronoi polygons produce a good partition of space by taking the sample's density into account. If a new measurement is done within its limits is likely to have a closer value to the center point. This method has the advantage of weighting points individually and not according to cells, like the cell declustering method. It also produces a representation of the area of influence of the samples. A disadvantage of the method, is that the spatial representation is not adapted to the sampling domain and therefore, does not take empty spaces into consideration.

When Voronoi polygons are built around sample points, there is no constraining limit on the border. This is not optimal when one wants to calculate an area of influence, since points on the border will be more influential. It is important, therefore, to constrain the limits of the Voronoi polygons to the sampling domain.

By giving more weight to a sample, its value is reproduced proportionally, and its influence is enhanced for the mean estimation. Therefore, as in the cell declustering method, isolated points become more important. Using a histogram of polygon areas, one can perceive a priori the distribution of weights and how they will affect data.

#### Cell declustering and the urban area

Because only cells containing at least one point are considered for the cell declustering weighting, many empty cells are put aside. In this way, the area created by the filled cells adapts to the shape of the sampled area. The cell declustering method creates a constrained domain by only considering cells with at least one point. This interesting property allows a comparison with the natural sampling domain, which is the urban area. It was also noted, that larger cell sizes should not be used for declustering because data inside a large surface could be very dissimilar and the transformation end up producing a high variance. So, a proper transform should make a compromise between the cell size and the space covering.

## 3.4 Moving Windows (MW) Statistics

### 3.4.1 MW statistics for spatial data partition

Moving windows statistics for a dataset are obtained by dividing the study area into equal sized squares and by calculating parameters for each cell out of the contained samples. Important parameters are the number of points, the mean, the median, the variance and the skewness. The size of the square must allow, on average, a sufficient number of points within the cell. For instance, it is not possible to obtain a variance value with just one point. A convenient procedure is to use overlapping windows to obtain more points per cell. On the other hand, the windows' sizes should not be so large that the local variations are not differentiated. MW statistics analyze local variations of parameters (i.e. the mean, the variance and others). It produces a scanning throughout the sampled area to detect major spatial differences of the variable under study.

MW analysis is proposed as a method to produce spatial partition of data in order to improve modeling. It is used to delineate homogeneous subzones within the study area. This is important for modeling and prediction purposes. A model-based prediction requires a model to be applicable to the entire study area. If the parameters of subzones are comparable to the whole area, a single model can be used to make estimations within this area with less error. Moreover, with the help of moving windows statistics, not only subzones but also an appropriate scale of analysis can be defined.

In the case of variance-model-based predictions (like kriging), modeling will benefit from a constant distribution of means (no trends present). Then, estimated values in any particular area will be as accurate as estimates elsewhere (31). What usually occurs with environmental data is that zones with a bigger mean also have more variance. This is known as the proportional effect and is a case of data heterostadicity.

If the local variability is roughly constant, then estimations in any particular area will be as accurate as estimates elsewhere; no area will suffer more than others from highly variable values. If a high variance is associated with a high mean, it is always possible that higher (unbiased) estimation errors will be produced simply because of high variability. When local variability changes, it is preferable to have a proportional effect with the local average in order to know if a predictable relationship exists. One must take into account that the proportional effect is very frequent for earth-sciences variables and is typical for lognormally distributed data (8).

As will be seen in the variography section, this is called non-stationarity. As discussed in (8), the proportional effect is not necessarily a deviation from global stationarity when it is correctly modeled. If this effect occurs within a constant local scale or a defined neighborhood $V_0$, the model will be proportional and the local variograms will differ by a constant value. Commonly, local variograms are proportional to the square of the local mean (31). In most cases, a perfect linear fitting of local sets with proportional effects is not possible and the areas with too high or low variance will generate more errors.

### 3.4.2 MW multiscale test of lognormal skewness

With the MW technique other relevant statistics can be locally calculated. In the research done by (76), it was found that Extreme Value Theory (EVT) density distributions fit better than lognormal distributions for global and local data. Not only the presence of extremes was identified, but it was also shown that the hypothesis of lognormal skewness can be rejected for large regions covering half of the Swiss territory. The first step for the lognormal skewness test is to transform data into logarithms. Then, the skewness parameter $(a_3)$ is tested against the lognormal null hypothesis $H_0$ stating that the skewness distribution is lognormal. $H_0$ is rejected at the $\alpha$ confidence level if and only if:

$$\frac{|a_3|\sqrt{n}}{\sqrt{6}} > u_{1-(\alpha/2)}$$

where $u_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution.

The lognormality of the data can be tested for a significant confidence level ($\alpha = 0.05$). In the present research, only the skewness test was done; in general it is more sensitive to

rejection in comparison with kurtosis. The kurtosis test are redundant and less sensitive.

## 3.5 Spatial continuity exploratory analysis

The concept of spatial continuity for data holds that samples close to each other are more likely to have similar values than those far apart. It is observed, with the help of quantile maps, how higher and lower values are concentrated in certain sectors. This has provided evidence for spatial continuity of data. The idea of continuity also implies a discontinuity that becomes greater with distance. Two basic methods can be used to explore continuity: neighborhood characterization with K nearest neighbors (KNNR) and variography.

### 3.5.1 Neighborhood characterization with KNNR

The spatial neighborhood can be characterized by finding the optimal $k$ number of nearest neighbors by leave-one-out cross validation (CV). This simple method consists of predicting a value for each sample point in a set from the average of $k$ number of neighbors closest to that point. The value of the sample point is 'left-out' and then, compared with the predicted value to obtain an error measure. The absolute error from all predictions is averaged in a Mean Squared Error (MSE) to obtain a global error for each $k$ number of neighbors. The prediction with KNNR having the lowest error is considered globally optimal. This optimization is represented in a graph of KNNR against MSE; when an optimum is attained, it appears as a curve with a minimum MSE.

### 3.5.2 h-scattergrams and the experimental variogram

A method of quantifying spatial continuity is to measure values' differences between a pair of points separated by a fixed distance $h$. This relation is made graphic using h-scattergrams (graphs of paired value dispersion). If the study variable (indoor radon) is expressed as $z$, its value at position $x$ is written as $z(x)$ and the paired value at distance $h$ becomes $z(x+h)$.

The experimental variogram is a function of semivariances at different distances that measures the dissimilarity of pairs of points. The semivariance is calculated as the moment of inertia of paired points from the bisector line of h-scattergrams. The semivariance $\gamma$ for a given lag distance $h_i$ was defined taking the concept that the dissimilarity of values' differences are proportional to the squared radius distance between the bisector line and the samples pairs. This radius is handled as a vector with a certain distance from the bisector (or inertia center) and perpendicular to a sample pair. In Figure 3.2, the graphical interpretation of the semivariance $\gamma(h_i)$ is shown (adapted from (23)).

The semivariance value at a given $h_i$ distance is calculated as the mean squared radius $r_p^2$ for all the $N$ pairs of sample values $(z(x_p), z(x_p+h))$. Where $r_p = |z(x_p) - z(x_p+h)| \cdot \cos 45°$.

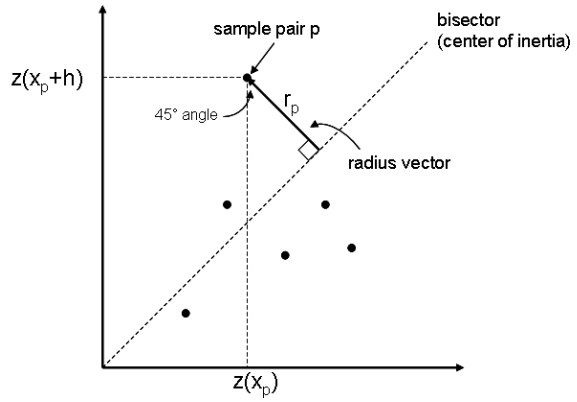



Figure 3.2: *interpretation of the semivariance value* $\gamma(h_i)$

This moment of inertia about the $45°$ line is then calculated as:

$$\begin{aligned}\gamma(h_i) &= \frac{1}{N}\sum_{p=1}^{N(h)}[z(x_p)-z(x_p+h)]^2[\cdot\cos 45°]^2 \\ &= \frac{1}{2N}\sum_{p=1}^{N(h)}[z(x_p)-z(x_p+h)]^2\}\end{aligned} \tag{3.8}$$

When a number of semivariances is calculated for an $i$ number of $h$ distances, we obtain what is commonly called as the variogram function $\gamma(h)$. What seems peculiar from the variogram function is that the right term in equation 3.8 is divided by 2. In fact, the proper name of such a function is semivariogram, while a variogram will retain the constant 2 and is expressed as $2\gamma(h)$. As will be later seen, this equation equivalence is convenient to derive relations between the variogram and the spatial correlation of pairs of points.

The spatial continuity property, the notion that neighboring points are related by their values, is described with variogram functions. Together with MW analysis, variography will provide hints on how to define the optimal scale to be used for modeling.

**The experimental semi-madogram and the drift**

Beside the semivariogram, other experimental measures of spatial variability are helpful to reveal spatial continuity (14). For example, in the case of the semi-madogram, instead of squaring the difference between $z(x_p)$ and $z(x_p+h)$, the absolute difference is taken:

$$\gamma(h_i) = \frac{1}{2N}\sum_{p=1}^{N(h)}|z(x_p)-z(x_p+h)| \tag{3.9}$$

Madograms are particularly useful for establishing large-scale structures (range and anisotropy). It is an especially robust measure regarding the presence of extreme values. In the case of

semivariograms, extreme values produce high semivariances when squared. The disadvantage of the madogram, is that it does not relate to the variance, and hence, it should not be used for modeling the nugget variance.

The drift is a useful measure to characterize global tendencies (trends). It describes spatial stationarity of data. In the case of an intrinsic RF, drift fluctuates around a value of zero (37). The theoretical formula is:

$$\gamma(h_i) = \frac{1}{N} \sum_{p=1}^{N(h)} [z(x_p) - z(x_p + h)] \tag{3.10}$$

It should be noticed, that the average per lag distance is divided by $1/N$ and not by $1/2N$, since the pair of points are counted only once. Pairs of points should also have a defined direction in order to analyze a consistent trend. If no direction is defined, in GEOSTAT software, pairs of points are taken by default for ascending values of x and y coordinates.

## 3.6 Comparative Spatial characterization between toyset1, toyset2 and set3

### 3.6.1 Comparisons for distances and Voronoi polygons

For the three toy examples, the average distance using equation 3.1 is 50 units for toyset1, 49.9 units for toyset2 and 968 m for set3. When using equation 3.2 distances are the same for the first toysets but are reduced to 855 m for toyset3. With equation 3.2, the anisotropy of the distribution is also taken into account. So, equation 3.2 is more recommendable to be used for real case studies.

In Figure 3.3 the series of Voronoi polygon graphs for the three sets are shown. A drastic change is visible from regular, to random and clustered spatial arrangements. A quantification of this clustering is also possible by computing a histogram of the polygon areas. The distribution of polygon areas allows the comparison of the level of clustering. Series of simplified histograms, are presented in Figure 3.4 for toysets 1 and 2 and set3.

In Figure 3.4a the histogram of polygons appears to be centralized in the mean area, while the random distribution and the real case already present a positive skewness (Figures 3.4b and 3.4c respectively). In other words, large numbers of small polygons are present when clustering exists. What is also interesting to point out is that spatial clustering is natural to random distributions (Figure 3.4b), but there are few extremely small polygons in the real case of set3.

The main reason for such a large clustering of the indoor radon dataset in set3, is that the sampling domain corresponds to built-up area, and in general these areas are naturally clustered.

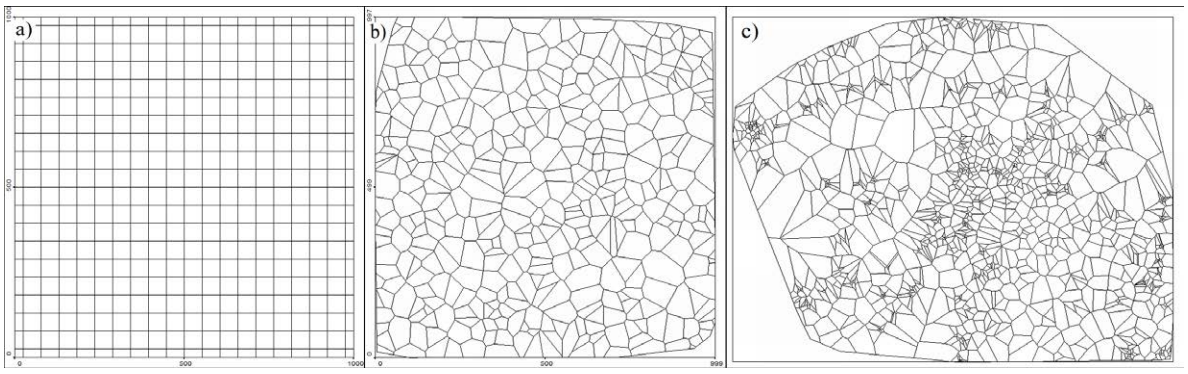



Figure 3.3: *Graphs of Voronoi polygons for a) toyset1, b) toyset2 and c) set3*

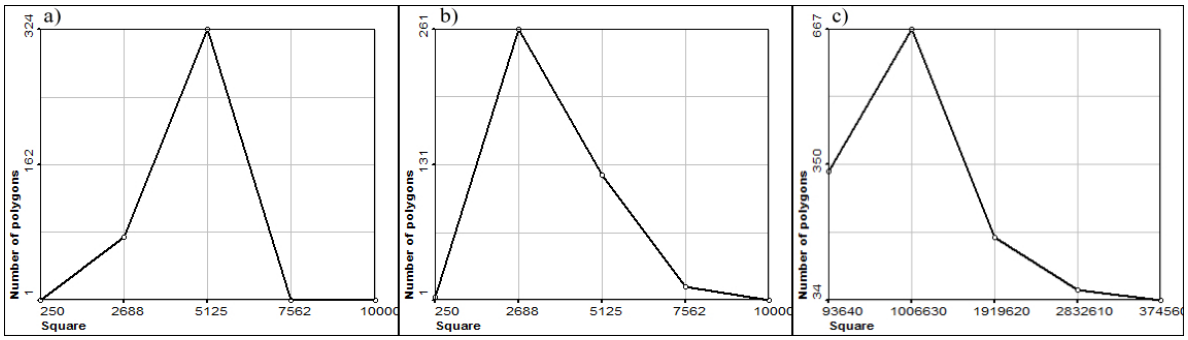



Figure 3.4: *Simplified histograms of Voronoi polygons for a) toyset1, b) toyset2 and c) set3*

### 3.6.2 Comparison for the sandbox and box-counting methods

Figure 3.5a shows a graph of the average number of points per cell size for toyset1. It starts from a size of 20 units, which is below the average distance of points and goes to the maximum box diameter size, which is 1421 units. The function that fits over points is more of the quadratic type with a sigmoid shape than the desired linear function. As mentioned in the methodological section this is due to a border effect of small and large cell sizes.

In Figure 3.5b the fractal dimension for toyset1 is shown, considering limits for dimensions of the fractal. The minimum cell size is limited to 50 units (the average distance between points) and the maximum to 500 units (half the shortest side of the bounding box). The number of steps must remain proportionate to the number of points to evade erratic values. A good rule of thumb is to divide the number of points by 10 to define the number of steps or cell sizes. For large datasets (over 1000 points), a maximum of 100 steps are enough to define the function. In any case, the function is better defined when using large datasets and as many counting cells as possible. In theory, a bigger number of cells with a random arrangement should be created in order to approach a real mean per cell size; but this becomes computationally time consuming for large datasets. With these considerations, 40 scales were calculated for toyset1 and a regression line was fitted for the logarithms of the mean points per cell size (radius). The $Df$ value (the slope of the line), calculated in this way, is near to 2.

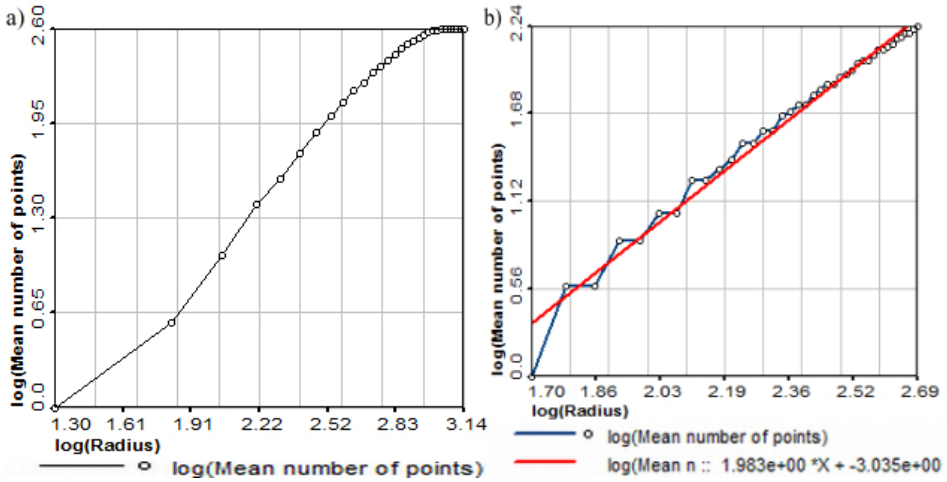



Figure 3.5: *a) Sandbox counting for a regular grid (toyset1) b) Sandbox counting and regression line for toyset1*

A comparison between fractality indexes for toyset1 and toyset2 was made to observe the differences between regular and random arrangements. While a regular arrangement presents a $Df$ of 1.98, the random arrangement has a $Df$ of 1.81 (Figure 3.6a). The $Df$ for set3 was 1.58 (Figure 3.6b), using the scale limits from 968 (average distance) to 15500 (half the shortest side of the bounding box). As a reference, the $Df$ was 1.37 for the box-counting method using the same lower and upper limits (968 and 15500) (Figure 3.7). Considering the results from sandbox or box-counting methods, important clustering of the training data from set3 become evident.

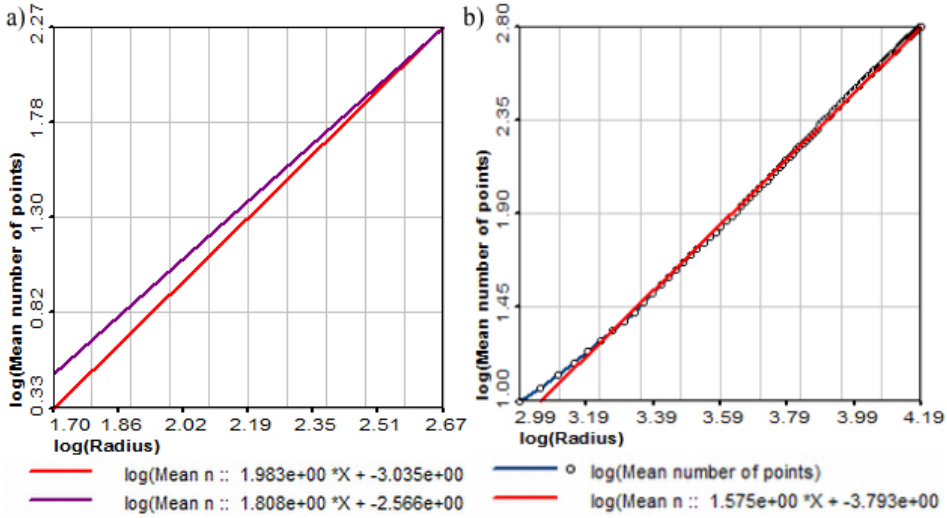



Figure 3.6: *a) Comparative of $Df$ by sandbox counting, between the toyset1 with regular distribution (redline) and the toyset2 with random distribution (purple line) b) Sandbox counting and $Df$ for set3 (redline is fitted)*

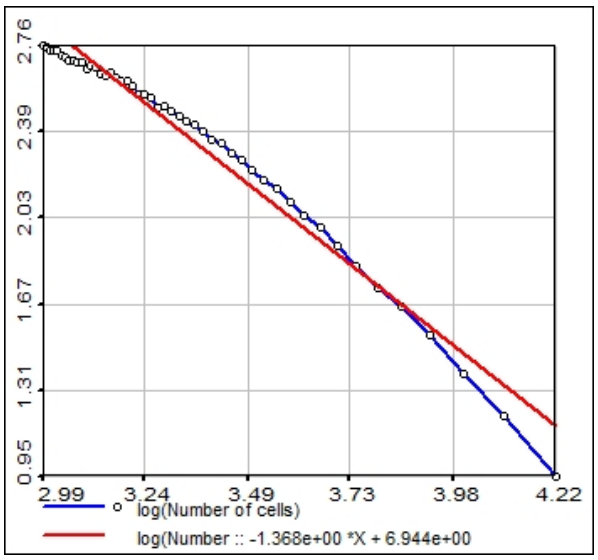



Figure 3.7: *Box-counting $Df$ for set3 (redline is fitted)*

From the inflexion of the curve in Figure 3.6b, it can be roughly said that the spatial arrangement of set3 is fractal on multiple scales. Two fractal dimensions can be depicted below and above an approximate scale of 1700 meters (logarithm of cell size equals 3.23). The sandbox counting tool can be used to characterize the fractality and clustering for a whole region and eventually to depict different fractal dimension behavior occurring per scale. The graphs have also shown an evident edge effect for small scales in the case of the box-counting method, while it is more marked on larger scales for the sandbox method.

The use of toysets in this series of analysis have made the condition of clustering in set3 evident, either by using Voronoi histograms or through the use of the sandbox method. The next analysis will concentrate only on the real case study, set3, and the use of functional clustering, moving windows functional scaling and continuity.

### 3.6.3 Comparison of neighborhood characterization with KNNR

As a test, the procedure was first launched for toyset1 and toyset2 on which a column with random values was added. The CV error from 1 to 40 $k$ neighbors was calculated to search the optimum $k$ number (Figure3.8)

It was not possible to find and optimal $k$ number for the values with regular and random spatial arrangements. In fact, it is possible to test with more Ks; however, the optimization curve will continue to decrease.

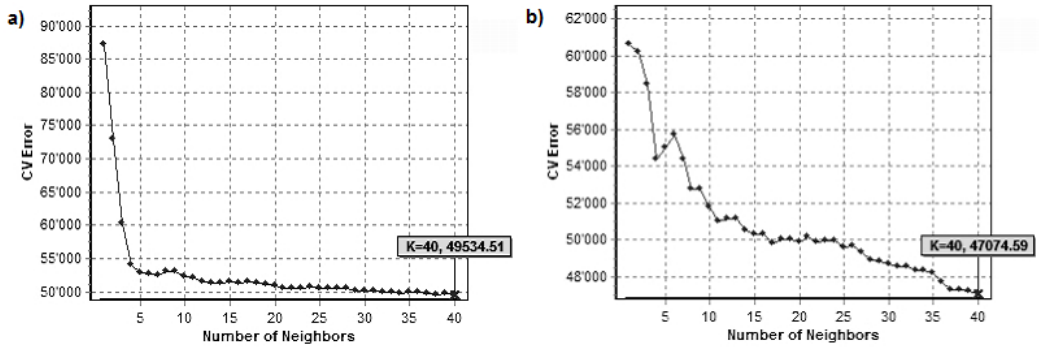



Figure 3.8: *KNNR optimization curves by CV for a) toyset1 and b) toyset2*

## 3.7 Clustering analysis of set3

### 3.7.1 Spatial clustering analysis of set3

A regular distribution of 1296 points in a lattice of 36x36 was produced to approximate the number of training samples of set3 (1310 samples). In addition, a random distribution of 1310 samples was created. The two datasets have been given coordinates within the same boundaries of set3. For the regular arrangement a $Df$ of 1.94 was found, while the random arrangement gave a $Df$ of 1.84. It must be recalled, for a sake of comparison, that the $Df$ for set3 was 1.58. If these fractal dimension results are compared with those of the toy datasets, we observe that they are very close for the conditions of regularity and randomness. The ability to compare datasets with different sizes and domains using a bounded clustering scale is a particular advantage of the sandbox method.

Meanwhile, the advantage of Morisita index is that it is possible to independently evaluate the clustering for each scale or cell size, making computations more efficient for large datasets. Figure 3.9 presents an MI diagram for the real case data (set3) compared with two MI diagrams for the regular and random distribution of points. The minimum cell size used for all cases was the average distance (around 970 meters for the three distributions). At this scale, the MI was 0 for the regular set, 0.99 for the random set and 6.43 for the real data.

In Figure 3.9 is possible to see how the diagram of MI values, for different point distribution, evolves to approximately the value of 1. Clustering, as done with the polygons histogram and the box-counting methods, becomes evident on smaller scales.

This analysis shows that it is possible to simplify the MI procedure. A cell size representative of clustering was used, which was the average distance.

### 3.7.2 Functional clustering analysis of set3

#### Quantile maps

Set3 was initially divided into two categories, below and above the median, to produce a median defined map (Figure 3.10).

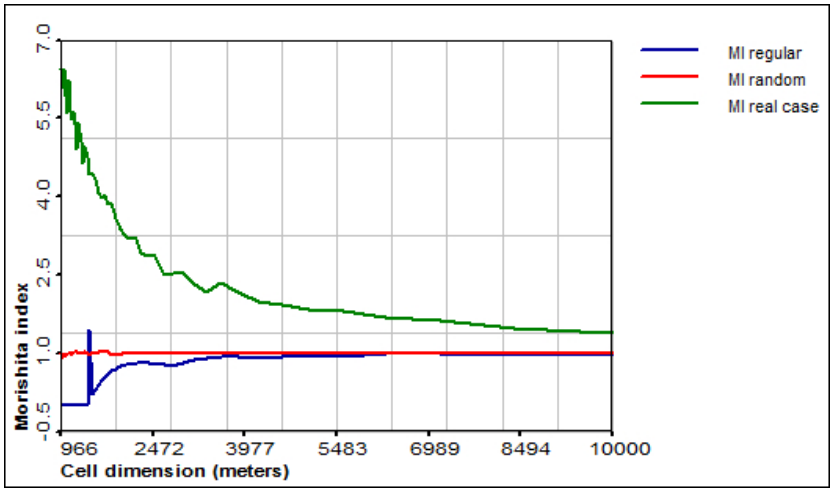



Figure 3.9: *Morisita index for radon measurements (green line) compared to a regular (blue line) and a random distribution (red line)*

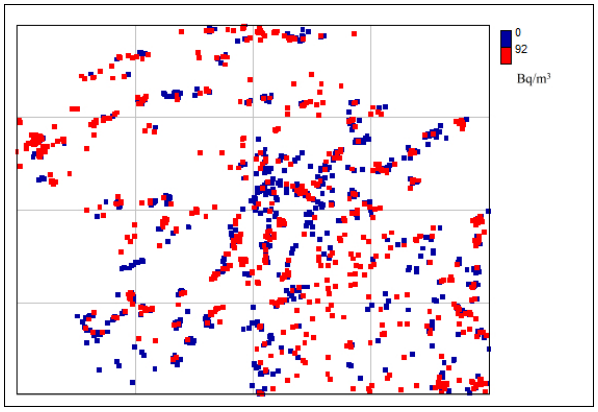



Figure 3.10: *Quantile map of indoor radon for set3 using median value as threshold*

The median map is a hard generalization of values to find out whether or not there are large sub regions with different concentrations. Differences in spatial distribution are not evident for this representation due to the presence of extreme values defining hot spots. A map with subdivisions into 10 equal-sized categories or deciles was build to visualize data distribution with a lower generalization level (Figure 3.11a). This decile map also shows how indoor radon level's categories are mixed-up in space. A separate representation of only the first and the last deciles (Figure 3.11b) show some tendency to form hotspots for high values in the NW direction and lower values in the central area.

**Functional box-counting**

The procedure of functional counting was applied to set3 using quartile limit values as thresholds $Tq$ (0, 58, 92 and 138 Bq/m$^3$). The series of graphs and their correspondent $Dfs$ are presented in Figure 3.12.

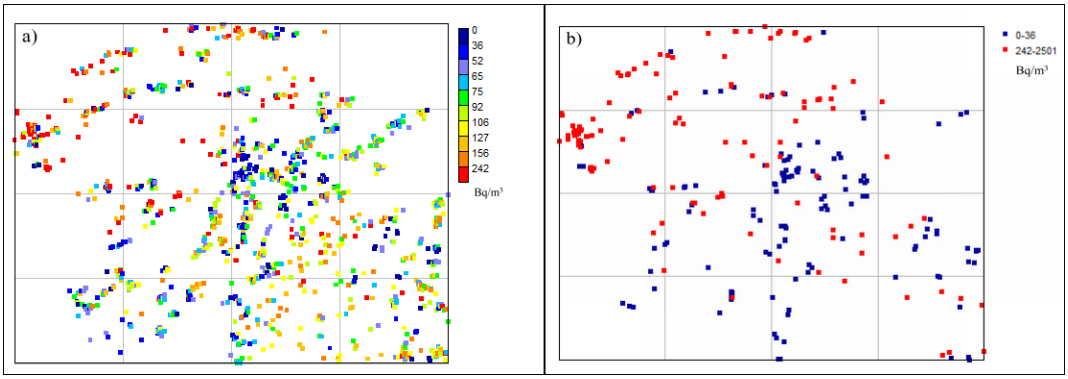



Figure 3.11: *Quantile maps for set3 a) considering decile thresholds and b) for first and last deciles*

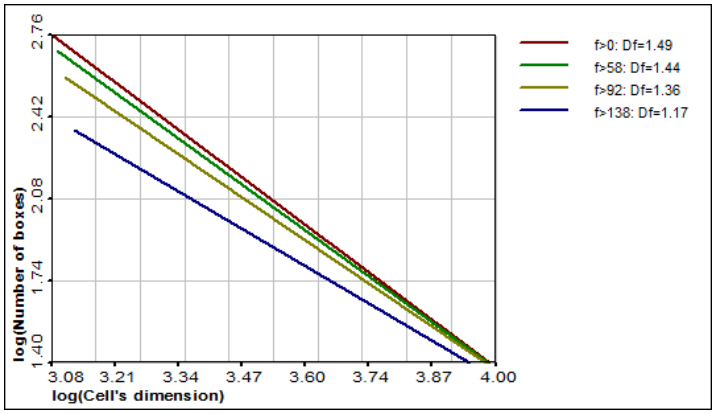



Figure 3.12: *Functional cumulative box counting diagrams for data over quartile thresholds*

The $Df$ for the whole dataset ($f(x) > 0$) is approximately 1.49; while for $f(x) > 138$ Bq/m$^3$, $Df \approx 1.17$. What this graph shows is how the dimensional resolution of points diminish for higher thresholds. At some point, the dimensional resolution is too low to find points.

### Quantile box-counting

As previously shown in Figure 3.12, the cumulative box-counting function is correlated with the number of points, and this can mask real clustering. Therefore, it is more pertinent to use quantile (i.e equal-sized) subsets, such as the ones used in quantile mapping. For the case of quartile thresholds $Tq$, the subsets are defined as $Q(Tq) = \{Tq_i < f(x) < Tq_{(i+1)}\}$, where $Tq = \{0, 58, 92, 138\}$ Bq/m$^3$. Figure 3.13 presents a series of fractal dimension diagrams using both, the sandbox (Figure 3.13a) and the box-counting method (Figure 3.13b) for quartile subsets from set3.

In Figure 3.13, the raw diagrams (instead of the regression lines) offer a better visual comparison of the methods. For every subset and method, $Df$ fluctuates around 1.5, and for both methods a lower $Df$ was measured for the first and the last quartiles (Q1 and Q4). The

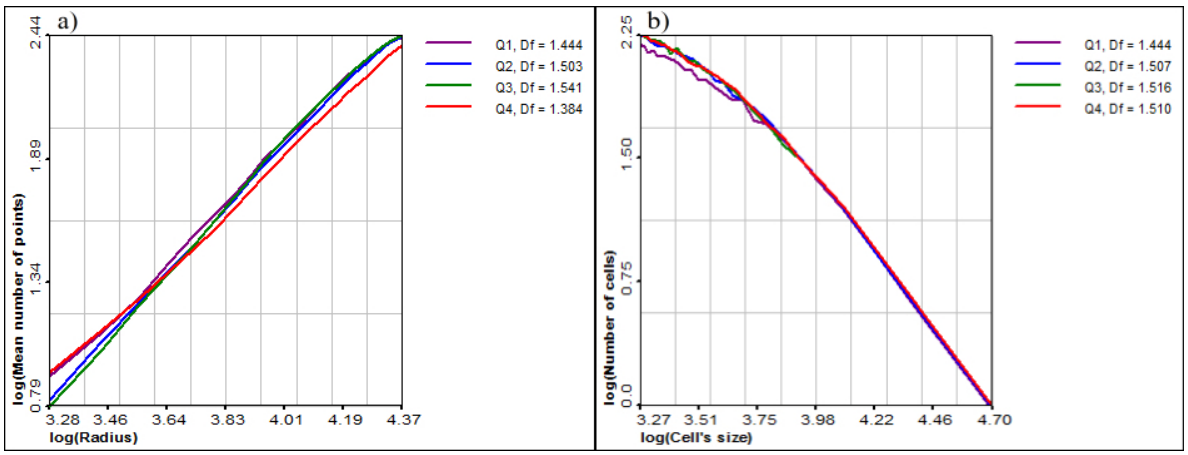



Figure 3.13: *D f diagrams for the quartile subsets from set3 using a)Sand-box method and b) Box-counting method*

sandbox method appears to be more sensitive than the box-counting in detecting differences. It can be concluded that more clustering is present for lower and higher values.

### Functional MI

Figure 3.14 presents the functional cumulated MI diagrams for subsets above thresholds from 50 to 1000 Bq/m$^3$. As with the box-counting method in Figure 3.12, MI shows a tendency of clustering for higher values. However, this is also due to the reduced number of points in subsets above high thresholds (400 and 1000 Bq/m$^3$).

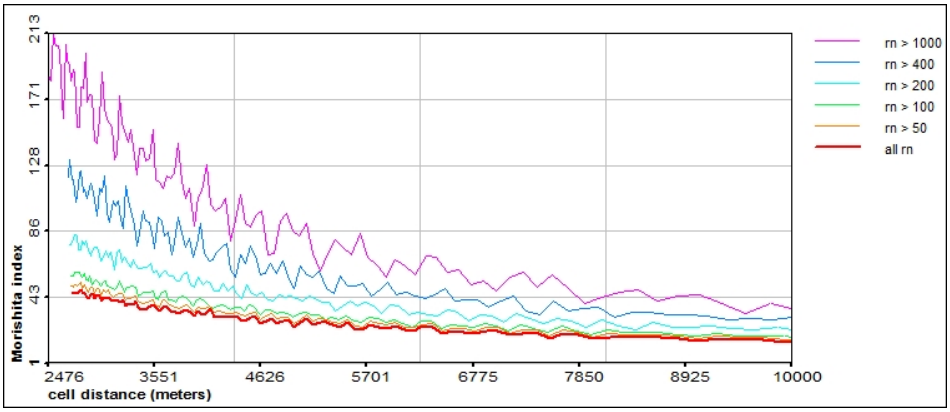



Figure 3.14: *Funtional MI for cumulate subsets for set3*

### Quantile MI

To avoid the class-size effect, the use of quantiles, equal-sized subsets defined according to quantile thresholds, has been proposed. This procedure is called quantile MI (QMI). Dia-

grams for quartile subsets for set3 using QMI method are presented in Figure 3.15.

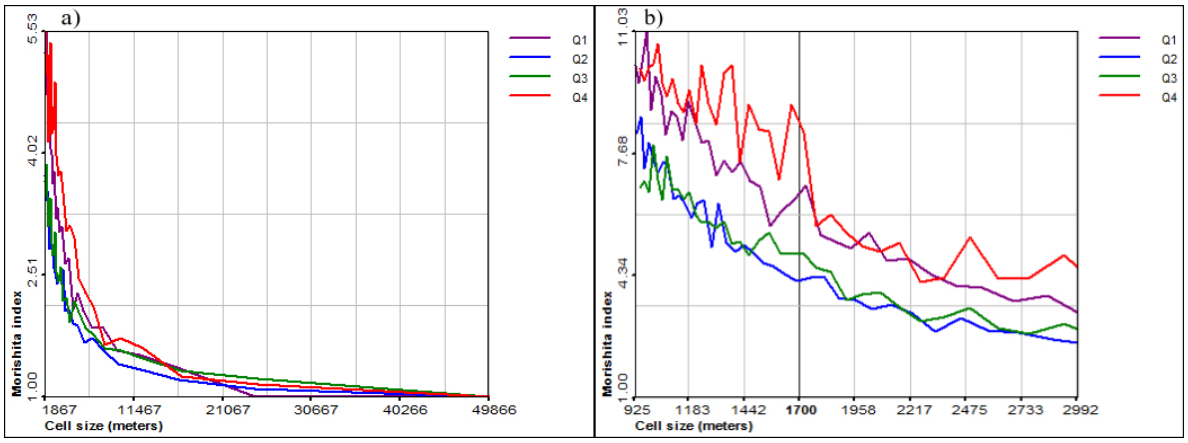



Figure 3.15: *QMI diagrams for quartiles subsets of set3 at a) global range b) Zoomed for cell sizes between 900 and 3000 meters*

Figure 3.15 shows that MI have the same pattern of clustering as the one measured with fractality methods. What can be also observed is that MI differentiates better clustering for smaller scales, while it tends to be one for larger scales (cell sizes). This is coherent in the sense that clustering is produced by concentration of points on smaller scales. It can be said, that if one set has a larger MI than another equal-sized set, at the same cell size, this set is effectively more clustered. Fluctuations can occur on intermediate scales and are measured using MI as well. In Figure 3.15 the QMI diagrams for quartile subsets of set3 are presented. Figure 3.15a shows diagrams for the whole range of scales. Figure 3.15b is a zoom, with scales ranging from the average distance to half the diagonal distance of set3. This zoom allows us to compare the results of QMI with the results of the sandbox method presented in Figure 3.6b.

The tendency of clustering is well defined and coherent with the results from fractality methods, as Q1 and Q4 appear more clustered. In the zoomed Figure (3.15b), it is remarkable that at a cell size of 1700 meters, an abrupt change in clustering was produced; something that was also detected with the sandbox method. The multifractality behavior from Figure 3.6b was also reproduced in the MI diagrams.

### Quantile MI Profiles

In Figure 3.16, a series of profiles from QMI quartile diagrams are presented for set3 for cell sizes $\Delta$ 1000, 1900 and 10000 meters. Quartile limits, as mentioned, are 0,58,92,138 Bq/m$^3$.

It is also interesting to calculate QMI for higher thresholds. To acheive this, set3 was subdivided into 15 quantiles. The corresponding QMI profile is presented in Figure 3.17.

In the example with set3, it is worth pointing out that the profile of the average distance $\Delta_{\text{Avr}}$ is also representative for small and medium distances (1000 and 10000 meters). The MI for Q1 and Q4 has a tendency to be higher in comparison to Q2 and Q3. In Figure 3.17,

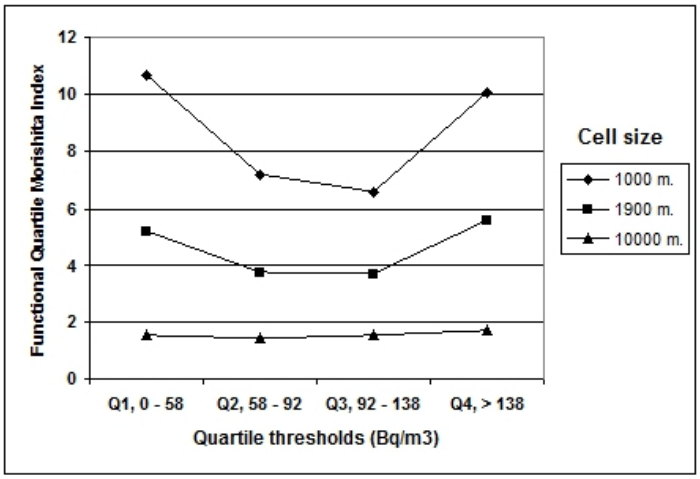



Figure 3.16: *Profile of the QMI diagrams for the set3 for quartile limits*

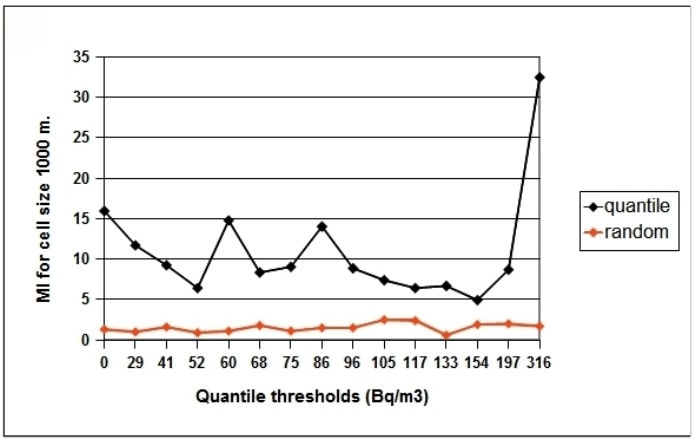



Figure 3.17: *Profile of the QMI diagrams for the set3 for 15 quantile limits and random distributions*

this tendency remains; however, clustering appears more accentuated for values above 316 $Bq/m^3$.

Furthermore in Figure 3.17, the MI for randomly distributed points was included (red line). In theory, the MI for random points has values around one. This value can fluctuate when there are a low number of points, which are shown as a red line in Figure 3.17. For the 15 quantiles case, each quantile has only 87 points and the MI fluctuates between 0.5 and 2.5 for several random sets. For quartiles, which have more points, the fluctuation of the MI is lower and remains close to one. In any case, the differences in MI between random distribution and actual values are clear.

It can be proposed that, statistically speaking, the most representative scale of interest for a spatial distribution is the average distance. The profile presented in Figure 3.17 is a fast characterization of QMI for a given spatial dataset, and can be used to analyze functional clustering in exploratory analysis. Preferential sampling can be depicted with the QMI

method as will be seen later.

## 3.8 Declustering analysis of set3

### 3.8.1 Cell declustering of set3

In Figure 3.18, a diagram of the mean of means per cell size for training set3 is shown.

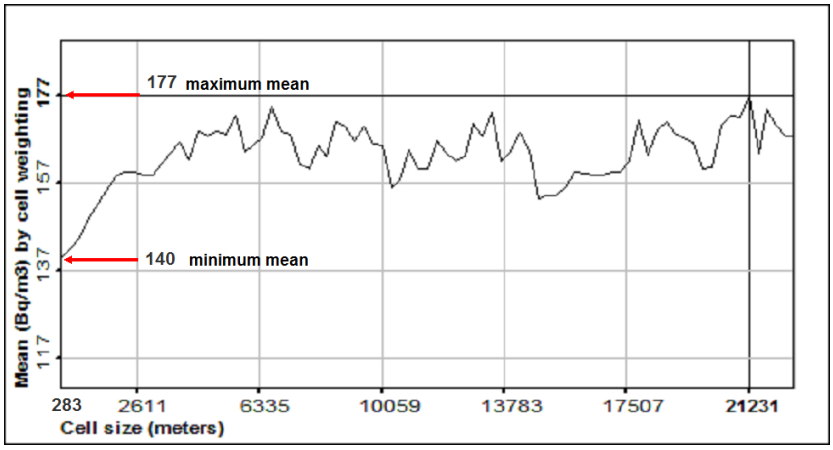



Figure 3.18: *Cell declustering mean diagram per different cell sizes for set 3*

In the diagram above, we observe that if declustering weighting is done using a cell size of 283 meters for set3, the estimated global mean is reduced to 140 Bq/m$^3$ (while the actual sample mean is 142 Bq/m$^3$). Inversely, an estimated maximum mean of 177 Bq/m$^3$ is obtained when conducting cell declustering for cells with a size of 21,231 meters. In this range of cell sizes, there are a variety of possible results.

In Figure 3.19a a map of declustering weights per point for the case when the mean results minimal (cell size of 283 m) is presented. In Figure 3.19b, the cell size is bigger and the mean reaches a maximum.

When using small cells, the weight values are lower than for larger cells. Because only cells containing at least one point are considered for the weighting, many empty cells are put aside. Then, the area considered for declustering, using a small cell size, also corresponds to a constrained spatial sampling domain.

For a dataset with heavy clustering, like set3, the differences in weights are remarkable. Moreover, due to the high skewness of indoor radon data, it is always possible to find high values that are isolated (not clustered). As a consequence, it is possible that high values are heavily weighted and that the total estimated variance is increased. In Figure 3.20, the mean and variance increments are presented graphically for five cell sizes. Set3 has an important spatial clustering, as indicated by the sandbox and MI methods. Even a certain threshold (1700 m.) appeared to be a breakpoint from less to more clustering. This means that the distribution of points is more regular under the threshold of 1700 m. Therefore, the cell

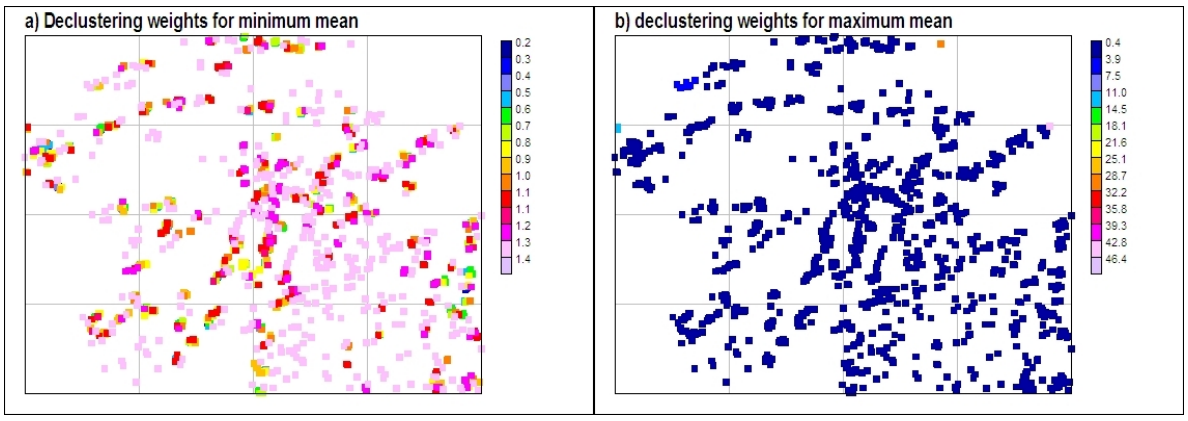



Figure 3.19: *Estimated mean and variance after cell declustering*

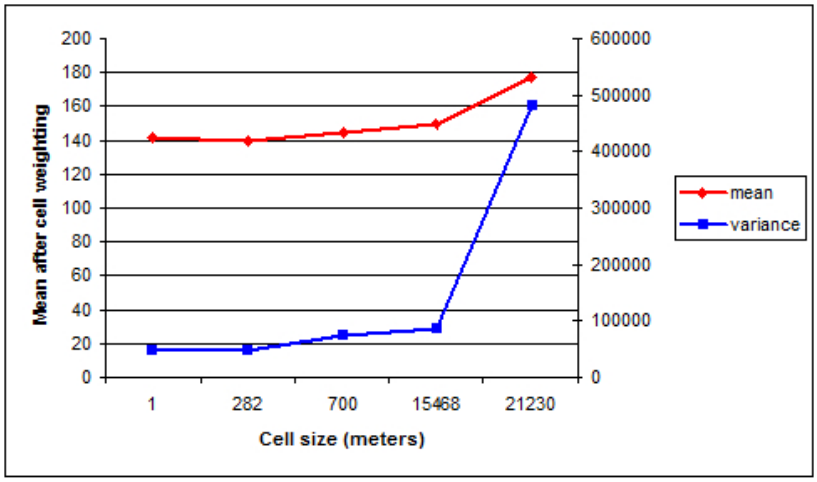



Figure 3.20: *Estimated mean and variance after cell declustering*

weighting beyond this limit will have more variability because values that are more isolated can be found.

Once the clustering of set3 was detected for both low and high values (Figure 3.16), the declustering task becomes more complicated. The main question is now, which kind of weighting can better represent the population mean? Should it be increased or reduced? Is this an indication that no preferential sampling was conducted? In any case, it is logic to interpret this as a preferential sampling for both low and high values. Under this premise, one can postulate that the sample mean is representative of the global mean. If this is the case, a minimum data transformation and the use of a constrained domain seem to be convenient for declustering. Nevertheless, the QMI detailed analysis of 15 quantiles indicates a tendency of extreme values to form hot spots by clustering. The questions remaining are, whether all hotspots were already sampled, and finally, which one would be the tendency towards the global mean? All of these questions are difficult to answer with the presence of extremes. Therefore, a space filling of the sampling domain was proposed as a criterion to find this

tendency.

### 3.8.2 Polygonal declustering method

In Figure 3.4, it was established that most polygons have a comparative small area and the distribution of areas is positively skewed. Therefore, a polygonal declustering weighting will probably also be skewed. Independently of the effect of polygonal declustering over the mean, a skewed weighting has an effect to increase variance. The combination of skewed distribution with skewed weights will enhance extreme values and increase the variance.

The limits of the surveyed area can be an appropriate constrained boundary, as is seen in Figure 3.21. For the case of set3, it is the administrative boundary that defined the surveyed domain (red line). This limit creates a convex domain that closes all the samples. In addition to this limit, the convex domain, created with external samples, can also provide a good definition for the area of influence (blue line).

For the particular case of indoor radon measurements, there is a second domain that constrains sample locations in a concave way, which is the urban area. As previously explained, indoor radon measurements are acquired inside buildings. Using the Swiss topographic charts at a scale of 1:25000, a buffer zone was defined around the representation of buildings to obtain an approximation of the urban area. In Figure 3.21, this area is represented with green polygons, and it was also used to constrain the sampling domain. For the polygonal declustering method, it is proposed that a constrained sampling domain can better correspond to a more realistic representation of the spatial and statistical distribution of the study variable.

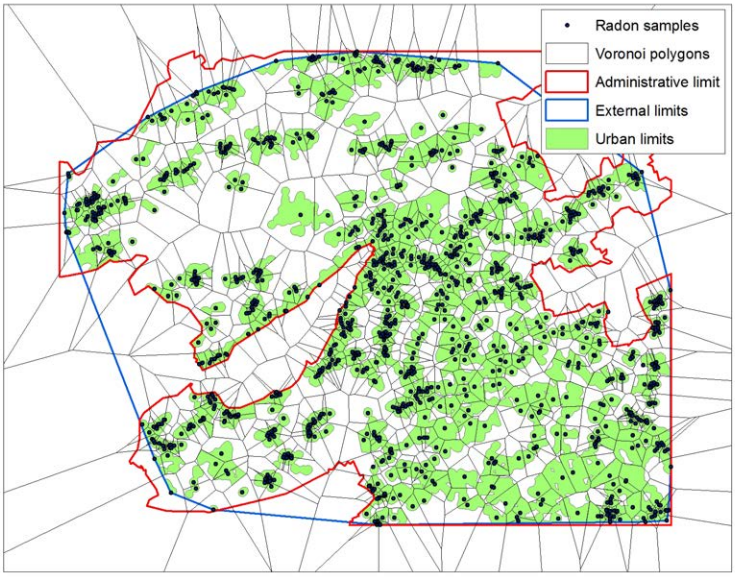



Figure 3.21: *Voronoi polygons constrained by the administrative boundaries and urban area spatial domains*

It is interesting to see how the statistical distribution of polygon areas varies depending on the different criteria for the domain constraining. In Figure 3.22, three boxplot graphs of polygon areas can be observed: Voronoi polygons constrained by administrative limits (in red), by external limits (blue) and by all the limits plus urban area (green). In addition, the distribution skewness is presented in the figure. Logically, when doing more constraining,

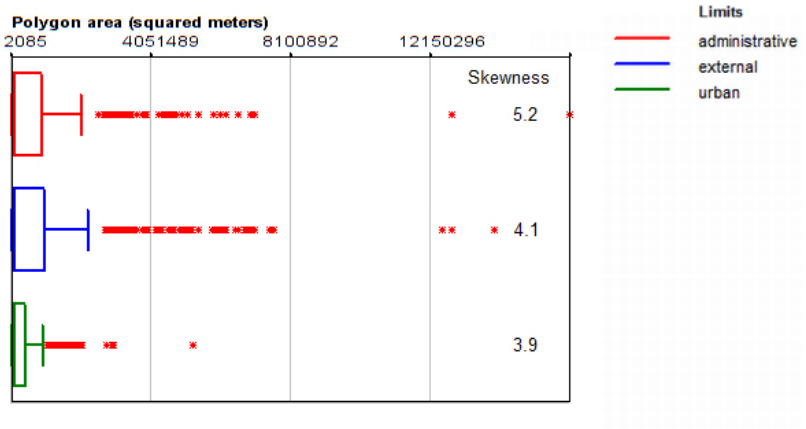



Figure 3.22: *Voronoi polygons constrained by the administrative boundaries and urban area spatial domains*

the sizes of polygons are reduced. It is important to observe that the skewness is slightly lower for the fully constrained urban domain, but in general, the high skewness pattern remains. It can be said, that the underlying sampling schema corresponds to an urban area that is fragmented, and consequently, the areas of influence for samples are irregular. This is clearly expressed by the skewness of the areas. It should be mentioned, that set3 corresponds to samples in inhabited buildings only. Such influence over the sampling domain definition is difficult to evaluate. This influence may be neglected if it is assumed that inhabited buildings are homogeneously distributed within the urban area.

Buffering parameters were also arbitrarily defined to provide better aggregation, and this resulted in large areas for certain isolated samples. A high skewness also indicates a suboptimal correspondence between samples and the theoretical sampling domain. As seen, the definition of this domain requires less accessible information (i.e. inhabited buildings) and proper knowledge of the urban area limits.

Areas of Voronoi polygons under different constraining schemas were used for polygonal declustering, and the results of the transformations were analyzed. In Table 3.1, the parameters of set3 and their estimation after declustering are presented.

Skewness of weights indicates the level of data transformation for global estimation. Polygonal weights obtained with a combination of constrained criteria have smaller areas of influence and therefore a lower skewness. The estimated global mean, after declustering, is closer to the sample statistics (yet not the skewness). The last statement does not mean that these global statistics are not a 'possible realization'. In fact, the only approximations to global parameters are those from the national dataset (section 2.3.3), with a mean value of 163

Table 3.1: *Estimated parameters of indoor radon after polygonal declustering*

| constrain | polygon skewness | estimated parameters after declustering | | | |
|---|---|---|---|---|---|
| | | mean | variance | skewness | maximum |
| **no declustering** | | 141.6 | 47098 | 6.08 | 2501 |
| **administrative (ad)** | 5.25 | 185.7 | 642559 | 16.40 | 21067 |
| **externe boundary (ex)** | 4.14 | 176.7 | 539784 | 15.24 | 18588 |
| **urban + ad + ex** | 3.95 | 144.4 | 121994 | 9.14 | 5529 |

Bq/m$^3$, a variance of 102,873, a maximum value of 15,045 Bq/m$^3$ and a skewness of 12.16. If we consider these parameters, declustering can occur using only administrative or external boundaries to constrain Voronoi polygons over a 'realistic' distribution. A weighting based on more constrained concave spatial domains of the Voronoi polygons can better approach the global statistical distribution of indoor radon.

Declustering is aimed to produce not only precise but also globally accurate estimations. In other words, not only must the individual predicted values be precise, but the results should also reproduce the global statistical parameters.

The usefulness of declustering can be tested with linear weighted estimation methods. Much care should be taken when applying declustering where non-preferential sampling has been identified. In the case of indoor radon, the necessary transformation seems to be minimal; with a very constrained domain giving weights close to one.

In the case of Polygonal declustering for indoor radon, it was observed that the underlying domain is not regular, nor is the sampling schema. The existence of an underlying sampling domain (the urban area) gives unrealistic weighting. A wiser option to improve the results, is to adjust declustering and other neighborhood parameters to the constrained urban domain.

### 3.8.3 Cell declustering and the urban area

The cell declustering method adapts better to a manifold surface created by the urban area and was tested using different cell sizes. Two cell sizes were used: one 400 meter cell, which produced a transform resulting in a mean of 141 Bq/m$^3$ and a variance of 61,112 and one with 1,000 m cell, resulting in a mean of 153 Bq/m$^3$ and a variance of 125,303. Figure 3.23a shows an approximation of the urban area in green with 400 meter cells on top. In Figure 3.23b, the same urban area with 1,000 meter cells on top is shown.

What can be observed, is that the 1,000 m cells produce a covering that is more in concordance with the urban area. It must also be recalled that this urban area was produced with a buffering zone and is probably larger than it is in reality. It can be also seen, that some central areas of the urban sampling domain lack samples. The level of transform produced with the 1,000 m cell is close to the constrained domain built with the intersection of Voronoi polygons, the urban area, the administrative area and the external boundary. Only the mean is higher when applying the 1,000 m cell for declustering. Other intermediate cell sizes are 600 m, with a mean of 146 Bq/m3 and a variance of 88,846 and 800 m with a mean of 149

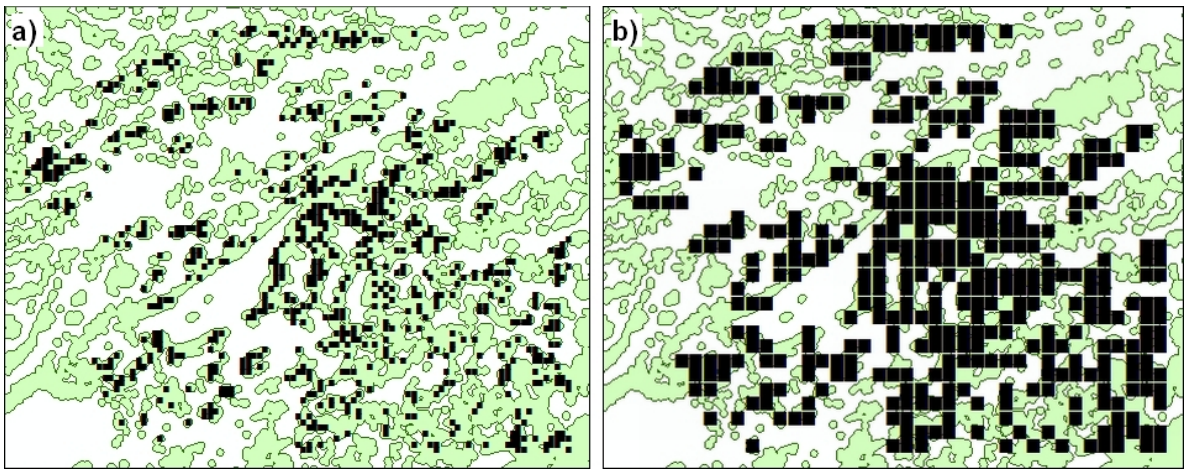



Figure 3.23: *Urban area and regular cells used for declustering with a) 400 m. size and b) 1,000 m. size*

Bq/m3 and a variance of 98,718.

Several options of transforms have been proposed with declustering methods, and it is not yet clear which one can best reproduce the unknown global distribution. The relative elevated clustering of high values can be interpreted as a preferential decision to sample areas where previous measurements were also high. In this hypothetical scenario, a complete survey of all dwellings will tend to result in a lower mean value. For instance, the sampling evolution analysis of data from canton of Ticino showed a decrement of the mean that gives some evidence of this behavior.

The relative elevated clustering of high values can be also explained by its location in isolated areas. So, it will be a matter of detailed analysis of the sampling domain and the actual value distribution to delineate a tendency towards the global distribution.

So far, it has been observed that a grid of filled 600 m cells provides a spatial covering that approximated the one produced by the build-up area domain. In this sense, weighting spaces should be constrained to the sampling domain in order to obtain transformations that are more realistic.

### 3.8.4 Gridding declustering and neighborhood characterization

As previously mentioned, adjusting data to a grid, or 'gridding', is a declustering technique that can help depict anisotropy and relate neighborhood parameters to distances.

To obtain a squared gridding for set3, 1296 samples were selected out of 1,310. Then, data were sorted by X coordinates and then subdivided into 36 sets in columns, with 36 samples each. Data within columns were then sorted to create a better correspondence to its original position in the grid. Thus, the resulting cell size in the gridded arrangement approximates the average distance of 968 m for set3. Maps of the original coordinates and the gridding are shown in Figure 3.24.

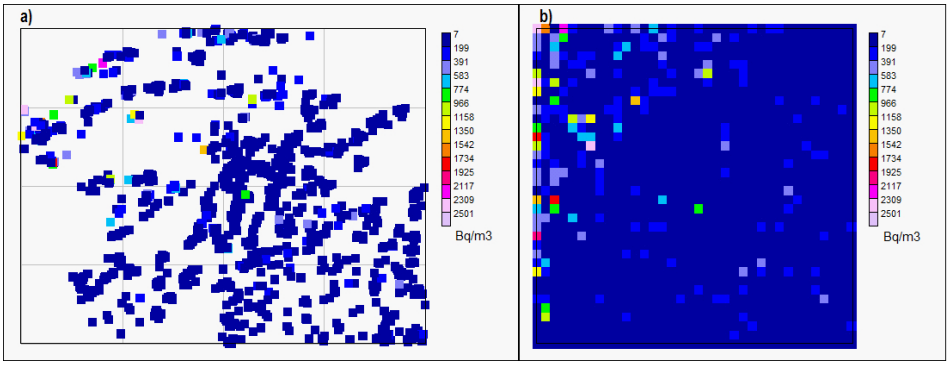



Figure 3.24: *map of the sample distribution for 1,296 samples of a)set3 and b)set3 adjusted to a squared grid*

With the help of this visualization tool, data anisotropy in the NE-SW direction is enhanced. The convexity of the sampled domain poses a difficulty for gridding since some data were largely displaced in order to fill empty spaces. A solution to this could be to constrain gridding to a convex bounded domain.

In order to use this gridding to relate neighborhood with distances, KNNR optima were calculated for different arrangements. KNNR optimization was then launched for set3 with the original spatial arrangement and the gridded arrangement. The corresponding KNNR optimization curves were calculated in the range from 4 to 40 neighbors and are presented in Figure 3.25.

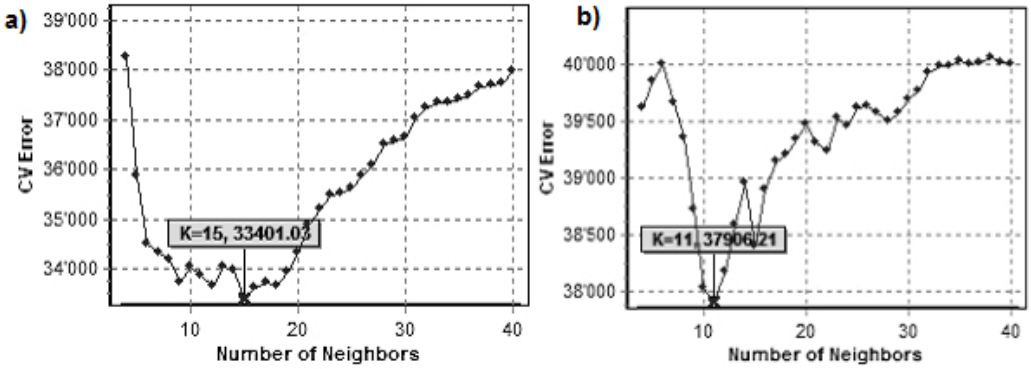



Figure 3.25: *KNNR optimization curves by CV for a) set3 and b) set3 adjusted to a squared grid*

In both arrangements, an optimum minimum KNNR was attained which is an indicator of the spatial continuity of set3. The optimum $k$ is 15 for the original arrangement and 11 when gridded. Despite the lower optimum $k$ for the gridded arrangement, error values are higher. It can be said, that a regular arrangement helps find an optimum, but the prediction errors are higher when the spatial arrangement is disturbed. To illustrate this last statement, data locations were randomly disturbed ('shuffled') to calculate the CV error and compare them with the other two spatial arrangements. In Figure 3.26, the CV optimization curve for

the shuffled arrangement is presented.

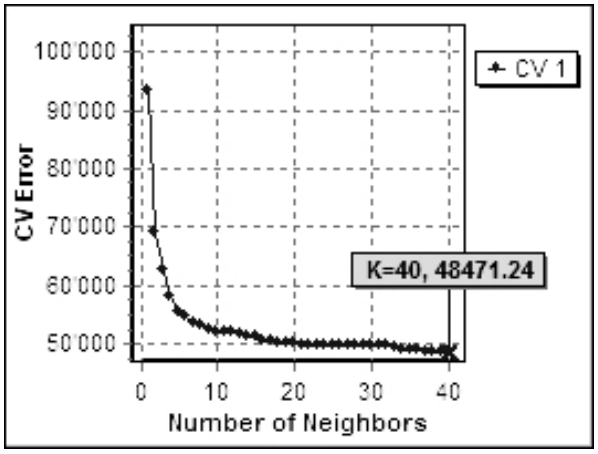



Figure 3.26: *KNNR optimization curve by CV for shuffled set3 data*

The optimum KNNR for set3 was obtained with an error of 33,401; when gridding, the minimum error was increased to 37,906. With a random disturbance of data, the minimum error was 48,471. In fact, no optimum was found for the random arrangement, and the error corresponds to the maximum tested range of 40 neighbors.

The next question is, how should this optimum of 11 neighbors for a gridded arrangement be interpreted? In the original arrangement, the number of nearest neighbors is more influenced by clustering conditions. A fixed $k$ number of points in a clustered sector will cover a smaller area than the area covered by the same $k$ points when they are scarcely distributed. If a fixed distance is adopted between neighbors, as is done with gridding, one can attempt to convert the neighborhood into distance ranges. For the set3 case study, 11 cells will cover an approximate area of 3,500 by 3,500 meters or a radius search of 1,750 meters. Although it is not the same calculation, one can attempt to compare this value with what was observed from the clustering methods. A value of 1,700 m was considered to be the range, below which higher clustering is present. Then, it can be suggested as a reference value to be considered as an optimal neighborhood.

## 3.9 Moving Windows (MW) statistics for set3

### 3.9.1 Local parameters of set3 with MW

MW statistics were calculated for set3 using windows with an approximated size of 6,500 by 6,000 meters. The resulting cells were filtrated so that only those with more than 6 samples were considered. In Figure 3.27, maps for the distribution of different MW parameters of set3 are presented.

This series of maps gives an quick overview of the local statistic parameters. In Figure 3.27a, we distinguish a zone with higher mean radon values to the northwest. Variance

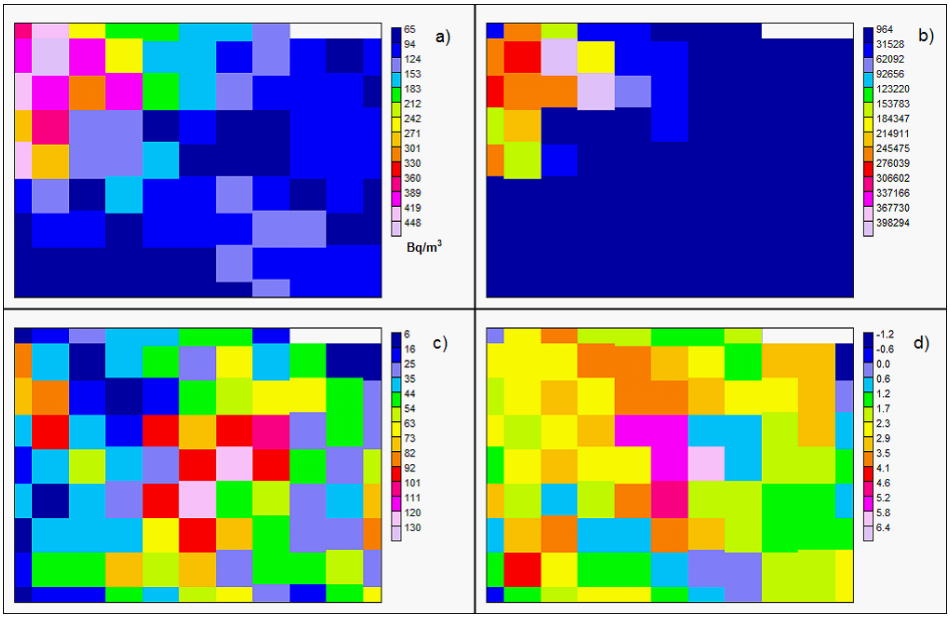



Figure 3.27: *Maps of moving windows statistics of set3 for: (a) the mean, (b) the variance, (c) number of points and (d) the skewness*

is also higher in this area, with a marked contrast in comparison to other areas. A detail that catches our attention, is that the cell in the far northwest direction has a proportional low variance in comparison to the mean for the same cell. The answer is perhaps in Figure 3.27c, where we see that the number of samples for this cell is only 6. The number of points per window gives a hint about the reliability of parameters. Skewness values appear to be more elevated in cells with more samples and at the center of the map, which appears to be a transition zone between areas with lower and higher radon. Skewness, being a third moment about the mean, is highly sensitive to extreme values. In transition areas like the central zone, extremes are more frequent.

The resulting question, is how do these local variations influence the modeling of spatial data distribution? For the purpose of prediction, it is preferable to have a constant mean and variance for subzones; which will result in a global stationarity for these parameters. If the mean is constant while the variance fluctuates, the interpolation method will fall into large errors for areas with high variance. However, if the mean for areas with large variance is also high, the estimate will differ less from the local mean.

### 3.9.2 Proportional effect of set3

When proportional effect exists, we expect to have a positive correlation between the local mean and the variance. Moving windows calculation provides such local parameters. If the local (the window) mean is plotted against the local variance, it is possible to compute the correlation from a linear function. The MW cells from set3 were used, considering the 6,500 by 6,000 distance, but selecting only the cells with more than 10 samples. When using

overlapping windows with this size, almost all data were used so that local statistics became quite representative. With a smaller windows size, many areas were discharged from the analysis.

Proportional effect will change depending on the scale of data. For instance, a logarithmic transform will reduce data skewness (extreme values) and the local influence of high values. The same can be said for transformation into nscores (transformation into a centralized and normalized distribution). Various local means to local variance plots are presented in Figure 3.28, in order to compare the proportional effect for raw data, log transformed and nscore transformed data.

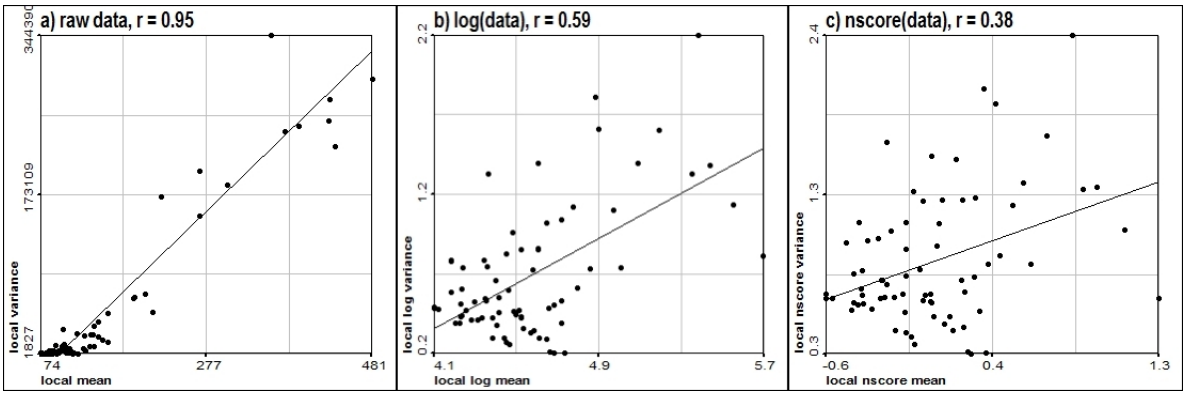



Figure 3.28: *Proportional effect calculated for set3 a) raw data b) logarithmic transformed and c) nscore transforms*

The correlation coefficient for the linear function (r) fitted on top of the mean to variance plot differs depending on transformations. It is observed that it is progressively reduced from $0.95$ for raw data (3.28a), to $0.59$ for logarithmic transforms (3.28b) and to $0.38$ for nscores (3.28c). Raw data clearly indicates a strong correlation or proportional effect, while nscores have reduced the influence of extremes and therefore local variability.

It is also interesting to observe that there are cells with a large regression error that differ very much from other cell values in Figure 3.28a. This is due to high local variances as shown in the map in Figure 3.27b. The dispersion of points in the graph also shows that there are few cells with higher means and variances, which enhance the proportional effect.

### 3.9.3 MW test of lognormal skewness

The test was conducted for each subset of data, after partition into 5 by 5 windows (or cells) and only for windows with more than 20 samples, for set3. The results are shown graphically in Figure 3.29 as a map of lognormality rejection.

The skewness test of lognormality has a spatial distribution that relates to the MW statistics of Figures 3.27a and 3.27d. Lognormality rejection relates to the combined effect of having a high skewness and a lower mean. These conditions are present, in what was called the

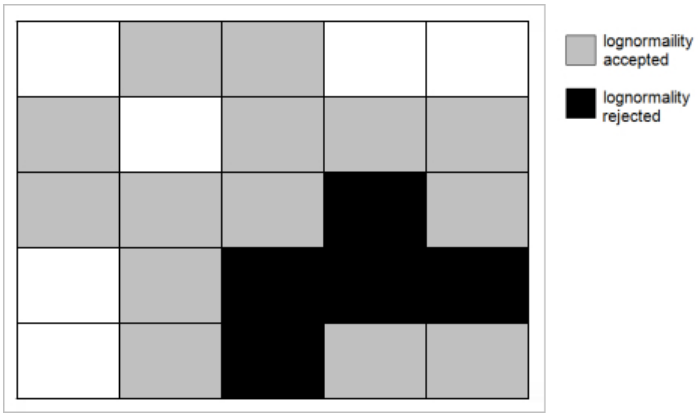



Figure 3.29: *test of skewness lognormality for moving windows for the set3*

transition zone, in the middle of the spatial domain for set3. The number of samples can also influence the test, since the number of points per windows is not homogeneous.

### 3.9.4 Spatial data partition of set3

One criterion to improve the spatial modeling of set3, is to produce subzones with a similar local mean. One idea on how to establish a boundary for data partition is to use MW mean statistics together with the identified transition zone. First, the local mean for a grid of 9x8 windows was calculated. This number of windows results in a mean of 21 points per cell, which is a minimum to consider for estimating parameters. For this partition, cells with a minimum of one point were also considered. In principle, the more points per cell available, the better statistics. Using these MW averages and considering the approximate position of the transition zone, a boundary was drawn to obtain a low and high radon area (sets A and B shown respectively in Figure 3.30).

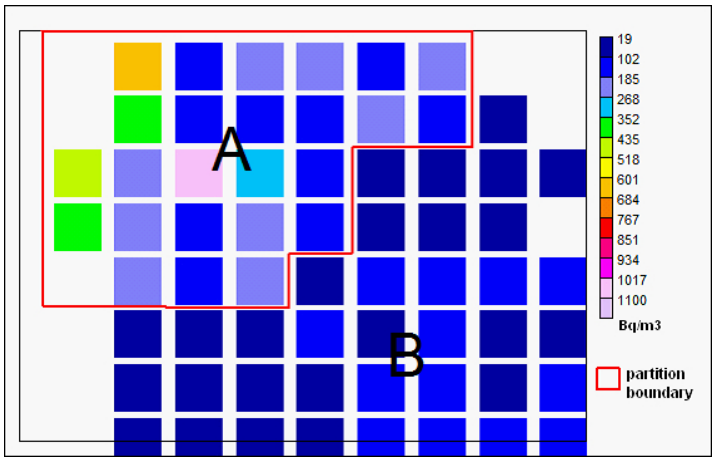



Figure 3.30: *Partition boundary based on MW mean for the set3*

In this way, MW was used to split the data of set3 to obtain a subset with stationarity of

the local mean (set3B).

The case of proportional effect for subset 3B is quite different from set3. As explained, set3B was reselected considering zones with a more constant mean and variance. This is reflected in the local mean to variance plot in Figure 3.31, where correlation is $0.43$ for raw data and $0.008$ for nscores.

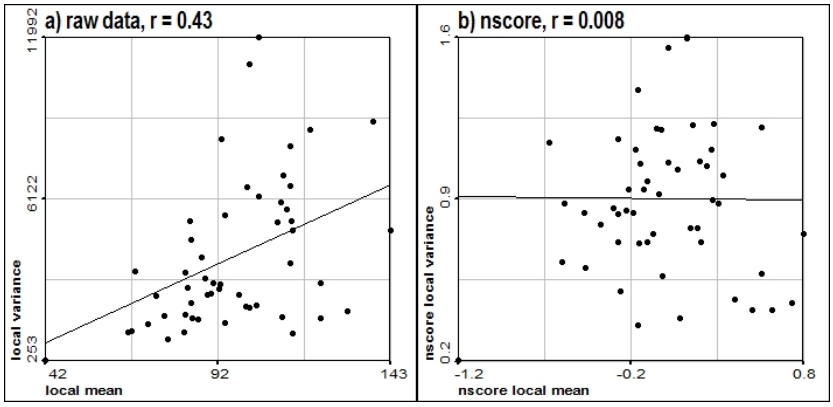



Figure 3.31: *Proportional effect calculated for set3B a) raw data b) nscore transforms*

## 3.10 Variography analysis for set3

### 3.10.1 h-scattergrams

Figure 3.32 presents h-scattergrams for set3 considering distances $h = \{100, 800, 1600\}$. Distance values are not precise but they lie within a range with a certain tolerance as points seldom have a precise gap between them. In this sense, the set of distances is better defined as $h = \{0 - 100, 700 - 900, 1500 - 1700\}$. These ranges correspond to short distances, the average distance between points and a distance over this average.

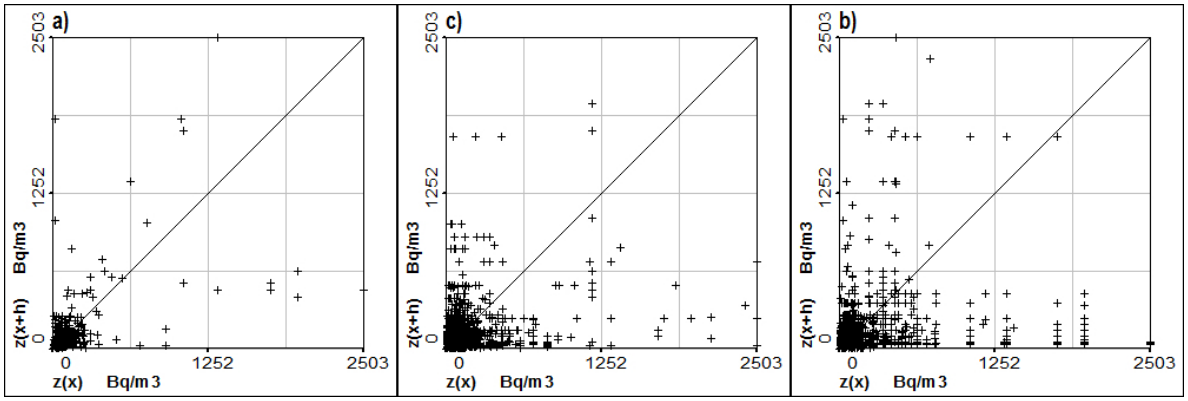



Figure 3.32: *h-scattergrams of the set3 for three distances a) h=100 m. b) h=800 m. and c) h=1600*

Differences between graphs are not as evident, with most points lying at the origin of the quadrant and some extremes with more scattering. Nevertheless, it is evident that the points at a mean distance of 100 meters are more correlated or, in other words, have less dissimilar values. On the graphs, a diagonal bisector line were also included to be used as a reference to calculate what is called the experimental variogram in geostatistics.

Using equation 3.8, the individual semivariance was calculated for $h = \{100, 800, 1600\}$ for set3 with the same tolerances $t$ used for the h-scattergrams $t(h) = \{50, 100, 100\}$. The semivariances for these distances are $\gamma(h) = \{33562, 67089, 25390\}$ and the number of pairs $N(h) = \{373, 1351, 1928\}$ respectively. It is interesting to observe that semivariance values increase until the distance of 800, and then, they decrease. The number of pairs logically increases with distance. If we compare the values with the sample variance $Var = 46929$, also called the *a priori* variance, we observe that it was surpassed at the distance of 800.

What is particular about this distance, and why is it lower at a 1,600 m lag? Some hints can be found with our referential distance of 1700 m. In fact, the semivariance for $h = 1700$ with a tolerance $t(1700) = 100$ results in $\gamma(1700) = 17746$, with $N(1700) = 1925$; which is the lower semivariance in the range of $h = 5000$. The variogram function for set3, within a range distances from 100 to 5000 m, is graphically displayed in Figure 3.33 together with the number of pairs of points and the *a priori* variance. These results show a relation between clustering and spatial continuity of values. Neighboring data appear more correlated and this occurs at a clustering limit of 1700 m for set3.

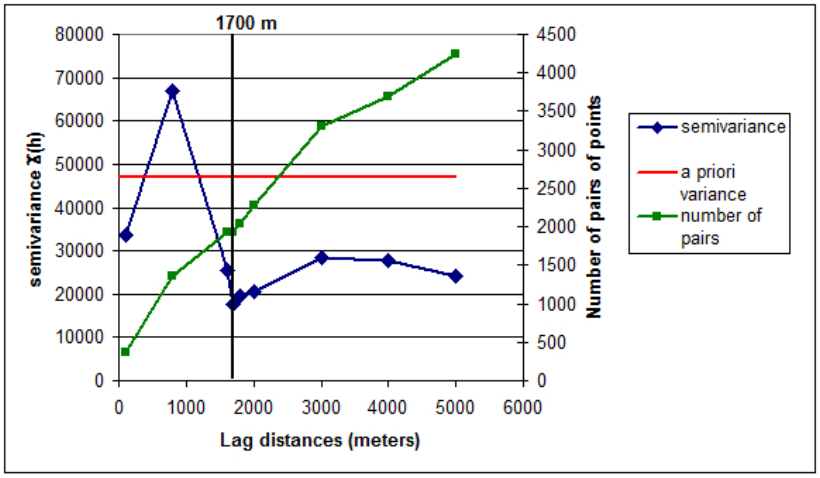



Figure 3.33: *Variogram function $\gamma(h)$ for set3 for distances between 100 and 5000 meters*

In practice, variograms are built considering a fixed lag and a tolerance distance until reaching an approximate distance range that is half the diameter of the bounding box. The initial distance can be as low as the data spacing and the set size permits. It is important to have enough pairs of points at any distance. As a rule of thumb (8), it is recommended that one have no less than 50 pairs of points to calculate the semivariance. The tolerance is usually half of the lag distance. Another variogram was built for set3 considering such hyper-parameters (see Figure 3.34).

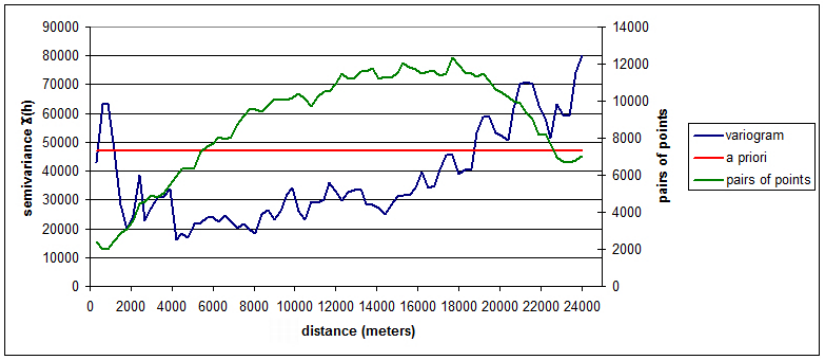



Figure 3.34: *Variogram function $\gamma(h)$ for set3 for distances between 500 and 25000 m.*

The variogram for set3 shows that semivariance values are very high for distances below 1700 m. The number of pairs of points at 300 m is over 2000, which must give an acceptable semivariance statistic. The number of pairs of points increases with distance until it reaches approximately 20,000 m; it then decreases from then on. The choice to limit the distance to half the bounding box diameter is based on the idea of always obtaining an acceptable number of pairs of points. It was also observed that, even for the 500 m distance, a good number of pairs exists. This allows a detailed variogram going from lag 50 till 2500 m to be built, in order to analyze the variations at very short distances (Figure 3.35).

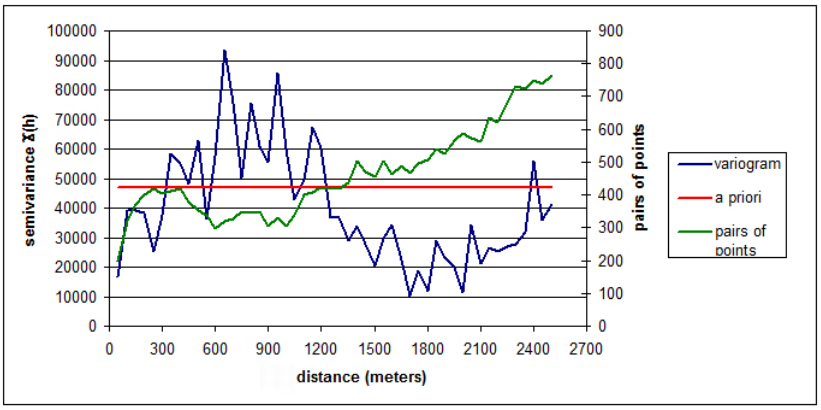



Figure 3.35: *Variogram function $\gamma(h)$ for set3 for distances between 50 and 2500 m.*

The number of pairs of points are considerably high for short distances because of the high spatial clustering. Nevertheless, it is remarkable that the number of pairs has a decrement, while the semivariance increases between 600 m and 900 m. Below 600 m, the variogram presents a defined spatial variance (or correlation) structure. Is this a dependency on the 'urban' spatial arrangement? Are there different sub-zone populations for which local variability are so different? To address some of these questions, a change in the scale of analysis using MW was subsequently performed.

The madogram and the drift for set3 data, were also calculated and represented in graph 3.36.

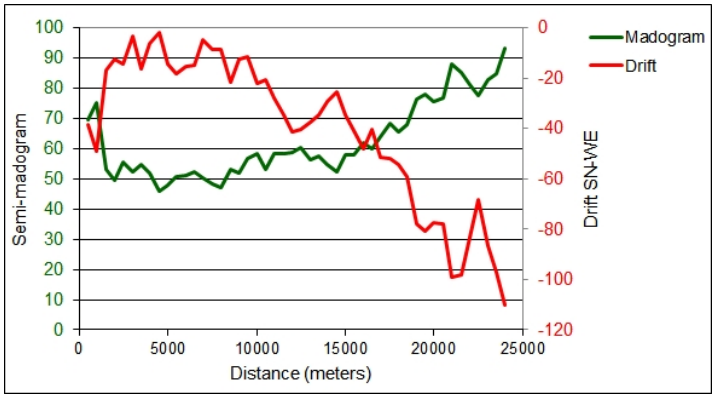



Figure 3.36: *Semimadogram function $M(h)$ and Drift for set3*

The madogram has a similar shape to the semivariogram. Although variances are significantly smoothed, it does not result in a useful measure to define a range of continuity. Moreover, the high local variability is always evident, as for the semi-variogram. On the other side, the drift line provides a clear indication of the data non-stationarity.

### 3.10.2 Variography using MW averaging

MW averaging was performed to observe the spatial correlation after the local variance had been eliminated. An important effect of MW averaging is that the local mean is assigned to a position into a grid, making the spatial distribution of data more regular. A series of variograms were built after averaging values within several window sizes. The number of samples logically decreases after averaging over larger windows and the distribution becomes more regular. To calculate the corresponding variograms, the lag distance must be increased as well. From a set size $n = 1310$ we end up with $n = \{1048, 875, 417, 234\}$ for $MW(h) = \{200, 300, 1000, 1700\}$. Four variograms, following MW averaging, are presented in Figure 3.37.

After MW averaging at 1000 m. (Figure 3.27d), it was possible to identify a continuous spatial structure going till a distance of 13000 m. Does this mean that subpopulations within the dataset exist? It was already seen that local variations exist and that skewness is higher on the transition zone between the north-west and the southern areas. Next, it will be analyzed if subpopulations within moving windows approach to a lognormal statistical distribution.

### 3.10.3 MW local variograms of set3

Statistics are not the same for all sub-regions in set3, and a proportional effect was also observed (Figure 3.28). It will be interesting to see how local variograms are influenced by these statistics and eventually whether some patterns can be recognized. Set3 was partitioned into 5 by 5 windows, and local variograms were calculated. The lag distance for the variograms is around 200 m with some variations depending on the dispersion of data. The results are presented in Figure 3.38.

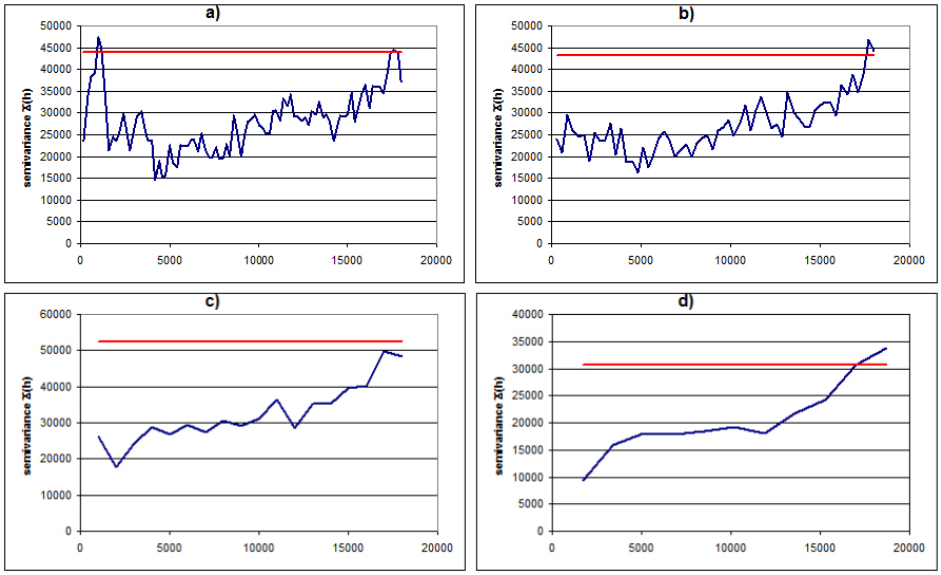



Figure 3.37: *Variograms for set3 after MW averaging at distances a)200 m. b)300 m. c)1000 and d)1700*

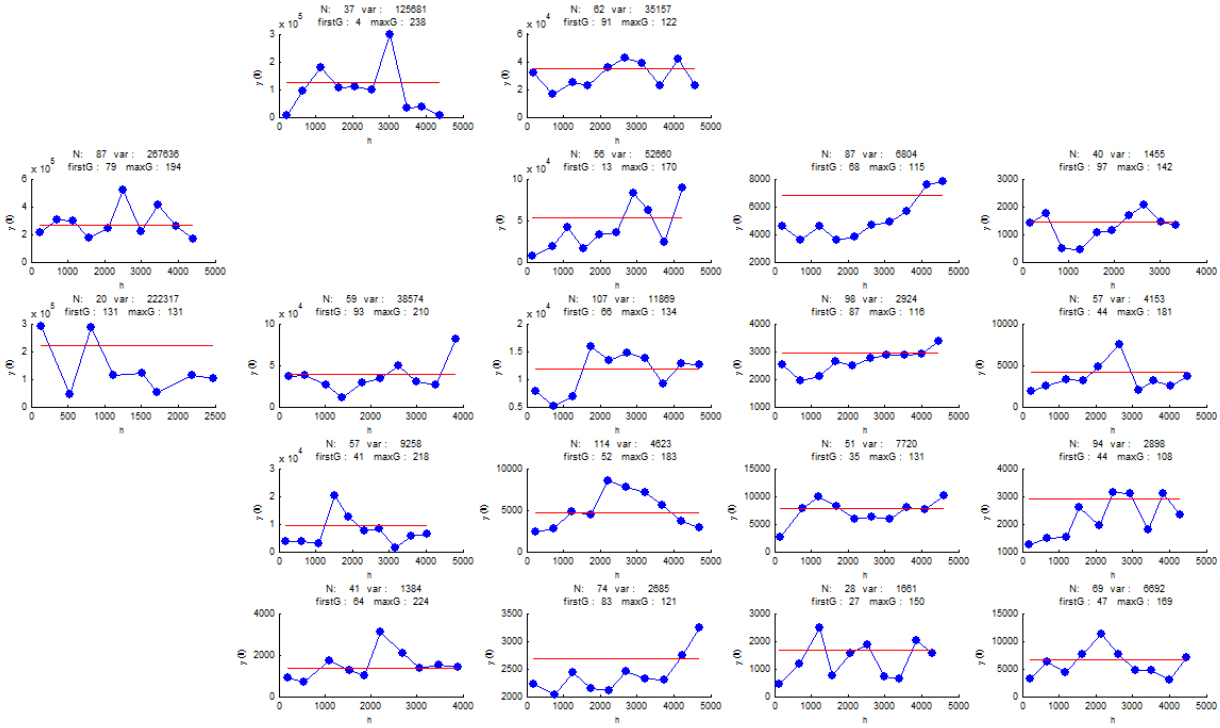


Figure 3.38: *Local variograms for set3 in 5 by 5 windows*

Some local variograms have the same pattern as the global variogram, in the sense that the semivariances at shorter distances are comparatively high. Some variance structures are

very short or simply not present. There are locations with structured variograms; in those cases, the semivariance increases with distance. To summarize this feature, a pattern map was built to indicate whether the local variogram presents structured semivariance or not (Figure 3.39)

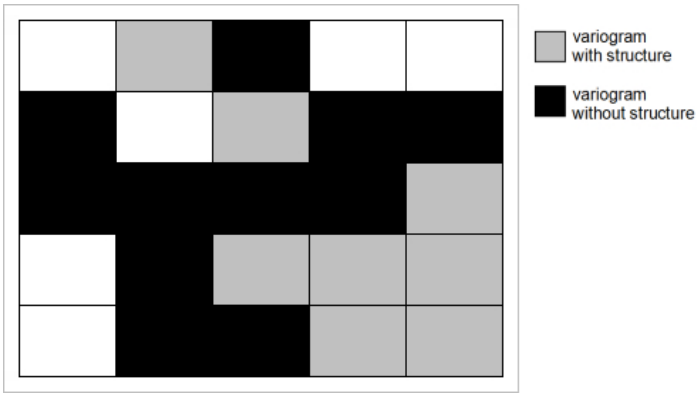



Figure 3.39: *Presence of structured semivariance in local variograms for the set3*

An important outcome from this analysis is that some local variograms appear structured; some local data has more spatial continuity, while the global variogram presented only a short structure (Figure 3.35). The so-called 'transition' zone presents more difficulties to be modeled. Local variograms have large semivariance values at the first lag distance ('nugget effect'). With the spatial tools analysis it was possible to depict a transition zone between a northwestern and an southeastern area. Hence, it would be wise to analyze these two areas separately.

## 3.11 Spatial characterization of the set3 subzones

After partition, two subsets (sets A and B) were created, compared and analyzed using the already presented tools. A rapid statistical and spatial characterization of the subsets was done, including statistical parameters, average distance, functional MI for average distance, MI diagrams, optimum KNNR and the experimental variogram. The first spatial indicator presented in Figure 3.40 is the quartile MI diagram for subsets A and B considering the average distance of points.

This characterization is very explicit, because it shows a completely different clustering behavior for the subsets. Subset A, as the total set3, has heavy clustering for lower and higher values while subset B has a light progressive clustering towards lower values. To illustrate, more in detail, what the spatial distribution of high values represents at different scales, MI diagrams for the fourth quartile are shown in Figure 3.41.

While the MI diagram for the fourth quartile (Q4) of subset 3B has more clustering at short distances, Q4 for subset 3A increases abruptly at a distance of 700 m. This clustering has already been seen for set3, but in this case, it is present at smaller distances. The fact

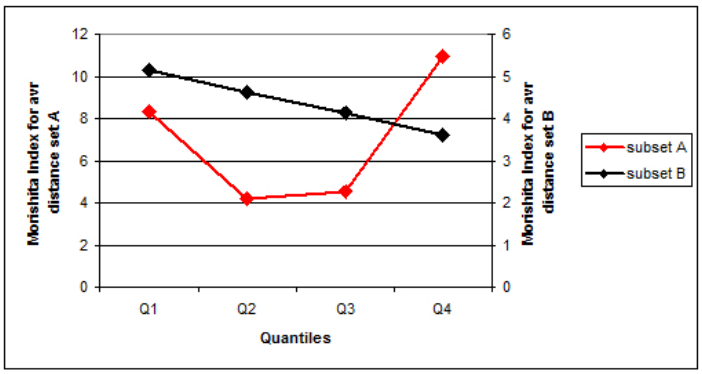



Figure 3.40: *Quartile Morisita Index diagram at average distance for: a) subset A and b) subset B*

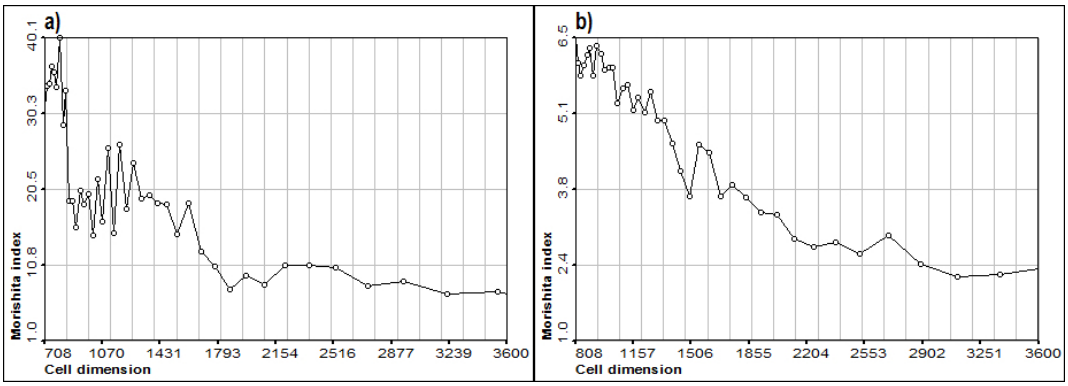



Figure 3.41: *Fourth quartile (Q4) MI diagram for: a) subset A and b) subset B*

that this occurs, also for lower values, has the effect of increasing local variability and makes it difficult to find spatial structures by variography. Subsequently, the experimental variograms for both sets were calculated considering a lag distance of 200 m, and a tolerance of 100 m over a distance of 5000 m (Figure 3.42).

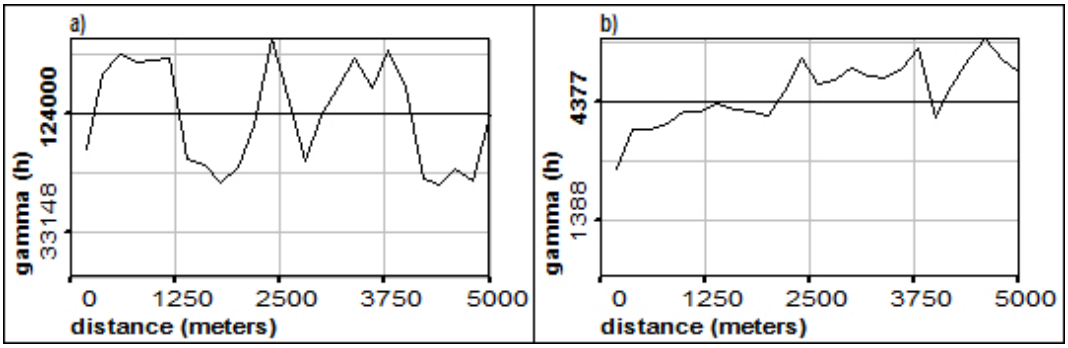



Figure 3.42: *Experimental variograms for a) subset A and b) subset B*

The variogram for subset 3B appears to have a semivariance spatial structure with lower local variability, while subset 3B appears unstructured or with very short structures. Perhaps

it is this local variability that is reflected in a lower optimal number of neighbors after KNNR analysis: 13 for subset 3A as opposed to 15 for subset 3B. In any case, globally, datasets have very dissimilar mean and variance values. A summary of these statistical and spatial characterization parameters can be seen together in Table 3.2.

Table 3.2: *Statistical and spatial characterization of set3 subsets*

| | **subset 3A** | **subset 3B** | **set3** |
|---|---|---|---|
| **N** | 419 | 891 | 1310 |
| **mean** | 238 | 97 | 142 |
| **variance** | 124014 | 4377 | 46929 |
| **skewness** | 3.59 | 3.42 | 6.08 |
| **Average distance** | 1178 | 967 | 968 |
| **QMI for Average dist** | Q1 and Q4 high | homogeneous | Q1 and Q4 high |
| **clustering rise** | c < 800 m | c < 1600 m | c < 1700 m |
| **optimum KNNR** | 15 | 13 | 15 |
| **KNNR minimum MS error** | 96660 | 4154 | 33801 |
| **variogram** | unstructured | structured | unstructured |

A last point to draw attention to is that skewness for both subsets are lower than for set3. This is due to the lower number of points and to subsets being more homogeneous in values. Partition seems, for the set3B case, to have a part in the success of modeling. Other solutions must be found to model subset A. A sub-partition, in this case, will probably not have the desired results since local means are more dissimilar (see Figure 3.43) and we can end-up with small datasets, which is not desirable for modeling.

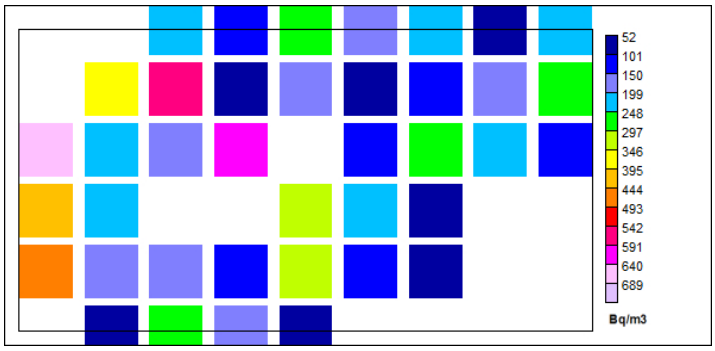



Figure 3.43: *Map of local means by MW averaging of the subset A*

As seen earlier, MW averaging is an option that reduces this somehow 'disturbing' local variability. On the other hand, MW averaging also produces a drastic reduction of sample size and prevents modeling at short distances. Averaging is a data transform that implies a loss of samples. Other transforms, like indicators, classification and n-score transforms, also provide solutions for modeling data having high variability. These methods will be explained later in the chapter dedicated to prediction.

## 3.12 Multiscale analysis and spatial partition

So far, different methods for spatial data analysis have been presented and applied for a reduced dataset of indoor radon measurements (set3). Set3 was used primarily to test spatial methods. It has been observed that clustering is present, and the spatial distribution of data is not the same for all quantiles. Data spatial variance is, therefore, concentrated at very short distances, which was an impediment for regional modeling. An important output from spatial analysis is that data partition into homogeneous regions can offer a solution to improve spatial modeling.

In accordance with the first and foremost objective of the present study, which is the analysis of the indoor radon on a national scale, data partition can be proposed as a method to find spatial continuity for subsets of the whole dataset. The next question to answer is, at which scale is it possible to obtain the most information from the data? Different scales must be analyzed.

Diverse possible scales of analysis exist for the Swiss radon dataset; some of them have been mentioned already. The starting point, of course, is the global scale of the Swiss national territory. Secondly, a physical factor approach, such as the natural regions scale, has been considered. Moving windows is a third multi-scale definition that has been taken into account. Finally, the administrative units' multiscale has provided a fourth scale definition.

To summarize, four multiscale definitions have been proposed for the Swiss indoor radon dataset:

1. National scale
2. Natural regions units
3. Moving windows scales
4. Administrative units

### 3.12.1 National scale analysis

As mentioned in section 2.1.1, the selected set for the Swiss indoor radon analysis consists of 41,787 samples. This selection corresponds to measurements in inhabited buildings on the ground floor. It is also the result of filtering out samples with misplaced coordinates (e.g. lakes) and samples from Liechtenstein. Values with the same coordinates were removed to avoid calculation problems during spatial analysis (that was a major data reduction). Finally the sampling points were superimposed with a layer of polygons representing the built areas in Switzerland. A reduced number of samples lying out of this domain were removed. This dataset has a mean value of 163 $Bq/m^3$, a median of 85 $Bq/m^3$, a standard deviation of 321 $Bq/m^3$,a variance of 102,873, a maximum value of 15,045 $Bq/m^3$ and a skewness of 12.16.

#### Swiss indoor radon decile maps

In addition to global statistics, it is important to represent the spatial distribution of data, in order to analyze whether there is an association of high or low value locations with clustering

of data. In other words, it is important to understand how values are distributed in space to eventually decide whether any declustering technique can be later applied. Decile maps show clustering of low or high values. The first and the tenth deciles were reselected from the dataset and plotted on a map (Figure 3.44). According to the map, there is spatial clustering for both, low and high values.

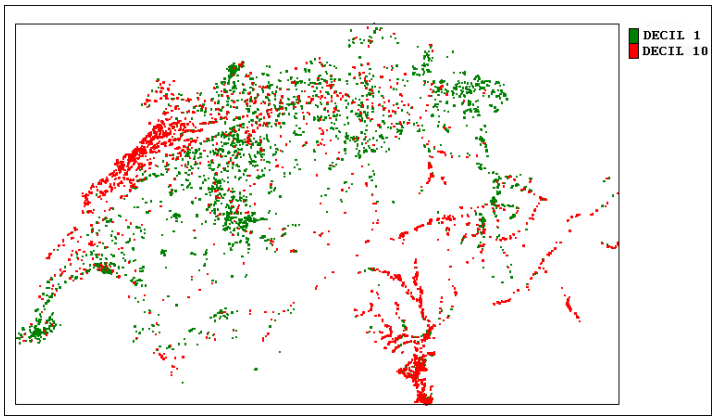



Figure 3.44: *Spatial distribution of first decile data (green points) and tenth decile (red points)*

**Swiss indoor radon fractal dimension by sandbox counting**

In order to analyze the status of clustering for the Swiss radon dataset, the fractal dimension using the sandbox counting method was computed. Clustering was analyzed for different spatial domains and for synthesized random point datasets. Spatial domains considered were a bounding box for the Swiss territory and the domain representing the built-up areas. For all clustering calculations the average distance was taken into consideration: 1339 m for original data, 1354 m for random points within bounding box and 1348 m for random points within built-up areas. The half distances of the shorter side of the bounding boxes are 108706, 110075 and 109916 meters respectively.

Clustering using the sandbox counting method was applied to a set of 41787 random points within the bounding box of the Swiss territory (Figure 3.45). Then, it was computed for the 41787 indoor samples from Switzerland (Figure 3.46). Finally, the same numbers of points were randomly distributed within build-up areas (Figure 3.47). In each case, a sandbox diagram shows the relation between the logarithm of the number of points $n$ and the logarithm of the radius $r$.

As seen in Figure 3.45, the random arrangement in the bounding box area has a $Df = 1.89$. The sandbox method gives a fractal dimension that is below 2 for the case of random points, due to border effects. Therefore, $Df = 1.89$ can be considered as a referent maximum fractal dimension.

In Figure 3.46 the fractal dimension is lower for original indoor random samples and is not uniform. The slope of the curve changes noticeably with scale. Fractal dimension can

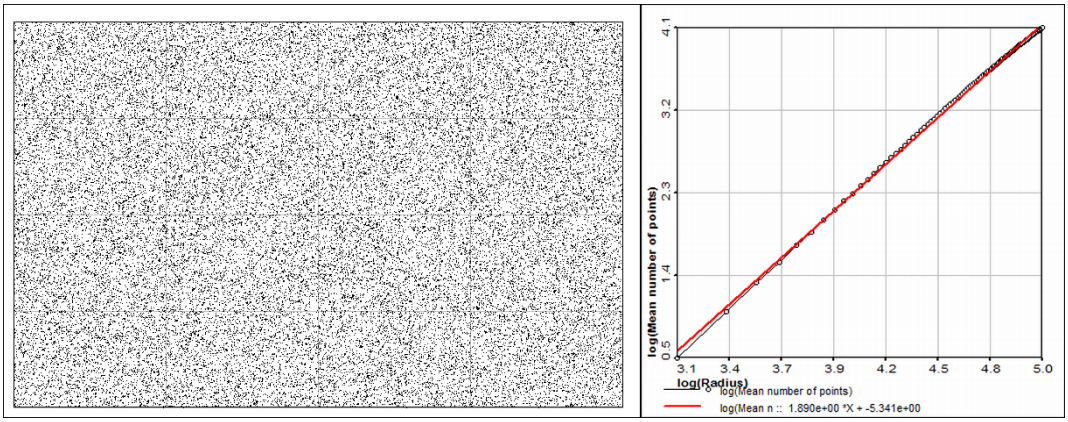



Figure 3.45: *Distribution of random points within the boundingbox of Swiss territory and the corresponding sandbox counting graph*

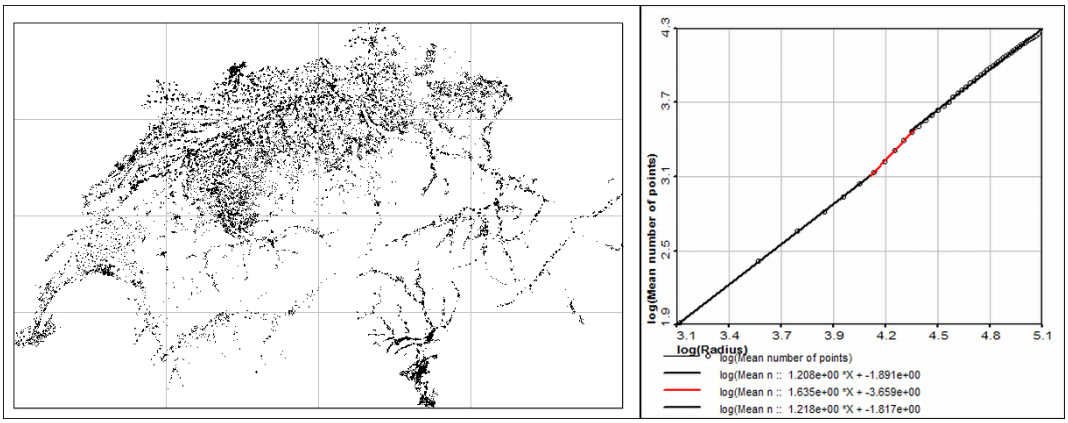



Figure 3.46: *Distribution of random points within populated areas of Swiss territory and the corresponding sandbox counting graph*

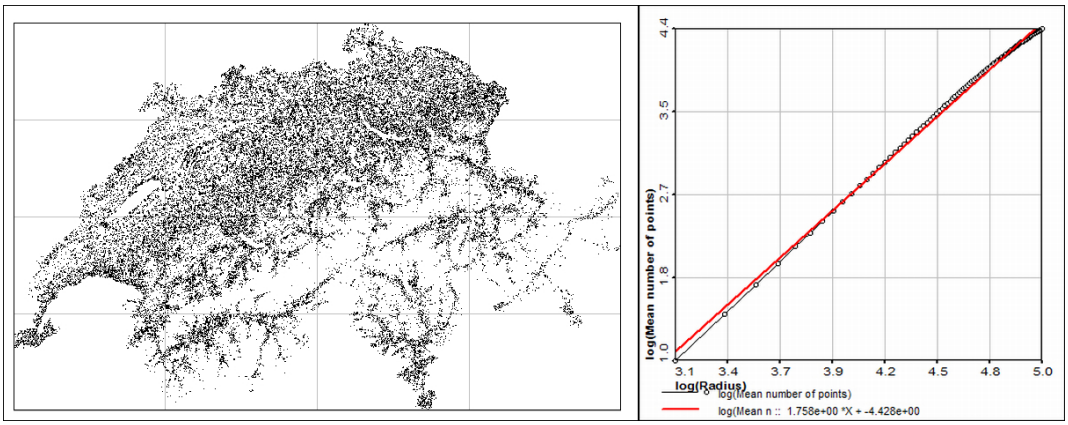



Figure 3.47: *Distribution of radon measurements within populated areas of Swiss territory and the corresponding sandbox counting graph*

be subdivided for distances below 14000 m (with a $Df = 1.21$) and between the range of 14000-23000 meters with a $Df = 1.63$. Clustering of data is then more pronounced below 14000 m. Fractality of random points within the built-up area have a little more clustering than that of the random points within the bounding box. With a $Df = 1.76$, it can be said that the built-up area proposed has a more homogeneous coverage within Switzerland than indoor radon samples (with a $Df = 1.30$).

There is a decrement of the fractal dimension at large distances in the sandbox diagram for random points within built-up areas. This decrement is due to the effect of the national border; cells in the border have progressively lower points, but at least one should be included in computations.

The three linear functions for the sandbox diagrams were joined into one comparative graph (Figure 3.48).

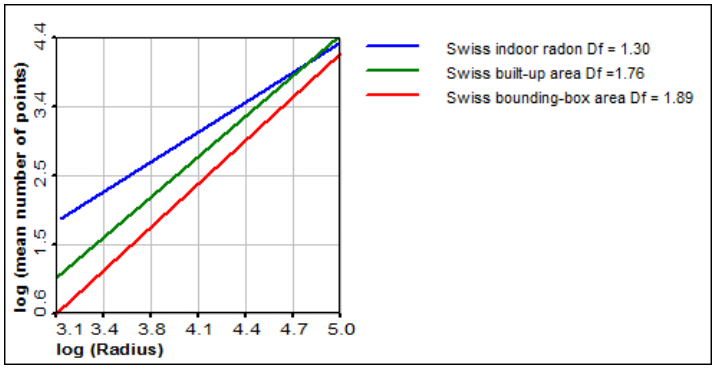



Figure 3.48: *Sandbox counting for random points and radon measurements*

It should be pointed out that the built-up area is certainly larger than it is in reality and this fractality dimension of $Df = 1.76$ is probably overestimated. In any case, it is an appropriate domain to be used for predictions.

#### Functional Morisita index on a national scale

Samples over 1000 Bq/m$^3$ were plotted to show their level of clustering. Map 3.49a shows data over 1000 Bq/m3 as purple crosses concentrated in the Jura and Ticino regions. To verify the sensitivity of MI compared to randomness conditions, the dataset was 'shuffled', so that measurements were randomly redistributed within sample coordinates. In Figure 3.49b, shuffled values have been plotted as blue crosses. Finally, a graph showing MI diagrams for the two datasets is presented in Figure 3.49c.

The MI diagram for shuffled data has quite the same trace as raw data since sample positions have not been hanged but partially re-selected. Taking the map and the MI diagram for data over 1000 Bq/m$^3$ into consideration, the high degree of clustering becomes more than evident. The question remaining is does this clustering have a tendency to increase for higher values, or (as seen for previous analysis of set3) does it also occur with lower values? This can be depicted with the functional MI at average distances (Figure 3.50).

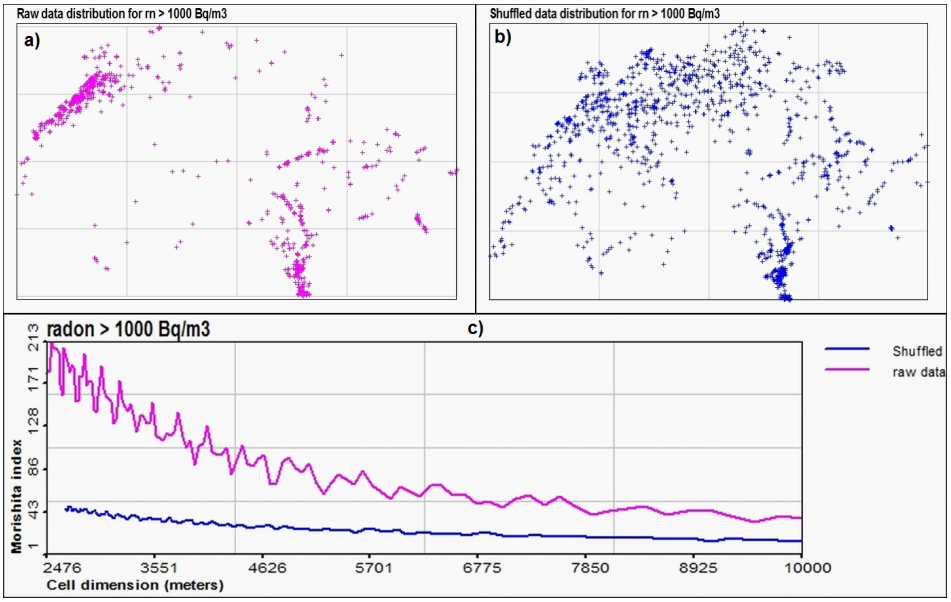



Figure 3.49: *Functional Morisita index for indor radon values over 1000 Bq/m3 a) map of indoor radon over 1000 Bq/m$^3$ b) map of shuffled data c) MI diagrams for indoor radon samples and shuffled data*

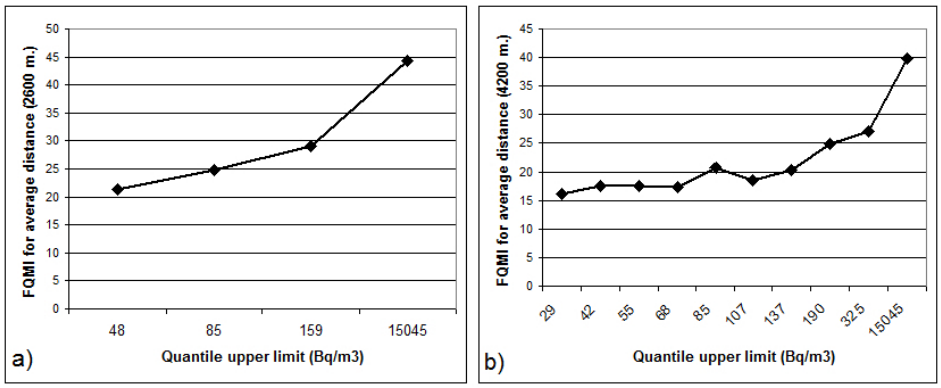



Figure 3.50: *Swiss indoor radon measurements Quantile MI profile at average distance for a) quartiles b) deciles*

In Figure 3.50a, MI was calculated for quartiles, while Figure 3.50b shows details at decile resolution. These images show a clear clustering tendency for higher values. It seems that the existence of the high concentration areas mentioned (Jura and Ticino regions) could be a cause for this clustering. It should be noticed that, for the last quartiles (samples over 159 Bq/m$^3$), clustering becomes exponential. This situation differs with set3, where clustering is present for low and high values and less marked. This indicates the presence of hotspots on a global scale and hence, enhances the continuity of values. It is also important to observe this continuity on shorter scales.

### Variography of the Swiss dataset

Several variograms were produced for the Swiss indoor radon dataset using the 41787 samples. The global variogram (Figure 3.51a) was calculated with a lag of 22 km up to a distance of 220 km. The intermediate variogram (Figure 3.51b) has a lag of 1000 m and a total distance of 50 km. The third variogram is more local (Figure 3.51c), with a lag of 200 m and for a distance of 10 km.

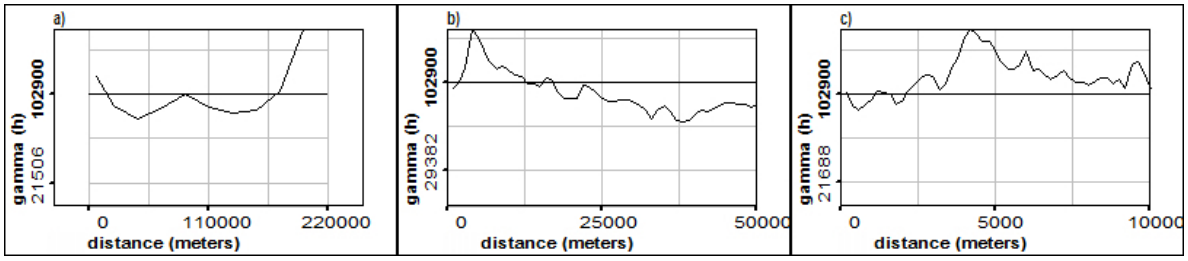



Figure 3.51: *Experimental variogram of raw data for indoor radon with lag distance a) 22 km, b) 1000 m. and c) 200 m.*

It is difficult to visualize spatial continuity for the Swiss dataset out of variograms. Globally, semivariances are high at distances below 22 km. When looking locally (below 5000 m), there is some continuity but with a high nugget effect upon departure and high semivariance values that decrease after 5 km. MW averaging was used to reduce local variability, and two variograms were produced to see the semivariance on global and local scales (Figures 3.52a and 3.52b). Due to the large dataset size, the minimum possible window size that could be used was 280 m.

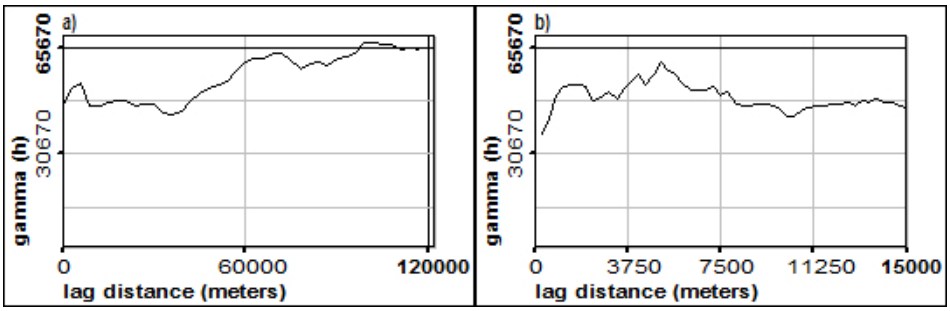



Figure 3.52: *Experimental variogram for MW 280 m. averages of indoor radon with lag distances a) 3 km, and b) 300 m.*

With MW averaging, local variance is drastically reduced regarding the total a priori variance. For the first 40 km, the variogram consists of a nugget effect, and it shows increments only after this distance. Very locally (below 5000 m), the averaged data present a steady semivariance increment. Once the local variance was filtered-out by MW averaging, it became possible to see that variations with spatial continuity occur at long distances only. Next, variography for MW averages with larger windows were tried. A variogram using

MW averaged indoor radon at the average distance (1339 meters) is presented in Figure 3.53. The lag distance considered was 1400 m.

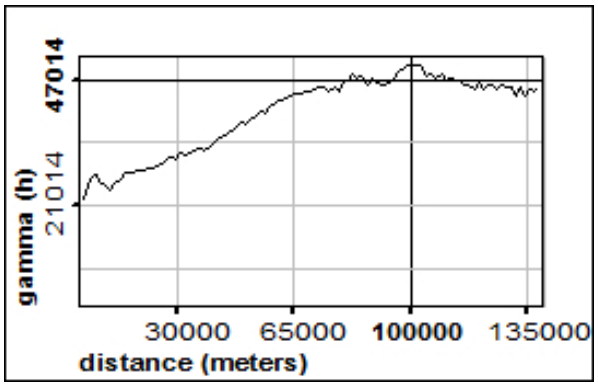



Figure 3.53: *Experimental variograms for indoor radon MW averages at 1339 m.*

This last variogram has a spatial continuity that can be modeled up to a distance of 100 km (range distance). The semivariance at origin (nugget) is well below the a priori variance (less than 50% of this). These conditions are acceptable for variogram modeling and prediction and are encouraging to proceed further with variography analysis. Unfortunately, they correspond to a coarse scale. This scale of prediction is less interesting because it does not benefit from the high volume of data and the local variability. By MW at 280 m. averaging the original dataset is reduced to 22389 values and to only 7235 values by MW at 1339 m. MW also modifies also the spatial arrangement and imposes a limit for the resolution of prediction nets. This modification can be perceived with a comparative of Morisita Index diagrams for raw data and MW averages (Figure 3.54).

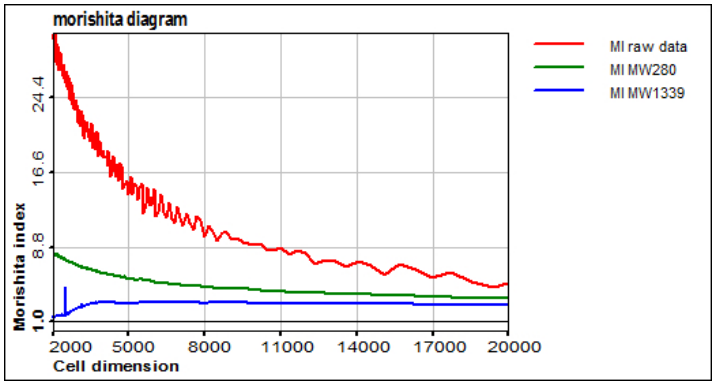



Figure 3.54: *Morisita index diagrams for raw data and Moving window averages at 280 and 1339 meters*

Not only clustering but also variance is reduced with MW averaging: from 102873 for the raw data, to 65670 for MW280 and to 47006 for MW1339. In fact, it is evident that much of the variance lies on a short scale because samples are taken inside buildings that are often located tenths of meters apart. Then, the scale of modeling and prediction depends on

whether maps on a national scale or on a more local scale must be produced. In order to obtain workable variograms on a more local scale, without modifying the values and the spatial arrangement, it would be wise to perform some data partition.

### 3.12.2 Natural regions

The first idea on how to define coherent regions for indoor radon predictions come from the results of Chapter 2, where interesting conclusions were given. After a multivariate analysis it became evident that at least two natural variables, geology and elevation, have a significant influence on indoor radon concentrations. For geotechnical units, the scale of analysis plays a crucial role in understanding the phenomena. The trends of the indoor radon spatial distribution in relation to geology were very clear on the global-national scale; while on a local scale the high spatial variation prevented distinguishing tendencies.

Actually, the Swiss natural regions are largely defined by the combination of the geological and elevation factors. The distinctive morphology of the Swiss landscape permits the division of the territory in three major regions: the Jura, the Plateau and the Alps region. It can be roughly said, that the Jura is an elevated formation with an important presence of calcareous conglomerates and rocks. The Plateau is the lower region and typically has sedimentary formations. The Alps is the region of high mountains, where igneous rock formations can be found. Natural region definition is certainly a broad division, but is also a coherent criterium of data partition that relates to the indoor accumulation physical process.

The map of the three Swiss natural regions is presented in Figure 3.55a. Figure 3.55b shows the location of indoor samples' points superimposed with the approximate region boundaries used to partition the global dataset.

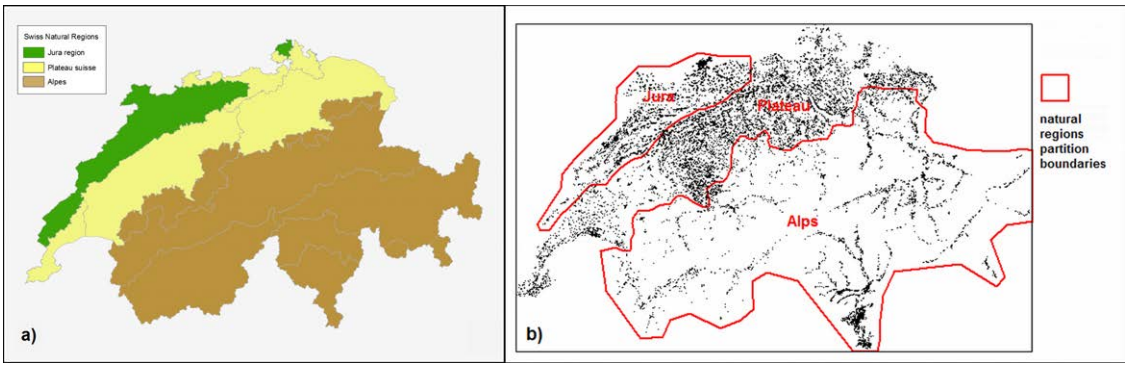



Figure 3.55: *a) Map of the Swiss natural regions b) Map of indoor radon samples partition according to natural regions boundaries*

A table of statistical and spatial subsets characterization were initially produced (Table 3.3).

It can be noticed that the Jura region subset tends to be effectively more homogeneous, with a skewness of 5.9 (despite having a high variance). The Plateau set has a lower mean

Table 3.3: *Statistical and spatial characterization of Swiss regions subsets*

| Indoor radon samples | Jura | Plateau | Alps | Swiss |
|---|---|---|---|---|
| **N** | 9342 | 16045 | 16400 | 41787 |
| **mean (Bq/m3)** | 235 | 97 | 187 | 163 |
| **variance** | 185107 | 13194 | 135964 | 102873 |
| **skewness** | 5.9 | 10.8 | 14.1 | 12.2 |
| **average distance (m.)** | 612 | 1111 | 1082 | 1339 |
| **bounding box min side** | 112533 | 183018 | 174848 | 217412 |
| **F quartile MI for Avr** | Q4 high | Q1, Q4 high | Q2 high | Q4 high |
| **optimum KNNR** | 13 | 42 | $> 50$ | |
| **KNNR minimum MS error** | 141419 | 12520 | $\sim 122515$ | |

and variance but appears to be not as homogenous (with a skewness of 10.8). Finally, the Alp set is the more heterogeneous with a skewness of 14.1. The level of variance is simply proportional to the high mean. Particularly interesting are the differences in skewness values, as were also seen for in the set3 analysis. Homogenous areas present lower skewness.

The diagrams for the functional quartile MI at an average distance for natural regions are also very dissimilar (Figure 3.56).

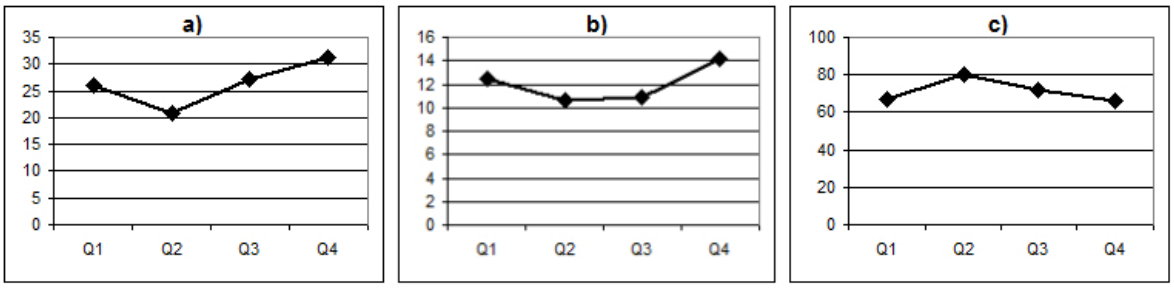



Figure 3.56: *Functional quartile MI for average distance diagrams for a) Jura region b) Plateau region and c)Alps region*

The Jura region shows a tendency of clustering of high values (MI for quartiles Q3 and Q4 are high). The plateau region has high local variability (Q1 and Q4 are high), as was seen with the set3 QMI profile diagram. Finally, the Alps region subset has a particular diagram where central values (Q2 and Q3) are more clustered (the inverse situation of the Plateau subset). Thanks to a larger number of points per dataset, MI values have fewer variations between quantiles. Spatial distribution becomes globally more homogeneous. The optimal KNNR for Jura is 13 neighbors, while Plateau and Alps have an elevated number of neighbors. From these parameters, it seems that the spatial variance is more structured for the Jura subset.

Experimental variograms were then produced for each of the subsets (Figures 3.57, 3.58 and 3.59). Two variograms were prepared for each subset, one global with a lag distance of 1000 m and one local with a lag of 200 m.

The Jura subset appears to have a more structured behavior, but at short distances.

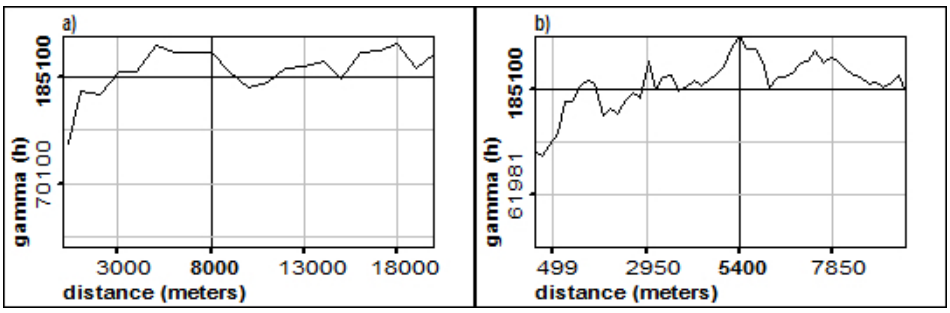



Figure 3.57: *Experimental variograms for Jura subset at scales a) global and b) local*

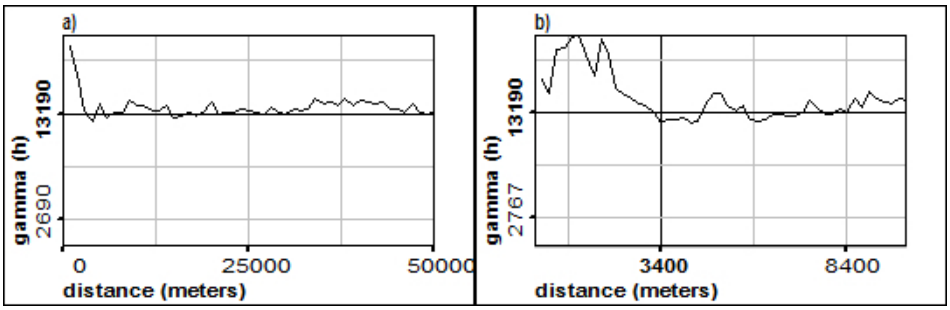



Figure 3.58: *Experimental variograms for Plateau subset at scales a) global and b) local*

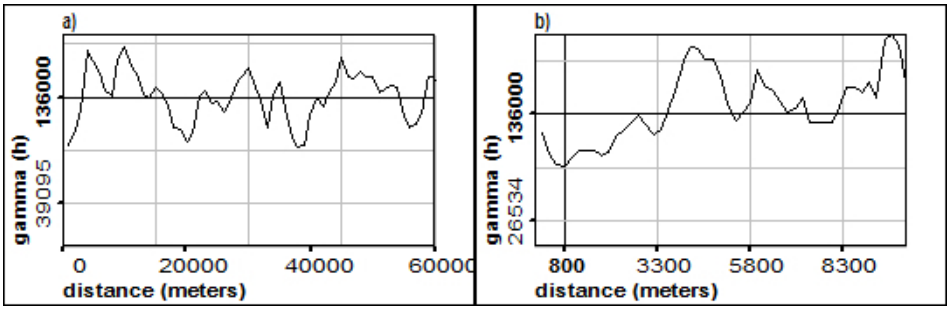



Figure 3.59: *Experimental variograms for Alps subset at scales a) global and b) local*

The Plateau subset has a high local variance, while the Alps data have cyclic semivariance changes with distance.

Results from the Jura region variography have given an important answer to our pending problem with the subset 3A modeling. This short subset was not able to be modeled due to high local variability (Figure 3.42), but once integrated into a higher subset, a global model becomes possible. As a statistical method, variography can be improved if the global mean is constant (no trend).

In the Plateau set, the influence of high local variability is clear. Using the MW averaging trick it will be possible to filter-out the high local variation and to observe variation structures. Variography of the Alps in turn, has what is called a 'hole effect': structures that are reproduced statistically, but they are far apart. The orography of the mountainous re-

gions, with valleys far apart from each other, is the cause of that. Conditions are reproduced within valleys that are distanced.

### 3.12.3 Moving windows multiscale analysis

Moving windows procedure is a simple partition of the sampling space into same-area cells that generates multiple scales of analysis. The subdivision of the space into cells with different sizes creates multiple levels of generalization. MW is an arbitrary division that can provide insight into the spatial and statistical behavior of data on global and finer scales.

#### MW statistics

For the multiscale analysis, 5 windows-sizes or scales were used by successive partition. The first scale is the global extension of the dataset and is contained in a rectangle of 345 by 218 km. The following scales consist of 4 grids that split the bounding box of data by cells with decreasing size. For the first partition (called grid5) the sampling space was divided into five columns and five rows, giving a total of 25 cells. Grid10 is a division into 10 by 10 cells and so on for grid20 and grid40. Grid40 corresponds to a 40x40 grid creating 1600 cells. Diagonal sizes of the cells for grid5 to grid40 are approx. 80, 40, 20 and 10 km respectively. Grid5 and grid40 partitions are presented as examples in Figure 3.60, superimposed to indoor radon sample points. For each grid, a number of analyses were carried out in order to evaluate statistical parameters and variogram features.

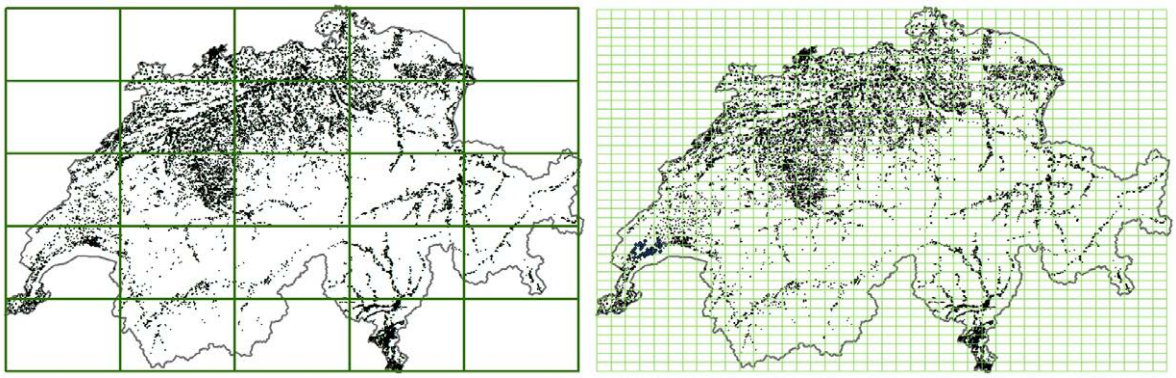


Figure 3.60: *Partition of sampling space using grid5 (left) and grid40 (right)*

The first analysis was the skewness test after lognormal transform. Only cells containing at least 20 points were considered valid for the analysis. Results are shown in Table 3.4.

The results for the statistical multiscale analysis shows that there is a reduction of the local mean variance and skewness as the resolution of windows decreases. The variation of the mean between scales is explained by the reduction of the number of samples, since a minimum of 20 points per windows was imposed. Isolated points, which are associated with higher values, are eliminated with this selection. The variance reduction, along with

Table 3.4: *Multi-scale lognormal skewness test for indoor data in Switzerland*

| Grid | Number cells | Mean points | Mean radon | Mean variance | Mean Skewness | Lognor skew rejection (%) |
|---|---|---|---|---|---|---|
| 5 | 23 | 1817 | 161 | 107137 | 6.37 | 82 |
| 10 | 65 | 641 | 172 | 106281 | 5.39 | 58 |
| 20 | 193 | 214 | 160 | 86888 | 3.69 | 48 |
| 40 | 447 | 88 | 150 | 65066 | 2.76 | 34 |

scale, is the result of having less points per cell. What is also interesting, is that rejection for the lognormal skewness has dropped from 82% for the net5 to just 34% for net40. By selecting samples within smaller windows, the presence of outliers is reduced. The spatial distribution of rejected cells (non-lognormal) and lognormal sets are represented in Figure 3.61.

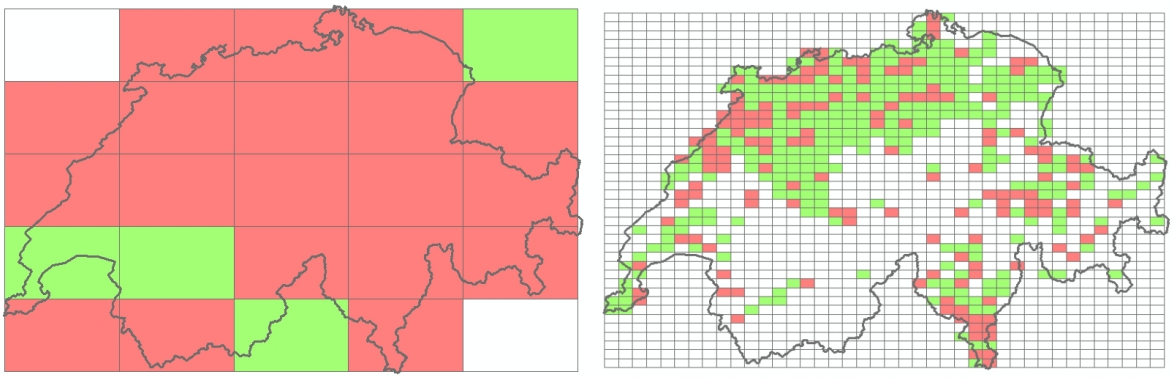


Figure 3.61: *Lognormal skewness test for scale grid5 (right) and grid40 (left), rejected cells are colored in red while accepted cells appear in green*

The local variance, the proportional effect and the proportional semi-variance at minimum distance (nugget effect) were also analyzed. As seen, the local within-cell variance diminishes with cell size, what is interesting to monitor, is the between-cells variance and how it is reflected in variograms. Variance and correlation coefficient of the proportional effect after MW averaging were calculated for cells having more than 5 points. An indicator for significant local variance can be obtained by dividing the semivariance, at a minimum distance, by the a priori variance. This minimum distance was set at 500 m, which is less than a half of the average distance. The semivariance was calculated for cells with more than 20 points. The results are presented in Table 3.5.

Table 3.5: *Mean values for variogram features at 4 scales of analysis, values are expressed in percentages*

| grid | Variance between-cells | correlation prop. effect | relative semivariance |
|---|---|---|---|
| 5 | 9720 | 0.91 | 72 |
| 10 | 15858 | 0.85 | 70 |
| 20 | 15926 | 0.82 | 68 |
| 40 | 19622 | 0.76 | 66 |

By reducing the size of windows, the between-cell variance increases, and the proportional effect decreases as the local variance increases. Finally, the semivariance for subsets remains, on average, very close. Behavior of statistical parameters after MW averaging is logical and predictable and reflects large global variations. On large scales, indoor radon measures for Switzerland have clear spatial variations. This is just a quantitative evidence of what was already observed. What is also clear by now is that the arbitrary spatial division of data within fixed windows cannot provide better local models. On the contrary, it was seen that the data partitioned considering environmental variables and mean stationarity after MW averaging produced better models.

**National data partition using MW averaging**

In section 3.12.2, it was seen that spatial variation of indoor radon for the Jura region presented a certain structured spatial continuity. How coherent is that with local averages and stationarity, and how can modeling be improved by MW averaging?. A MW of 4 by 4 km can show a level of generalization that preserves somehow the spatial distribution of data for Switzerland. Considering the limits of the 'Jura partition' boundary, another partition was proposed by selecting a homogenous zone with high local mean values. In Figure 3.62a, an indoor radon map after MW, together with the Jura region limit and the MW optimized partition boundary, is presented. In Figure 3.62b, the corresponding variogram of the selected subset within this new boundary is shown.

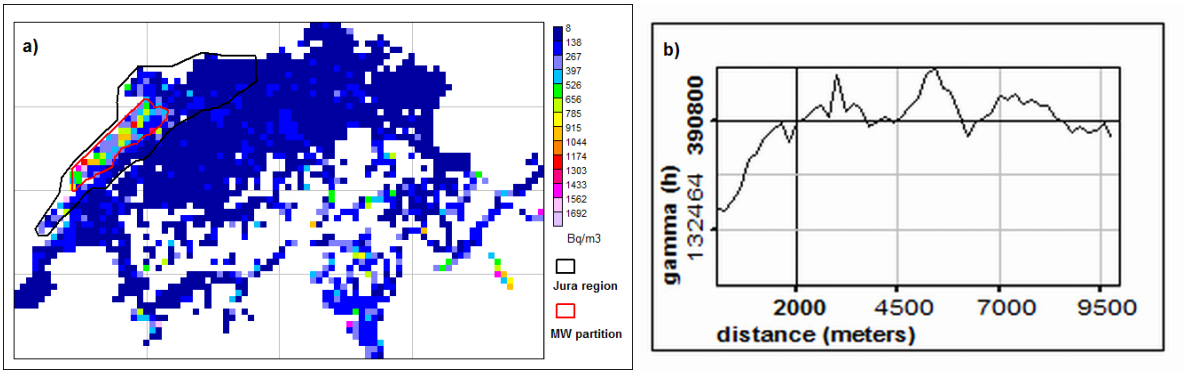



Figure 3.62: *a) Map of MW averages for indoor radon in Switzerland and partition boundaries b) Experimental variogram of MW data partition in Jura region*

The reselection of data considering MW averages has considerably enhanced the variogram model for the Jura region. Nevertheless, it covers a smaller area and makes use of fewer points (3482 against 9342 for the Jura region). In addition, delimitation using MW was a bit coarse because of the squared limits of windows. In this sense, an interpolation using the minimum curvature spline method to obtain a smooth map for indoor radon was done. The numeric interpolation from this method is not so accurate because of the high data variability, but the generalization helped to demarcate homogenous areas. This can be seen after superimposing the MW Jura partition boundary over the spline map (Figure 3.63).

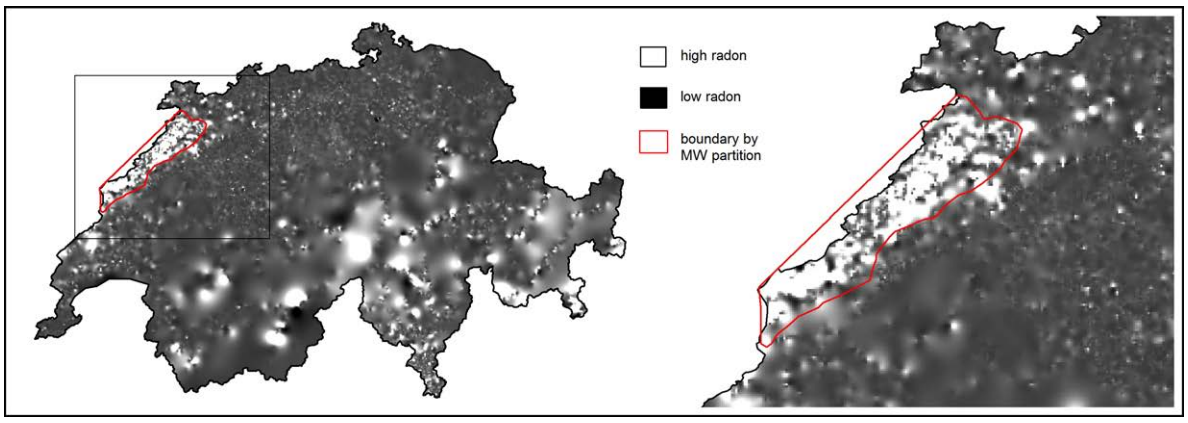



Figure 3.63: *Map of indoor radon spline interpolation and Jura delimitation using MW*

By applying more contrast over the spline map, the limits of the zone appear clearly and it is possible to propose modifications. The next goal is to find other homogeneous areas with defined spatial continuity, with the help of the spline map. Nine sectors were roughly delimited on this map (Figure 3.64) and the corresponding variograms per sector were computed (Figure 3.65)

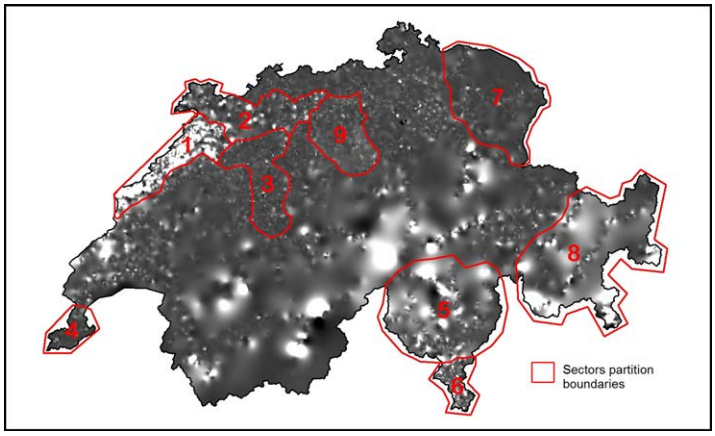



Figure 3.64: *Sectors with spatial continuity defined over an spline map*

From this collection of sectorial variograms it can be said that sector 1, 7 and 9 present some spatial continuity with a short-range structure. Variograms 4 and 6 are slightly structured but have an important nugget effect. Variograms 2, 3 and 8 indicate strong local variability. Sector 5 is a particular case because it has an important local variability followed by some continuity and has the inverse behavior of data from sector 8. It seems that these two sectors were wrongly delimited; they are simply not homogeneous, and they present a spatial trend of the mean. The smooth spline map also shows how different the regions within Switzerland are in terms of data density (clustering) and spatial variability. This is partially due to the differences in the sampling schemas adopted by the cantons.

In any case, data generalization, by either MW averaging or other interpolation methods

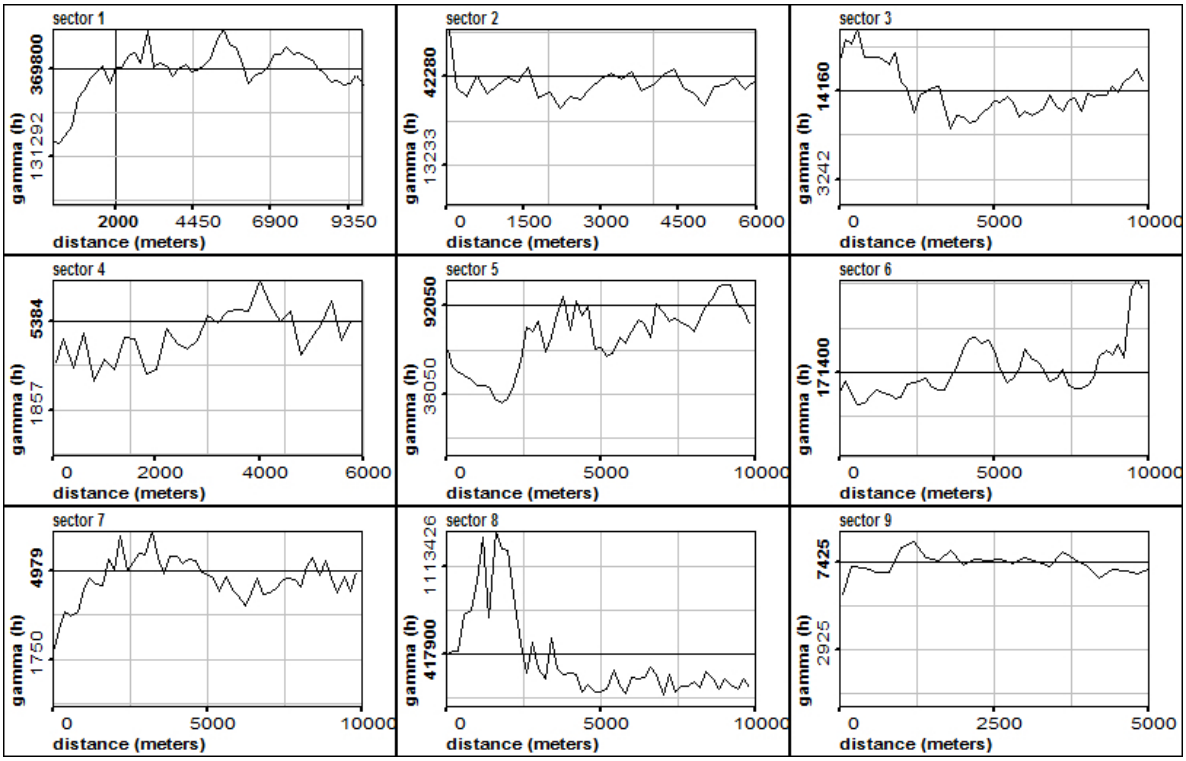



Figure 3.65: *Variograms for data corresponding to 9 defined sectors*

resulted in an effective method for data partition considering mean stationarity. As it will be seen in the next chapter, this condition is essential for regional linear interpolation by kriging. It will be addressed again in chapter 5, where non-linear methods are proposed to improve variography modeling.

### 3.12.4 Administrative scale

The Swiss territory is politically divided into cantons, districts and communes. The Federal Office of Public Health uses this subdivision as the spatial support to consolidate the measured values of indoor radon, as indicated in chapter 1 and 2.

Since the campaigns and the laboratory treatment of detectors are under cantonal responsibility, cantons could have different sampling designs. Some cantons are more active in promoting measurements for all houses while others could have preferential sampling.

Moreover, the political division of Switzerland has some disadvantages as a spatial support for interpolation. Cantonal borders are irregular in shape and in the number of samples. Some regions have few records compared to others. More important, is that the spatial distribution of values within a canton can be more or less clustered depending on landscape changes.

**Preferential sampling analysis by clustering on a cantonal level**

In the univariate analysis of chapter 2, it was observed that the mean of samples for the canton of Ticino has decreased from 2006-2008 in comparison to samples taken up to 2005. On the contrary, the mean of samples, in the same period, has increased for canton of Neuchatel. Is it possible to determine whether this change corresponds to a change of strategy, targeting urban instead of isolated areas? Is it possible to use a single measure of clustering to do this analysis?

The proposed method is to use the Quantile Morisita Index (QMI) to compare different campaigns. The first analysis was conducted at a national level by creating two datasets for the main survey periods 1982-2005 and 2006-2008. They are simply named sets 2005 and 2008. The sets correspond to measurements taken in inhabited buildings on the ground-floor. It was reduced to include just one measurement per location if several rooms had been surveyed from the same dwelling.

The set 2008 include 9,099 samples and had a MI of 180 at average distance. To compare the level of clustering with 2005, the same number of samples was selected out of the 2005 dataset, consisting of 33,228 samples. This 2005 comparison set had a MI of 16, which indicates that later samples are more clustered than older ones on a national scale.

This clustering is due to increased sampling in certain cantons with high radon-prone areas. It will, therefore, be interesting to analyze the level of clustering within these cantons. In fact, using the same procedure, it was observed that the level of clustering for the cantons of Ticino and Neuchatel was lower in 2005 than in 2008. In Ticino the MI for 2005 is 43 while for 2008 it is 82. In Neuchatel MI changes from 20 to 27.

It is clear that resampling campaigns are more clustered, especially in Ticino. Nevertheless the mean indoor radon for the 2005 campaign is 234 $Bq/m^3$, while it is lower for 2008 (211 $Bq/m^3$). This indicates that the resampling for this canton was done for relatively low radon-prone areas. In fact, the strategy in Ticino is to survey as many dwellings as possible, regardless of local conditions of vulnerability. This has increased the number of samples within highly urbanized areas and with lower exposition. Previous samples may have also been preferentially clustered to high concentration areas.

On the contrary, in Neuchatel the mean for the 2005 campaign is 314 Bq/m3, while the samples taken in the 2008 campaign have a mean of 462 $Bq/m^3$. This is the opposite of Ticino because preferential resampling was carried out for high radon areas. Probably, the purpose of the sampling strategy in this case was to find the most exposed dwellings or areas.

For the purpose of interpolation, it would be preferable to have the same degree of clustering for all quartiles, which will be traduced on a good spread of points. A good sampling strategy can help with this purpose. A certain optimization for the samples placement will later benefit interpolation. Nevertheless, in practice, surveying isolated areas will always pose more difficulties.

### Variography at cantonal level

Subsets per canton were selected from the Swiss indoor radon data. This national dataset differs from the dataset used in chapter 2 for the sampling evolution analysis (section 2.2.2). Table 3.6 presents some statistics and spatial parameters per canton from the standardized dataset with 41787 samples.

Table 3.6: *Statistics and spatial parameters per canton*

| canton | samples | mean | variance | skewness | average distance | fractal dimension |
|---|---|---|---|---|---|---|
| Ticino | 8905 | 199 | 132578 | 13.8 | 661 | 1.00 |
| Bern | 7746 | 145 | 68471 | 8.6 | 1241 | 1.46 |
| Grisons | 4130 | 217 | 222282 | 12.6 | 1719 | 1.33 |
| Neuchatel | 2277 | 321 | 281294 | 4.6 | 845 | 1.21 |
| Zurich | 2568 | 98 | 13627 | 11.6 | 1000 | 1.68 |
| Aargau | 2672 | 99 | 15189 | 18.5 | 1005 | 1.56 |

In Figure 3.66, graphs of quartile Morisita index at average distance (QMIAVR) for some cantons are included.

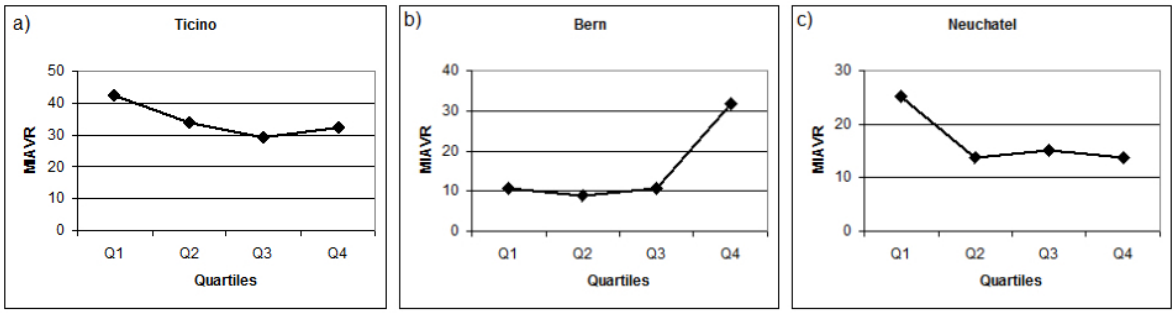



Figure 3.66: *Quartile Morisita index profile diagrams at average distance for different cantons*

This measure shows different degrees of spatial clustering for the three cantons. In Ticino, the values are more homogeneously distributed with a certain tendency towards low value clustering. The clustering of low values is more evident in Nechautel's samples. The canton of Bern has an opposite functional clustering arrangement because high values appear to be concentrated, while low values are better distributed. An interesting observation is that clustering for canton of Neuchatel has evolved between campaigns with a change to preferential sampling, as can be seen in Figure 3.67.

Until 2005, samples had an even spatial distribution, while the 2006-2008 campaign has lower values clearly clustered in comparison to higher values.

The corresponding variograms for 6 cantons are shown in Figure 3.68. Out of these variograms, it seems that the Ticino dataset has some spatial structure.

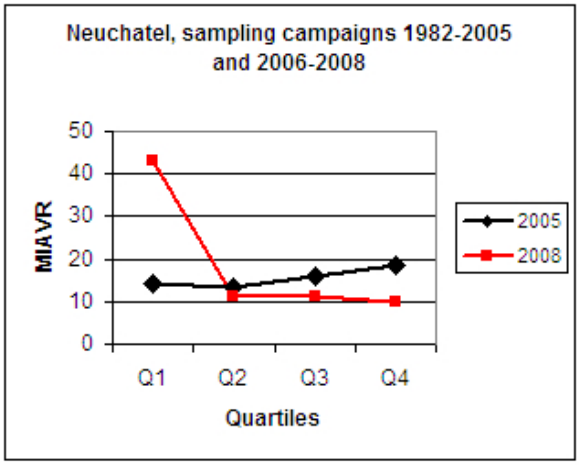



Figure 3.67: *Quartile Morisita index diagram at average distance for canton Neuchatel for 1982-2005 and 2006-2008 campaigns*

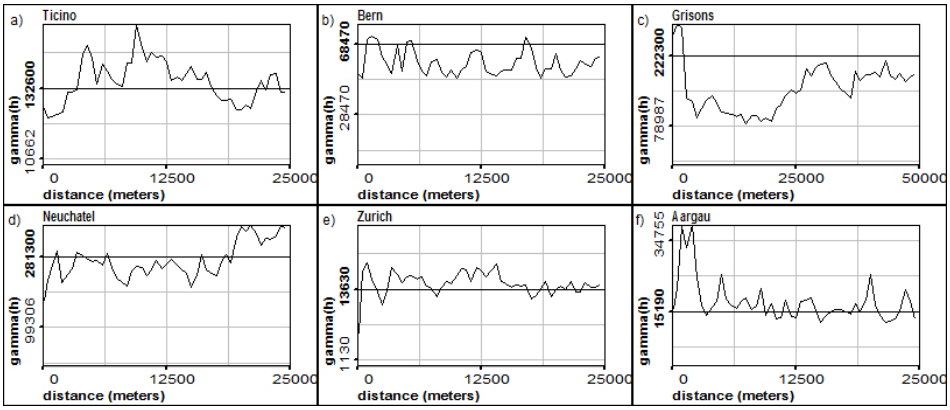



Figure 3.68: *Indoor radon raw data variograms for 6 cantons*

## 3.13 Conclusions about the spatial analysis

This chapter includes a comprehensive analysis of the spatial distribution of data points with the aim of helping with the optimization of interpolation parameters. The first spatial characterization of the set3 case study was the average distance between sampling points. Although it is a very basic parameter, it proved to be important for the QMI profiles, MW averaging and variography calculations.

The high level of clustering of set3 was quantified by different methods like Voronoi polygon statistics, sandbox counting, box-counting and Morisita index. With the use of validity domains including a build-up constrained domain, the relative clustering was measured. On the national scale, the sandbox method has indicated that the actual national indoor radon sampling is more clustered (with a $Df = 1.3$) in comparison to the build-up area ($Df = 1.76$). Meanwhile, the build-up area has a clustering coefficient similar to a random distribution of points. The high degree of spatial clustering for indoor radon samples

is mainly due to preferential sampling of certain areas and partially due to the sampling domain, which is the build-up zone in Switzerland.

The idea of using quantiles was introduced in order to make comparative clustering calculations for the functional box-counting method and the Morisita index. The so-called Quantile Morisita Index (QMI) analyzes the spatial distribution of equal-sized subsets defined according to quantile thresholds for a certain scale. The QMI diagram showed, more clearly, that the results obtained with sandbox and box-counting methods for set3 present higher spatial clustering both for low and high values at the average distance.

The cell and the polygonal declustering methods were applied to approximate the global unknown statistics of indoor radon. It was observed that data transformation is more coherent when using weights with lower variations. This was obtained by limiting the cells and polygons covering a more constrained sampling space. It was reaffirmed that the build-up area is not only the optimal but also the natural sampling domain to be used. A very constrained Voronoi polygons coverage or a cell declustering, using windows of around 600 m, can fit within the sampling domain. In particular, the cell declustering method is the one that can better adapt to a manifold domain such as the build-up area.

Methods used for interpolation, like KNNR or variography, can be also used in his modeling phase as exploratory tools for spatial properties. The convexity of the KNNR cross-validation (CV) optimization curve gave a first measurement of spatial continuity. Variography over set3 showed high local variability, which prevents data from being fit into a structured model.

Moving Windows (MW) averaging was proposed as a method to reduce local variability. Based on the coherent results of clustering characterization using different methods, the windows size used for the national set was 1700 m. With MW, a local mean was calculated and adjusted to a regular configuration. This procedure helped reduce the high local variability and find variography structures on a national scale.

The mean and the skewness parameters after an MW analysis of set3, have indicated the presence of a transition zone in the middle of the area. The presence of this transition zone was also cleared with a lognormality skewness test. The recurrent indication of a transition zone for set3 lead to the idea of doing a spatial partition of data based on local statistics.

An important output from the spatial data analysis is that data partition into homogeneous distribution areas (in particular for high values) can provide a solution to improve modeling. As this is a matter of using different scales and domains, a multiscale and partition analysis was proposed. The scales of interest defined for the study of indoor radon in Switzerland were the national domain, the natural regions domain, the Moving windows scales, and the administrative units domains.

During the multivariate analysis, it was concluded, that geotechnical units and elevation play an important role in explaining indoor radon spatial distribution. Natural regions in Switzerland have distinctive characteristics based on these two variables. Therefore, they constitute an important domain for a regional analysis of indoor radon. Indoor radon samples for the Jura region presented a structured variogram, while spatial continuity for the

Alps region was not clear. The Plateau region had clustering for low and high values. This clustering feature, reflecting high local variability, was also seen for set3.

As for the administrative domain, the cantons of Ticino, Bern and Neuchatel have a clustering of either low or high values. The resulting variograms for these cantons present some structure, with Bern being the least structured. The historical analysis of samples for the canton of Neuchatel show an evolution from an even distribution before 2005 to clustering of high values during the 2006-2008 campaign. This information can be used to approach the global statistics on a cantonal level.

# Chapter 4

# Spatial interpolation with regression methods

Tuning model parameters and hyper-parameters properly, is almost entirely the task of interpolation. The optimization of parameters becomes more difficult when also considering hyper-parameters, as for example, scale or neighborhood. In the previous chapter, the issue of scaling was considered and some solutions were proposed, in particular for variography modeling. In addition, neighborhoods were characterized with the use of the KNNR method. Other methods are more robust for handling data without stationarity assumptions.

This chapter will be dedicated to optimizing modeling for different interpolation methods. It is an attempt to compare and integrate methods to achieve a common goal, which is the proper parameter definition. The basic assumption is that neighborhood and continuity are common properties of interpolation methods, but they are expressed in different ways.

It can be said that methods can be differentiated based on their orientation to express a certain spatial property. Neighborhood or the level of similarity of values in space can be depicted with a simple method such as K Nearest Neighbors Regression (KNNR). Variography aims to describe the continuity of values in space. Density functions, describing the distribution of points around a central point, are used in the General Regression Neural Networks (GRNN).

In the present chapter, linear and deterministic models, like KNNR and IDW, are presented. Then, the linear geostatistical models (the kriging family) are applied, followed by the non-linear GRNN regression method. These methods are referred to as regression interpolation, since in one way they aim to reduce an error measure in order to obtain a general model: the optimum $k$ for KNNR, unbiased estimators for kriging or a generalized regression model for GRNN. The multiscale analysis of chapter 3 is complemented here with the MW analysis for GRNN.

On the one hand, the goal is to find the method that adapts best to the data characteristics and on the other hand one must examine how suitable data are to make estimations. These are the method robustness and the data consistency analysis addressed at the end of the chapter.

## 4.1 Estimations and Predictions

### 4.1.1 Notes on terminology

The final objective of data sampling is to calculate new values for a dependent variable. In space, the calculation of a new value is made for an unvisited location with a defined position. When this is done within the limits of the sampling spatial domain, it is referred as interpolation, and when it refers to the external case it is known as extrapolation. In statistics, the calculation of a new value based on samples is referred to as a prediction, while when speaking of parameters the term estimation is used. Spatial interpolation is often referred to as an estimation to differentiate it from temporal predictions. This is particularly the case with the geostatistical jargon (31). As will be seen, the estimation of parameters is the essential part of geostatistics, and hence, the results of interpolation using kriging family interpolators imply estimating values and the corresponding uncertainty. For the case of simulations, the estimation concept is even clearer since the parameters of a local distribution are estimated for each unsampled location. Within the following text the term estimation will be used to refer to interpolations of unsampled points when involving statistical modeling; as it is done in geostatistical literature.

In contraposition to methods that made use of statistical modeling there are deterministic methods, which are simpler. Even when it is understood that most earth science processes are uncertain and cannot follow a well-defined physical model, the term 'deterministic' is used for methods where this uncertainty is simply not taken into account. In literature, the term spatial prediction is preferred for the results of deterministic methods. As will be seen, the results of both types of methods can have both indistinctively deterministic and probabilistic interpretations. Regardless of this interpretation, the interpolation methods can be also distinguished by their use of either linear or non-linear models.

### 4.1.2 Error measures

A first error measure to be considered is the training error that is obtained with cross-validation as explained in chapter 3. It is an error measure for the model itself and not necessarily valid for predictions. The training error for subsets can be an effective measure of the statistical consistency between samples and global distribution.

Additionally, there are measures of model uncertainties particular to a certain method. For instance, kriging provides a local measure of training error called the kriging variance. GRNN provides a measure of data density. Both measures are indicators of areas prone to generating high estimation uncertainty due to the lack of data.

Errors can be represented in different ways; the basic measure is the difference between the predicted value $\hat{z}$ and the real value $z$. These real values constitute the validation set and are independent samples that were not used for modeling or prediction. A good measure of

error is the mean of the squared errors (MSE) and is represented as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i - z_i)^2 \tag{4.1}$$

It is a convenient measure because it relates to statistics of error distribution: MSE = variance + bias$^2$. The bias is the mean of the error distribution and it indicates a tendency of the prediction model to produce either underestimates or overestimates. The prediction error variance can be used as an indicator of model adjustment to data. Prediction model parameters are tuned based on the error reduction concept. For the sake of precision, it is always good to include other error measures, such as the correlation between real and predicted values.

## 4.2 Deterministic interpolation methods

### 4.2.1 K Nearest Neighbors Regression (KNNR) Interpolation

The KNNR method is a simple linear combination of $k$ neighboring $z_i$ values to predict a value at an unsampled location $z_0$. It can be represented by the basic formula:

$$z_0 = \frac{1}{k}\sum_{i=1}^{k} z_i \tag{4.2}$$

As seen in section 3.5.1 of chapter 3, an optimal number of neighbors can be determined by the leave-one-out cross-validation (CV) analysis.

### 4.2.2 The KNNR CV mean filter

In the spatial analysis chapter, it was clearly observed that not only high spatial clustering, but also high local variability is characteristic of indoor radon samples. The local variability appeared in variograms as a high nugget effect. By conducting MW averaging, the effect of the local variability was very much attenuated, and spatial structures in variograms were revealed. Nevertheless, this averaging implied disturbing the original spatial distribution and limiting the model to scales up from the windows size.

The alternative solution, proposed to filter local variability while preserving sample locations, is to develop a local filter. The KNNR interpolation provides a basis to perform this because a set of $k$ neighbors create a constrained 'vicinity' that relates locally to a central point. An optimal number of points define this relation. It is of course difficult to measure the number of points that relate 'best' to the central point. Instead, the cross-validation (CV) optimal $K$ gives a 'collective' or model measure.

What is especially appealing with CV calculations is that the global optimal is obtained from all possible local optimal vicinities. If this information is used in an inverse sense, so that the vicinity information is optimally averaged at each location, then it will be possible

to reduce local variability while preserving spatial distribution. The data mean is preserved even though this data transformation will cause a loss of variability due to averaging.

It should be also mentioned that such a filter would have a limited use for cases where there are large local differences in vicinity, in others words, when vicinity is not stationary. A way to predict how optimal is the $K$ parameter consists in verifying the convexity of CV optimization curve. The main idea behind this filter is to regularize data for the cases were extreme values exist. Subsets of data, such training and validation set, can be then been more comparable by ensuring not only statistical but also spatial consistency.

Another important application of the KNNR CV filter is the option to optimize data visualization. The presence of extreme values in the indoor radon data poses a problem for their visualization in a map. Independent of the chosen scale of visualization, extremes are often not perceived because they constitute a reduced amount out of the total samples. Likewise, the mass of samples cannot be conveniently differentiated because the maximum of the scale is missing. Even when playing with the scale thresholds, graphic representation and mapping are difficult. By smoothing data locally, visualization is improved.

In summary, the method consists of local averaging using a single kernel and a convolution procedure. The kernel is the KNNR mean using $K$ neighbors, which is convoluted over sample locations. It is also a mean filter in the sense that it reduces local variability by averaging when data is deemed contaminated by outliers.

### 4.2.3 Inverse Distance Weighting (IDW)

Interpolation with IDW works like the KNNR method but includes a distance parameter. Neighboring values are weighted inversely proportional to its distance from the point being estimated. Additionally, this distance can be powered with a $p$ parameter to increase or decrease its weighting influence:

$$z_0 = \frac{\sum_{i=1}^{k} \frac{1}{d_i^p} z_i}{\sum_{i=1}^{k} \frac{1}{d_i^p}} \tag{4.3}$$

When $p = 0$ the weights are the same for all distances and the algorithm works exactly as KNNR. The best value for $p$ can be tuned considering the training CV errors.

## 4.3 Geostatistical interpolation methods

### 4.3.1 Statistical parameters

Geostatistical interpolation methods made use of statistical parameters of sample data in space to obtain estimated values for any unsampled location. The statistical parameter can be the variance between pair of values for points separated by a certain distance. As a statistical method, it is assumed that some probabilistic mechanism is generating an attribute random variable (RV) at location $x$. Then, we no longer speak about a variable $z$, but about a random variable $Z$ with $x$ infinite locations within an area $A$, so that $Z(x), x \in A$ (14).

The two model parameters most commonly used in probabilistic approaches to estimation are the mean or expected value of the RV ($E\{Z\}$) and its variance (Var$\{Z\}$). These two parameters can provide a complete description of the distribution, in the case of normal distributions. In other cases, they provide useful information on how the RV behaves (31).

The mean calculated from a finite number of $z$ observed outcomes is denoted as $m$. When it is chosen to conceptualize these values as outcomes of some RV, the mean of the corresponding RV will be denoted by $\tilde{m}$ and considered equal to $E\{Z\}$.

The variance of an RV is the expected squared difference from the mean of the RV:

Var$\{Z\} = \tilde{\sigma^2} = E\{[Z - E\{Z\}^2\}$

It also has an alternative expression that can be derived by expanding and using the sum property of expected values:

$$\begin{aligned} \text{Var}\{Z\} &= E\{Z^2 - 2ZE\{Z\} + E\{Z\}^2\} \\ &= E\{Z^2\} - E\{2ZE\{Z\}\} + E\{E\{Z\}^2\} \\ &= E\{Z^2\} - 2E\{Z\}E\{Z\} + E\{Z\}^2 \\ \text{Var}\{Z\} &= E\{Z^2\} - E\{Z\}^2 \end{aligned} \tag{4.4}$$

A third statistical parameter used in geostatistics is the covariance (Cov or C) between two random variables (say X and Y). Similar to the variance, it can be defined in terms of expected values and presented with an alternative expression:

$$\begin{aligned} \text{Cov}\{XY\} = \tilde{C}_{XY} &= E\{(X - E\{X\})(Y - E\{Y\})\} \\ \tilde{C}_{XY} &= E\{XY\} - E\{X\}E\{Y\} \end{aligned} \tag{4.5}$$

Another important expression of RV's is the variance of a weighted linear combination. The variance of a RV that is created by a weighted linear combination of other RV's is represented by the following equation:

$$Var\{\sum_{i=1}^{k} w_i Z_i\} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i \cdot w_j \cdot Cov\{Z_i Z_j\} \tag{4.6}$$

### 4.3.2 Geostatistical parameters under stationarity hypothesis

#### Second order stationarity

Stationarity is a basic assumption of statistical modeling and particularly for Gaussian processes. Geostatistical methods are mostly based on Gaussian processes and their functions relations. If the attribute variable is considered as an RV $Z$, its random function at any location $Z(x)$ will have the same expected value $E\{Z\}$ for any location $x$. This independence of the first moment parameter or the mean $m$ regarding location is commonly named the first order stationarity assumption. It is expressed for points $x$ and $x + h$ as:

$$E\{Z(x)\} = E\{Z(x + h)\} = m \tag{4.7}$$

Under first order stationarity, the variance and the covariance functions of $Z(x)$ will also be stationary. If the covariance is defined as in equation 4.5, the spatial covariance function of RVs $Z(x)$ and $Z(x+h)$ under the stationarity condition will led to:

$$\begin{aligned} \text{Cov}\{Z(x), Z(x+h)\} &= E\{Z(x) \cdot Z(x+h)\} - E\{Z(x)\}E\{Z(x+h)\} \\ &= E\{Z(x) \cdot Z(x+h)\} - m^2 \\ C(h) &= E\{Z(x) \cdot Z(x+h)\} - m^2 \end{aligned} \tag{4.8}$$

Where the RV $Z(x)$ can be any position in the space domain and the covariance function $C(h)$ depends solely on the separation lag $h$. This independency of the second order moments from position $x$ is called the second order stationarity hypothesis (equations 4.7 and 4.8) (8). This covariance must be a positive definite function.

**Intrinsic hypothesis**

In geostatistics, the particular interest is to define spatial continuity. This was first done empirically by defining the so-called experimental variogram. In order to work statistically with increments in space, Matheron defined an intrinsic hypothesis of second order stationarity for a RF of the difference of a pair of values $Z(x+h) - Z(x)$ (79) (78). An intrinsic model was proposed by satisfying two conditions (8). The first condition states that the drift of the increment is zero:

$$E\{[Z(x) - Z(x+h)]\} = 0 \tag{4.9}$$

This led to a second condition, that is the statistical form of the variogram $2\gamma(h)$, starting from the variance of the differences $[Z(x) - Z(x+h)]$, as defined in equation 4.4:

$$\begin{aligned} \text{Var}\{[Z(x) - Z(x+h)]\} &= E\{Z(x+h) - Z(x)]\}^2 - [E\{Z(x+h) - Z(x)\}]^2 \\ E\{Z(x+h) - Z(x)]\}^2 &= 2\gamma(h) \end{aligned} \tag{4.10}$$

The variogram function $2\gamma(h)$ will depend on lag distances $(h)$ only and not on the absolute position $x$. This hypothesis mean that different parts of a modeled region are statistically similar. The number 2, in equation 4.10 will help establish an equivalence with the covariance function as will be seen later.

### 4.3.3 Relations between stationary random functions

As mentioned, for the stationary random functions there are a number of models most frequently used in the practice of geostatistics. The variance model $Var\{Z(x)\}$, the covariance function model $C(h)$ and the variogram model $2\gamma(h)$ can provide the same information in a slightly different form. Under second order stationary assumptions, these three parameters are related by a few simple expressions. A first relation to be deduced is the case when positions $x$ and $x+h$ coincide, so that h equals 0 ($h = 0$). In this case the covariance function $C_0$

ends up equal to the variance of the RV Var$\{Z(x)\}$, also called the a priori variance. Taking the covariance from equation 4.8:

$$\begin{aligned} C(0) &= E\{Z(x) \cdot Z(x+h)\} - m^2\} \quad \text{if h=0 then:} \\ &= E\{Z(x)^2\} - m^2\} \quad \text{(is the variance 4.4)} \\ &= \text{Var}\{Z(0)\} \end{aligned} \tag{4.11}$$

Using the above relations and expressions, it is finally possible to establish a relation between the variogram function $2\gamma(h)$, the correlation of pairs of points $C(h)$, and the RV variance Var$\{Z(x)\}$:

$$\begin{aligned} 2\gamma(h) &= E\{[Z(x) - Z(x+h)]^2\} \\ &= [E\{Z^2(x)\} - m^2] + [E\{Z^2(x+h)\} - m^2] - 2[E\{Z(x)Z(x+h)\} - m^2] \\ &= \text{Var}\{Z(x)\} + \text{Var}\{Z(x+h)\} - 2\text{Cov}\{Z(x), Z(x+h)\} \\ &= 2[\text{Var}\{Z(0)\} - C(h)] \\ \gamma(h) &= \text{Var}\{Z(0)\} - C(h) \\ \gamma(h) &= C_0 - C(h) \end{aligned} \tag{4.12}$$

From this last equation, we deduce that the variogram and the covariance have a symmetric relation that depends on a constant, which is the variance. These assumptions are of course for the theoretical model of random functions under stationarity assumptions. In practice, variograms will reach an asymptotic sill value that is not necessarily equal to the variance. In other cases, they will simply not reach a sill or constant value and will continue to increase with distance.

### 4.3.4 Variogram model parameters and restrictions

Variogram and covariance, as defined in the relation 4.12, are functions of the vector $h$ (length and direction). When that function depends only on the length of the vector $h$, the model is said to be isotropic. When it depends also on the direction of the vector $h$, the model is said to be anisotropic. Directional variography is a modeling refinement that is often necessary for geological and other natural processes. The variogram function models commonly used can be described by a series of parameters that were already mentioned when conducting empirical variography. Now that a mathematical relation has been defined for the variogram, these parameters can be better described as:

- $C_0$, is the covariance function at distance $h = 0$, commonly called the nugget effect, which provides a discontinuity at the origin. - $a$, is commonly called the range, which provides a distance beyond which the variogram or covariance value remains essentially constant. - $C_0 + C_h$, is commonly called the sill, which is the variogram value for very large distances, $\gamma(\infty)$. It is also the covariance value for $h = 0$, and (under stationarity conditions) the variance of the modeled RV, $\tilde{\sigma}^2$.

If variogram and covariance can express the same spatial relations, why is the variogram used in geostatistics modeling? Two reasons are given by Chileés (8). The first one is that the

variogram is a more general structural tool than the covariance, since it is an Intrinsic RF and therefore it includes stationary RF. The second reason is practical: the variogram, unlike the covariance, does not require the knowledge of the mean. In practice, this mean is not known and has to be estimated from the data.

Although the spatial continuity is defined through the variogram, the Kriging systems are solved using the covariance (31). The relation between the variance of a linear weighted combination of values (e.g. kriging) and covariance values was given in equation 4.6. A variance cannot be negative and this includes the variance of any linear sum of random variables:

$$\text{Var}\{\sum_{i=1}^{k} w_i Z_i\} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i \cdot w_j \cdot \text{Cov}\{Z_i Z_j\} > 0 \tag{4.13}$$

The variance is necessarily non-negative, thus, the previous expression must be non-negative whatever the choice of the $k$ weights $w$, possibly negative. The variance is zero only if all weights are zero, assuming that none of the component RVs are exactly constant. Thus the covariance function $C(h)$ or the covariance matrix $\text{Cov}\{Z_i Z_j\}$ must be such to ensure the positivity of the variance operator 4.13, whatever the weights signs are. Then, the covariance matrix for any number of points is positive semidefinite. That is to say that for a matrix of order $n$ its determinant and all its principal minors are positive or zero (79). In like manner, the variogram as established in the relation 4.12 must be negative semidefinite.

### 4.3.5 Authorized variogram models

The need for a variogram model comes from the fact that, for estimation purposes, a variogram value is needed for distances for which the sample variogram does not have a value. The use of a model also guarantees the uniqueness of the covariance matrices for each distance $h$ used to calculate estimation weights. To attempt this condition of uniqueness, only positive definite covariance models must be used. The positive definite condition is as a guarantee that the variance of RV formed by a weighted linear combination of other RVs will be positive.

One way of satisfying the positive definiteness condition, is to use only a few covariance functions that are known to be positive definite. The basic variogram models can be conveniently divided into two types; those that reach a plateau or sill and those that do not. Variogram models of the first type are often referred to as transition models. The plateau they reach is called the sill and the distance at which they reach this plateau is called the range. Some of the transition models reach their sill asymptotically (like the Gaussian model). For such models, the range is arbitrarily defined and corresponds to a distance at which 95% of the sill is reached. In the group of transition models, the most common ones are the pure nugget, the spherical, the Gaussian and the exponential models.

**Pure nugget model (nug)**

As seen from sample variograms there is an obvious spatial discontinuity at the origin. While the variogram value for $h = 0$ is strictly 0, the variogram value at very small separation distances may be significantly larger than 0 giving rise to a discontinuity. We can model such discontinuity using a discontinuous positive definite transition model that is 0 when h is equal to 0 and 1 otherwise. This is the nugget effect model and its equation is given by:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0, & h \neq 0 \end{cases}$$

**Spherical model (sph)**

Perhaps the most commonly used variogram model because of its validity in 3D space, its well-marked range, and its ease of calculation (8). Its equation is:

$$\gamma(h) = \begin{cases} C_0 + C \cdot (\frac{3h}{2a} - \frac{1}{2}(\frac{h}{a})^3), & \text{if } h \leq a \\ C_0 + C, & \text{if } h > a \end{cases}$$

Where $a$ is the range. It has a linear behavior at small separation distances near the origin but flattens out at larger distances, and reaches the sill at $a$. In fitting this model to a sample variogram it is helpful to remember that the tangent at the origin reaches at the origin reaches the sill at about two thirds of the range. Up to the range distance correlation exists. The spherical variogram reaches its sill with zero derivative; this should be the statistical (a priori) variance.

**Exponential model (exp)**

Its equations are:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + (C - C_0) \cdot (1 - \exp(\frac{-h}{a})), & \text{if } h \neq 0 \end{cases}$$

This model reaches its sill asymptotically, with the practical range $a$ defined as the distance at which the variogram value is 95 % of the sill. Like the spherical model, the exponential model is linear at very short distances near the origin, however it rises more steeply and then flattens out more gradually. It is helpful to remember that the tangent at the origin reaches the sill at about one fifth of the range.

**Gaussian model (gauss)**

The Gaussian model is a transition model that is often used to model extremely continuous phenomena. Its equation is

$$\gamma(h) = C_0 + C \cdot (1 - \exp(\frac{-h^2}{a^2}))$$

Like the exponential model, the Gaussian model reaches its sill asymptotically, and the parameter $a$ is defined as the practical range or distance at which the variogram value is 95% of the sill. It has a parabolic behavior near the origin, and is the only model with an inflection point. This model is associated with a high regularity of variations. It is often used to model extremely continuous phenomena and is hardly encountered in the earth sciences (8).

**Description of the variogram models**

Throughout the present thesis a coding was adopted to describe semivariogram parameters. It follows, in some way, the sequence of parameters adopted in the GSLIB software (14). Nested variogram structures are joined with a + sign. Each variogram structure is initially described by an abbreviation of the model type; for example, nug stands for nugget or gauss for Gaussian. Beside to the model, comes the contribution of the corresponding structure. For example, nug0.5 indicates a nugget model with 0.5 variance.

If the model is anisotropic, the ranges of the semivariogram structure are specified for an azimuth angle with a maximum range and another direction for the minimum range. The direction with maximum continuity and range is called the principal range ($R$) and the orthogonal direction is called the secondary range of continuity ($r$). With the use of a variogram rose graphic (37), the direction having the most continuity can be better identified.

The maximum range is indicated following an uppercase R and the minimum range after a lowercase r. For example the coding sph0.3(90)R1000r500 indicates that the structure in the 90 degrees azimuth has a range of 1000 meters, while the orthogonal direction has a range of 500 meters. When the structure is isotropic, only the maximum range is indicated with an R. For example, the coding: nug0.5+sph0.3R1000+sph0.2R5000 represent an isotropic semivariogram model with a nugget model and 2 nested structures.

### 4.3.6 Linear regression and Simple Kriging (SK)

The basic idea of linear regression is to estimate an unknown value $z_0$ by a linear combination of $k$ known values $z_i$, $i = 1, \ldots, k$ (33). In the same way, kriging is a linear estimator method that uses the increasing semi-variance of points along distance as estimation weights $w$. Then the unknown value $z_0$ is estimated by a linear combination of $k$ data plus a shift parameter $r_0$. The bias term $r_0$ is introduced and assumes that the estimator is biased to obtain an expression for it:

$$\hat{z}_0 = \sum_{i=1}^{k} w_i z_i + r_0$$

Where $\hat{z}_0$ is the estimate. To build a linear model, the known values $z_i$ are interpreted as the outcomes of $i$ RV's $Z_i$. Correspondingly, the linear combination of the $k$ RV's $Z_i$ will also give an RV $Z_0$ called the estimator:

$$\hat{Z}_0 = \sum_{i=1}^{k} w_i Z_i + R_0 \tag{4.14}$$

Here, $R_0$ is a global bias and no longer a local bias $r_0$ and is expressed as the error or difference between estimators $\hat{Z}_0$ and real values $Z_0$: $R_0 = \hat{Z}_0 - Z_0$. In order to have good estimations, one should ensure that an unbiased (U) estimator is obtained. This can be achieved if we assume that the expected global bias is zero: $E\{R_0\} = 0$. If this expectation expression is decomposed as follows:

$$E\{\hat{Z}_0 - Z_0\} = E\{\hat{Z}_0\} - E\{Z_0\} = 0$$

It is possible to replace $Z_0$ from equation 4.14, in order to introduce the bias term $R_0$:

$$\begin{aligned} E\{\hat{Z}_0 - Z_0\} &= E\{R_0 + \sum_{i=1}^{k} w_i Z_i\} - E\{Z_0\} \\ &= R_0 + \sum_{i=1}^{k} w_i E\{Z_i\} - E\{Z_0\} \\ &= R_0 + \sum_{i=1}^{k} w_i m_i - m_0 = 0 \end{aligned}$$

Then, we end-up with an expression for the bias, under expected global unbiasedness that corresponds to:

$$R_0 = m_0 - \sum_{i=1}^{k} w_i m_i \tag{4.15}$$

By reintroducing this bias term 4.15 in the estimator expression 4.14, we obtain the equation for an unbiased estimator. Additionally, the terms $m_i$ and $m_0$ will be replaced by a single term mean $m_0$ assuming that this is the local mean. By doing this we obtain what is called the simple kriging (SK) estimator equation:

$$\begin{aligned} \hat{Z}_0 &= \sum_{i=1}^{k} w_i Z_i + m_0 - \sum_{i=1}^{k} w_i m_i \\ \hat{Z}_{SK} &= \sum_{i=1}^{k} w_i Z_i + [1 - \sum_{i=1}^{k} w_i] m_0 \end{aligned} \tag{4.16}$$

It is important to observe that the SK equation proposes to be globally unbiased and that local estimates are obtained assuming that the mean is known and is valid for all localities or neighborhoods $i = k$. It should also be pointed out that the SK estimator requires this mean value to be provided. An interesting point about spatial stationarity is that, although it is a model assumption and data should not necessarily hold for it, it depends on the scale and domain of analysis.

As mentioned, stationarity is a model assumption that is often not held by the data to be modeled. An approach made in the present research is that estimations can be improved, not only by good model selection but also through data selection to fit the model hypothesis. For the case of geostatistical modeling, a proposal was made in the sense of optimal spatial data partition to attain some stationarity. In addition, a KNNR filter can also be considered an alternative to enforce local unbiasedness or consistency in spatial terms.

### 4.3.7 Ordinary Kriging (OK) estimator

A limitation for the SK estimator, frequently observed for real data, is the difficulty to assume that the global mean $m$ is representative of the local mean $m_0$ at each estimation point $x_0$. The samples in practice do not hold for the mean stationarity assumption and this could lead to important estimation errors or effective bias. Therefore, the SK estimator is sometimes called just a BLE estimator since it is locally biased. The solution to this was the development of an estimator with intrinsic stationarity that enforces local unbiasedness. This was called ordinary kriging (OK).

The first idea is to work with the minimization of the expected estimation error again, $R_0 = \hat{Z}_0 - \hat{Z}$, however, taking into consideration that the estimator $\hat{Z}_0$ does not have the error term $R_0$ (is locally unbiased). That means, if the estimator $\hat{Z}_0 = \sum_{i=1}^{k} w_i Z_i$ is included in the expression of the expected error, we obtain:

$$E\{R_0\} = \sum_{i=1}^{k} w_i E\{Z_i\} - E\{Z_0\}$$

In this way, the weights variables $w_i$ were introduced as wished. To ensure unbiasedness, we assume that the expected estimation error is zero $E\{R_0\} = 0$ so that:

$$\sum_{i=1}^{k} w_i E\{Z_i\} = E\{Z_0\}$$

If the intrinsic stationary assumption is affirmed (no trend is present), by assuming that $E\{Z_i\} = E\{Z_0\}$, then:

$$\sum_{i=1}^{k} w_i = 1 \tag{4.17}$$

The first conclusion is that restricting the sum of estimation weights to one can ensure local unbiasedness under stationarity conditions. When including this unbiased condition in the SK estimator 4.16, the OK estimator is obtained simply as:

$$\hat{Z}_{OK} = \sum_{i=1}^{k} w_i Z_i \tag{4.18}$$

It should be noticed that a condition for the estimator to work is that the sum of weights should result in 1. This condition must be ensured when solving weight equations. It is possible that weights and estimations may end up being negative. Some proposals are given in literature to correct this kind of error (13) (81).

### 4.3.8 The OK equation system by variance minimization

In order to obtain the estimations, what remains to be determined is the kriging weights $w_i$. These weights will be obtained by minimizing the variance of the estimation errors. Kriging

interpolation is a type of Best Linear Unbiased Estimator (BLUE). It has been seen, how the unbiased condition of the estimator can be obtained by minimizing estimation errors. Within kriging, we intend to also have the minimum of the estimation error variances to obtain the best weights. This entitles the Best (B) estimator property to the method.

If the estimation error is defined as $R_0 = \hat{Z}_0 - Z_0$, then the estimation error variance ($Var\{R_0\}$) can be expressed in terms of covariances using equation 4.6. For the simple addition of two random variables $Var\{\hat{Z}_0 + (-Z_0)\}$, equation 4.6 becomes:

$$\begin{aligned} Var\{\hat{Z}_0 - Z_0\} &= Cov\{\hat{Z}_0\hat{Z}_0\} - Cov\{\hat{Z}_0 Z_0\} - Cov\{Z_0\hat{Z}_0\} + Cov\{Z_0 Z_0\} \\ &= Cov\{\hat{Z}_0\hat{Z}_0\} - 2Cov\{\hat{Z}_0 Z_0\} + Cov\{Z_0 Z_0\} \\ &= Var\{\hat{Z}_0\} - 2Cov\{\hat{Z}_0 Z_0\} + Var\{Z_0\} \end{aligned} \tag{4.19}$$

Having this expression, it is possible to introduce the weight $w_i$ variable for which we want to deduce an equation. The estimator of $Z_0$ can be replaced by the weighted form: $\hat{Z}_0 = \sum_{i=1}^{k} w_i Z_i$. Also the terms $Var\{\hat{Z}_0\}$ and $2Cov\{\hat{Z}_0 Z_0\}$ in equation 4.19 can be changed using the equations 4.6 and 4.5 respectively:

$$Var\{\hat{Z}_0\} = Var\{\sum_{i=1}^{k} w_i Z_i\} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j Cov\{Z_i Z_j\} \tag{4.20}$$

$$\begin{aligned} 2Cov\{\hat{Z}_0 Z_0\} &= 2Cov\{(\sum_{i=1}^{k} w_i Z_i) Z_0\} \\ &= 2E\{\sum_{i=1}^{k} w_i Z_i Z_0\} - 2E\{\sum_{i=1}^{k} w_i Z_i\} \cdot E\{Z_0\} \\ &= 2E\sum_{i=1}^{k} w_i \cdot E\{Z_i Z_0\} - 2E\sum_{i=1}^{k} w_i \cdot E\{Z_i\} \cdot E\{Z_0\} \\ &= 2E\sum_{i=1}^{k} w_i \cdot Cov\{Z_i Z_0\} \end{aligned}$$

By replacement into equation 4.19

$$Var\{R_0\} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j Cov\{Z_i Z_j\} - 2E\sum_{i=1}^{k} w_i \cdot Cov\{Z_i Z_0\} + Var\{Z_0\}$$

Now that we have expressed the estimation error variance in terms of weights $w_i$, $w_j$ and RV $Z$ covariances, it is possible to minimize it by calculating the $k$ partial derivatives of this equation with respect to $w_i$.

$$\frac{\partial(\sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j Cov\{Z_i Z_j\} - 2E\sum_{i=1}^{k} w_i \cdot Cov\{Z_i Z_0\} + Var\{Z_0\})}{\partial(w_i)} = 0$$

With this, the normal system of equations can be established, also known as linear regression equations and called the kriging equations system for its use:

$$\begin{aligned} 2\sum_{i=1}^{k} w_i C_{ij} - 2C_{i0} &= 0 \\ \sum_{i=1}^{k} w_i C_{ij} &= C_{i0} \end{aligned} \tag{4.21}$$

It can be rewritten in matrix notation as:

$$C \cdot w = D \quad or \quad w = C^{-1} \cdot D \tag{4.22}$$

Where $C_{ij}$ is the covariance matrix between the $k$ neighboring samples used to obtain the estimation, $w_i$ are the weights for each of those samples and the $C_{i0}$ matrix (called D distance matrix) contains the covariance values between the unknown value and each of the samples. An interesting property of kriging systems can be seen in this equation. The D matrix is a type of inverse distance weighting in which the distance is the corresponding spatial covariance or statistical distance defined in the variogram for the actual geometric distance.

What distinguishes kriging from other interpolation methods is the roll of the C matrix. The C matrix records covariance distances between each sample and every other sample, providing the kriging system with information on the clustering of the available sample data. If two samples are very close to each other, this will be recorded by a large value in the C matrix and vice versa. The multiplication of D by $C^{-1}$ adjusts the raw inverse statistical distance weights in D to account for possible redundancies between samples. This is often mentioned as the screening or declustering property of kriging.

Because OK constrains the sum of the weights to one, a Lagrangian variable $\mu$ must be included in order to solve the constrained problem. For the correct calculation of weights $w_i$, the OK equation system will now consist of $k$ plus one equations:

$$\begin{aligned} \sum_{j=1}^{k} w_j C_{ij} + \mu &= C_{i0} \\ \sum_{i=1}^{k} w_i &= 1 \end{aligned} \tag{4.23}$$

The $\mu$ variable is the Lagrange parameter or the dummy variable that helps solve a $k+1$ equations system for a constrained problem. The matrix form remains, as for the normal system of equations, by including the $\mu$ variable in the weights matrix $w$ and unity multiplicators in the covariances matrices C and D.

## 4.4 General Regression Neural Networks method

### 4.4.1 The GRNN estimator

The General Regression Neural Networks (GRNN) is a non-linear regression method that can be used for spatial interpolations. It is based on Bayesian probability theory and non-parametric kernels (42). The objective of GRNN is to build an underlying regression function given a limited number of training samples with a number of input variables $x_i$ and a target variable to be estimated, called $z_i$. Then, the $N$ available samples are a set of the type $\{x_i, z_i\}_{i=1}^{N}$.

A regression function $f(x)$ can be estimated as a general function for all density functions of $z_i$ given an $x$ neighborhood. In other words, $f(x)$ will be a conditional mean of $z$ given $x$ (i.e. the regression of $z$ on $x$). Expressing $Z$ and $X$ as random variables, its conditional expectation can be written as:

$$f(x) = E[Z \mid x] = \int_{-\infty}^{\infty} z f_Z(z \mid x) dz \tag{4.24}$$

Then the regression function $f(x)$ used to make estimates will be the integration of the conditional probability density function (c.p.d.f.) $f_Z(z \mid x)$ for all $z$ given $x$. These c.p.d.f can be expressed in Bayesian probability terms as:

$$f_Z(z \mid x) = \frac{f_{X,Z}(x, z)}{f_X(x)} \tag{4.25}$$

$f_{X,Z}(x, z)$ is the joint pdf of x and z, and $f_X(x)$ is the marginal *pdf* of x. By introducing equation 4.25 into 4.24 and by expressing both the numerator and the denominator as integrals over dz:

$$f(x) = \frac{\int_{-\infty}^{\infty} z f_{X,Z}(x, z) dz}{\int_{-\infty}^{\infty} f_{X,Z}(x, z) dz} \tag{4.26}$$

The marginal *pdf* of x is expressed in terms of the joint *pdf* $f_{X,Z}(x, z)$, so that this is the only term to be estimated. To solve this, the method proposed by Nadaraya and Watson, and later by Specht is to use the Parzen-Rosenblatt density estimator of the joint *pdf* $f_{X,Z}(x, z)$ in the form:

$$\hat{f}_{X,Z}(x, z) = \frac{1}{\sigma^{p+1} N} \sum_{i=1}^{N} K\left(\frac{x - x_i}{\sigma_x}\right) K\left(\frac{z - z_i}{\sigma_z}\right) \tag{4.27}$$

This estimator is non-parametric in the sense that it does not use unique parameters like the mean or variance but instead it uses all sample data $x_i$ and $z_i$. It produces a symmetrical shape around the origin where it attains its maximum value. This shape is obtained through a group of kernel functions $K$. This kernel can be modified to better-fit with sample data. It has a smoothing parameter sigma $\sigma$ called *bandwidth*, which controls the size of the kernel

$K(x)$. It is quite equivalent to the bandwidth used to define histograms. This means that a small $\sigma$ results in a more detailed distribution function of the joint *pdf* while a large $\sigma$ tends to smooth it. $\sigma$ always has positive values. An important property of the Parzen estimator 4.27 is that it has a unique $\sigma$ parameter.

By integrating the Parzen estimator 4.27, with respect to $z$ in equation 4.26, one obtains the following estimate of the isotropic regression function $f(x)$:

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} z_i K(\frac{x-x_i}{\sigma_x})}{\sum_{i=1}^{N} K(\frac{x-x_i}{\sigma_x})} \tag{4.28}$$

The common term $\frac{1}{\sigma^p N}$ resulting after integration, has been eliminated. The estimator in the presented form is called the generalized regression (GR) estimator proposed by Nadaraya and Watson (NWKRE). It can be presented as a neural network (GRNN), with a layer of inputs $x$, a hidden layer of kernel centers and a target layer $z$.

As mentioned, a variety of kernel functions such as Gaussian, reciprocal, triangular, rectangular and Epanechnikov functions exist. Their graphical representation is shown in Figure 4.1. The Gaussian shape has inflexion points describing a high probability around the central point, which vanishes to both tails. On the contrary, the Epanechnikov function is a concave form without tails. The reciprocal function is more asymptotic than the Gaussian, while the triangular and rectangular are not as asymptotical as the Epanechnikov.

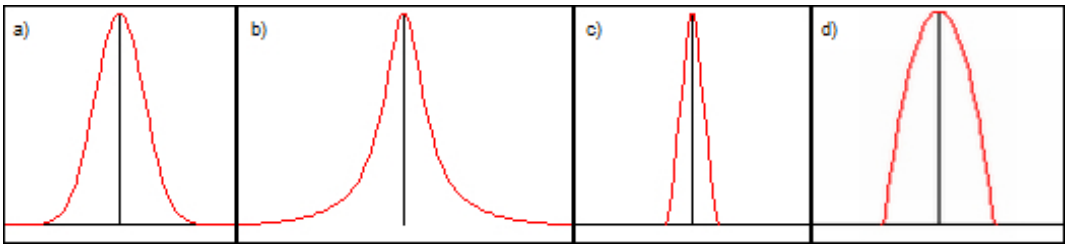



Figure 4.1: *Some kernel functions used for GRNN, a) gaussian, b) reciprocal, c) triangular, d) Epanechnikov*

Most often, the multivariate Gaussian function is used as a kernel:

$$K(x) = \prod_{i=1}^{p} K_i(x_i) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x\|^2}{2}\right) \text{ with } \quad x = (x_1, \ldots, x_p)$$

In this case, $\|x\|$ means the norm defined in the p-dimensional space. Centering the kernel on a data point $x_i$ and scaling the width with the smoothing parameter $\sigma$, the following form is obtained:

$$K\left(\frac{x-x_i}{\sigma}\right) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \quad i = (1, \ldots, N) \tag{4.29}$$

With the Gaussian kernel 4.29 the NWKRE estimator 4.28 is the one proposed by Specht:

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} z_i \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)}{\sum_{i=1}^{N} \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)} \tag{4.30}$$

For the case of spatial prediction, GRNN learns from input coordinates $(x_1, x_2)$ of points in a 2D space and from output values $z$ in the training dataset. $\hat{f}(x)$ is the prediction value, and $\|x - x_i\|^2$ is the distance norm from the predicted point $x$ to the training data $x_i$.

To optimize the sigma parameter, a starting value is required. The starting sigma has to be related to the density of points; the same value used for the lag distance used in variography can be applied. The starting sigma cannot be too small of a value, since errors of division arise during calculations. Since the algorithm calculations involve distances between points, data spatial density is an important issue. High spatial clustering can result in a higher sigma because smoothing is the best solution for such problems.

### 4.4.2 GRNN for validity domain definition

As mentioned in (42), the GRNN algorithm can produce a useful result, which is the density estimation of input variables. To obtain this, one should only compute the denominator term in equation 4.30 normalized by the constant term $(2\pi)^{p/2}N$ as follows:

$$\hat{f}(x) = \frac{1}{(2\pi\sigma^2)^{p/2}N} \sum_{i=1}^{N} \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{4.31}$$

The input density, according to equation 4.31, is a good indicator of the sampling schema in the case that input variables are spatial coordinates. Knowing the spatial density of samples can help define the validity of the sampling domain for interpolation methods. As a consequence, it can help improve the sampling schema and define resampling campaigns. It can be also helpful in improving the configuration of fixed locations for measurement, as in the case of monitoring networks.

As discussed earlier in chapter 3, the sampling schema of indoor radon fits to the urban area. Any improvement to the sampling must be, therefore, constrained to this domain.

## 4.5 Validation, methods robustness and data consistency

A panoply of interpolation methods exist for mapping; the question is, how can one decide which one is performing better? The usual procedure to test methods is to reserve part of the samples in order to make a blind validation of the estimator. In this way, the best model can be chosen. But what about the hypothesis and data requirements behind the methods, do they have some influence on validation results? For instance, it was mentioned that the performance of kriging would depend on stationarity and that GRNN requires a minimum volume of data. Methods perform differently depending on data conditions and

consequently the validation error can vary. It can be probed that the optimal model parameters by cross validation vary depending on data selection. It is therefore important to analyze how validation errors vary in function of methods (robustness) and in function of the samples consistency.

This analysis is restricted only to the consistency between the training data (used to tune the model) and the validation data, which are both obtained from the available samples. This analysis is not 'realistic' in the sense that does consider the consistency between samples and the unknown global data. The goal of sampling is to obtain values that are as representative as possible to the variable under study. Statistically speaking, that means that sample parameters should be consistent to global parameters. Unfortunately, we do not know the global parameters which would allow us to determine whether samples are consistent; this is a matter of good sampling design.

The presence of extreme values makes consistency analysis more complicated. The inclusion of certain extremes in a set of data can largely modify the statistical parameters. It is therefore convenient to work with relatively large datasets. It is also important to select subsets several times in order to be more confident about the results. The jackknife procedure is a repetitive selection of data with the purpose of modeling. Such a procedure was adopted to validate interpolation methods under different conditions of data consistency.

### 4.5.1 Data consistency by Jackknife procedure

Jackknife is a resampling technique, without replacement, used to calculate statistical parameters from a set of samples. It can be used to test the parameter distributions of samples and the models used for prediction. The idea is to take a subset of samples (resampling) and to check how much the parameters vary. When the variations are not significant, it can be said that these sets are consistent regarding the larger dataset and other subsamples. Of course, this procedure must be repeated with a number of splits in order to obtain some statistics.

The self-consistency of the samples is, ultimately, an indicator of consistency with respect to the underlying population, which is unknown. Large variations can indicate that samples are not representative of the process or that an important amount of noise is present.

With jackknife splitting, it is also possible to test the robustness of interpolation methods. That is, the model parameters are tested for variations, and the estimator performance is also tested through blind validation. For this purpose the remaining data from the jackknife procedure is used as a validation dataset each time. A random generator seed must be used to ensure the heterogeneity of the resampling.

### 4.5.2 Statistical and spatial consistency after data splitting

Using spatial data analysis methodologies such as fractality and clustering measures, the indoor radon samples were labeled as highly clustered. This clustering has a counter-part of high spread of isolated samples, which makes data spatial distribution heterogeneous. The uneven spread of samples can be also a drawback for data resampling because large isolated

areas will result uncovered. Therefore it is necessary to verify after data splitting that the subsets are not only statistically consistent but are also spatially consistent.

### 4.5.3 Splitting unbias optimization

The lack of statistical consistency between datasets can be primarily perceived when their mean and variance parameters differ. This bias for the two main statistical parameters can already cause erratic results in estimations. When doing a relaxed random splitting of data, all sorts of situations can be expected. It is possible that, by chance, both sets are statistically similar or quite different. An optimization criterion was introduced to the random splitting in order to reduce the bias between the large dataset, the validation set and the training set.

After launching a large number of random splits the mean bias of the mean and variance can be computed. An iterative searching procedure was then implemented to to make splitting with a multiple criteria of minimal bias of four parameters. The four biases to be minimized were the mean and variance for both validation and training with respect to the large set. By forcing the selection of only unbiased sets, we ensure the statistical consistency between subsets.

To ensure that sets have comparable parameters, exactly the same number of samples must be selected for validation and training sets during successive jackknife repetitions.

### 4.5.4 Data splitting by random declustering

In order to obtain subsets, not only with statistically consistency but also with spatial consistency, it is necessary to improve data selection using a spatial criterion. In (42), the use of the spatial declustering technique is proposed as a means of proper splitting. To perform this, a good compromise between the cell size and the number of selected samples should be found. For data splitting into training and validation sets, it is wise to resample the validation set first (usually smaller) leaving the rest for training. The size of the cell should not be so small that the single point from an isolated area is taken for one of the sets. A moving window (MW) statistic of the number of points helps to determine the best cell size in advance.

The idea is to consider the spatial distribution of points to perform an equivalent partition according to data density. The first step is to calculate the number of points per window after MW partition, so that each window hat at least 2 points. Then, data within windows are split following the selected proportion for training and validation sets. The proportion is rounded so that at least one point per window is selected for the validation set. The selection within windows is done at random.

This spatial optimization selection can be combined with an unbiased selection in order to have statistical and spatial consistencies.

## 4.6 Deterministic modeling and predictions of sets 3A and 3B

### 4.6.1 KNNR model validation of set 3A

KNNR was performed using the optimal $K$ parameter for set3A, which is 15, and the results were compared with validation data to evaluate the model. The validation set3A consists of 144 samples. The validation MSE for set3A is 91207, with a mean error of 15.6 and an error variance of 91624. In Figure 4.2a, a plot of true values vs. predicted values is presented, with the mean indicated for both. Out of the predicted mean, it appears that some overestimation has occurred.

The first thing to notice is that the mean of predictions (238 Bq/m3) is higher than the mean of validation points (222 Bq/m3). The training dataset toy3A contains higher values than the corresponding validation set, and this is reflected in predictions. A second point to observe in Figure 4.2a, is that the range (between minimum and maximum) of predicted values is shorter than the true values from the validation set. This reflects the smoothing effect of averaging neighbors values to obtain a prediction. It is also noticeable that for some low validation values, high predictions were obtained. This is a result and an expression of local variability.

Regarding the distribution of errors, a plot was made with predicted values against prediction errors (Figure 4.2b). This graph shows that out of the mass validation points, a majority was overestimated. This is reflected by a median error of 47.5 (the mass of points are over 0). The mean error is also a positive value of 15.7, which indicates that errors are slightly biased to high values. The presence of extreme underestimations and overestimations is also noteworthy.

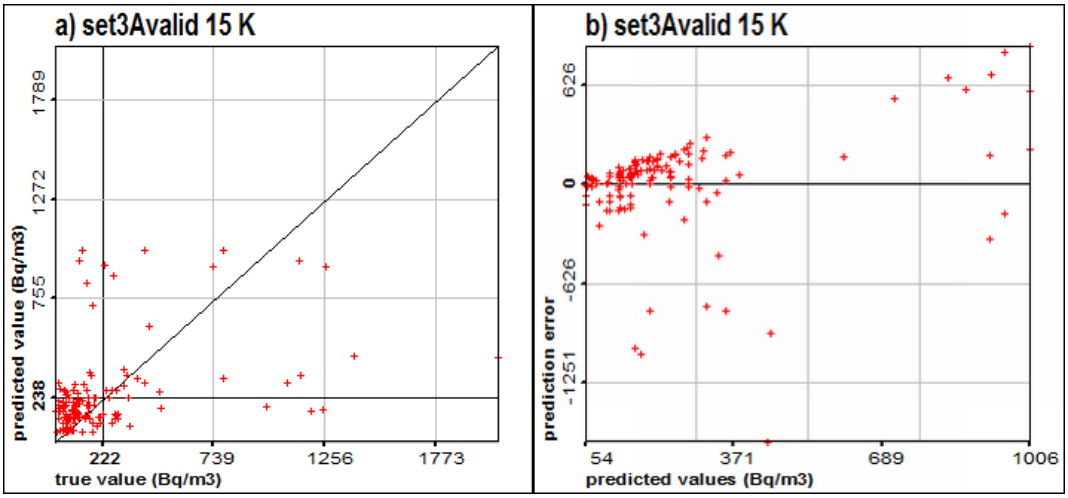



Figure 4.2: *Plots of a) true values against predicted values and b) predicted values vs. predictions errors or the validation set3A*

### 4.6.2 KNNR model validation of set 3B

The spatial modeling of set3A can be difficult, even using a simple method like KNNR because of the presence of extreme values. Then, for the purpose of methods comparison, it is interesting to focus on set3B, for which some spatial continuity structures have been identified. The sensibility of the optimal parameter $K$, which is 13, was also tested against 1 and 26 $K$.

The corresponding validation set contains 256 samples. For set3B, the KNNR prediction of the validation values using 13 NN has resulted in an MSE of 3126, a mean error of 2.6 and an error variance of 3140. When using 1 $K$ for validation, MSE increases to 6908, the mean error is 9.42 and the error variance 6873. With 26 $K$ neighbors, the MSE is 3339; the mean error is 2 and the variance 3348. After refinement, it was found that the lower MSE is obtained using 12 $K$ with a MSE of 3110 as can be seen in Figure 4.3.

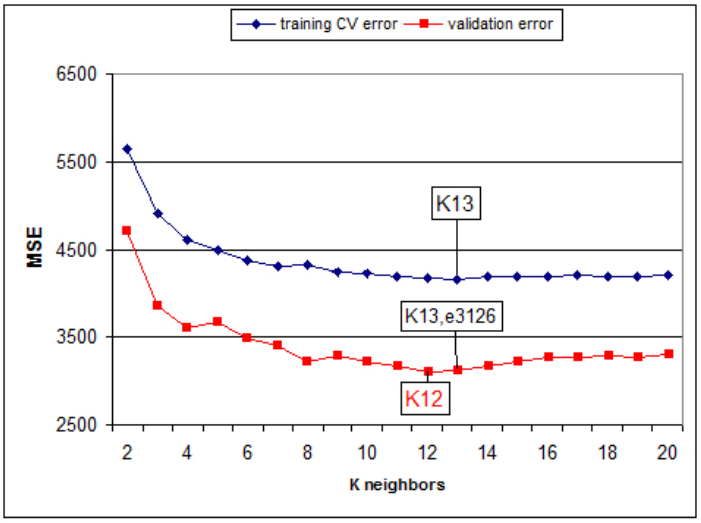



Figure 4.3: *MSE training and validation curves for the set3B using KNNR method*

The optimal CV training parameter of 13 $K$ was not far from the optimal validation $K$ which indicates that training and validation seem to be comparable sets in terms of distribution. Validation data are highly sensitive to 1 $K$ in comparison to the use of 26 $K$. For training data with high local variability, the more averaging that is done, the less high errors result. This statement can be confirmed by looking at plots of true vs. predicted values in Figure 4.5.

The distribution of errors for 1 $K$ has a tendency to increase, together with values, and more errors are committed for high values. For 13 $K$ and 26 $K$, the distribution of errors is more homogeneous because of the smoothing effect. It is also important to point out that predictions result less skewed (2.5) than the skewness (3.1) of the true validation values .

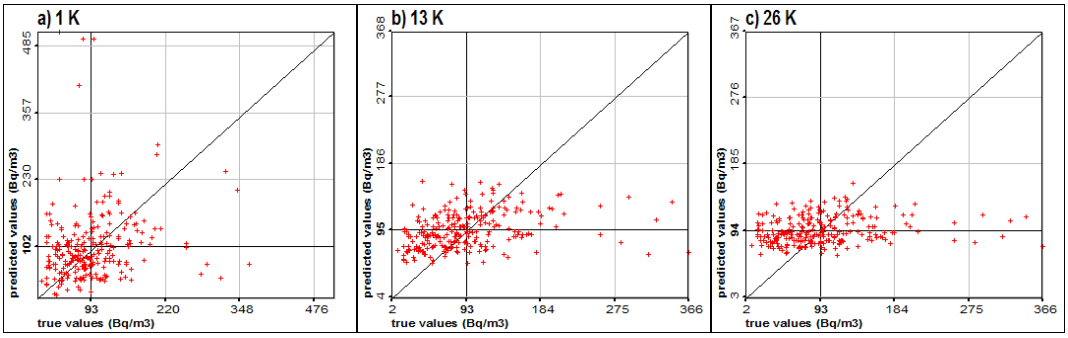



Figure 4.4: *Plots of true values against predicted values for the validation set3B using a) 1 K b) 13 K and c) 26 K*

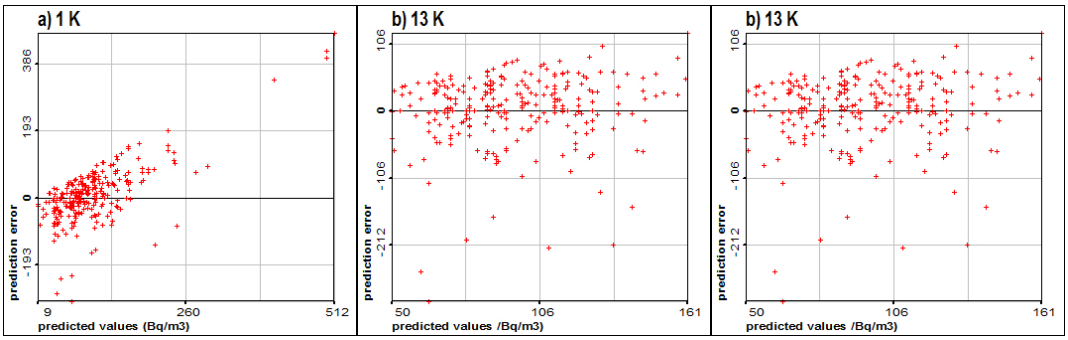



Figure 4.5: *Plots of predicted values against prediction error for the validation set3B using a) 1 K b) 13 K and c) 26 K*

### 4.6.3 Data filtering of set 3B with KNNR CVMF

The KNNR cross-validation mean filter (CVMF) filter has been applied to the set3B using 13 $K$. The result is an enhanced visualization of data with an optimal scaling. The results provide a better data visualization as obtained with MW averaging. In Figure 4.6, the plot of radon samples is compared with the MW averaging results. In Figure 4.7, a plot of indoor radon samples before and after the use of KNNR mean filter is presented.

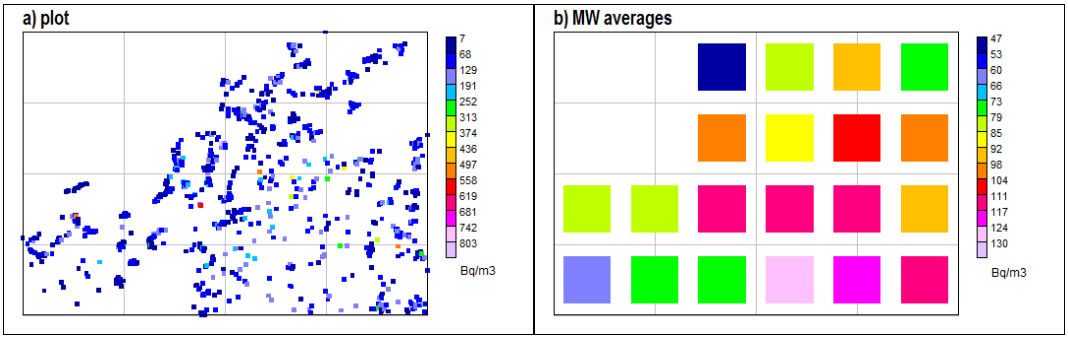



Figure 4.6: *Plot map of samples before and after MW averaging*

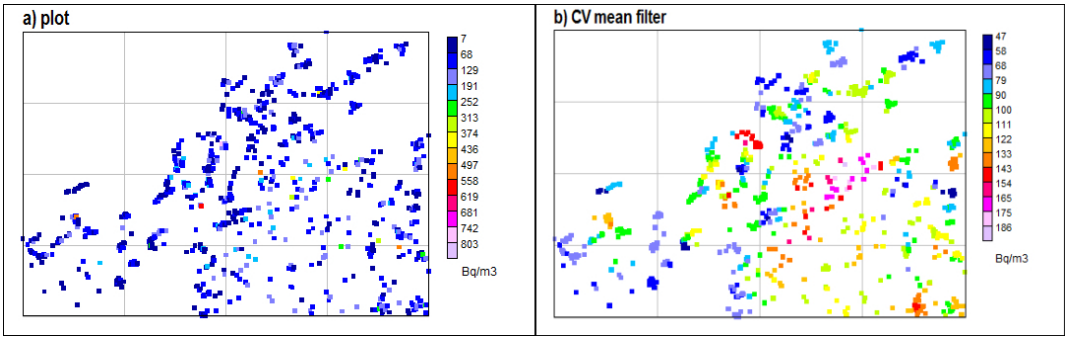



Figure 4.7: *Plot map of samples before and after KNNR mean filter transformation*

The use of such a KNNR filter is useful to identify areas with relatively higher values, as well as for the physical interpretation of the process due to the enhanced visualization. The effect of the KNNR filter on the local variability was measured by doing the KNNR cross-validation once more over the filtered data. For set3B, the optimal number of $K$ after filtering decreases from 13 $K$ to 8 $K$.

The KNNR CVMF has produced a data transformation that has reduced variance from 4377 to 706. This is an expected effect of averaging. Skewness was reduced as well, from 3.4 to 0.6, and kurtosis decreased from 23 to 0. Meanwhile, the mean value of 96 Bq/m3 remains. The resulting data distribution is tempting for variography modeling, even knowing that original values were modified. The variograms for radon measurements before and after a KNNR CVMF transform are presented in Figure 4.8.

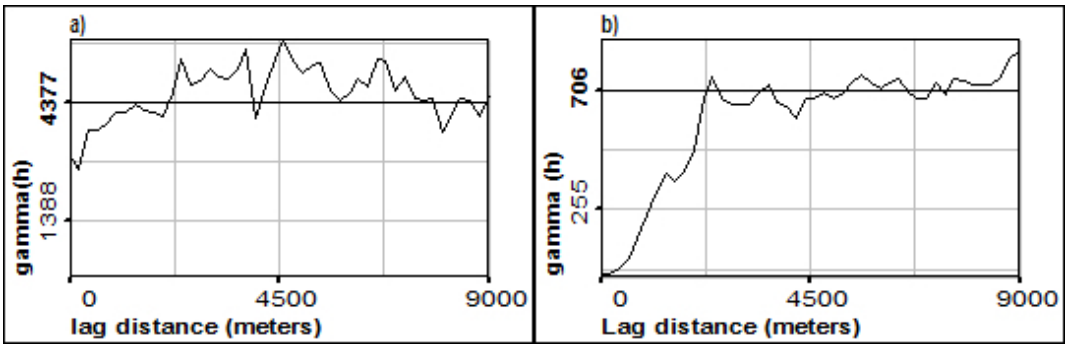



Figure 4.8: *Variogram of set3 for raw values (a) and transformed with KNNR CVMF (b)*

Variograms using a KNNR CVMF present ideal features, such as zero semi-variance at origin and a bounded semi-variance (not exceeding the a priori variance) at a certain range-distance. With these conditions, it is easy to fit a stationary model. Is this filtering also useful for interpolation purposes, knowing that the original values were modified? The advantage of easy modeling is what makes it worth trying to use this filtering data for interpolation purposes, as will be later addressed.

### 4.6.4 Inverse Distance Weighting (IDW) modeling of set 3B

For the IDW procedure, the lowest training error was obtained with an optimization curve for the power parameter. The best number of neighbors was assumed to be constant and was obtained from KNNR (13 $K$). In fact, after testing different number of neighbors and power values, it was seen that the optimum $K$ from KNNR is also optimum for IDW. The quadrant selection hyper-parameter was also tested, but this does not result in improvements for the present dataset. The optimal power was 0.5, as is presented in the optimization curves in Figure 4.9.

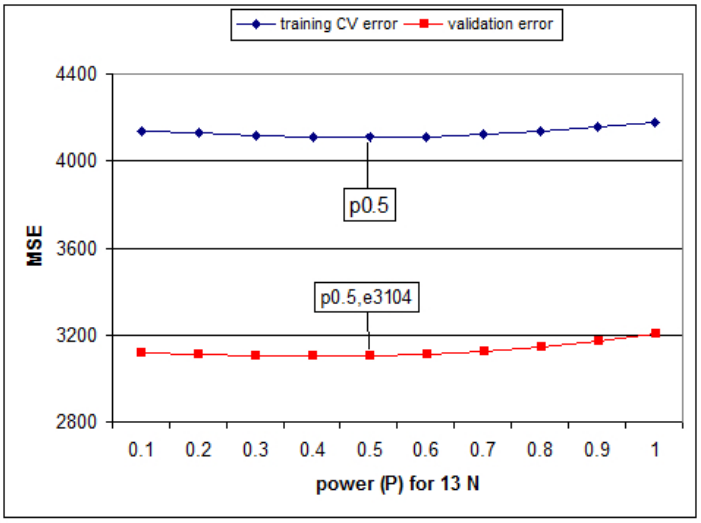



Figure 4.9: *MSE training and validation curves for the set3B using IDW method*

The validation error using the best training parameters of 13 $K$ and $p = 0.5$ was 3104. The MSE validation curve and a plot of predicted values against real values are presented in Figure 4.10.

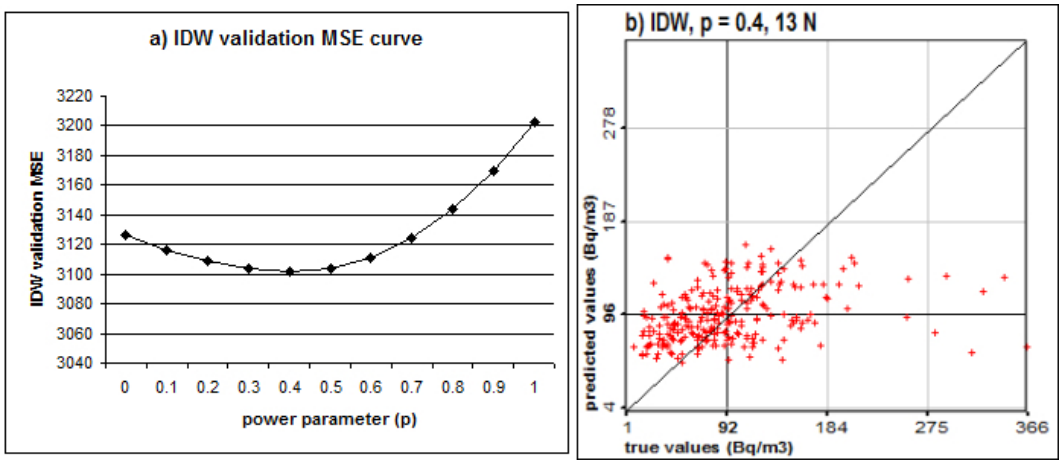



Figure 4.10: *a) MSE validation curve for the set3B and b) true versus predicted values plot*

The validation error curve shows that a lower error can be obtained with p=0.4. Of course, the training parameters are rarely optimal for all validation sets, but in this case, they closely approach to an optimal validation, which indicates consistency of data. The true vs. predicted value plot has the same range of values at the two axes to make evident how much predictions are smoothed compared to the true values. A bisector was also included in the graph to show that predictions are slightly higher.

Compared to KNNR (MSE=3126), the IDW method has improved the validation results (MSE =3104). The lowest MSE was also obtained with 13 neighbors and a power distance lower than 1. In general, it seems wise to use the optimal $K$ parameter for departure and then, to test the $p$ parameter. In general,the more local variability exists the more neighbors are needed for prediction, and likewise, the distance property becomes less significant. Once the number of neighbors defines the optimal limits of local variability, a slight improvement is obtained by weighting distances.

## 4.7 Geostatistical modeling and Kriging of set 3B

### 4.7.1 Variography and parameters modeling

The first step when working with kriging is to produce an experimental variogram where a variogram model can be fitted. In Figure 4.11, the variogram for set3B, with two models fitted on top, is presented. In Figure 4.11a, the model corresponds to a variogram that is not exceeding the a priori variance of 4377. It is composed of a nugget model accounting for 2377 of total variance plus an exponential model with a variance of 2000. The anisotropic range is $R = 1700$ meters. In short, the variogram model can be coded as nug2377+exp2000R1700m. It is an isotropic model so that only distances $h$ and no directions were modeled. A second variogram (Figure 4.11b) was built by exceeding sample variance in order to better fit the sample variogram (it is coded as nug2500+exp2500R3300m). The use of an exponential model reflects the high local variability because it accounts for more variance modeled at short distances.

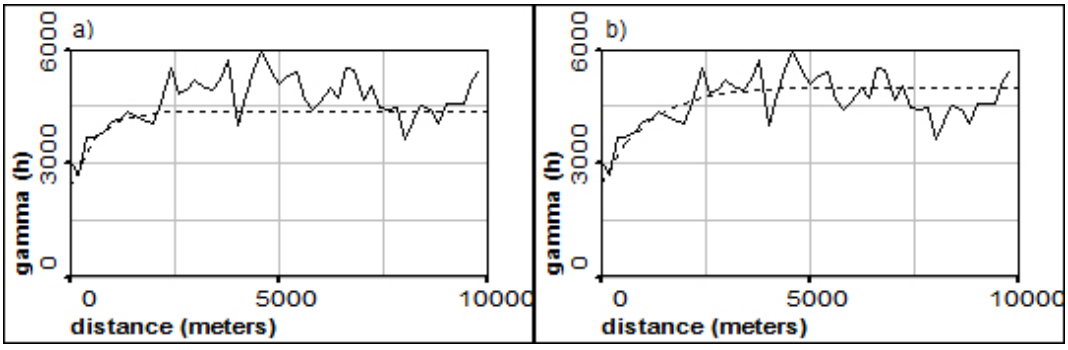



Figure 4.11: *Varigram models fitted on top of the experimental variogram for set3B a) bounded to the apriori variance b) exceeding apriori variance*

The second step is the training procedure by cross validation to test parameters. SK and

OK methods were tested, as well as the number of neighbors. The search radius was given a high value and has no major influence. In theory, all sample points within the variogram range can be used for estimations since they have common random functions. The number of neighbors and the variogram defined the neighborhood. In practice, the range is short in comparison to the domain, and the influence of the variogram decreases with distance. So, a limited number of neighbor samples are enough for estimations due to local variations.

The curves of training optimization for the two kriging estimation methods and different numbers of neighbors indicate a set of optimal parameters (see Figure 4.12)

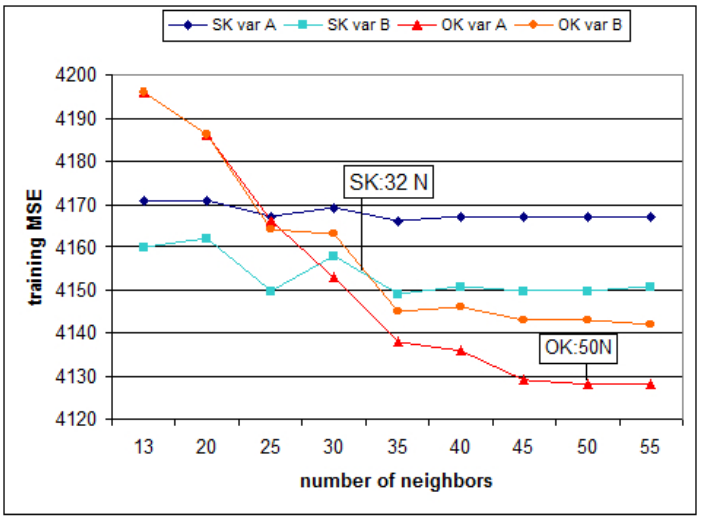



Figure 4.12: *CV training optimization curves using SK and OK with different number of neighbors and two variograms for the set3B*

The optimal model is obtained with OK considering a stationary model (variogram A) and 50 neighbors, and then comes OK with variogram B and 55 neighbors. The SK method with variogram B has less MSE training errors than variogram A. For SK the optimal number of neighbors for each variogram after refinement is 32.

### 4.7.2 Kriging validation

Using the training parameters, SK and OK methods were launched for the validation set, as was done for KNNR and IDW. In this case, the results do not follow the tendency seen during the training procedure. The lowest validation error was obtained with SK and variogram A with an MSE of 3126 (r=31). SK holds the second position with variogram B and an MSE of 3175. The validation MSE values for the OK were 3209 and 3235 for variograms A and B respectively.

It is always possible that the best training model does not give the lowest validation errors, as is here the case. In fact, it seems that the OK method falls into over fitting due to the high local variance. In general, OK performs better due to its local unbiasedness assumption. It seems that the high local variability of indoor radon data, in this case, has

exceeded the generalization capabilities of OK. SK, on the other hand, is a simplest method that does not rely so heavily on local variations but on global stationarity.

What is particular to the set3B dataset is that a spatial data selection was done in order to attain global mean stationarity. Actually, a central idea in geostatistics is to define a statistical model valid for a variable within a region, which was called *variable regionalizé* by Matheron. Hence, trend analysis and a data selection is a recommended step before using kriging. In most cases, this spatial partition is not possible due to a limited amount of samples. An advantage from the indoor radon dataset is its large number of samples, which allows making an effective partition in regard to trends.

To check whether the training parameters approach what will be optimal to validate, the two error curves were put together in Figure 4.13. This was done for the SK method by testing a different number of neighbors (Figure 4.13). The variogram model is assumed to be optimal. The two CV curves look alike and indicate some consistency of the training set with

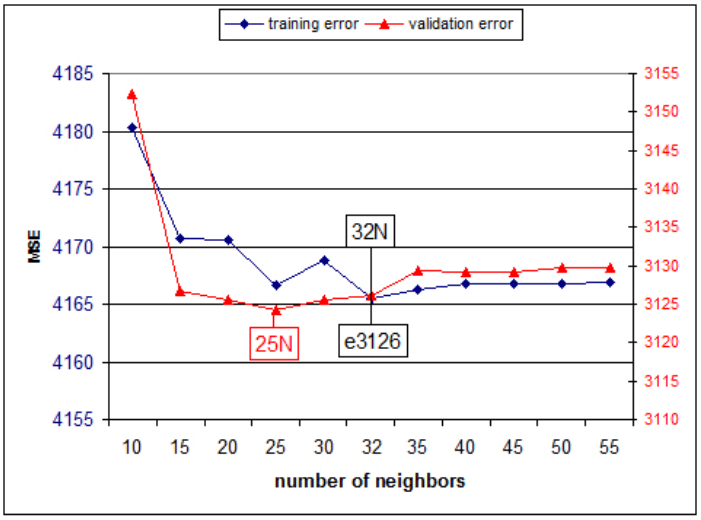



Figure 4.13: *Training and validation error CV curves using SK with different number of neighbors for the set3B*

respect to the validation set. The curve does not have a smooth shape as it does in KNNR or IDW methods, most probably due to the influence of local discontinuities of the experimental variogram with respect to the variogram model. These shifts are visible in the experimental variogram (Figure 4.11). Variogram modeling and training of hyperparameters (e.g. the number of neighbors) is a time consuming task in kriging. This modeling effort does not always pay-off, as seems to be the case for this example data. Nevertheless, it always gives details about the spatial continuity of data.

### 4.7.3 Modeling of variogram anisotropy

An additional refinement was done considering the possible influence of anisotropic structures. Anisotropy is not evident for the dataset, but small tendencies can de identified with

the help of a variogram rose (Figure 4.14). At an approximate distance of 2500 m. an ellipse was drawn in white to indicate what seems to be a tendency to the northeast direction. With this information, the experimental variograms in a direction of 20° appear to have the most continuity as presented in Figure 4.15a. The corresponding orthogonal at 110° has a shorter range of continuity. Both directions were modeled with a variogram with a nugget of 2277, a spherical model with 2100, a principal range of 1800 meters in direction 20° and an orthogonal range of 550 meters. This information can be coded as nug2277+sph2100(20)R1800r550.

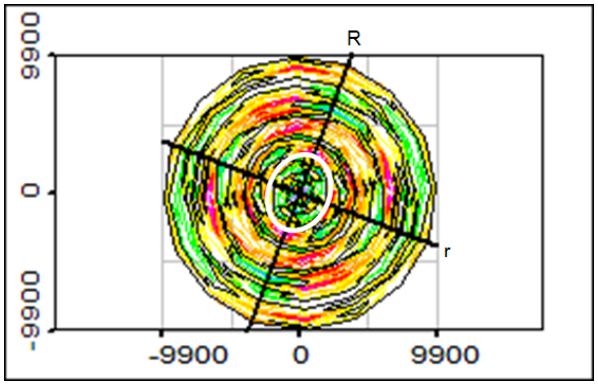



Figure 4.14: *Variogram rose graphic used to represent anisotropy of data, the larger continuity seems to be at 20 degrees direction as indicated by the ellipse*

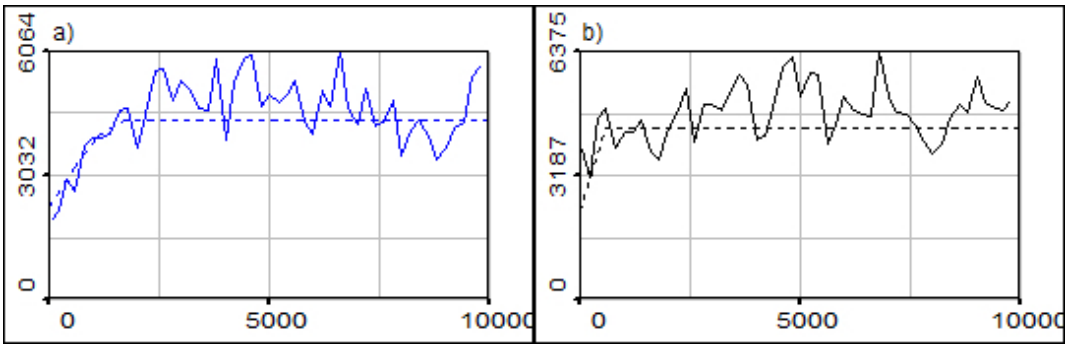



Figure 4.15: *Directional variograms a) at* 20° *b)* 110° *with models fitted on top*

The optimal parameters obtained after training using an anisotropic model indicates that the optimal number of neighbors for SK is 35 while 50 for OK. These optimal numbers are similar to those in the isotropic model. The MSE for the validation set using SK and OK were 3335 and 3409 respectively. In the case of the anisotropic modeling, again, the SK method performed better. The anisotropic modeling for this dataset does not seem to give any advantage to estimations. On the other hand, the slight advantage for the SK method can be, as already mentioned, due to the 'de-trended' condition of set 3B.

The validation results with kriging (MSE of 3126) are worse than those obtained with the IDW method (with a validation MSE of 3104) and KNNR (MSE of 3126). Kriging being

a model-based method, is hardly dependent on the integrity of the model. The presence of a high nugget model and a short range is not optimal for geostatistical methods. The high local variability of data was a major impediment to model spatial continuity at short distances. Moreover, the short range of the variogram indicates that spatial structures have low continuity.

### 4.7.4 Kriging of KNNR CV mean filter transformed data

As remarked earlier, kriging was not very successful due to data limitations in adjusting to continuous models. The KNNR CVMF method, however, used to improve visualization, has effectively reduced local variability and allowed a clear model-able experimental variogram. The question is, whether data filtered in this way and its model can be used to make estimations. How much information is lost after filtering? Could this loss be compensated by the fact that the variogram model is well structured?

A trial was conducted by adjusting a model to the set3B KNNR filtered data. As shown in Figure 4.16a, the best model is a pure Gaussian one with a range of 4800 m or in short, gaus706R4800m. Next, the optimal parameters were obtained for SK and OK methods (Figure 4.16b).

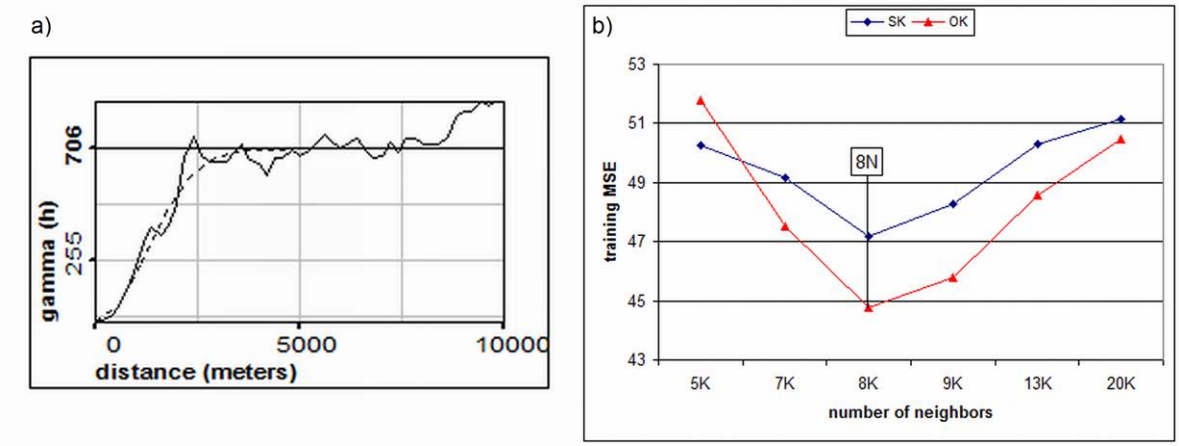



Figure 4.16: *a) Variogram model of gaussian type with a range of 4800 meters for the CV mean filtered set3B data, b) CV optimization curves using SK and OK with different number of neighbors for filtered set3B*

Finally, simple kriging was performed using the same raw data validation set used for previous methods in order to compare the results. The MSE for SK and for OK was 3085. The correlation between estimations and true values was 0.33. The results are better compared to other methods but are still low for estimations. It is remarkable that the local variability of this indoor radon dataset is such that an averaging by filtering can improve estimations. One must be conscious that this local filtering is not necessarily a solution for every kind of dataset, but it seems to be helpful for the particular case of indoor radon measurements in Switzerland.

### 4.7.5 Kriging comparison with KNNR and IDW

As a matter of comparison, set3B filtered with KNNR CV was used for interpolations with KNNR and IDW methods. The training procedure indicates that the optimal number of neighbors for KNNR is 8 and the validation MS error is 3208. For IDW the optimal power was 1.4 and the validation MSE 3149. In Figure 4.17, the optimization curves for the three methods using filtered data are presented together.

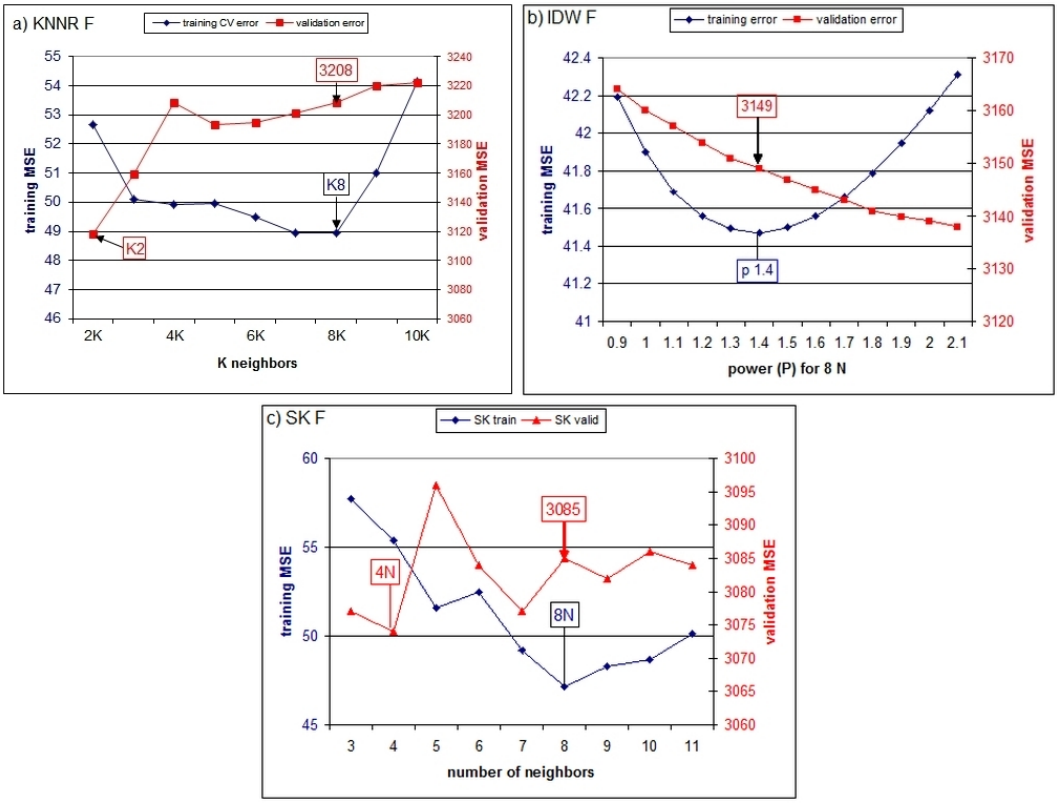



Figure 4.17: *Training and validation CV MSE curves for filtered set3B using a) KNNR method b) IDW method and c) simple kriging (SK)*

The smoothness of filtered data gives no advantage to the IDW or to the KNNR method. Validation results are even worse with filtered data than when using raw data. For SK there is a slight improvement mainly because of the well-structured variogram and the reduced local variability.

The optimization curves show that for SK, the training parameters are closer to validation parameters with respect to the other methods. As a statistical method, kriging requires consistency between global parameters as the mean (which is the same for the training and validation set after filtering). On the contrary, for local interpolators as KNNR and IDW, local consistency has more influence.

The main inconsistency between training and validation sets, in this case, is the variance. Filtering has drastically reduced the local variance and hence, the global variance in

the training data. On the contrary, the use of a structured variogram (without a nugget model) and a larger range has reproduced the variance from the training data. What can be concluded so far is that much of the local variability is working as white noise. It is part of the process, but it prevents modeling. Mean filtering helps to obtain a model, although it is somehow an imposed model.

## 4.8 GRNN modeling and estimations for the set 3B

GRNN was employed to estimate indoor radon values for the set3B. At first, the training set was used to obtain an optimal sigma value by cross-validation optimization. Different kernel functions were also tested. Next, estimations were done over validation points to have an independent measure of the modeling performance.

The first trial was conducted with a Gaussian kernel as it is the most commonly used. Figure 4.18a shows the CV optimization curve and the optimal sigma value for training data, along with the validation error curve. In Figure 4.18b, the same comparison is made using the reciprocal kernel.

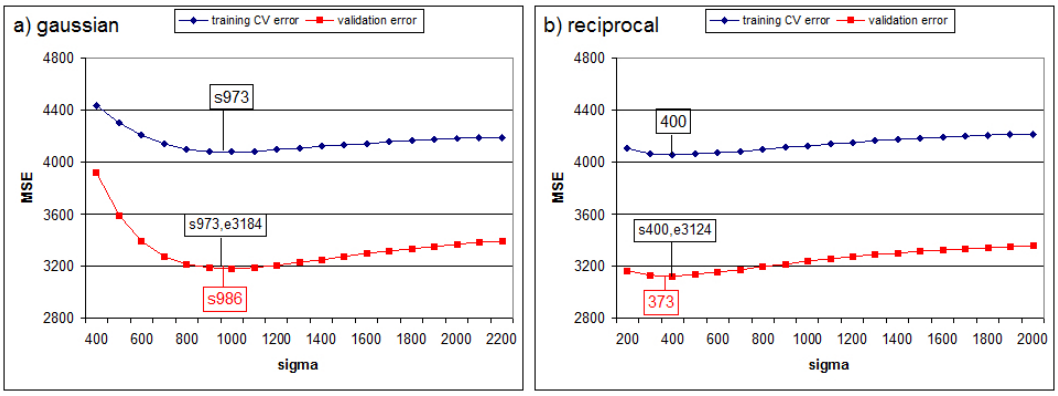



Figure 4.18: *Crossvalidation training error curves and validation error curves using a) a gaussian kernel and b) and reciprocal kernel function*

The optimal sigma obtained with the training CV procedure is the one that also gives the lowest validation error. The training and validation error curves are similar in shape and have the same optimum, which indicates that they are representative of the same population and spatial distribution. For the case of the Gaussian kernel, an optimal sigma of 973 was obtained with a validation MSE of 3184. With the reciprocal kernel, the optimal sigma was 400 and the validation MSE was reduced to 3124. The same procedures were done for non-asymptotical kernels as the triangular and Epanechnikov. The corresponding graphs are shown in figure 4.19.

For the case of the triangular kernel, the optimal CV sigma by training is in the order of 3205, giving a validation MSE of 3209. The lowest validation MSE was obtained with a sigma of 2500, which deviates from the optimal training. This kernel is too linear and proves some

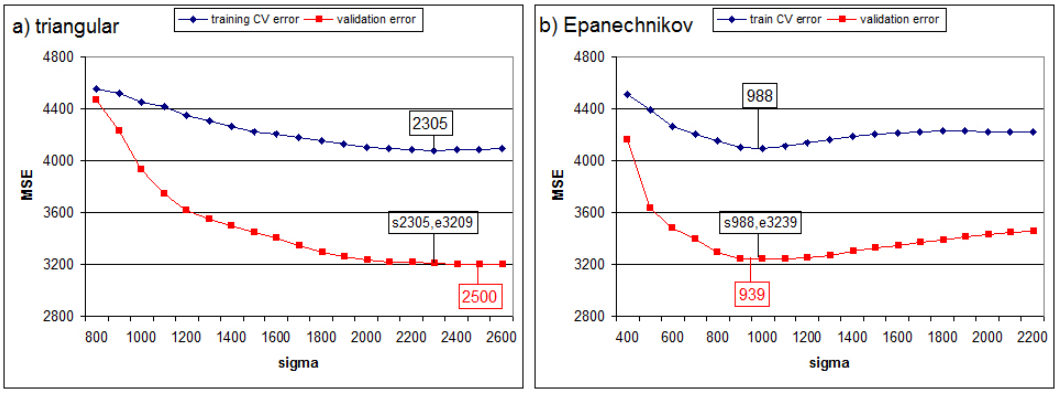



Figure 4.19: *Crossvalidation training error curves and validation error curves using a) a triangular kernel and b) and Epanechnikov kernel function*

dissimilitude between the distribution of the training and the validation set. This feature is even more noticeable with the rectangular kernel. The Epanechnikov kernel gave an optimal sigma of 988, which is not far from the optimal sigma for validation (939). Nevertheless, the validation error was higher, with an MSE of 3239. The low sigma value obtained with the reciprocal kernel indicates a better fitting of data. In general, GRNN does not improve validation results in comparison to other interpolation methods, such as KNNR, IDW and kriging. Next, it is proposed that the method should be applied to the KNNR CVFM filtered data.

### 4.8.1 GRNN Estimations using KNNR CVFM filtered data

The best results of indoor estimations, up until now, were obtained with the kriging methods using KNNR CVMF filtered data. Is the smoothness of filtered data also convenient for the GRNN? GRNN is already a smoother algorithm on a global scale; so, it is interesting to see how it behaves with locally smoothed data after a CVMF filter. The optimization curves of training parameters show which sigma value is the best one to be used for different kernels and what the corresponding validation error is (Figure 4.20).

The optimal sigma with the reciprocal kernel was approximately zero; therefore, no optimization curve was drawn. The reciprocal kernel can easily fall into over fitting and particularly with smoothed data as in the filtered set3B. The reciprocal function provides a density model with a large base, which expresses already high smoothing. The filtered data, which is highly smoothed, fits very well with such kinds of functions, and this is reflected by the very low sigma values obtained by the GRNN CV optimization.

Using a very low sigma value with the reciprocal kernel the validation error is in this case 3118, which is lower than when using a Gaussian kernel (3144) or an Epanechnivov kernel (3184). The MSE validation error using a Gaussian kernel is somewhat lower than when using the same kernel with raw data.

Slightly better results were obtained with the reciprocal kernel. While the training step

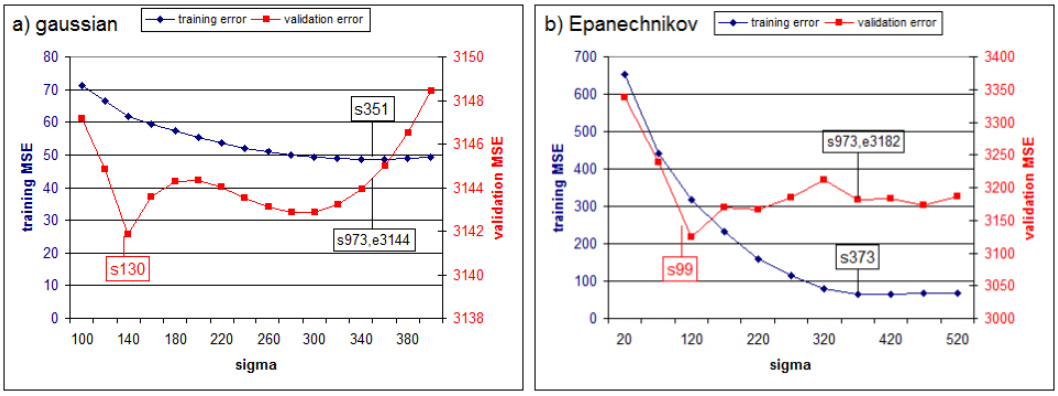



Figure 4.20: *Training and validation error curves for filtered set3B using a) a gaussian kernel and b) and Epanechnivok kernel function*

may appear to be successful, the estimations applied to validation data do not reflect an improvement in comparison with other methods. The best correlation of estimates and real values is 0.296. In general, the results with GRNN are comparable to the simple regression methods of KNNR and IDW. As for the kriging methods, it is noticeable that GRNN works equally well with filtered data despite the statistical differences between the training and validation sets.

### 4.8.2 GRNN for moving windows multiscale analysis

As was done with the skewness-test and the experimental variography in chapter 3, the GRNN method can be used to give insights of the data multiscale behavior. It was observed how skewness and variance reduces with data's size. It was also observed that the arbitrary partition of data with moving windows does not help with variography modeling but helped rather with the regional partition. The optimal sigma parameter from GRNN indicates the degree of spatial variability of the dataset; a high variable dataset will require larger sigma values or bandwidths to define their distribution function.

For the MW multiscale analysis, a GRNN optimization by leave-one-out cross-validation (CV) was launched. The Gaussian kernel is used for its simplicity and because it adapts to most kinds of data configuration. The sigma parameter with the lowest CV mean squared error (MSE) was recorded for each dataset enclosed by each of the cells for the resolution grids of analysis (namely grids 5, 10, 20 and 40). The algorithm is sensitive to the initial settings of sigma parameter; the starting sigma should be small enough to consider cases of high fitting. It was found that an appropriate starting sigma value was the average distance between points (equation 3.1).

In addition, a maximum sigma must be provided as a reference for the range of optimization. This maximum was computed as half of the diagonal of the bounding box. In practice, when the optimal sigma is found beyond this value the model becomes highly smoothed. A refinement of the sigma value is completed to give more precision to the results. A refinement is also conducted when there is a tendency of the optimal sigma to be

below the initial minimum or above the maximum values. In cases where the optimal sigma tends towards infinity, the dataset was considered unstructured. Grid cells having less than 20 points were filtered-out.

Table 4.1 presents the mean number of points, the mean variance, the mean of the half diagonals, the mean optimal sigma for GRNN, and a referential percentage of cells for each grid with sigma values under 80,000 (more structured).

Table 4.1: *Mean values for GRNN features at 4 scales of analysis*

| MW grid | mean n points | mean variance | mean half diagonal | mean optimal sigma | % structured datasets |
|---|---|---|---|---|---|
| **5** | 1817 | 107137 | 69099 | 2576 | 100 |
| **10** | 641 | 106281 | 35510 | 3152 | 89 |
| **20** | 214 | 86888 | 17187 | 2485 | 76 |
| **40** | 88 | 65066 | 8508 | 2183 | 71 |

To calculate the mean value of optimal sigma per grid, only the subsets with an optimal lower than 80,000 were considered. When a very large sigma is obtained, the optimization curve is no more convex and tends towards infinity. It should be taken into account that sigma relates to spatial distances since the input variables are spatial coordinates, but the sigma value itself is not a distance. Sigma is the smoothing parameter of the density function and reflects primarily how well it adjusts to the kernel function.

For grid5, 100% of the window subsets attained an optimal sigma in the range of the starting minimum and maximum sigma proposed; 7992 was the maximum optimal sigma obtained. It has been considered that these sets are more or less structured. For the grids with smaller moving windows and smaller datasets, there were cases where no convex curve of optimization was attained and therefore they were considered as spatially unstructured. The percentages of 'structured' sets were 100, 89, 76 and 71 % for MW grid5, grid10, grid20 and grid40 respectively. In map 4.21a, two subsets with an optimal sigma over 7000 were highlighted with red while the other 2 sets with small sigma values were highlighted in green. The corresponding CV optimization curves using GRNN are presented a-side.

As seen in Figure 4.21 the optimization curves have different shapes depending on whether a tendency towards fitting or smoothing exists. For the cells G5C7 and G5C10 a large sigma was found optimal. The common denominator for these two sets is that they both demonstrate significant spatial clustering. For cell G5C11, spatial clustering appears to also be important, but the relatively large number of samples seems to play a role in resulting in a low sigma. For the case of cell G5C25, the number of samples is low (only 108) but they appear to be well distributed spatially. In this case, more than one local optimum can be found. When using a starting minimum sigma higher than 3000, the optimization curve can fall in a local minima with a very large sigma. When forcing the search to lower sigma values, a local minimum of 888 was found. As mentioned before, GRNN optimization is sensitive to initial parameters.

The number of samples and spatial clustering appear to play a major role when it comes to train the optimal sigma. Having more samples in a set tends to produce a unique solution.

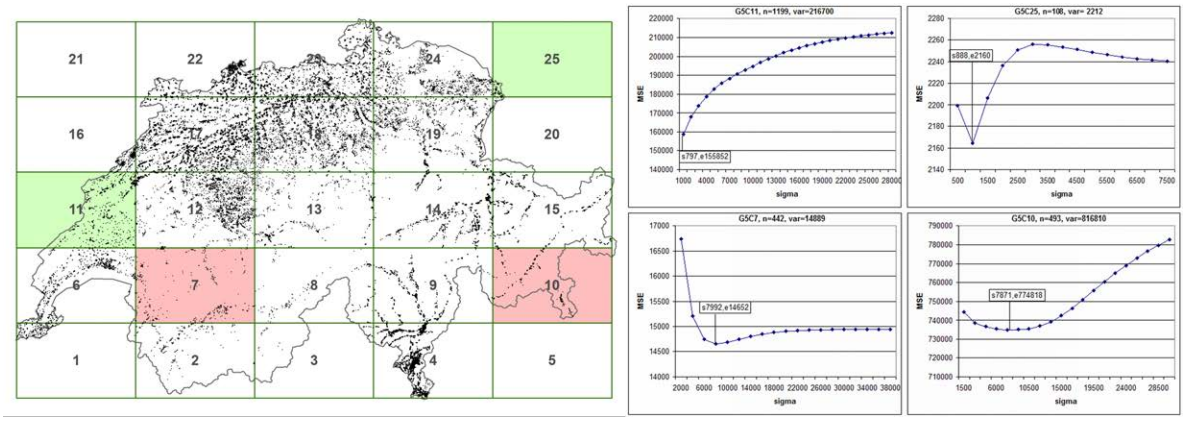



Figure 4.21: *Map with MW grid5 and some indoor radon GRNN sigma optimization curves with large and short sigma*

This tendency is seen in the results of Table 4.1 where the datasets of finer grid cells (G10, G20 and G40) are more unstructured. The more data available, the better GRNN defines the (non-parametric) spatial distribution function.

In Figure 4.22, a map with the MW cells of the grid10 is shown. Some cells are highlighted with colors according to the results of the GRNN sigma optimization. Cells with a sigma exceeding the bounding box half-diagonal distance are highlighted in red. The cells in orange color represent subsets with an optimal sigma between 5000 and the half-diagonal distance. The cyan colored cells are those that were discharged because they have less than 20 sample points.

From the previous figure, it can be noted that when datasets are smaller, there is a tendency for some subsets to have large sigma values. When data are heavily clustered (with large empty spaces) and with high local variability (as cell 83 from grid10), sigma values are high. The GRNN multiscale analysis is useful in identifying areas posing difficulties to modeling. At first glance, the areas with lower information are logically difficult to be modeled. Some areas with high clustering will prevent an adequate use of GRNN. The high local variance will primarily affect kriging methods as GRNN is more robust in this regard.

It is interesting to compare the general results from MW experimental variography with those of the MW GRNN. On the one hand, the number of samples to build a variogram model are not very constraining, but the stationary conditions are indeed critical for variography. GRNN, on the other hand, suffers with the lack of samples but it is robust to local variability and noise.

In Figure 4.23, the boxplot of optimal sigma values for the GRNN analysis on different scales are shown. Sigma values become skewed for subsets in grid40 because there are a significant number of extreme sigma values. In fact, the lowest optimal sigma values obtained for the datasets of grid40 were somehow forced to a local minima by constraining the optimization search to lower initial values. When data resulted absolutely unstructured, sigma values were very large.

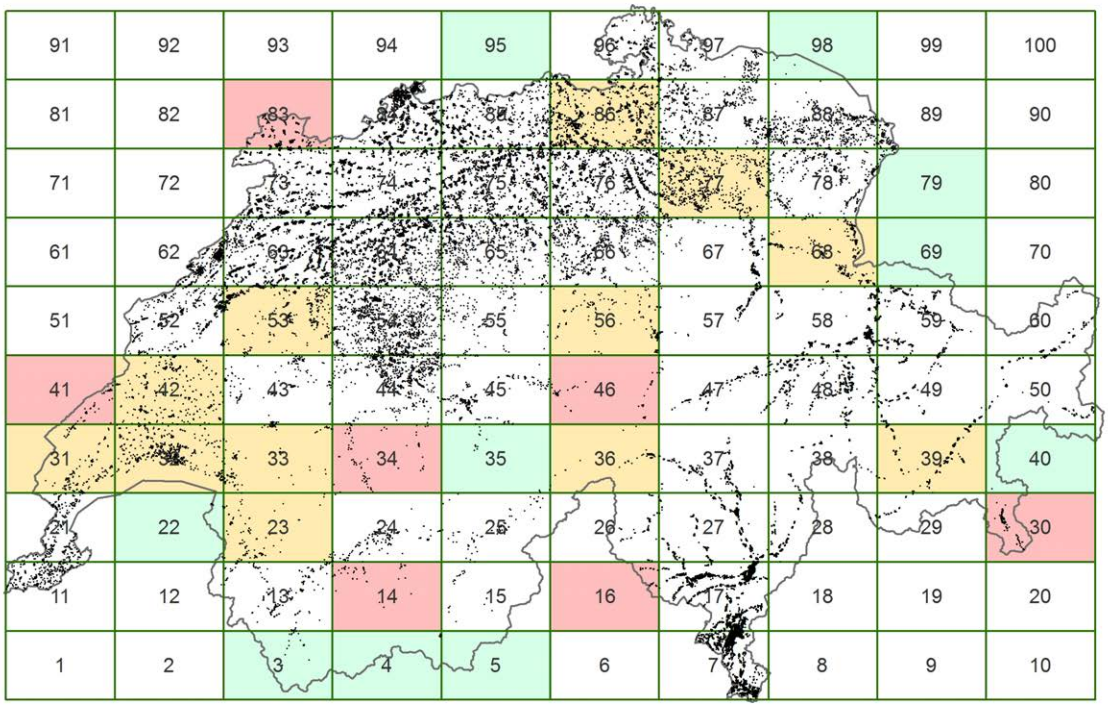



Figure 4.22: *Map of higher grnn optimal sigma values. Cells with larger sigma are in red and with medium sigmas in orange. Filtered-out cells are highlighted with cyan color*

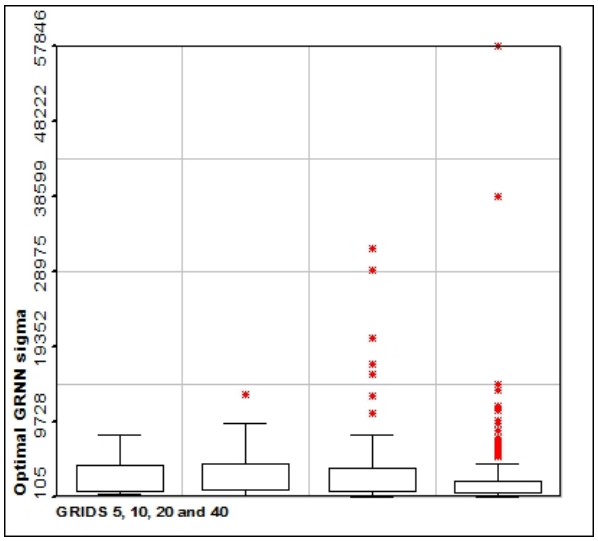



Figure 4.23: *Boxplots of optimal sigma values for grid5, 10, 20 and 40 (from left to right)*

If a very large sigma value is used to make predictions, the result is simply a generalized estimator or the mean of training points, as was initially expressed in equation 4.24.

### 4.8.3 Spatial density characterization with GRNN

As mentioned, the input density function of GRNN, expressed by the equation 4.31 provides a characterization of spatial density. Some data subsets of grid5 were used to compute the input density function, and the results are presented graphically in Figure 4.24

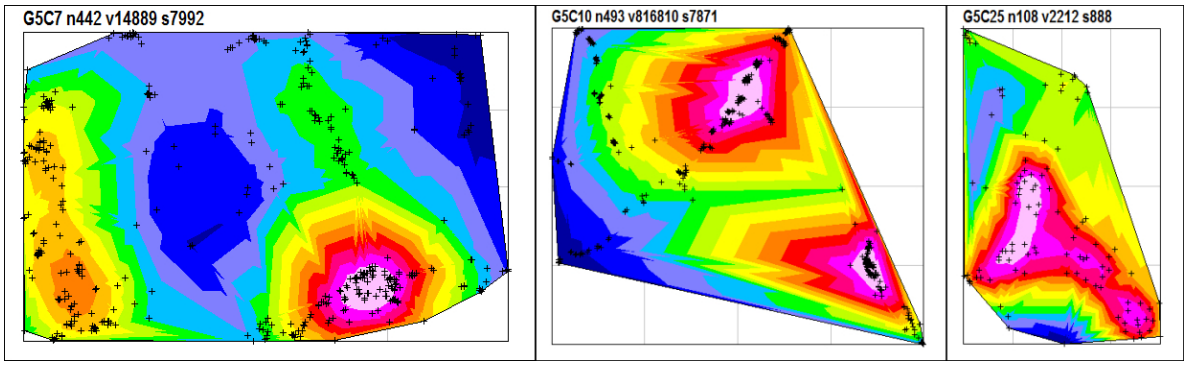



Figure 4.24: *GRNN input density maps for cells 7, 10 and 25*

These density maps were superimposed with the sample locations. The highly sampled areas present high-density values. What is particularly important is the extension of the areas with high-density values. They give a delimitation of a domain where the kernel function can be better applied for predictions.

## 4.9 Mapping inter-comparison for the set 3B

### 4.9.1 Indoor radon mapping of 3B raw data

To summarize the results of regression methods for set 3B, the lowest validation error for set3B raw data was obtained with the IDW method (3104), followed by GRNN (3124), KNNR (3126) and SK (3126). Using filtered data as estimators, the validation error was better using SK (3085), followed by GRNN (3118), IDW (3149) and KNNR (3208). The validation errors for the different methods are so close, that they cannot strongly indicate which one can perform better. Data interpolation over a grid will produce a continuous map and a better visual impression in order to observe the results. A series of maps were prepared using the interpolation methods and the already mentioned parameters.

A grid with a resolution of 200 meters was prepared and a legend scale defined. For the legend, a scale with 13 intervals of 25 Bq/m3 each, going from 0 to over 300 Bq/m3, was used. In fact, the training set has a range of values ranging from 7 to 803 Bq/m3, but values after interpolation have a lower range due to smoothing. The use of a common scale is important in order to compare the results. Another way to made a visual comparison is to produce a more realistic image of data. The KNNR using 1 neighbor or simply called Nearest Neighbor (NN) method is a simple visualization of data. For the first map, NN method was used for the whole dataset (meaning training and validation together). This

first map contains more comprehensive information, which is what a realistic map should look like.

Figure 4.25a represents the NN whole dataset map. Figure 4.25b is the KNNR map with 13 neighbors. In Figure 4.26, there are two maps, one for the simple kriging (SK) method (figure 4.26a) and another for the ordinary kriging (OK) method (Figure 4.26b). Finally, the maps for the methods with lowest validation errors using raw data are shown, IDW in Figure 4.27a and GRNN with reciprocal kernel in Figure 4.27b. GRNN with a Gaussian kernel was also used to produce a map (in Figure 4.28).

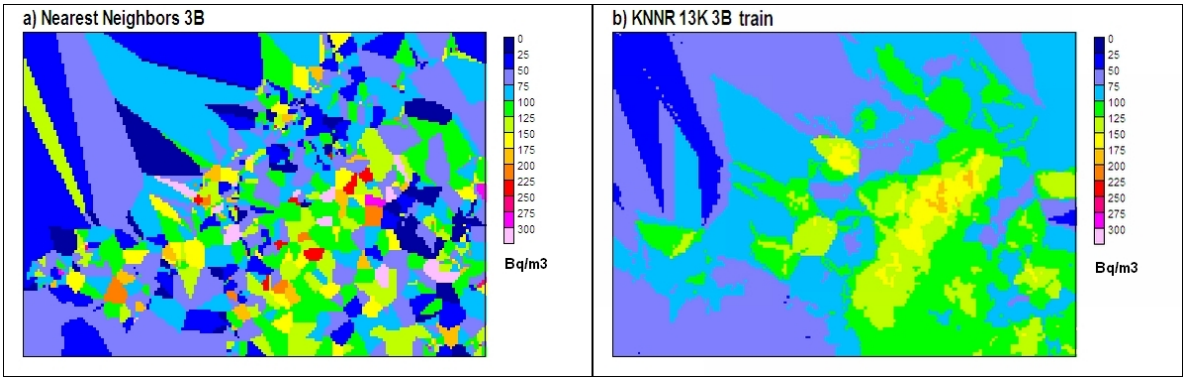



Figure 4.25: *a) Map for the set 3B using NN b) Map for the training set 3B using KNNR*

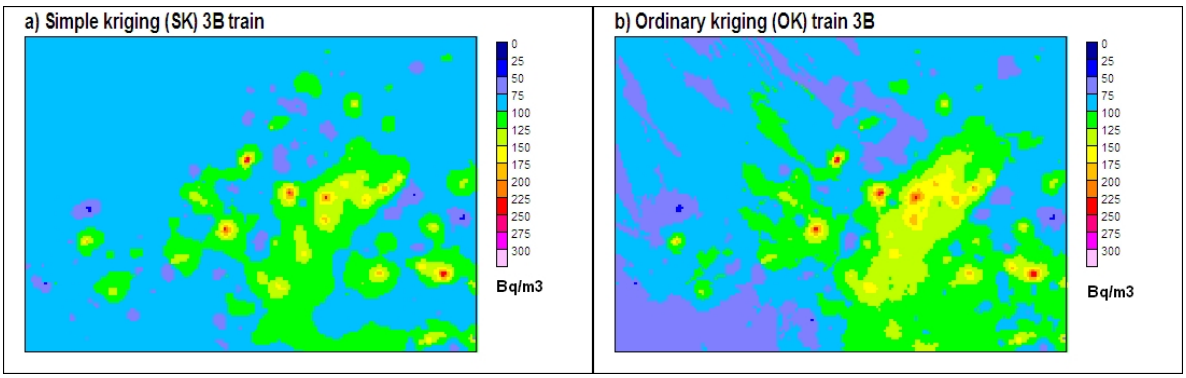



Figure 4.26: *Maps for the training set 3B using a) Simple kriging and b) Ordinary kriging*

There are visual differences among maps that are also expressed by their statistics. In Table 4.2 a summary of the range of values (minimum and maximum), the mean and the variance for the maps is presented.

The map statistics indicate that significant smoothing occurs with all methods. The maximum values cannot be reproduced and the variance is much lower, which is normal for regression methods. The GRNN with reciprocal kernel methods have a particular elevated smoothing effect. It should be noticed that the northwest region is not covered with samples and that the methods provide distinct results. Depending on the interpolation results in this area, the map variance shows large differences between methods. If we speak strictly about

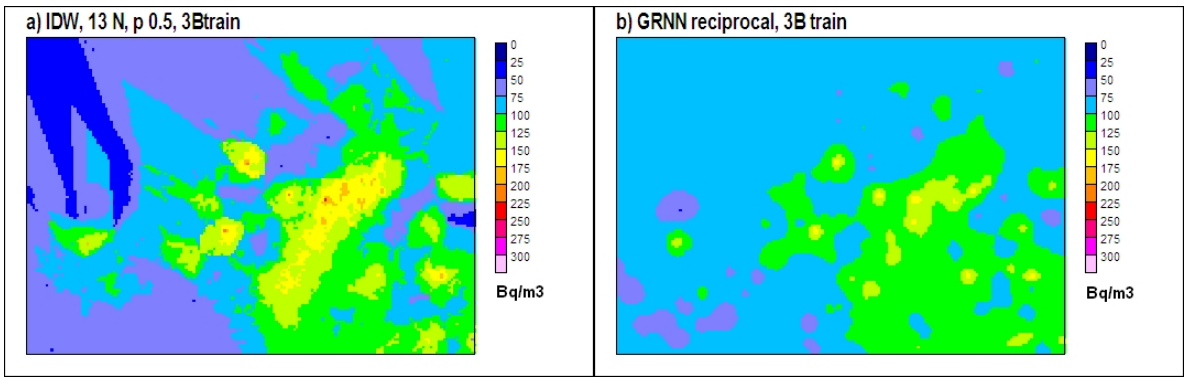



Figure 4.27: *Maps for the training set 3B using a) Inverse Distance Weighting and b) GRNN with reciprocal kernel*

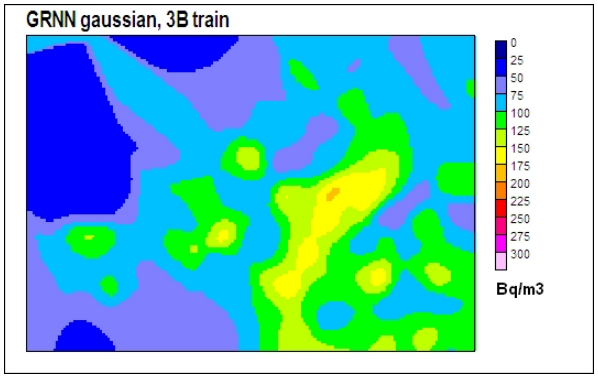



Figure 4.28: *Map for the training set 3B using GRNN with Gaussian kernel*

Table 4.2: *Validation errors and statistical parameters for the training set3B's maps (in Bq/m3)*

| Method | valid. error | range values | mean | variance |
|---|---|---|---|---|
| **train data** | | 7 - 803 | 96 | 4377 |
| **NN all 3B** | | 7 - 803 | 84 | 3565 |
| **KNNR 13K** | 3126 | 35 - 187 | 89 | 787 |
| **SK 32K** | 3126 | 32 - 261 | 97 | 208 |
| **OK 50K** | 3209 | 32 - 262 | 93 | 488 |
| **IDW 13K p0.5** | 3104 | 32 - 228 | 88 | 822 |
| **GRNN recip.** | 3124 | 50 - 170 | 95 | 182 |
| **GRNN gauss.** | 3184 | 34 - 178 | 87 | 861 |

statistics reproduction, the NN method provides the best approximation because it is a copy of the training dataset. However, the validation error for the NN method was the highest (with an MSE of 6908), as seen in section 4.6.2.

The best method in reproducing maximum values was kriging. For variance reproduction, the Gaussian kernel GRNN produced a better-contrasted map. The best compromise between mean and variance reproduction was finally produced with the IDW method. If the validation results are revised once more, we can observe that smoothing has an advantage

for prediction because of the high local variance conditions.

### 4.9.2 Indoor radon mapping using 3B KNNR filtered data

A proposed method to deal with local variance was to use the KNNR CV filtered data. The validation results were comparatively close to those using raw data. In this section, the corresponding maps will be presented for the methods tested, to see if there are visual and statistical mapping differences. In Figures 4.29, 4.30 and 4.31, the maps for KNNR, IDW, SK, OK and GRNN methods using filtered data are presented.

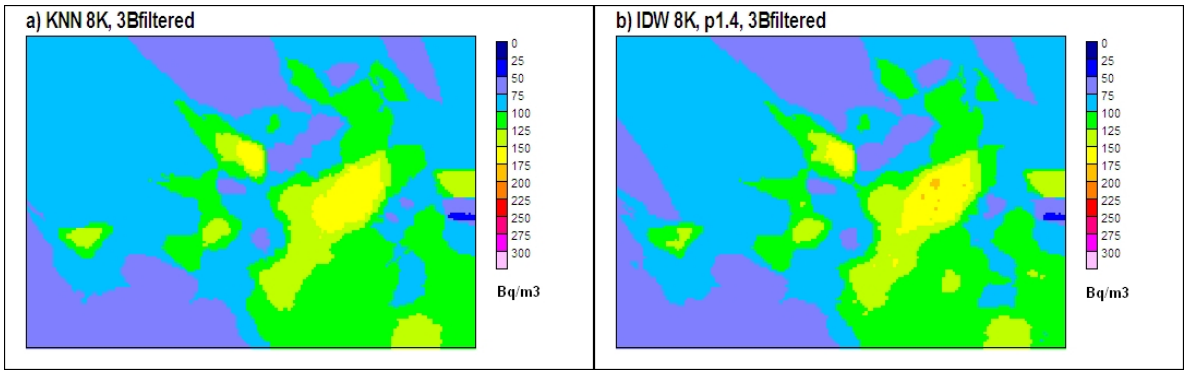



Figure 4.29: *Maps for the filtered set3B using methods a) NN and b) KNNR*

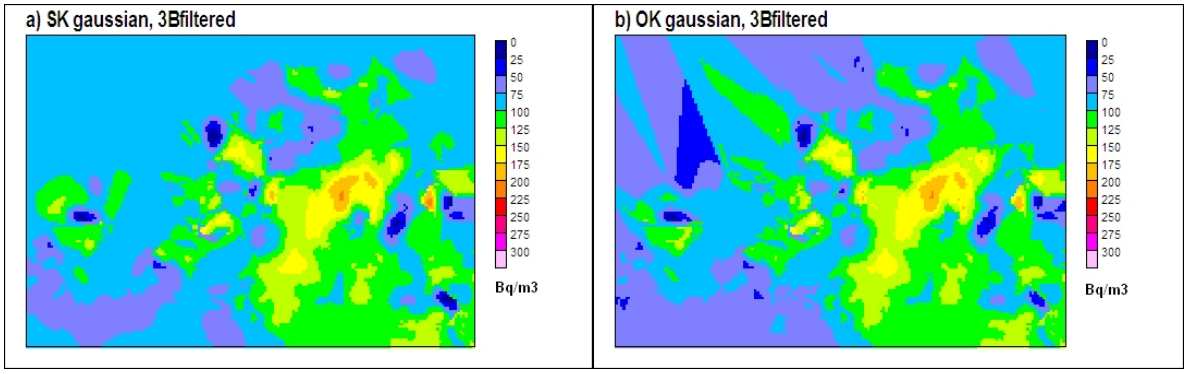



Figure 4.30: *Maps for the filtered set3B using methods a) SK and b) OK*

The mapping statistics can be found in Table 4.3 as well.

The CV filtering was particularly convenient to reach a sound variogram model for simple kriging. Some hot spots (areas with high values) appear clearly defined. Reproduction of the maximum values is better achieved with SK, and the maximum variance was obtained with OK. In general, the lowest validation error corresponded to the SK method. This can be particular to the used training set but indicates that filtering combined with kriging can produce results as good as IDW or GRNN.

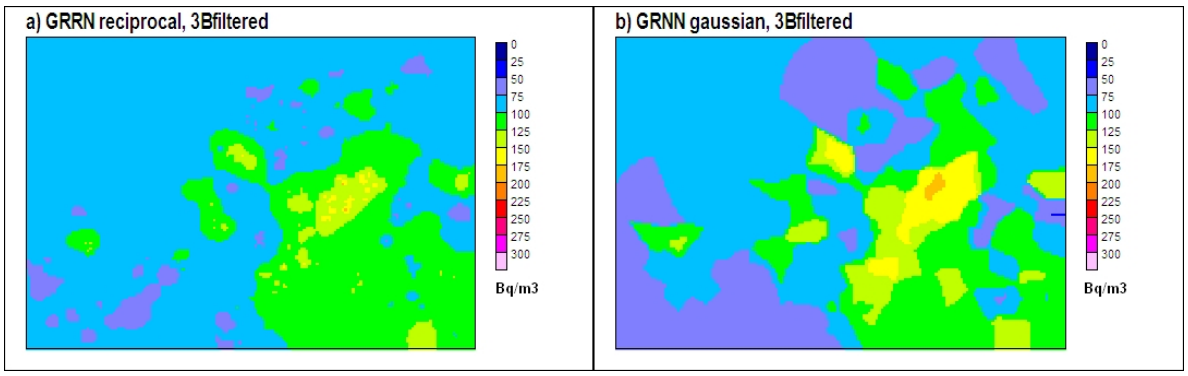



Figure 4.31: *Maps for the filtered set3B using the GRNN method for a) reciprocal kernel and b) gaussian kernel*

Table 4.3: *validation errors and statistical parameters of estimation maps (in Bq/m3) using different methods for filtered set 3B*

| Method | valid. MSE | range values | mean | variance |
|---|---|---|---|---|
| **CVF train data** | | 47 - 186 | 96 | 705 |
| **KNNR 8K** | 3208 | 49 - 173 | 93 | 529 |
| **IDW 8K p1.4** | 3149 | 48 - 183 | 93 | 544 |
| **SK 8K** | 3085 | 2 - 213 | 98 | 531 |
| **OK 8K** | 3085 | 25 - 202 | 91 | 759 |
| **GRNN recip.** | 3118 | 51 - 179 | 96 | 190 |
| **GRNN gauss.** | 3144 | 49 - 184 | 94 | 576 |

## 4.10 Mapping inter-comparison with other methods for set3B

In order to extend the comparative analysis of regression, other methods implemented by commercial software, like ArcGIS and Surfer, have been applied. From the list of methods, asides from the ones already used, nearest neighbor, natural neighbor, triangulation, Shepard's method, polynomial regression, local polynomial and spline methods are also widespread.

The Nearest Neighbor (NN) method is the simplest existing interpolation method and is equivalent to a KNNR using just one neighbor. It can be improved by defining an ellipse search or other search methods. The Nearest Neighbor gridding method assigns the value of the nearest point to each grid node using a defined anisotropy. The Natural Neighbor or polygonal method uses a weighted average of the neighboring observations, where the weights are proportional to the area of the Thiessen polygons formed with neighboring locations. The triangulation method is an implementation of linear interpolation considering three neighboring points. Together with the NN and the polygonal method these are fast methods; however, they generate coarse results and are more convenient for evenly spaced data.

The polynomial method is a generalization of the linear interpolation, where inter-

polants are replaced with polynomials of higher degree. Since this method suffers from high complexity and artifacts at borders, it was enhanced with spline methods. Spline interpolation uses low-degree local polynomials in each interval of a linear function. While global polynomials fit a polynomial to the entire surface, local polynomial interpolation fits many polynomials, each within specified overlapping neighborhoods. This is defined by the search neighborhood's hyper-parameters. Thus, local polynomial interpolation produces surfaces that account for more local variation. In fact, lower validation errors were obtained with local polynomials than with global polynomials.

None of the mentioned methods provided better results than kriging and GRNN methods for set3B. A great limitation of methods based on the fitting of a geometric surface like polynomials and their derivates, is the high spatial clustering of indoor radon. Simple methods, such as IDW, can give comparatively better results. The validation MSE was reduced until 3160, using local polynomials without distance weighting, and without considering anisotropy. This method was mainly affected by a border effect. A map of indoor radon for set3B using local polynomials is presented in Figure 4.32.

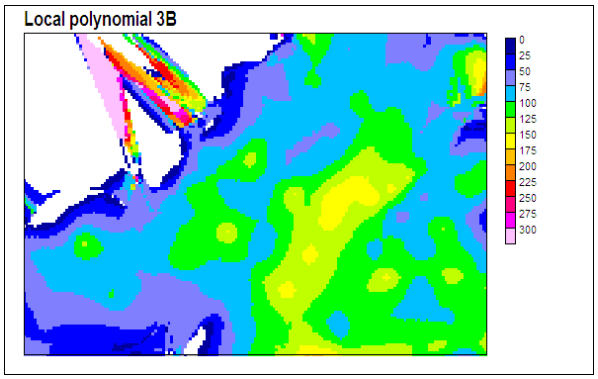



Figure 4.32: *Map for the set3B using the local polynomials method*

## 4.11 Mapping inter-comparison for the set 3A

As remarked earlier, set3A presents a different statistical and spatial distribution of samples in comparison to set3B. With higher a priori variance and very high local variability, the variography modeling for this data is arduous. A validation procedure was launched using the KNNR, IDW, OK, SK and GRNN methods with raw and filtered data.

### 4.11.1 KNN and IDW methods for the set3A

In Figure 4.33a the optimal KNNR for the training and validation procedures were obtained for the indoor radon toy3a raw dataset. In Figure 4.33b, the same tuning with CV is performed but for the filtered data using CVMF.

The optimal number of K neighbors after training was 15, while the lowest CV error for validation data was obtained with 27 neighbors. This is indicating that perhaps data splitting

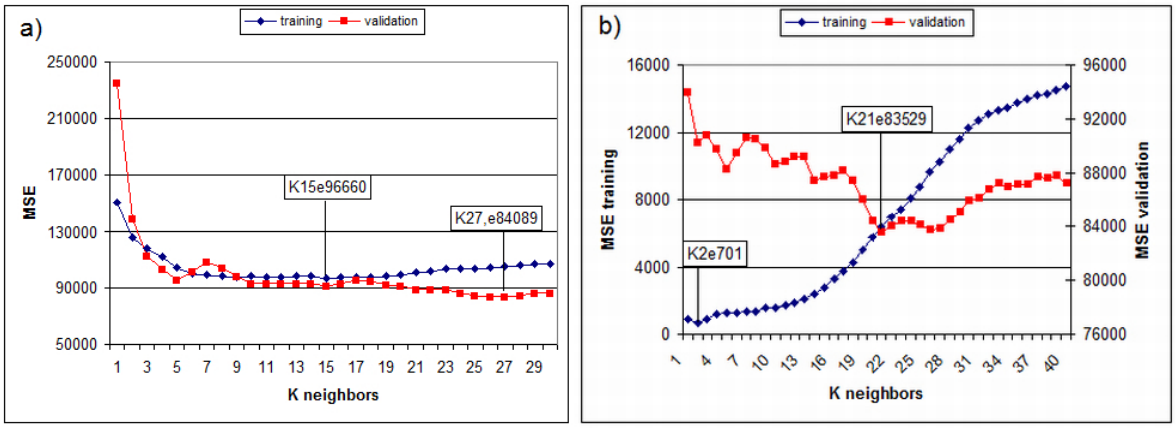



Figure 4.33: *KNNR training and validation optimal curves for a)raw data b) CVMF filtered data*

has not produced a representative subset. In addition, for KNNR CV filtered data, there is a difference in the optimum for training and validation. In fact, the training datasets, after filtering, are quite different from the validation sets. Because local variability was reduced, the optimal training K was just 2, while for validation was 21. Using 15 K, the KNNR validation error was 91,207. For filtered data the error was lower (MSE = 90,172).

As seen before, the parameters obtained with KNNR are also useful for the IDW method. Considering already 15 neighbors as the optimal, a weighting power of $p = 1$, produced the lowest training error. With these parameters the validation error was 106,190 for the raw data. For the filtered data, the optimal training power was 2.1 and the validation error was 92,266.

### 4.11.2 Kriging methods for the set3A

The raw variogram for the training set is shown in Figure 4.34a. The best fitting was obtained with a Gaussian variogram model defined as nug54000+gauss70000R550m. The Gaussian model had a better fitting than the spherical one. Additionally, a hole model was tested because of the sinusoidal shape of the variogram, but the validation results indicated that it was not adequate. Another feature to consider during the modeling step, to improve results, is that the model should not exceed the a priori variance. Even if the fitting appears to be good considering a certain trend, it is better to assume it is stationary.

For the CVMF filtered data the best-fitted variogram is a type of Gaussian with no nugget, defined as nug0+gauss35150R1400m. The variogram for filtered data is presented in Figure 4.34b.

As seen in Figure 4.34b, the variogram after KNNR CV filter had better features. Local variability is reduced but still gives a variogram with discontinuities at the middle range. The training and optimization curves for SK and OK are presented in Figures 4.35a and 4.35b respectively. The same optimization curves, but for the filtered data, are shown in Figures 4.36a and 4.36b.

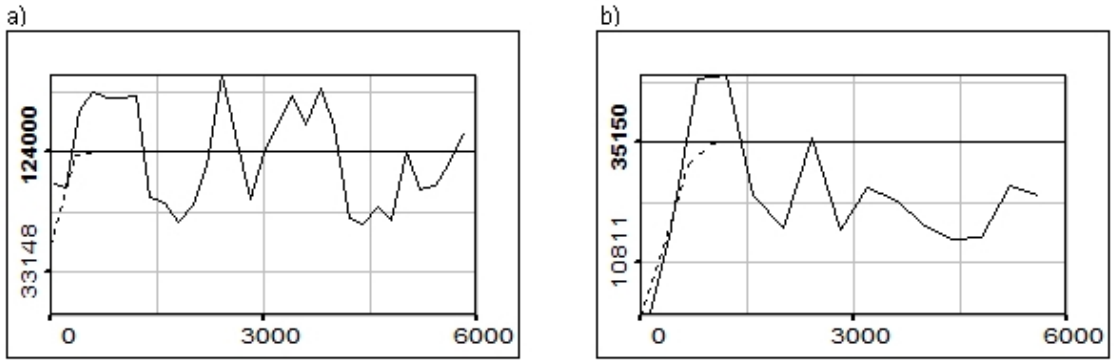



Figure 4.34: *Variogram models for the 3A training set for a) raw data and b) CV filtered data*

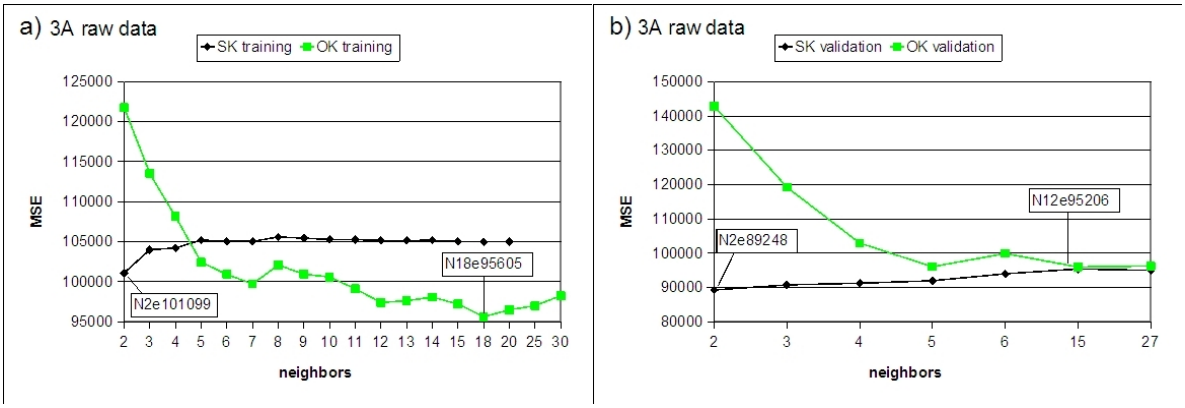



Figure 4.35: *Set3A raw data optimization curves for SK and OK for a) training and b) validation*

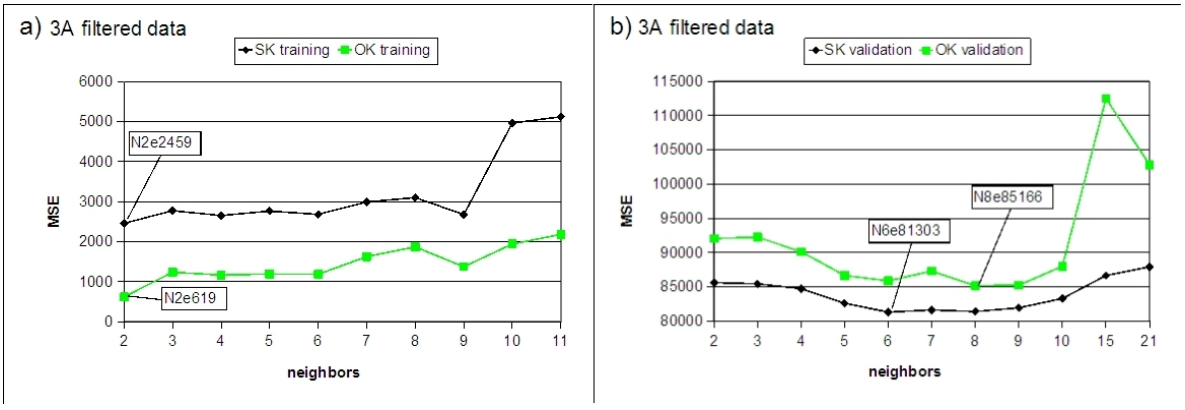



Figure 4.36: *Set3A KNNR filtered data optimization curves for SK and OK for a) training and b) validation*

The optimal number of neighbors $n$ for SK after training is 2 (Figure 4.35a). The validation error using this parameter is 89248. In Figure 4.35b, the validation error curve shows that this is the lowest possible validation error that can be obtained. For OK, the optimal

training number of neighbors is 18 and the validation error using this parameter was 96542. In shape, training and validation curves are similar for both methods. A very noticeable difference is that the optimal number of neighbors for SK is much lower.

It appears that the high local variability is preventing the assumption of local stationarity and forcing the closure of a neighborhood search for SK. OK can better manage this local variability, but this has a consequence on the validation result, which is a global indicator. In fact, the proposed variogram model has a very short range in comparison to the domain (just 550 m.). This corresponds to a close neighborhood.

For the KNNR CV filtered data, the schema changes because the optimal training $n$ is the same for SK and OK (Figure 4.36a). With this parameter ($n$ =2), SK which has the lowest validation error in comparison to OK (85610 for SK against 92064 for OK). The results of kriging using filtered data are more comparable to KNNR and IDW. The number of optimal neighbors in the case of raw data was different for either SK or OK. SK performed better with just 2 neighbors while OK required 18 neighbors to reduce the validation error. In summary, SK performs more locally than OK and combined with CV filtering, seems to be a valid method for indoor radon interpolation.

### 4.11.3 GRNN method for the set3A

In Figures 4.37 and 4.38, optimization curves for training and validation using the GRNN method for raw data and filtered data are shown.

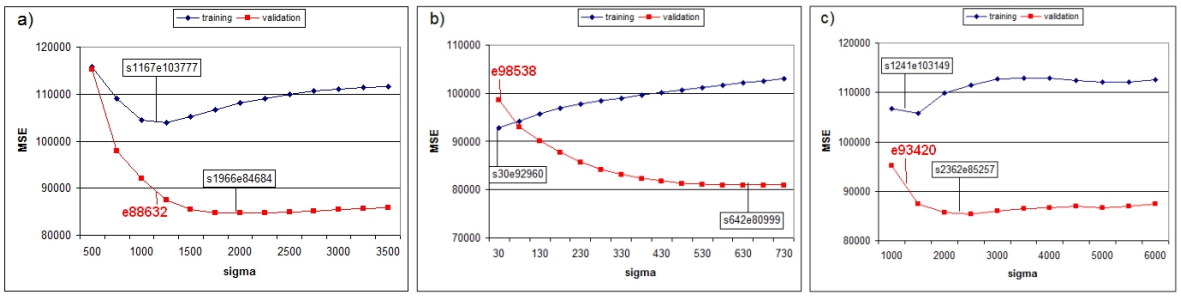



Figure 4.37: *GRNN training and validation curves for set3a with different kernels: a) Gaussian b) Reciprocal and c) Epanechnikov*

The lowest MSE for raw data (88632) was achieved with a gaussian kernel. For filtered data, the Epanechnikov kernel performed better producing the lowest MSE (79411) among all methods and parameters tested with the toy3a set.

It is important to notice that for set3A the optimal parameters obtained with training procedures do not always coincide with the optimal parameters for validation. This is especially true for the raw data, while with filtered data parameters are more akin. For the GRNN method with an Epanechnikov kernel, the training and validation minimum errors are very close when using filtered data.

For kriging methods the parameters optimization can be tedious because of the complexity of kriging modeling. The GRNN method has the advantage of being more automatic

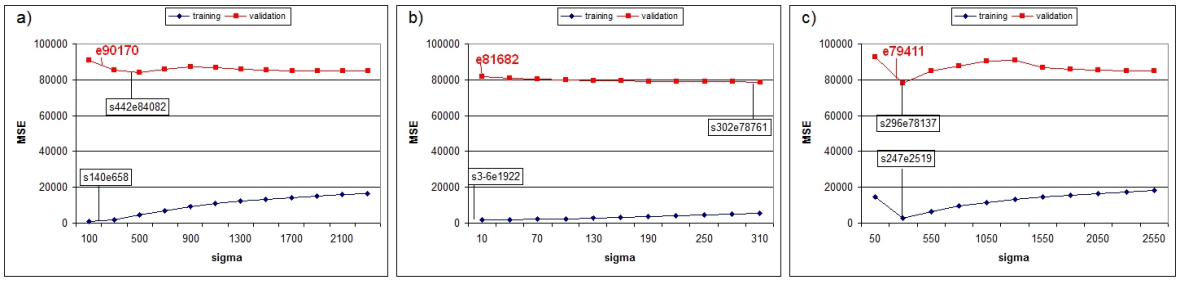



Figure 4.38: *GRNN training and validation curves for filtered data of set3a with different kernels: a) Gaussian b) Reciprocal and c) Epanechnikov*

and also facilitates parameter testing. Nevertheless, the optimization of sigma values is sensitive to initial values.

### 4.11.4 Indoor radon mapping for set3A using regression methods

A series of six maps were produced using different interpolation methods for set3A, for raw and filtered data. In Figures 4.39, 4.40 and 4.41, are the maps for the interpolation methods that produced the lowest validation errors.

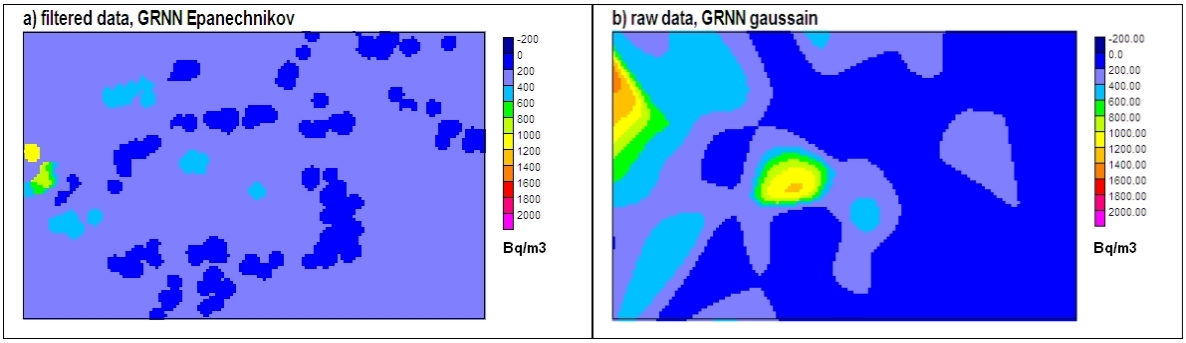



Figure 4.39: *Maps for the set3A using GRNN methods with a) Epanechnikov kernel for filtered data, and b) Gaussian kernel for raw data*

The mapping statistics for set3A are shown in Table 4.4,

Table 4.4: *Validation error and statistical parameters for set3A maps (in Bq/m3)*

| Method | valid. MSE | range values | mean | variance |
|---|---|---|---|---|
| **train data** | | 7 - 2501 | 238 | 124014 |
| **GRNNFIL Epa.** | 79411 | 37 - 1006 | 222 | 5220 |
| **GRNN Gauss.** | 88632 | 56 - 1462 | 252 | 43372 |
| **SKFIL 2N** | 85610 | 15 - 1645 | 222 | 2938 |
| **OKFIL 2N** | 92064 | -12 - 1935 | 221 | 31947 |
| **KNNFIL 2K** | 90172 | 38 - 1006 | 221 | 31664 |
| **KNNR 15K** | 91207 | 35 - 1027 | 230 | 28723 |

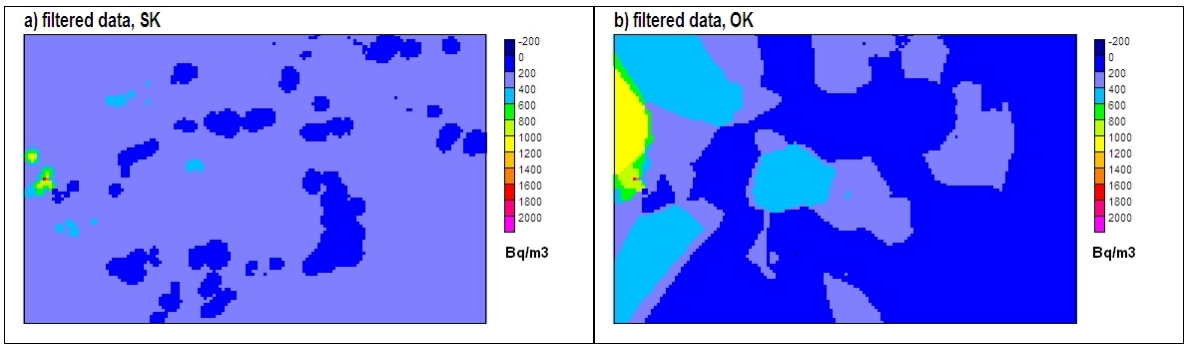



Figure 4.40: *Maps for the filtered set3A using kriging methods with a) Simple kriging with 2N, and b) Ordinary kriging with 2N*

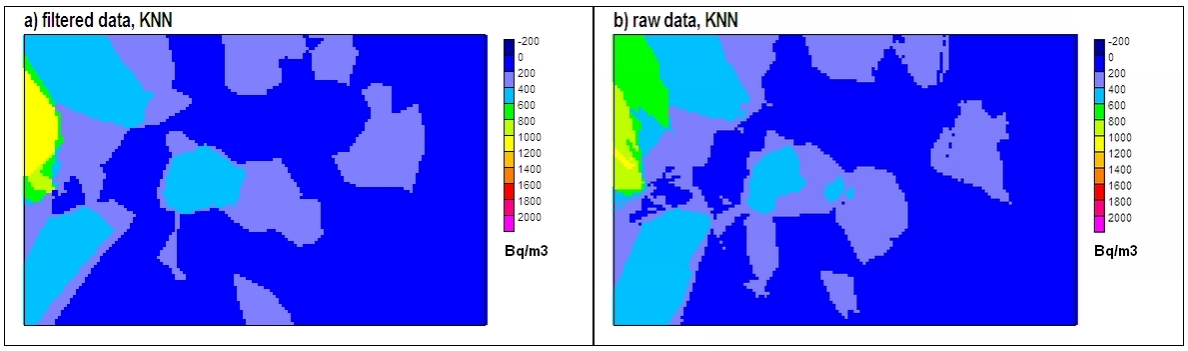



Figure 4.41: *Maps for the set3A using KNNR method for a) filtered data with 2N and b) raw data with 15N*

In the range of estimation values in Table 4.3, the OK method has some negatives values. This negative value is at a single location that is visible in the corresponding map as a white point (Figure 4.40b). As mentioned in the methodological section, this is due to an error calculation that arises because of the constrained condition of OK for the sum of weights. Possible solutions to this are either to force negative weights to be zero or to adjust the variogram model.

## 4.12 Method robustness and data consistency of set3B

After the many training and validation routines performed on set3B some question have arisen. The optimization curves have shown that training data appears to be more or less consistent regarding the validation set.

### 4.12.1 The easy and difficult tasks

An important feature to be noted is the influence of the bias sign of the validation and training sets in the validation error. When the mean and the variance from the training set are higher than those from the larger set (positive bias), the validation biases will be logically

negative. This occurs because larger values drop into the training set during splitting; then, the training mean is higher than the validation mean. In such cases, the validation error is always lower.

Furthermore, in such cases, the estimation task has been simplified because all regression estimation methods apply smoothing or averaging of values. If the training set has a more 'realistic' large data distribution, the validation values are included and modeled by such distribution. This situation can be labeled as an 'easy task' estimation. On the contrary, a validation set with large values and high variance will produce larger validation errors and will represent a difficult estimation task.

In either cases, with the easy or the difficult task, the validation error is not realistic because training and validation sets are different. In addition, with respect to the global unknown distribution, we suppose that the sample distribution is a fair approximation of it. With the same logic, the training set must preserve the properties of the larger set, and bias, with respect to this should be avoided.

For instance, the random splitting of set3 used so far, has produced a validation set with a mean of 92.6, which produces a negative bias of 3, with respect of set3. Meanwhile, the training set has a positive bias of 0.9. The variance bias for validation is -742, and for the training set it is +214. So, it can be said that this training-validation splitting of set3B was an easy task interpolation.

### 4.12.2 Statistical consistency and bias of set3B after data splitting

For the purpose of the jackknife procedure, the previous training and validation sets were merged into a single dataset with 1147 samples. First, a comparison between a purely random or relaxed split against a random optimized split was performed. A series of 1000 splits were performed with both splitting methods to observe the differences between validation and training sets.

After splitting, the bias of the mean and variance were compared with the correspondent large sample set values, which are 95.6 Bq/m3 as mean and 4163 as variance. A large number of split repetitions were made to calculate the mean bias of the mean and variance for validation and training set, with a splitting proportion of 20% and 80 %. It was observed that the bias is the same for the simple random method and the MW random selection. The mean bias of the mean for the validation set fluctuated around 3, while the mean variance bias fluctuated around 900. For the training sets, the mean and variance fluctuated around 0.75 and 245 respectively.

When doing 100 repetitions, the fluctuations are already reduced and they are comparable to doing 1000 or 10,000 repetitions. Therefore, there is not much advantage in doing a very large number of repetitions to calculate bias statistics because fluctuations are always present for random selections. What is good to know, is that the selection is comparable for the simple and MW random. The mean bias thus obtained is used as a reference to indicate if the splitting is more or less biased.

### 4.12.3 KNNR and GRNN for statistical consistent splits

It was observed that if validation and training sets are biased, the validation errors are not realistic. They can be either too good or too bad just because of the data splitting. Of course, in practice, the splitting is done just once and the methods are compared using the same splitting regardless of whether it is an easy or difficult task. This is a logical approach if the idea is just to select the best method but results will not be realistic. To improve test methods, the splitting can be optimized by allowing only training and validation sets to have low biases (unbiasedness).

Unbiasedness must be focused on training sets because they must be representative of the large set. Splits for set3B were repeated until the training variance bias was below 1. This value is much lower than the average bias measured with the jackknife procedure (bias=3). 20 optimized splits were done to observe the behavior of interpolation parameters. For the interpolation part, the KNNR and GRNN methods (with Gaussian kernel) were launched. The interpolation parameters, such as the optimal $K$ for KNNR and the optimal sigma for GRNN were automatically obtained by training cross-validation.

The validation MSE for the random splitting has more fluctuation between repetitions than for the unbiased splitting. Nevertheless, they have the same average errors for both methods combined. Out of the combined 40 repetitions for both splitting methods, the average MSE for KNNR was 3915 and 3863 for GRNN. There are split cases where KNNR have outperformed GRNN (30 % of cases) but GRNN appear to give the lower validation errors. Both splitting methods, with and without unbiasedness resulted, on average, in the same validation errors with differences observed only between interpolation methods.

### 4.12.4 Spatial consistency of the set 3B after data splitting

The random selection within MW will not reduce statistical bias but the question is whether it gives some spatial consistency that is useful for spatial predictions. To check this, a simple estimation method as KNNR, was launched for 100 data splits for random selection with and without MW. On average, no differences of the KNNR validation error between the two splitting methods were noted.

The next approximation proposed is to compare KNNR and GRNN with a greater number of repetitions by using total random and MW random splits. The idea is to test which method can better perform under different conditions of spatial selection. In other words, one wants to find out which method is more robust to fluctuations of spatial distribution of data.

Considering the 100 repetitions, there are no significant differences between the mean of validation MSE for splitting methods. With respect to methods, GRNN has, on average, a lower validation error. The mean MSE validation error of the combined 200 repetitions is 3756 for KNNR and 3698 for GRNN.

Within the repetitions, cases of 'easy tasks' were observed where KNNR outperformed GRNN. Also, for the cases where splitting was done with bias minimization, other interpo-

lation methods can give comparatively lower validation errors than GRNN. From the 100 splitting repetitions and the correspondent interpolation trials it was observed that GRNN has lower validation errors 79 times for the random splitting and 67 times for MW splitting. This can indicate that GRNN is more robust regarding data splitting at random and in general better than KNNR.

It was observed that spatial consistency of indoor radon subsets was difficult to attain throughout MW splitting because of the high spatial clustering.

### 4.12.5 Robustness of estimation methods

The MW splitting method and the unbiasedness optimization appear to have no major effect on the validation error for KNNR and GRNN. There is no evidence of improvement by using statistical or spatial optimization of splitting. The validation error varies according to the 'easy' and 'hard' tasks. Therefore, the statistical unbiased splitting has less fluctuation in error values. Even if on average GRNN is slightly better than KNNR, the fluctuation of values of the validation error prevents to conclude that the method performs better regarding data consistency. The fact that GRNN perform better than GRNN is somehow predictable because it is more robust regarding noise. To complete the robustness analysis, a comparison with other methods was made. For GRNN just the Gaussian kernel was tested because of simplicity but the reciprocal kernel seems to be better adapted to indoor radon data. Finally, IDW and kriging methods as well as data filtering were also compared. 10 splitting sets were selected for the methods comparison, 5 for the case where GRNN was better than KNNR and 5 others where KNNR was the best. When KNNR performs better, IDW does also. That resulted in cases where either GRNN other KNNR was challenged against the other methods. In Table 4.5, is a comparison of statistical and interpolation spatial parameters for 10 split sets. The sets result from all type of splitting methods.

Table 4.5: *validation errors for splits of set3B using regression methods*

| | **IDW challenge** | | | | | **GRNN challenge** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method/par | ran4 | ranun5 | MW1 | MW5 | MWOP4 | ran5 | ranun2 | ranun4 | MW4 | MWOP5 |
| **KNNvalE** | 4825 | 3867 | 5280 | 3969 | 3494 | 5475 | 3914 | 3981 | 4408 | 3805 |
| **GRNNgvalE** | 4834 | 3924 | 5299 | 4002 | 3663 | 5258 | 3775 | 3848 | 4265 | 3725 |
| **IDW05valE** | 4788 | 3865 | 5157 | 3893 | 3450 | 5380 | 3808 | 3945 | 4362 | 3703 |
| **GRNNrvalE** | 4832 | 3825 | 5305 | 3943 | 3626 | 5199 | 3737 | 3905 | 4166 | 3737 |
| **SK15valE** | 4945 | 4073 | 5301 | 3995 | 3764 | 5181 | 3871 | 4158 | 4482 | 4372 |
| **OK15valE** | 4882 | 4155 | 5151 | 3938 | 3669 | 5215 | 3954 | 4061 | 4503 | 4335 |
| **OK30valE** | 4955 | 4044 | 5194 | 3980 | 3709 | 5146 | 3815 | 4006 | 4503 | 4314 |
| **SKF15valE** | 4932 | 3959 | 5312 | 4085 | 3561 | 5548 | 3815 | 4169 | 4574 | 3857 |
| **OKF15valE** | 4925 | 3971 | 5309 | 4096 | 3567 | 5579 | 3820 | 4161 | 4576 | 3861 |
| **BEST** | **IDW** | **GRNNr** | **OK15** | **IDW** | **IDW** | **OK30** | **GRNNr** | **GRNNg** | **GRNNr** | **IDW** |

The validation errors are very close for this series of examples. In general, when IDW or GRNN performed better, it was difficult to challenge them with other methods but the results were very close. Out from the 10 repetitions, 4 have lower errors using IDW, 4 using GRNN and 2 with ordinary kriging. The use of the KNNR filtering method has not given any advantage in any of these cases. The GRNN appears to have a very small advantage over the KNNR method and performs equally as well as IDW.

When looking at parameters obtained by cross-validation for the whole 3B dataset it was observed that both IDW and GRNN have closer results (Table 4.5). The jackknife splitting provided a measure of statistical bias of data. Is there any association between bias and interpolation performance? Table 4.6 presents the corresponding mean and variance (M and V) bias for the validation and the training sets (MV, VV, MT and VT) for the same splits tested in Table 4.5. With these series of splitting is also possible to show the influence of the bias on the validation results.

Table 4.6: *Mean and variance bias for the validation and training subsets for different types of splits*

| | **IDW challenge** | | | | | **GRNN challenge** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method/par | ran4 | ranun5 | MW1 | MW5 | MWOP4 | ran5 | ranun2 | ranun4 | MW4 | MWOP5 |
| MVbias | 1.60 | -0.28 | 5.23 | 3.44 | -0.10 | -4.76 | 0.25 | -0.27 | -2.04 | 0.15 |
| VVbias | 1095.0 | -5.0 | 1647.0 | 203.1 | 8.8 | 1254.2 | 0.7 | 9.4 | 405.0 | 8.3 |
| MTbias | -0.43 | 0.06 | -1.38 | -0.91 | 0.02 | 1.24 | -0.08 | 0.06 | 0.53 | -0.05 |
| VTbias | -282.1 | 5.9 | -434.4 | -52.4 | 2.3 | -330.3 | 4.4 | 2.1 | -102.5 | 2.4 |
| **mean valE** | 4459 | 3746 | 4825 | 3855 | 3442 | 4749 | 3677 | 3795 | 4073 | 3819 |
| **BEST** | **IDW** | **GRNNr** | **OK15** | **IDW** | **IDW** | **OK30** | **GRNNr** | **GRNNg** | **GRNNr** | **IDW** |

In Table 4.6, the influence of the negative bias for the training set is shown. It is clear that to a higher validation variance will correspond a higher validation error. This in turn corresponds to a lower training variance and vice versa. Of the group of splitting tested, it is remarkable that OK performed better under the conditions called difficult task (when the training variance has a negative bias). Meanwhile, GRNN appears to perform better for the called 'easy tasks' or the more unbiased splits.

### 4.12.6 Jackknife procedure for set3A

Set3B before analyzed has a certain global stationarity and lower variability. Set 3A is somehow the opposite because it has a global trend and very high values producing an overall high variance. It is therefore interesting to analyze the data consistency and the methods robustness regarding this kind of data.

On a first run, 100 random splits were done to obtain a training and a validation dataset and to analyze statistical and spatial parameters. The whole set3A, with the first training and validation sets merged, has a mean of 234 and a variance of 117702. The bias of the training mean from the 100 splits was around 3.2 and 1300 for the variance. For the validation set (which is just 20% of data), the mean bias was 12 and the variance bias was 4900.

For set3A, the spatial analysis gave a different feature than set3B. With the resampling done at random, the KNNR method performed 62 times better than GRNN out of the 100 splits. In this case, KNNR appears to be more robust than GRNN. The data is less consistent regarding the larger differences that can appear between training and validation sets.

If statistical unbiased split forces consistency, then a success rate of 50 to 50 was obtained between KNNR and GRNNN. From previous results, it can be deduced that IDW will improve the results obtained with KNNR. Therefore, deterministic methods such as KNNR and IDW show to be more robust to handle trended and biased data than GRNN. The performance is similar between methods when the training and validation bias are reduced.

## 4.13 Conclusions about regression methods

In chapter 4, several spatial estimation methods were employed to make a comparative validation over sets 3A and 3B. First the deterministic methods were revised and then, those involving statistical modeling. Both methods have indistinctively deterministic or probabilistic interpretations.

A method to deal with the high local variance was proposed based on the use of KNNR cross-validation. The called KNNR CV mean filter (KNNR CVMF) was first used to improve data visualization by means of smoothing using the optimal K neighbors' parameter. Because the mean and the spatial distribution are preserved after CVMF, a trial was made to do variography with such data transformation. It resulted in a variogram with almost zero nugget, large range and a stationary model fitted on top. A drawback of this filtering is the high reduction of variance, which deviates from original sample statistics. Nevertheless, the mean was preserved and modeling parameters were similar for training and validation curves. The proximity of spatial parameters after the KNNR filtering, such as the optimal K, was interpreted as a spatial consistency of data resulting from the filter.

The next method revised was the inverse distance weighting. For this method the selection of the power parameter also requires CV. A proposed way to speed-up IDW tuning was to use the optimal K for neighbors' parameter that was also optimal for IDW. These parameter and hyper-parameter tuning procedures went in the direction of automatic mapping.

The application of probabilistic methods originates from the understanding that earth science processes cannot always follow a well-defined physical model and therefore uncertainty should be included in the model. Methods using spatial statistics are a first approach to this uncertainty modeling by considering random variables and second order statistical parameters. The basic theory was revised, including the important stationarity assumptions, which relate closely with the data spatial selection proposed to improve modeling.

Another important regression method revised was the general regression neural network (GRNN). It was employed not only as an interpolator but also as a spatial characterization and spatial domain definition tool. First evidence of spatial continuity was obtained by looking at the convergence of CV curves to an optimum. This was particularly simple for KNNR. For GRNN, an optimum sigma parameter gives an indication about spatial structure, data completeness and clustering. GRNN optimization is sensitive to initial sigma parameters and some automatic settings were tried in this regard.

GRNN was used for a MW multiscale analysis to test the method parameters in relation to the indoor radon behavior. GRNN benefits from large datasets in order to build a distribution function. Having more samples in a set tend to give a unique solution. Heavy clustering is sometimes expressed as over-fitting or by local minima in the GRNN optimization curves. GRNN is robust in the sense that it always provides a solution regardless of noise. A drawback is that the solution can tend to simply be a generalized estimator or the mean of the training points. The sigma value is therefore an indicator of the level of fitting between model and data. It also indicates the level of smoothing that will be applied by the interpolator.

To summarize what was done with regression methods for set3B, the lowest validation error for raw data set3B was obtained with the IDW method (3104) followed by GRNN (3124), KNNR (3126) and SK (3126). Using filtered data as estimators, the validation error was better using SK (3085), followed by GRNN (3118), IDW (3149) and KNNR (3208). The validation errors for the different methods are so close, that they cannot indicate strongly which can perform better.

Mapping visualization and statistics is another criteria used to evaluate regression methods. For raw data, the best method for reproducing maximum values was kriging. For the variance reproduction, the Gaussian kernel GRNN gave a better-contrasted map. The best compromise between mean and variance reproduction was finally provided by the IDW method.

The CV filtering was especially convenient to reach a sound model for simple kriging and to produce maps. Some hot zones appeared clearly defined, while over-smoothing with other methods prevented zones of high concentration from being identified. Reproduction of the maximum values was better done with SK and raw data while the maximum variance was obtained with OK combined with the CVMF. Even though OK is a preferred kriging method, SK also performs well provided stationary conditions.

The split of set3B between training and validation subsets was done at random for the initial tests. The training set resulted in a higher mean and variance than the validation set. This was typified as an 'easy task' prediction for which the comparative performance of methods was the same. Therefore, it was important to conduct further interpolation analysis after spatial and statistical unbiased splitting, the purpose being to test the robustness of methods under different conditions of data consistency. Unbiased splitting and MW splitting techniques were analyzed for set3B.

After successive splitting and interpolation with KNNR, validation errors presented fluctuations, but in general no differences were noted between the two splitting methods. Unbiased splitting should be preferred because validation will result more realistic. When comparing KNNR with Gaussian GRNN for a large number of splits, it was seen that GRNN results, on average, in lower validation errors compared with KNNR for set3B. This is due to the generalization capabilities of GRNN with Gaussian kernel and the lower variance of set3B. A disadvantage of reciprocal and Epanechnikov kernels is their tendency towards over-fitting, which results in local artifacts within maps. The CV filter contributed to data generalization as well.

In order to enlarge the comparison, all methods were contrasted to a series of splitting sets where either GRNN or IDW outperformed. For these cases, the performance of GRNN, IDW and kriging are comparatively the same. GRNN and IDW have the advantage of operating easily under automation than kriging methods. Optimization of parameters for these two methods can be done easier than for kriging. For set3B, it was found that the validation errors were more sensitive to the data used to validate that to the interpolation method used. For the case of set 3A, deterministic methods, such as KNNR and IDW, have proven to be more robust to handle trended and biased data than GRNN.

# Chapter 5

# Probabilistic mapping methods for indoor radon

The application of regression methods over different datasets in chapter 4 has shown that estimation is a smoothing process. While they aim to be unbiased, results tend to approach the mean, which results in a reduction of the original data variance. This indicates that the information used to make estimates is not entirely reproduced. In some cases, regression methods can perform well locally, but the model cannot be generalized over a large area. In such cases, the mapping results in a spotty visualization. In other cases, the methods tend to over-generalize, producing a coarse visualization.

The two challenges of the estimations are then, to reproduce not only the statistical but also the spatial data distribution. In the case of statistical distribution, the major difficulty is that the true underlying population distribution is unknown. This is more a matter of good sampling as we saw in the previous chapters. It is desirable that the available samples, in deed, represent (are consistent) with the underlying population but this cannot be taken for granted. Data consistency can be checked by the change of parameters between campaigns (as done in chapter 3 for the cantons of Neuchatel and Jura) and by successive splitting to observe subsets bias.

If the sample distribution is deemed to represent the population distribution, simulation processes can reproduce it. Stochastic simulations provide the most complete information because not a single but a set of realizations are given for every point. One such method is the sequential Gaussian simulation (SGS).

Within this method, the problem of modeling also has a solution through data transformation into normalized scores or nscores. By doing simulations, many realizations can be obtained for the same point, from which the probability to exceed a certain threshold can be derived. As the data distribution is reproduced, the goal of the method is to obtain not the single estimation but the complete probabilistic distribution for each location. The indicator kriging (IK) method also works in the probabilistic line of thinking. In this case, data transformation is based on thresholds, so that the input data represent already the probability to be below or above a certain value. This probability is calculated for other locations with the use of kriging methods.

Other estimation methods producing probabilistic results are classifiers, as KNN, PNN and SVM (37), (42). The first step to apply such methods consists of a data transformation into classes to allow easier modeling. The task is then reduced to a classification problem having a probabilistic interpretation as a byproduct.

In the present chapter, emphasis will be made in the application of simulation methods to comb with the particular properties of indoor radon. The critical problem of spatial and statistical distribution will be addressed, as well as the parameters and hyper-parameters definition. The feasibility of classification methods will be evaluated and an inter-comparison of all probabilistic methods will be made.

## 5.1 Sequential Gaussian Simulation and Multigaussian Kriging principles

Sequential Gaussian simulation (SGS) is a geostatistical method that aims to obtain exhaustive information using real and simulated data (14), (8), (23), (37). A joint distribution for random variables $Z_i$, is assumed and reproduced with simulations. A joint multi-Gaussian distribution is then chosen as a model because it is easy to reproduce. The process requires only local mean and variance parameters, which can be obtained through kriging methods.

First a joint probability density function $F_{(N)}$ of $N$ random variables (RVs) $Z_i$, conditioned to $n$ samples is defined:

$$F_{(N)}(z_1, \ldots, z_N|(n)) = \text{Prob}\{Z_i \leq z_i, i = 1, \ldots, N|(n)\} \tag{5.1}$$

This function $F_{(N)}$ can be decomposed into several univariate distributions with an increasing level of conditioning:

$\text{Prob}\{Z_i \leq z_i, i = 1, \ldots, N|(n)\} =$
$\text{Prob}\{Z_1 \leq z_1|(n)\} *$
$\text{Prob}\{Z_2 \leq z_2|(n+1)\} *$
$\vdots$
$\text{Prob}\{Z_2 \leq z_2|(n+N-1)\}$

Then a simulated realization of this joint distribution function, meaning a set of the kind $\{z_i, i = 1, ..., N\}$ can be obtained by sequential sampling of the $N$ univariate functions $F(z_i)$ for $i = 1, \ldots, N$ (14) (8). Being $N$ equal to the number of nodes within a simulated grid or image. To verify that the simulated realizations (images) corresponds to the conditioned distribution function, the spatial correlation (the variogram) and the histogram of sample data must be reproduced.

Taking the set of simulated values for a certain node, a local probability density function (*pdf*) is approximated. This *pdf* is then used to define probabilities to exceed cutoff values. An advantage of stochastic simulations is that they reproduce distribution functions from which the probability of being above (or below) a certain value is calculated.

Since samples are limited in number and many factors affect the distribution of the indoor radon pollutant, there is a high uncertainty that must be considered in spatial distribution modeling. Therefore, the probability of occurrence is a convenient and realistic way to indicate the level of pollutant in unsampled areas.

### 5.1.1 Sequential Gaussian simulation procedures

There are many aspects to address before launching the SGS algorithm. First, an exhaustive exploratory data analysis must be done in order to identify major problems related to the reproduction of a statistical distribution. The sample distribution should approach a certain global distribution. Because samples are the main information available regarding the study variable, they are often assumed as the departure point to calculate global parameters. Eventually, it becomes possible to apply some declustering techniques if they enhance statistics or if there is clear proof that preferential sampling has been committed. Next a trend analysis must be performed to see if impediments to fitting stationary models exist.

A requirement for the method is to validate the assumption that the joint distribution of RVs $Z_i$ approaches a multi-Gaussian model. The next important step is variogram modeling; variography is performed using the nscore's transforms. When a multi-Gaussian model is assumed (the case with the SGS method), only positive definite models such as the spherical, the exponential or the Gaussian must be used. To hold the stationary condition, the sum of the nugget and the sill must not exceed the a priori variance of the population. For the SGS method, the theoretical a priori variance corresponds to a standard Gaussian distribution. Then, the variance of data after the transformation to nscores should also approximate to 1. The chosen nscore variogram is then used for the kriging procedure during the simulation.

### 5.1.2 Nscore transformation and bigaussian test

The procedure with SGS (14) starts with a transformation of the data into nscores to ensure univariate Gaussian distribution. Transformation into a centralized and normalized distribution is important for variogram modeling, because it allows structures to be identified. The reverse procedure, i.e. the nscore back-transformation, should guarantee the reproduction of the sampled data. Nscore transform has the advantage of preserving ranking order and first order *pdf* parameters when applying linear interpolations such as kriging.

To use SGS we should ensure that the joint distribution of random variables is multigaussian. In practice, only joint bi-Gaussian distribution can be checked. Theoretical methods check whether the covariance function for nscores $Y$ at lag distances $(h)$ correspond to a standard bivariate normal distribution (14). An empirical test is possible by comparing the ratio $\sqrt{\gamma(h)}/madogram$ with $\sqrt{\pi}$ at all lag distances (37).

Additionally, it is possible to propose a distribution that better approaches the global distribution during back-transform. Other methods propose the use of a direct transform during simulation.

### 5.1.3 Multi-Gaussian kriging MGK

Multi-Gaussian kriging (MGK) can be seen as a simplification of the SGS method. It is based on the nscore transform to a Gaussian distribution so that all one-point RVs are deemed normally distributed. Then, the two-point distribution of any pairs of RVs is also normal and fully determined by the covariance function $C(h)$, and finally, any linear combination (e.g. kriging) of the RV components will result also normal.

The idea is to consider kriging of nscores as a Gaussian distribution and consequently to do a back-transform. It can be seen as a single realization of the multi-Gaussian distribution. It can be used as a rapid-mapping tool for data having a structured variogram for nscores.

### 5.1.4 Simulation process

As mentioned in section 5.1, the idea of SGS is to reproduce the joint *pdf* of RV $Z(x)$ on a grid of $N$ nodes, conditioned (or not) to the $n$ samples. Conditional distribution for RV $Z(x)$, being multi-Gaussian, can be fully characterized by its conditional expectation (the mean) and conditional variance (the variance) of the random function (RF) (Equation 5.1).

Sequential gaussian simulations take samples from local distributions using the simple kriging estimate (Equation 4.16) and the kriging variance (Equation 4.20). Predictions are made using not only sampled conditional data but also previously simulated nodes. If predictions are made sequentially for the $N$ nodes, the level of conditioning will increase to $|(n-1+N)$ (14).

In theory, simple kriging (SK) must be used for the SGS procedure because it is assumed to be globally unbiased. Another theoretical reason is that simulation aims to reproduce the global mean and variance distribution. In SK, the global mean is taken explicitly into account, while the nscore transform provides a defined global mean that approximates 0. Then, if the SK estimator equation is recalled from chapter 4:

$$\hat{Z}_{SK} = \sum_{i=1}^{k} w_i Z_i + [1 - \sum_{i=1}^{k} w_i] m_0 \tag{5.2}$$

Where the local mean $m_0$ is assumed to be equal to the global mean (for SGS equals to 0), the estimator equation is reduced to:

$$\hat{Z}_{SK} = \sum_{i=1}^{k} w_i Z_i \tag{5.3}$$

This equivalent form assumes local unbiasedness.

A sequential path for visiting all locations on the prediction grid is randomly defined to avoid artifacts. Solving one kriging system is necessary for each grid node. To have a conditional procedure, we must ensure that the prediction in the node corresponding to a data sample returns the same sample value. Prediction is done by the simple kriging of nscores. Monte Carlo simulation is applied, in which a value is randomly drawn from a

Gaussian distribution using the kriging mean and variance for each node and adding the new value to the data set. The SK variance is, in fact, the error variance locally defined in Equation 4.20, that can be calculated in the function of the interpolation weights $w$ and spatial covariances:

$$Var\{\hat{Z}_0\} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j Cov\{Z_i Z_j\}$$

The procedure continues from node to node until all nodes in one image are completed. To simulate another image, the mechanism is repeated using a different random path.

Results vary according to different parameters such as the variogram range, the neighborhood search radius and the proportion of original and simulated data used for conditioning.

### 5.1.5 Variogram model and parameters validation

Simulation includes the selection of many parameters that need validation. To validate the results of the simulation, it must be verified that the global histogram and variogram have been reproduced. Every image produced with SGS is a possible realization that tries to do such reproduction. When working with simulations, a large number of realizations are required to fit a model. The validation used in practice relies on graphic analysis of ergodic fluctuations in variogram and histogram reproductions (23). A large number of simulations may be produced if ergodic fluctuations are excessive. As stated by Goovaerst (23), when departures from model statistics are deemed too important, the realization can be discarded and another realization generated.

### 5.1.6 Post processing of simulations and probability mapping

For probability mapping, it is important to have the largest number of simulations possible, which are used to represent the local probability of exceeding guideline thresholds. Using the many simulated values for each location, a local *pdf* can be approximated. When considering a single value as the critical threshold, the goal is to estimate the local probability of being above this cutoff using the generated *pdf* of the random variable $Z(x)$:

$$P_{\text{above}} = \text{Prob}\{Z(x) \geq z_k|(n)\} = \int_{z_k}^{+\infty} \text{pdf}(z)dz \tag{5.4}$$

## 5.2 MGK and SGS at the Swiss national scale

As discussed in chapter 4, various scales of analysis are pertinent to indoor radon mapping in Switzerland. For the use of the SGS and MGK methods, the national and the local scales have been considered with a downscaling strategy. As will be seen, the problem of modeling diverge depending on the scale of analysis, and assorted solutions are to be proposed. The

reduced size of the local set3 will allow a deeper analysis and a larger experimentation of the methodological options. When applying SGS, it must be taken into account that many realizations of the model are required, and therefore, it is a computer-time-consuming algorithm.

Essentially, in order to start with Gaussian simulations, the nscore transform must be performed and a structured variogram for this transforms obtained. The transformation helps, in some cases, to reveal spatial continuities affected by high variance noise. In other cases, variograms end up being less structured because of trends or high local variability. It is therefore wise to do the analysis using different sets and scales.

### 5.2.1 MGK for the global set of Switzerland

To perform MGK on the indoor radon dataset for Switzerland, a transform to nscores was first done. A consequence of adjusting data to a Gaussian distribution is that data is centralized to a distribution with a mean of 0 and a variance of 1. Hence, skewness is reduced to 0 and the influence of the extreme values is attenuated. This, of course, also has an influence on the resulting variograms.

The variogram for the Swiss raw data was already presented in Figure 3.51 of chapter 3, with the indication that the spatial structure was not clear. Later, in Figures 3.52 and 3.53, the use of MW averages to transform data was proposed, and an improvement in variography modeling was obtained. With this in mind, nscore transforms were done for raw data and for the MW averaging at 280 and 1339 meters to see whether the variography could be further improved (Figure 5.1).

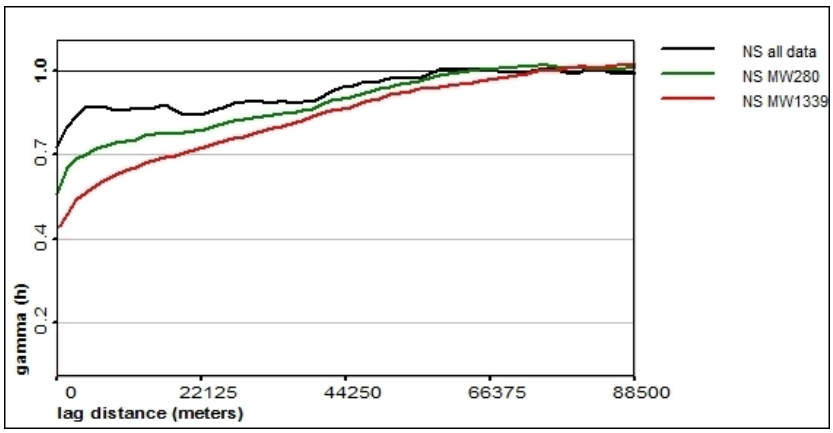



Figure 5.1: *Experimental variograms for nscores tranforms of raw data and MW 280 m. and 1339 m. averages of the Swiss indoor radon dataset*

Compared to previous variograms, nscore transform allowed raw data to reveal some spatial continuity. Nevertheless, the sill or the portion of explained variance is only 20% of the total variance for the raw data. The variogram of nscores for raw data shows also more than one structure. For the MW averaging, variograms appear smoother for short distances,

while the structure is essentially the same. With MW at 280 m, the nugget effect seems to be around 60%, while MW averaging at 1339 m reduces it to nearly 45%. Most important, is that when using MW, experimental variograms can finally be adjusted to valid models.

Then, the advantages of using all data but having a high nugget and a short range should be evaluated against the disadvantage of using a short and smoothed dataset, but with a structured variogram. Using the variogram for nscores of MW 1339, a MGK was first launched as a rapid-mapping method. The variogram model fitted consists of the nugget model and two spherical models with a maximum range of 88000 m. The corresponding formula is nug0.45+sph0.13R5000+sph0.42R88000 (Figure 5.2).

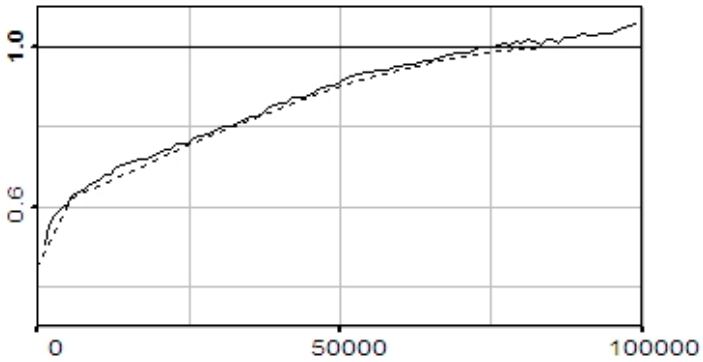



Figure 5.2: *Experimental variogram for nscores tranforms of MW1339 m. averages of the Swiss indoor radon dataset and variogram model fitted*

Based on the MGK optimization curve, through cross-validation (Figure 5.3), the number of neighbors used were 25. In theory, all training points can be used because the variogram model is a statistical regional one, and eventually, kriging weights become 0 for large distances. An optimal number of neighbors indicate which effective range minimizes the training error. Because experimental models are not perfectly stationary, they attain a limited range.

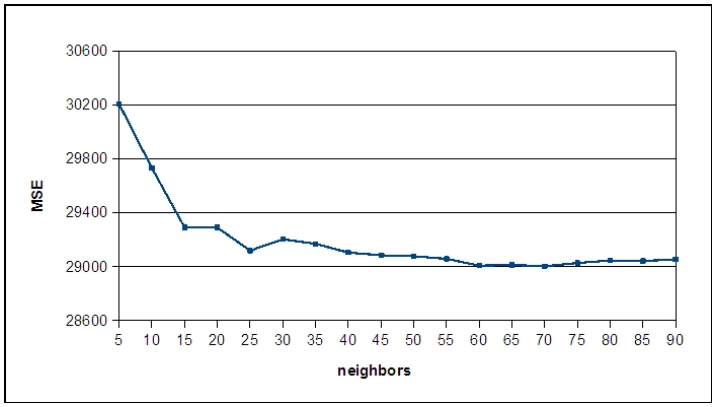



Figure 5.3: *MGK neighbor CV optimization curve for the MW1339 dataset*

In fact, the optimum in the curve is 70 neighbors, but the error is almost the same with 25. The low influence of additional neighbors is because the variogram model progressively has very low weight for points far apart. The random distribution of points within the urban area (with 41,787 points created) was used as an interpolation grid to produce a more realistic view of the results (Figure 5.4).

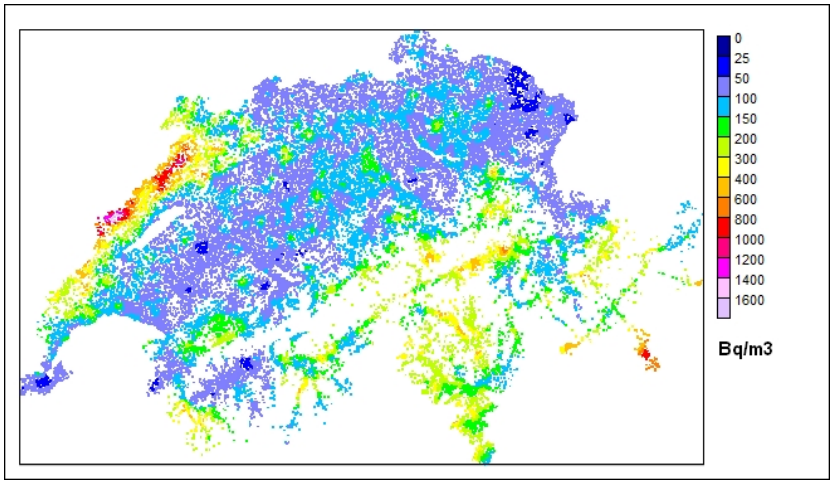



Figure 5.4: *MW1339 dataset interpolation with MGK method, using an urban area grid for Switzerland*

Statistics for map 5.4 indicate that the range of values reproduced varies from 34 to 1646 Bq/m3, the mean is 136 Bq/m3 and the variance is 12481. Meanwhile the MW1339 dataset has a range between 5 and 4602 Bq/m3, a mean of 148 Bq/m3, a median of 90 Bq/m3 and a variance of 47006. As mentioned, a crucial component of kriging calculations is the so-called kriging variance. It is the uncertainty component of interpolation and also used to produce the many realizations within SGS. A map of the distribution of kriging variance after MGK for MW1339 appears in Figure 5.5. Kriging variance after MGK has a scale that relates to the nscore transform (For the MW1339 set, it goes from -3.8 to 3.8).

The kriging variance map clearly shows the areas with more uncertainty (high variance). They are directly related to zones with a lack of samples. In Switzerland, the areas least covered are the canton of Fribourg and some mountainous areas. Mountainous areas appear to be undersampled in some cases, simply because the building density is lower in these areas. High kriging variance is an expression of spatial clustering and the boundary effect. In this particular result, variance is not so high on external borders because of the interpolation domain used.

Emery (18) has shown that the nscore back-transform can lead to errors, in particular for cases where the sample variogram does not reach the theoretical sill of 1. The causes are due to sampling configurations and trends. It is, therefore, important to verify that stationarity conditions are fulfilled or to perform post-transform corrections as indicated by the author.

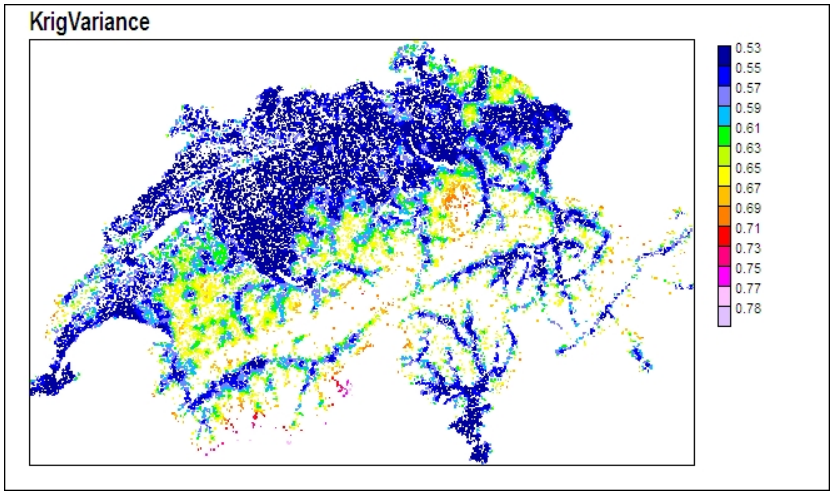



Figure 5.5: *Simple kriging variance map after MW1339 interpolation using MGK*

### 5.2.2 SGS for the national set of Switzerland

For SGS a series of 4 realizations were launched, the maps are presented in Figure 5.6, and the box plots' distributions for these realizations are shown in Figure 5.7.

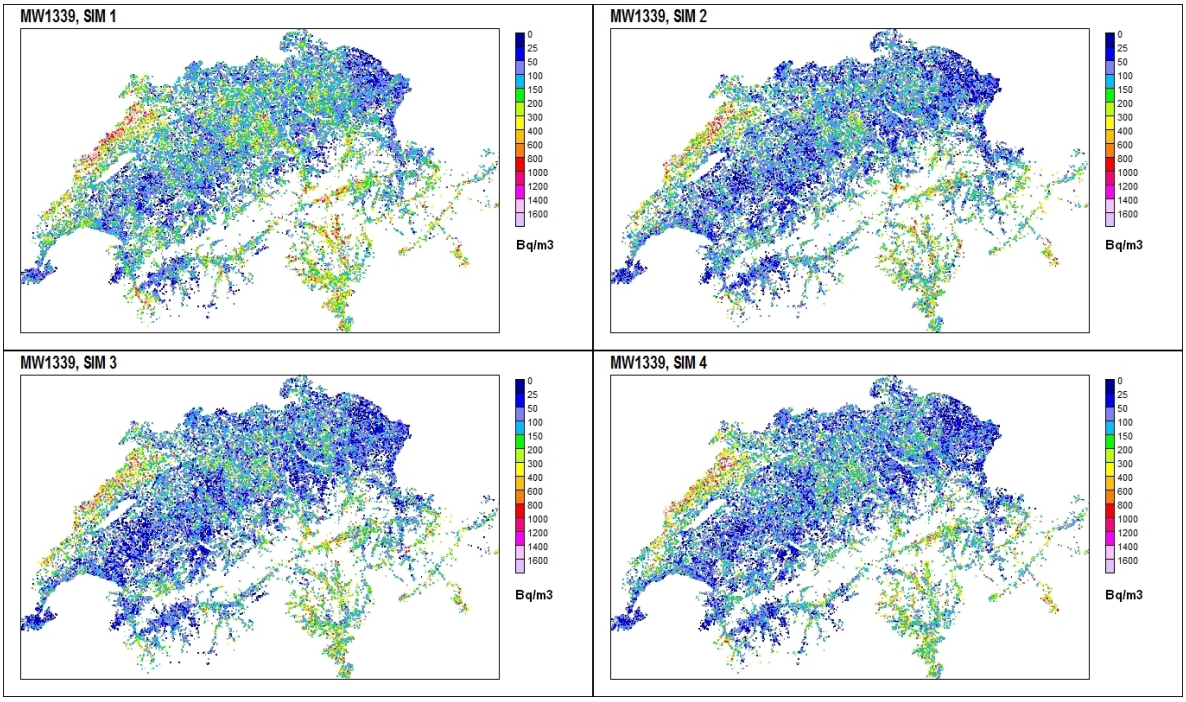



Figure 5.6: *Series of four SGS realizations maps for the MW1339 Swiss indoor radon dataset*

The box plots for the realizations show that the range of training values (between 5 to 4602 Bq/m3) is always reproduced. The mean of each realization is always around 134 Bq/m3, which is close to the sample mean. The median values of the simulations are 82, 83,

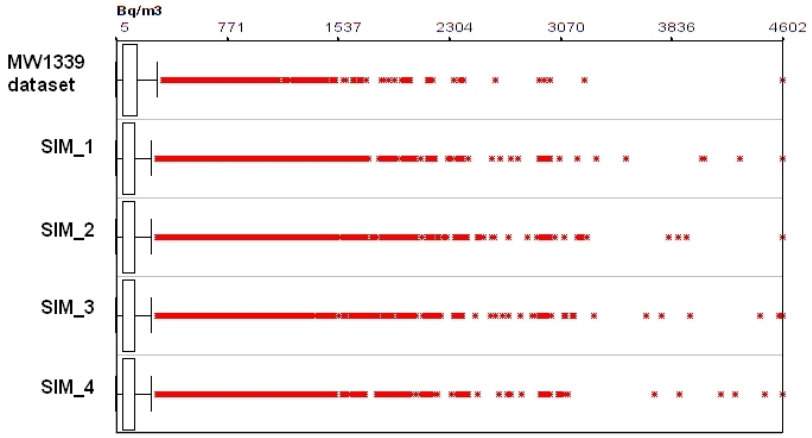



Figure 5.7: *Boxplots of the statistical distribution of the MW1339 dataset and the four SGS realizations*

83 and 84, which is a bit lower than the sample's median (90 Bq/m3). This slight difference can be seen in the box plots (Figure 5.7). The simulation variances for the four realizations shown are 42143, 39509, 42097 and 44900. Compared to the variance of the MW1339 set (var=47006), these simulations have a tendency to centralize estimates. In fact, the nscore transform adjusts to a centralized distribution and that influences the back-transformed values.

It is also interesting to present a box plot to compare the original MW1339 data distribution, not only with a simulation but also with MGK, OK and SK (Figure 5.8). When looking at this box plot, the smoothing effect of regression methods is evident.

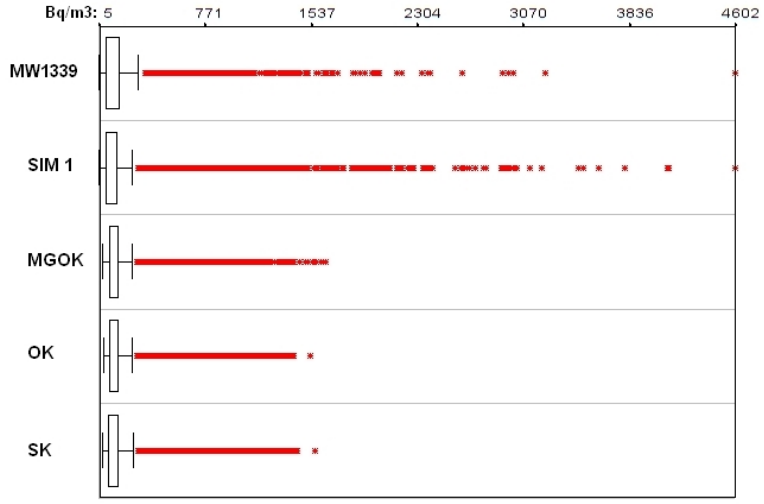



Figure 5.8: *Boxplots of the statistical distribution of the MW1339 dataset, one SGS realization, MGK, OK and SK estimations*

In Figure 5.9 the variograms for each of the four realizations and the variogram for

the MGK estimates are presented together. The main objective of the SGS method is to reproduce the statistical distribution of the samples and the spatial variance (the variogram). Variograms were roughly reproduced at each realization; variability appears increased in the

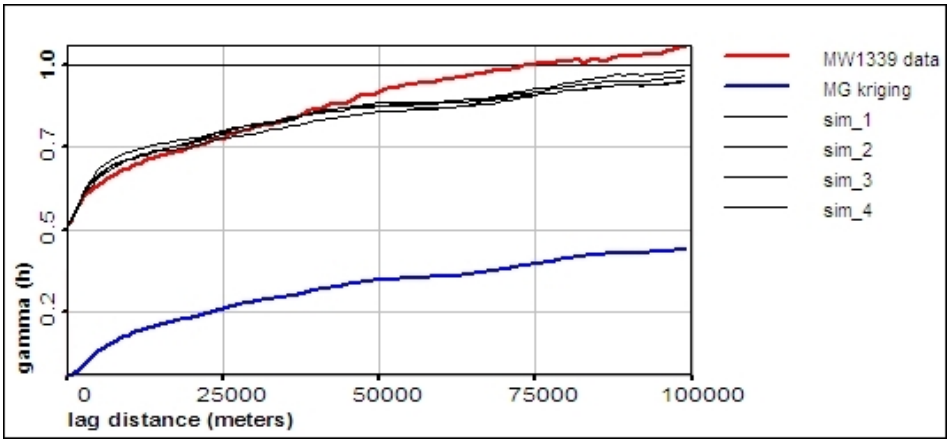



Figure 5.9: *Variogram reproduction of the four simulations and the kriging estimate by MGK*

first portion of the range while variance appears underestimated at the longest range. The simulation variogram analysis helps improve modeling; in this case, more attention should be paid to the variances at longer distances. Especially because variograms at distances longer than the 88 km range exceed variance = 1; this is a portion of variance that cannot be modeled under stationary hypothesis. Two other variogram models with small modifications were tested to observe their reproduction after simulation: v2 nug0.4+exp0.17R1500+sph0.43R80000 and v3 nug0.45+sph0.1R3000+sph0.45R80000. Variogram 3 shows a better reproduction at short and long range (Figure 5.10). There are other factors linked to neighborhood that also influence variogram reproduction, and these will be analyzed using local data.

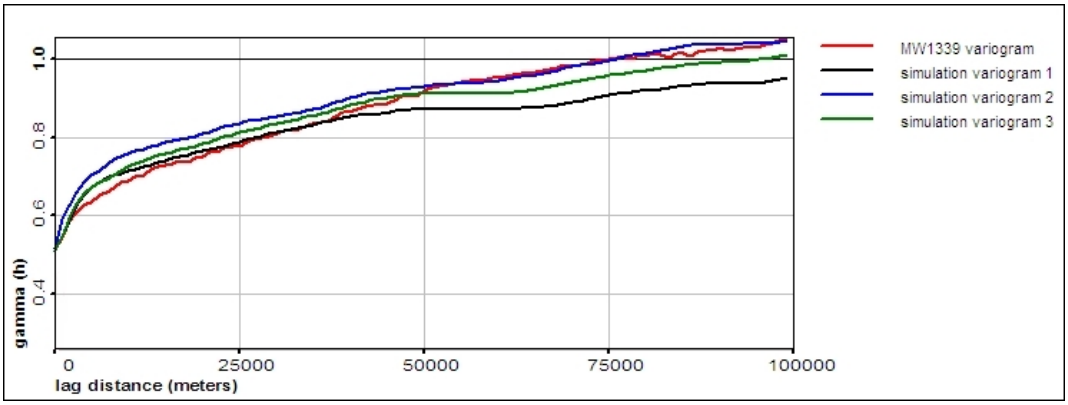



Figure 5.10: *Simulation variogram reproduction using variogram model v3*

For the case of MGK, it is clear that variance of estimates is drastically reduced and therefore the variogram appears smoothed in comparison to simulations. MGK estimates are equivalent to the central values or the means of all simulations at each point. What is

important to highlight, is that simulations have reproduced variability on a whole, including a precise reproduction of the nugget effect, which is an inherent part of data uncertainty. That makes simulations' results more realistic than any other method.

In order to obtain a probability map, a large number of realizations are needed. For the Swiss dataset, the computer-time will be very high considering a good number of realizations and the need for a more dense interpolation net for a detailed map. Besides this, the original dataset was transformed and reduced in order to delineate a variogram. Under these conditions, the national map is just an approximation, and as stated in chapter 3, mapping of subsets can be a better strategy to increase precision.

## 5.3 SGS neighborhood parameters and variogram reproduction for the set3

As sequential simulation is a process of adding information within a net, there is a gradual change in the neighborhood with each new estimation added. Therefore, it is essential to investigate the influence of trends, clustering and neighborhood in the variogram reproduction (68).

### 5.3.1 Statistics, trend, clustering and simulation net of set 3

A basic statistical and spatial data analysis for set3 was already done in chapter 3. In this section, we are interested in knowing how the nscore transform affects the spatio-functional properties of data. The influence of clustering and the simulation net are also discussed. For instance, in chapter 3, a spatial trend to the northwest direction using MW analysis was observed. The clarity of this trend had helped make a spatial partition into two distinctive zones. Is this trend always present after nscore transforms?

The directional diagrams of the indoor radon values (Figure 5.11) present in detail the spatial trend already identified by the plot maps. The trend is clear in the NW direction, but the diagrams also show that this is mainly due to the influence of a few high values. The regression line fitted on top (in red) does not have a steep slope, indicating that it is only a few (but consistent) sets of values that are creating this trend. After nscore transforms, the new scale attenuates the high values, but the trend is preserved (Figure 5.12).

During the spatial analysis, it was established that the high clustering of samples is a consequence of the fragmentation of the urban coverage. The cell to urban comparison in chapter 3 (Figure 3.23) showed that samples within the limits of the urban domain have a quite homogeneous distribution in the set3 zone. It can be accepted that the set3 samples are spatially representative of the population in this zone despite clustering. It is also certain that samples represent a reduced percentage of the existing dwellings in the area. As mentioned in chapter 3, the use of constrained spatial domains appeared coherent to approach global statistics. In order to approach a global histogram using stochastic simulations, declustering-weighting methods should be tried out. In the case of SGS, this procedure must be done with raw data before doing an nscore transform.

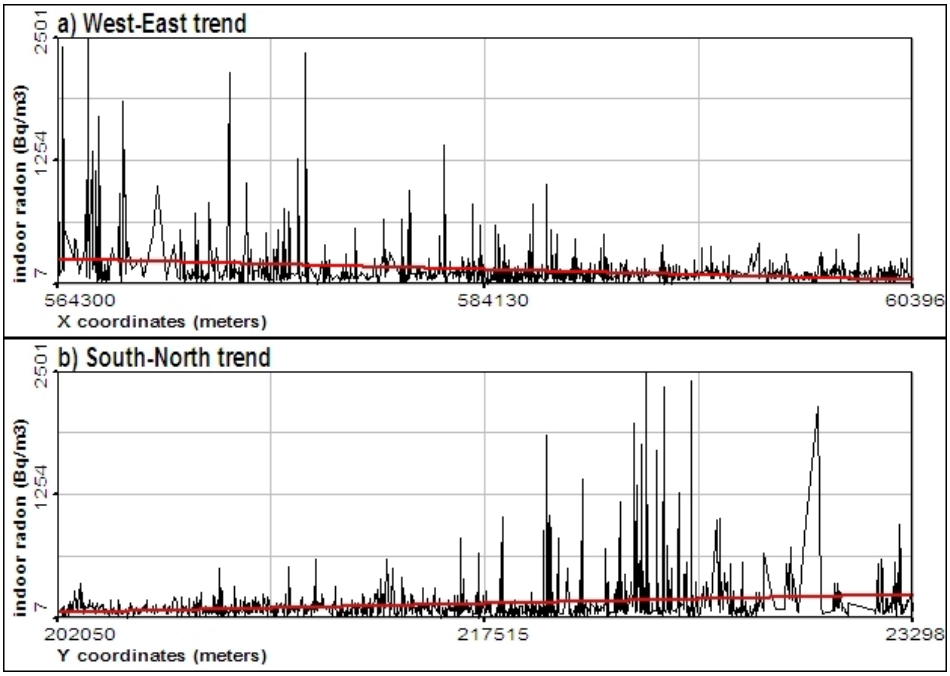



Figure 5.11: *Indoor radon values directional diagrams on a) WE and b) NS directions*

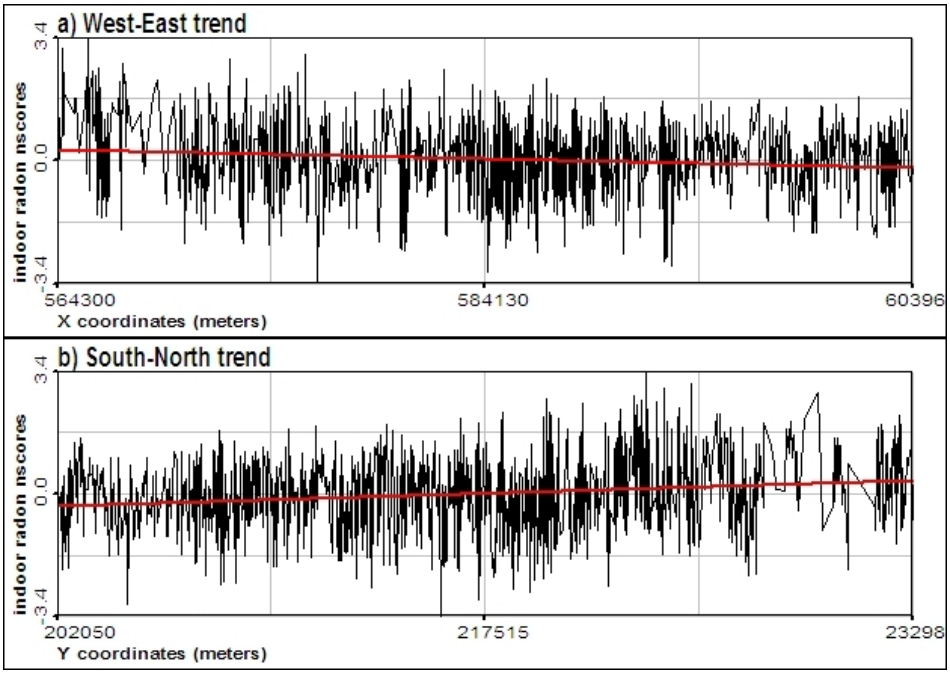



Figure 5.12: *Indoor radon nscore transformed values directional diagrams on a) WE and b) NS directions*

The influence of the spatial distribution of points and the neighborhood tuning for the method is thereafter analyzed. The possible influence of the constrained urban domain is taken into consideration. Besides the classic simulation net used for mapping, which consists

of a rectangular arrangement of points within the bounding box, a net within the external limits of the urban constrained domain was created for the purpose of comparison (Figure 5.13). Points in both nets are separated by a distance of 500 meters.

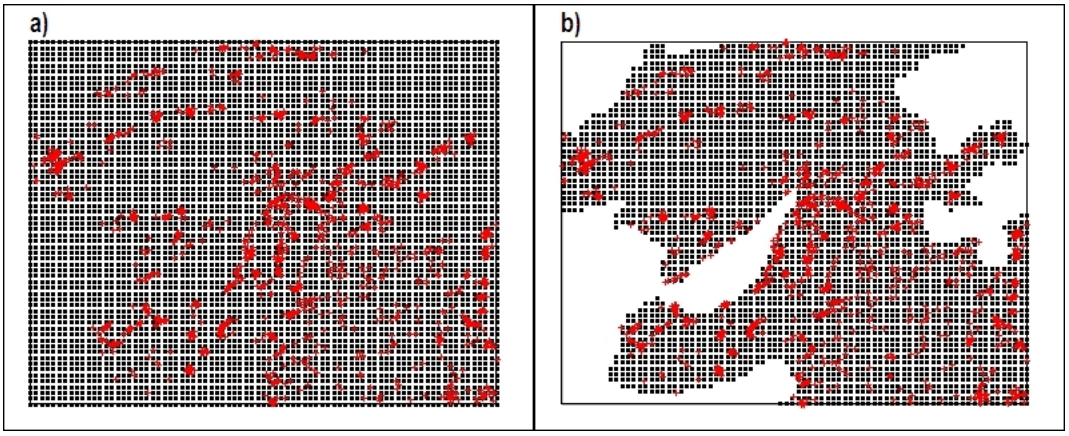



Figure 5.13: *Simulation nets for the set3 limited by a) rectangular bounding box b) external polygon*

### 5.3.2 Nscore variography of set3

The anisotropic experimental variogram of nscores was calculated with a lag of 300 m and a high tolerance of 500 m in order to create a smoother variogram. This regularization of the variogram proposed in (6) helps with variography modeling. The idea is to facilitate variogram modeling by doing some smoothing at short distances so that local variability is a bit masked. A spherical variogram model with two nested structures was fitted on top. The nugget effect accounts for 0.55 of the variance, the remaining variance is divided into a first structure having a range of 850 m and a second structure, going up to 18,000 m (nug0.55+sph0.26R850+sph0.19R18000).

As mentioned, the behavior of indoor radon depends on many factors and contributes to data uncertainty on small scales. The high nugget effect and the short range of the first structure reflect this local variability. The first range of variance occurs at less than 1 km of distance, which approximates the average distance between points (968 meters). The second structure is intended to model the influence of global environmental factors. Variogram modeling becomes more difficult for indoor radon because of the complexity of the physical cumulation process. However, the long range can be assumed to be an effect of lithology and soil properties. To calculate the total range, it is reasonable to think that most of the spatial variance can be detected through half of the diagonal of the bounding box. Figure 5.14 shows the experimental variogram of nscores and the chosen model.

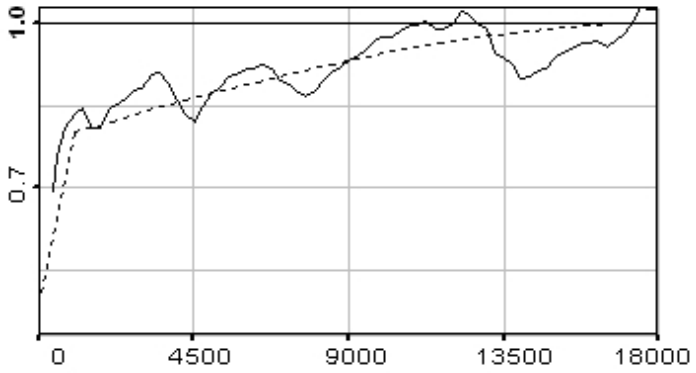



Figure 5.14: *Experimental variogram (continuous line) and variogram model (dashed line) for nscores of set3*

### 5.3.3 Influence of the simulation net and the number of neighbors for set3

#### MGK optimal neighbors and kriging variance

MGK allows rapid mapping with nscores and was used to obtain the optimal number of neighbors for the kriging procedure. Because there is not a unique solution with CV for SGS, the solution from MGK was used as a reference for the neighbor search optimization. Figure 5.15 presents the CV curve for neighbors optimization. For MGK, the CV error is slightly improved beyond 20 neighbors. This parameter was further considered as a reference for the minimum number of neighbors to be used.

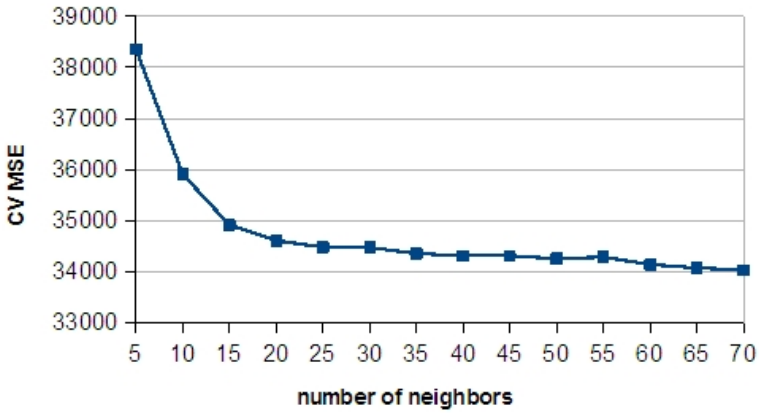



Figure 5.15: *CV neighbors optimization curve for set 3 by MGK of nscores*

MGK also shows that the kriging variance is lower within the constrained net in comparison to the rectangular net (Figure 5.16), thus, indicating a possible advantage with the use of the constrained net.

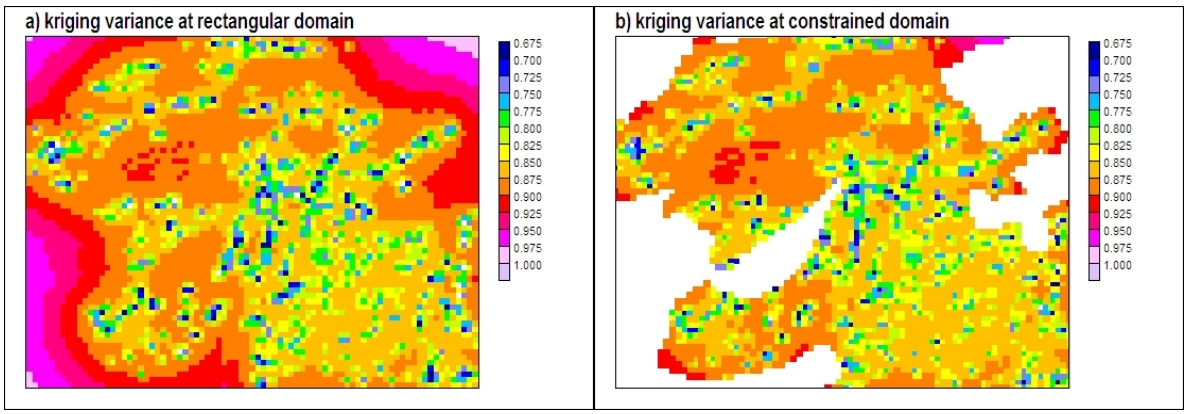



Figure 5.16: *MGK kriging variance distribution for the set 3 using a) a rectangular net and b) a constrained net*

**Combined effect of neighbors and nets on variogram reproduction**

The objective of the proposed tests is to verify whether, as required for SGS, the variogram model proposed was reproduced after simulation. The variogram reproduction was evaluated under different parameters of numbers of neighbors and types of simulation nets used.

The previous inquiries, have shown that a minimum of 20 neighbors might be required for the simulations. The next thing to check is whether there is an optimal maximum or whether the use of specific ranges of neighborhood can influence variogram reproduction. Hence, the neighbor search parameter was tested using an option with a minimum of 20 and a maximum of 30 neighbors. The second option was an interval ranging from 40 to 60 neighbors. It is expected that the use of more neighbors will produce more smoothing. These tests were combined with the two types of simulation nets.

For the first run, the number of minimum neighbors was fixed at 20 and the maximum at 30. Figure 5.17a shows the simulated variogram using the constrained simulation net (dashed line) and the variogram when using the rectangular net (continuous thin line) on top of the experimental variogram (bold line). For the comparison, a single simulation with the same random seed was used. We see that a better reproduction of middle range was obtained using the rectangular grid, while the short range seems to be better reproduced using the constrained grid (Figure 5.17b).

For the second run, the number of neighbors was increased to a minimum of 40 and a maximum of 60. The results in Figure 5.18a show a better reproduction of the variogram at a long range for both nets and almost the same reproduction at a short range. The use of a defined interval of neighbors has a clear influence on variogram reproduction. In theory, all points can be used for a kriging system but weighting decreases exponentially for points far apart. Hence, is important to limit neighbors to the optimal maximum necessary in order to reduce time calculations.

According to these results, a fairly accurate reproduction of the variogram during simulations can be obtained with a minimum of 40 neighbors. It is also important to verify that

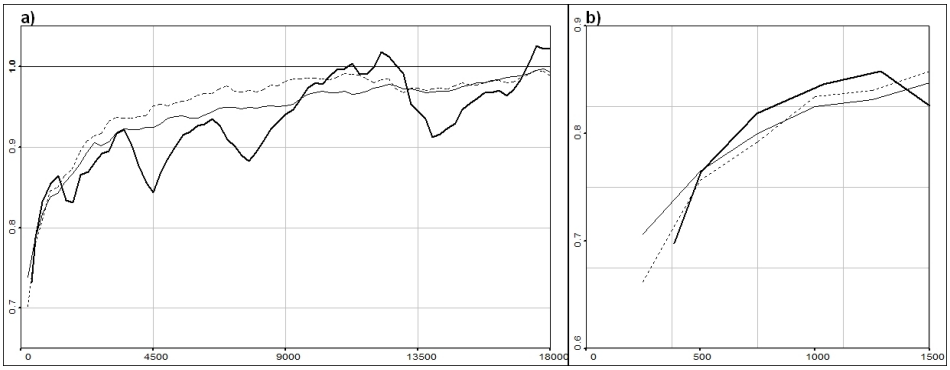



Figure 5.17: *Variogram reproduction using a constrained net (dashed line) and a rectangular net (continuous thin line) for a minimum of 20 and a maximum of 30 neighbours for a) the long range and b) the short range*

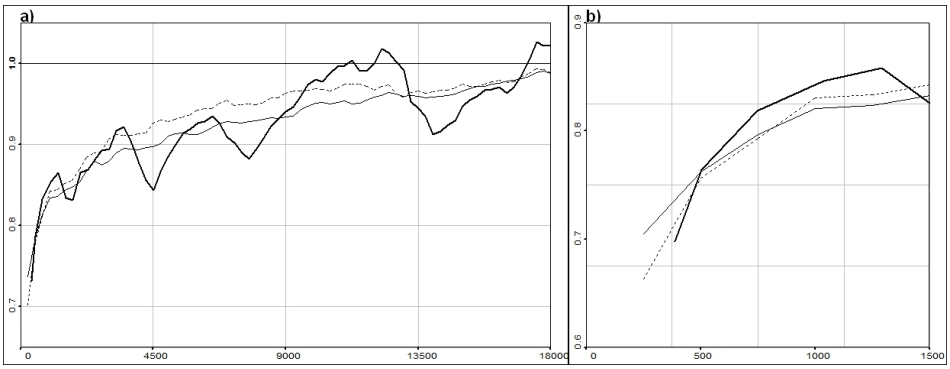



Figure 5.18: *Variogram reproduction using a constrained net (dashed line) and a rectangular net (continuous thin line) for a minimum of 40 and a maximum of 60 neighbours for a) the long range and b) the short range*

these simulated variograms have fluctuations that do not exceed the model. In Figure 5.19a, the fluctuations of 5 simulations are shown using a minimum of 40 neighbors and the rectangular net. In Figure 5.19b, the corresponding reproduction of the data histogram is also shown.

While the influence of the number of neighbors seems to be clear, the effect of the simulation net is not. After thoroughly observing the shape of variogram reproduction in Figure 5.17a, some similarities are seen to exist because of the same seed generator was used. Nevertheless, having a difference in the number of points between simulation nets may produce a different spatial distribution of values. The use of a single reproduction was useful to detect the effect of neighbors because it was consistent throughout the nets but not for the net testing itself. Fluctuations (as shown in Figure 5.19a) may be larger than the effect of the net influence, even though variograms have been smoothed by using a large lag tolerance.

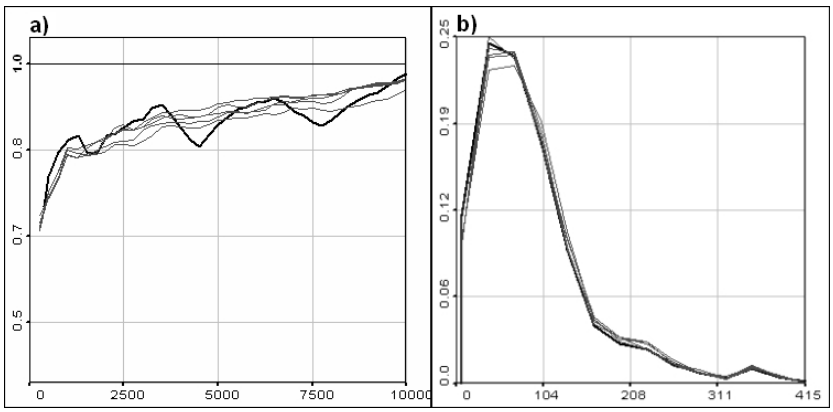



Figure 5.19: *Fluctuations (thin lines) of the a) variogram model and b) the histogram (bold line) for 5 realizations by sequential gaussian simulation*

**Optimized simulation net and variogram fluctuations**

A third simulation net was built as a compromise between the rectangular and the constrained net. A constrained net considering interior empty spaces and a buffer zone around training points was created (Figure 5.20). The buffer area has a radius of 1500 meters, and this will increase the number of points in the net, while empty interior spaces will reduce them. In the end, there are more points in this net than in the first constrained net.

Figure 5.20: *Constrained net with empty interior spaces and buffering zone of 1500 meters*

A series of SGS simulations were launched for neighborhood ranges of 1 to 20, 20 to 50 and 40 to 60 n (Figures 5.21 and 5.22) using this constrained net. The series of 5 simulation variograms in Figure 5.22b are superposed to the smoothed training variogram in order to compare them with the results in Figure 5.19a.

The fluctuations adjust better to the model when using a neighborhood between 40 and

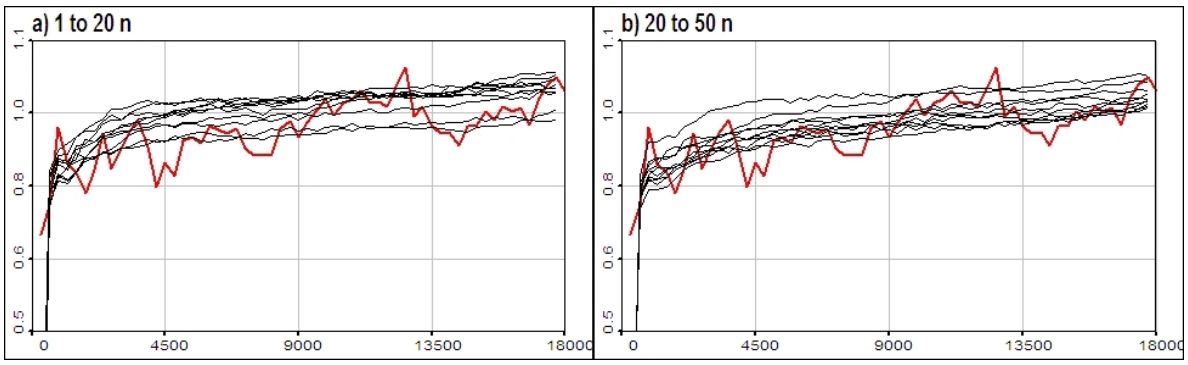



Figure 5.21: *Fluctuations (thin lines) of the variograms using a) 1 to 20 neighbors and b) 20 to 50 neighbors for 10 realizations by SGS*

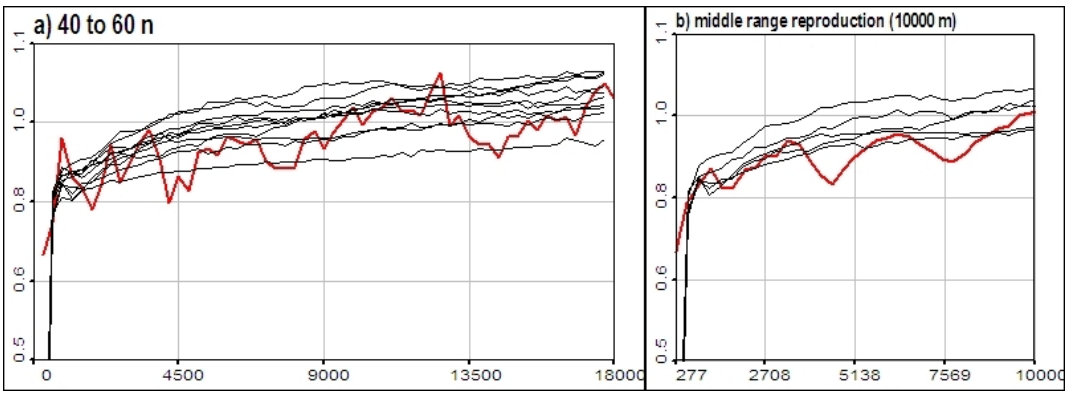



Figure 5.22: *a) Fluctuations (thin lines) of the variograms using 40 to 60 neighbors and b) detail of the short range reproduction*

60 n. In general, the reproduced variograms do not adjust at the middle range using the constrained net with a buffer zone. It seems that the sequential performs a better variogram reproduction when a bounding box net is used. The presence of filling points may contribute to a reproduction at all ranges.

It can be concluded that the reproduced variogram does not adjust the experimental variogram at some ranges. The question is wheter the reproduced variogram is wrong or whether it represents a possible realization. The point here is that variogram reproduction also depends on the simulation net. In the case of indoor radon, this net provides some realistic information, which is the global sampling domain. Therefore, it is acceptable to think that the reproduced variogram reflects, to a certain extent, the spatial distribution of the variable.

A model was fitted over a simulated variogram in order to analyze the posterior spatial distribution. The variogram has a coding of nug0.55+exp0.24R850m+exp0.21R24000. The main difference with the training variogram model is that instead of a spherical model the posterior one will fit better to an exponential. The exponential model proposes that higher variance exists at middle ranges (between 3000 and 10000 meters). This can reflect higher

differences between localities. It can also be an artificial effect created from the net; in any case, the urban domain used for this test is only an approximation, which still requires some refinement. The use of more accurate domains will be stressed in following analysis for set3B.

### 5.3.4 Probability maps for set3 with SGS

Considering a minimum of 40 neighbors and the previously defined simulation nets as hyper parameters, 100 simulations were run for the training data. Each of these simulations produced a prediction map, which are represented in Figure 5.23. A set of four maps for the first simulations after back-transformation from nscores is presented. These images are possible realizations of the joint Gaussian distribution obtained with the sequential mechanism. A large number of realizations are then required to build a complete *pdf* for every location in the simulation net. From this *pdf*, it is possible to calculate the probability of exceeding a certain threshold value. In Figure 5.24, the map of probability exceeding 200 Bq/m3 is shown using the rectangular and the constrained nets. Visually, the probability maps using both nets look alike.

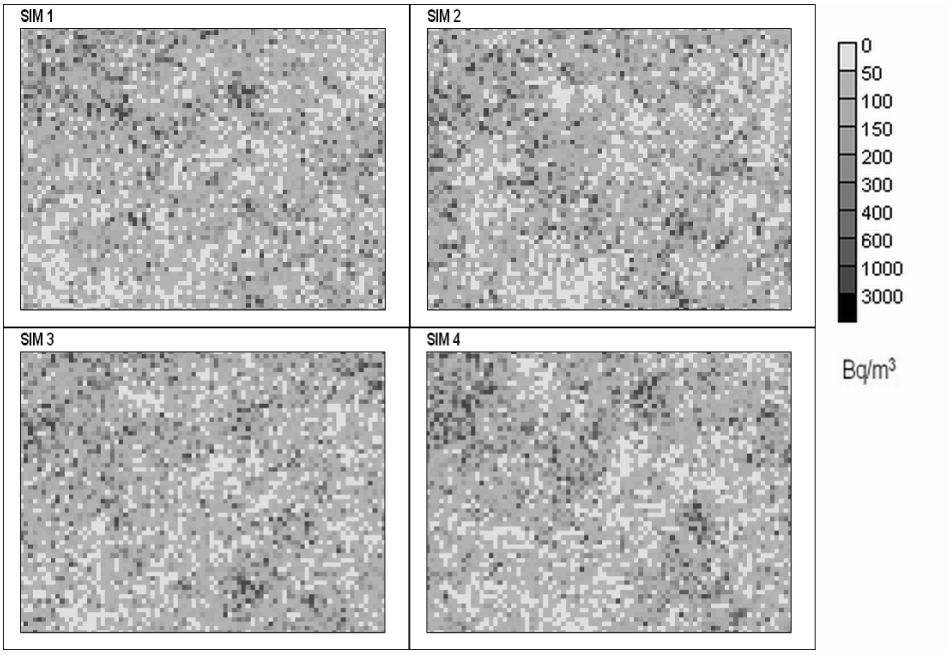



Figure 5.23: *Four simulation maps of the joint distribution realization using SGS with rectangular net*

The measure of estimation error or uncertainty is a useful information for decision making to be added to probability maps. For simulations, the uncertainty can be expressed by the variance of simulated values at each location. It is calculated for each point from the set of realizations. It is an expression of the fluctuations and it mainly depends on the conditional data. The proportional effect between local mean and variance, analyzed in chapter 3,

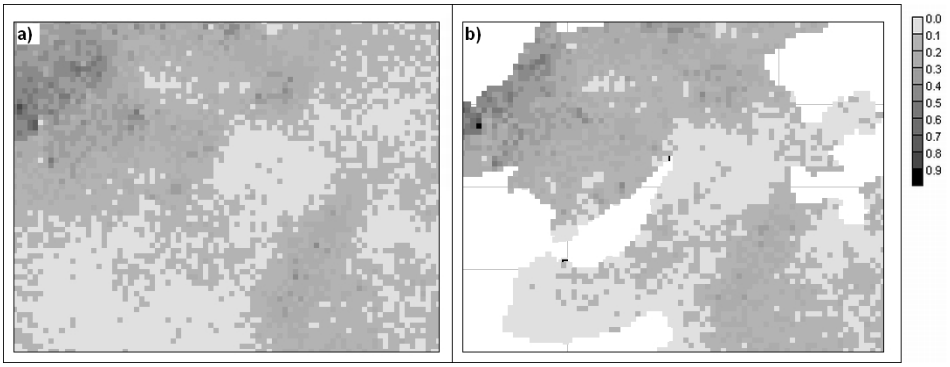



Figure 5.24: *Probability maps for the 200 Bq/m3 threshold using SGS method with a) a rectangular net and b) a constrained net*

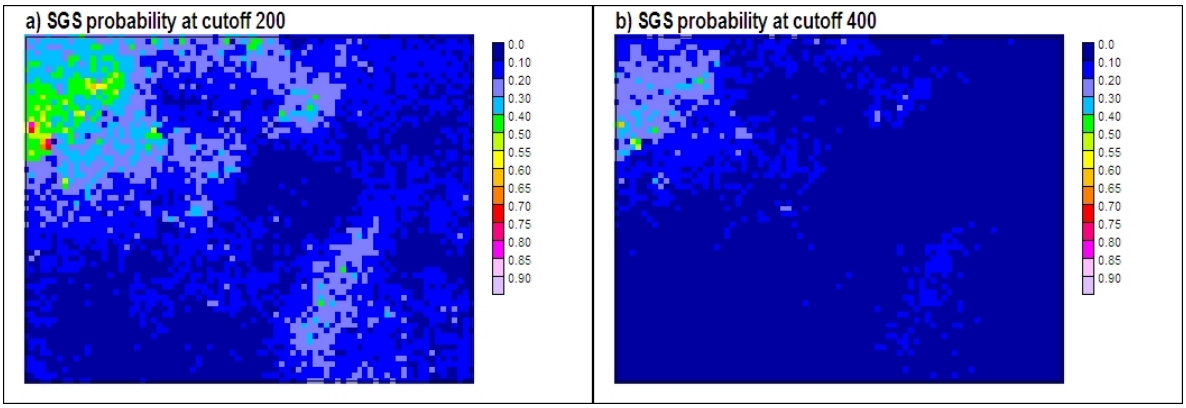



Figure 5.25: *probability maps with SGS for cutoff values of a) 200 and b) 400 Bq/m3*

is appearing here once more. It is clear that higher local variances will generate more fluctuations of the estimates. Here, it is proposed that maps should be produced by combining the probability information with the simulations' uncertainty.

In Figure 5.26a, the p-values map for the 200 Bq/m3 threshold, using a simplified categorization, is presented. Next to it, is a map of the kriging variance displayed with proportional symbols (Figure 5.26b). Larger dark circles correspond to a larger variance. Then, the probability map can be filtered by superimposing the proportional variance representation, as shown in Figure 5.27.

This mixed cartography pretends to proportionally mask the areas where uncertainty is more elevated. In the case of set3, we observe that uncertainty is mainly influenced by the conditional data and is higher in the northwest area.

## 5.4 SGS scenarios modeling for the set 3B

Applying SGS to the Swiss data demonstrated how the reproduction of variograms depends on the model itself. Deviations of the model from the experimental variogram were magni-

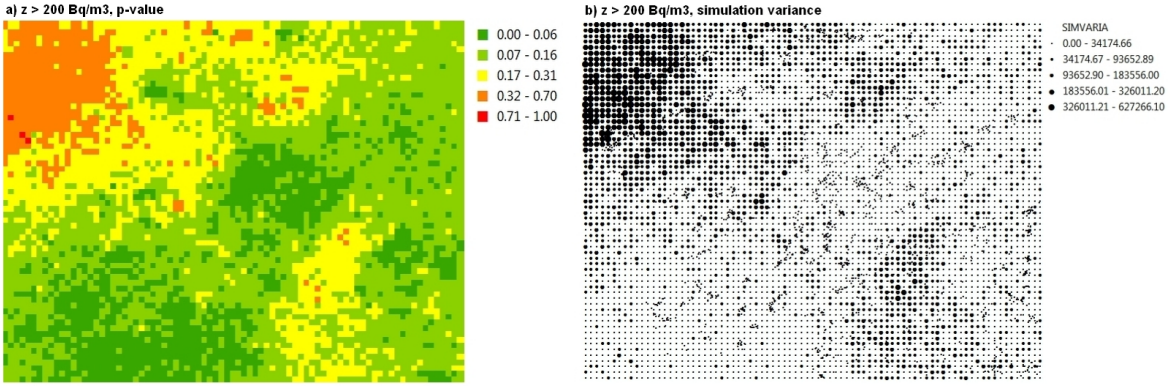



Figure 5.26: *a) SGS probability map for cutoff value of 200 Bq/m3 and b) Proportional symbology map of simulations variance*

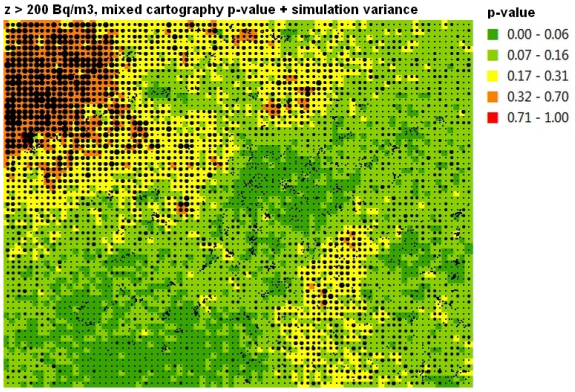



Figure 5.27: *Mixed cartography between probability maps and simulation variance for cutoff values of 200 Bq/m3*

fied during simulations. During tests with set3, it was observed that the simulation net and the neighbor search also influence this reproduction. It is also certain that a coupled effect between variogram definition and neighbor search exists, since they are all neighborhood expressions. Some deviations in variogram reproduction were also observed when using a constrained net as opposed to the more continuous rectangular net. Nevertheless, it is impossible to affirm that this is not a potential realization of set 3B spatial distribution. Moreover, it can be suggested that a constrained net represents the inherent spatial discontinuity of indoor radon in regards to the spatial distribution reproduction.

In previous tests, emphasis was put on variogram reproduction, and some alternative spatial distributions were proposed. The advantage of stochastic simulations is to reproduce a statistical distribution. For the purpose of analysis, the histogram reproduction goal was left as a simple fitting to the existing estimators or sample data. In order to make realistic simulations, the task should be enlarged to a simultaneous reproduction of a variogram and a histogram. The question is, as stated for variography, which histogram should be reproduced? It can be argued that, in addition to the sample distribution, alternative statistical

distributions exist to describe the global unknown distribution.

In this section, the objective has been to compare two possible simulation scenarios that are either based on the available samples or on an alternative histogram. If a more conservative way of thinking is adopted, then the information at hand (the samples) and the derived models are assumed to be sufficient to explain the process. If reality provides some plausible evidence to argue the validity of such models, then alternative models for spatial and statistical distribution based on this evidence are needed. Subset 3B was used for these tests because it is statistically less variable than the whole set3 and is more pertinent for testing histogram reproduction.

For spatial distribution reproduction the element introduced in the alternative scenario will be a refined simulation net constrained to the urban area (as shown in Figure 5.28b), opposed to a rectangular net, which is used for the conservative scenario (Figure 5.28a)

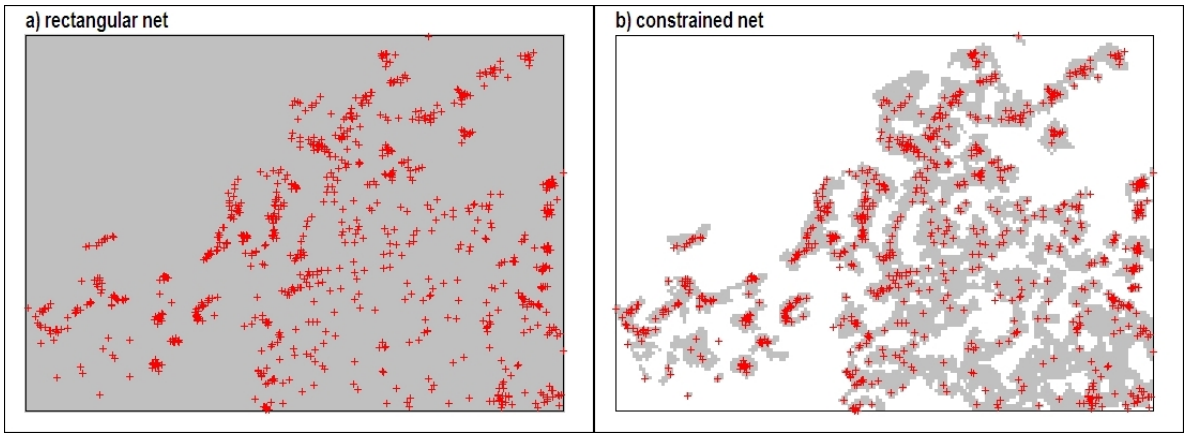



Figure 5.28: *simulation nets used for the set 3B simulation scenarios definition a) rectangular net and b) constrained net*

### 5.4.1 The sample-based simulation scenario

#### Empirical bigaussian test

As stated, it is possible to verify the bigaussian distribution of pairs of points using an empirical test. The concept is to compare the constant value $\sqrt{\pi}$ with the ratio $\sqrt{\gamma(h)}/madogram$, which in theory have the same value. In Figure 5.29 the test was performed for the sets 3 and 3B in order to compare them.

The test indicates that most deviations from the bigaussian distribution occurs below 3000 m. Therefore, we expect to have more difficulties for modeling and reproducing the spatial distribution at these distances. Figure 5.29 also indicates that the set 3B deviates comparatively less than the set 3.

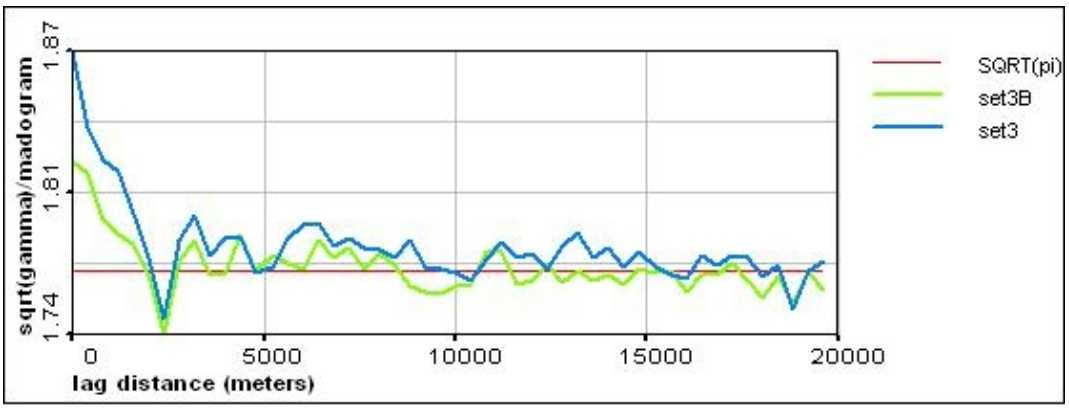



Figure 5.29: *Empirical test comparison of bigaussian distribution of pair of points for sets 3 and 3B*

**Neighborhood parameters testing and mapping using MGK**

The first element to be introduced into a classical simulation scenario is the regular net for the $L$ image. As a reference, the sequential play performed well using a regular space filling for set3. The next element is an optimal number of neighbors obtained from the MGK optimization curve. Previously, the error curve indicated that an optimal minimum number of neighbors would help reduce the extreme influence of local variability. The maximum number of neighbors has an effect of computation efficiency. MGK can also be used to refine variogram modeling. The influence of all these neighborhood parameters makes SGS a complex process; therefore, a good compromise between parameters must be found in order to obtain valid simulations. Figure 5.30a presents the variogram for nscores transforms of the training data, and Figure 5.30b shows the neighbor optimization curve using MGK.

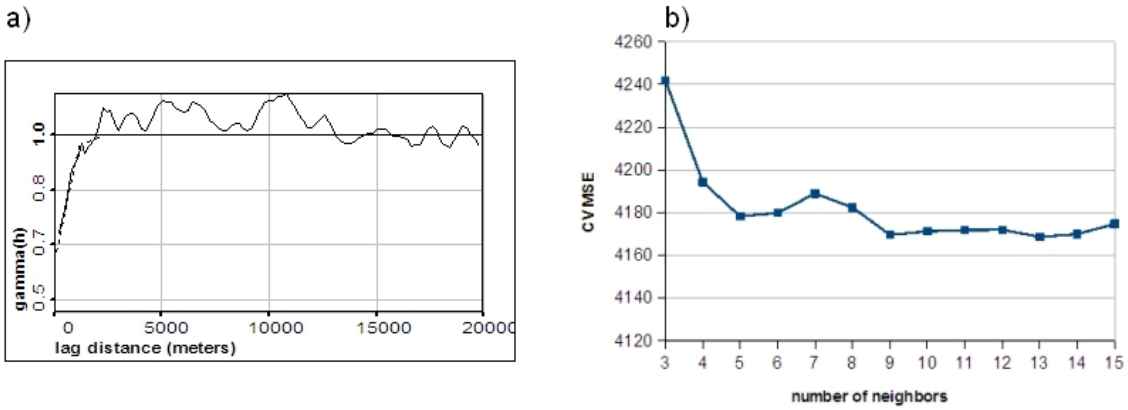



Figure 5.30: *a) experimental variogram and model of 3B nscores transforms b) MGK neigbors optimization curve*

The variogram model for nscores (nug0.65+sph0.35R1500) is less structured than the one for raw data (nug0.55+sph0.45R1700). In particular, the nugget is higher and the range is shorter for the nscore variogram. The nscores' transform has not contributed, in this case,

to revealing a better structure. An alternative option to be tested is the use of KNNR CVMF filtered data. The nscore transform variogram for filtered data and MGK curve are shown in Figures 5.31.

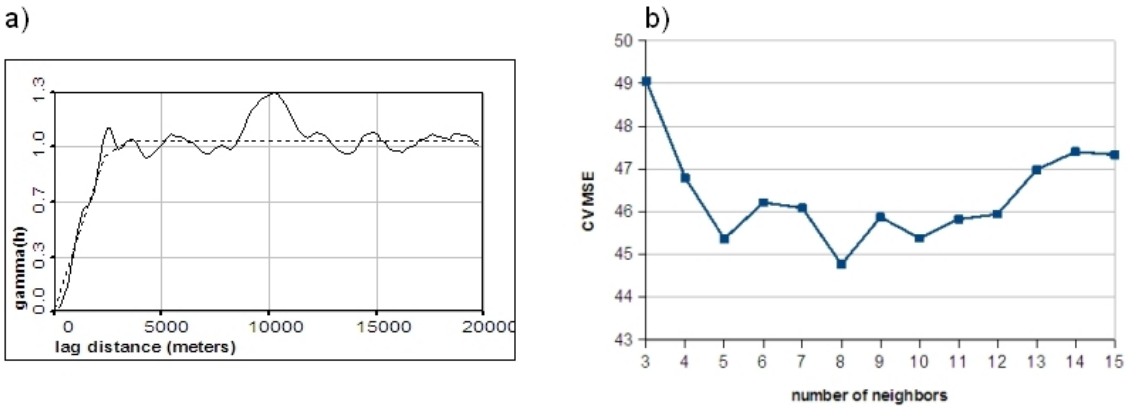



Figure 5.31: *a) experimental variogram and model of 3B CVM filtered data nscores transforms b) MGK neigbors optimization curve*

The variogram model for nscores of CVMF filtered data has a formulation of a pure Gaussian model: nug0+gaus1R4500. The filtered data has a variogram model without a nugget even when transformed to nscores.

Based on the optimization curves for non-filtered and filtered data, a minimum of 5 neighbors seems to create less error. The maximum number of neighbors can be set to 32 (as done for the SK estimation in chapter 4), since the curve shows a steady error level with the increment of neighbors. Filtered data is different because the error increases when using more than 12 n.

The validation MS error using MGK with 5 to 32 n for non-filtered data was 3095. For filtered data using 5 to 12 n, the error was 3088. To test the sensitivity of filtered data in regards to the number of neighbors, MGK was launched using a maximum of 8n and 30n. When more neighbors are used (30 n) the error increases to 3102 and by limiting the maximum n to 8, the error decreases to 3079. The differences in validation errors are small, but they give an indication of the optimal neighborhood to be used for SGS. In particular, neighbors should be limited when using filtered data. Additionally, these MGK validation errors are similar to the ones obtained with other methods for set3B in chapter 4.

Regarding map statistics, nscores of raw data resulted in a mean of 97 Bq/m3, a range of values between 32 and 225 Bq/m3 and a variance of 170. For nscores of filtered data, the mean is somewhat higher (99 Bq/m3) and the range lower (30 to 197 Bq/m3), but mapping variances are higher than raw nscores: 424. This can be depicted by maps of nscores of filtered data (Figure 5.32b); the zones appear more aggregated than, while the raw nscores maps look spotty (Figure 5.32a). In general, the use of filtered data offers some improvement for mapping visualization.

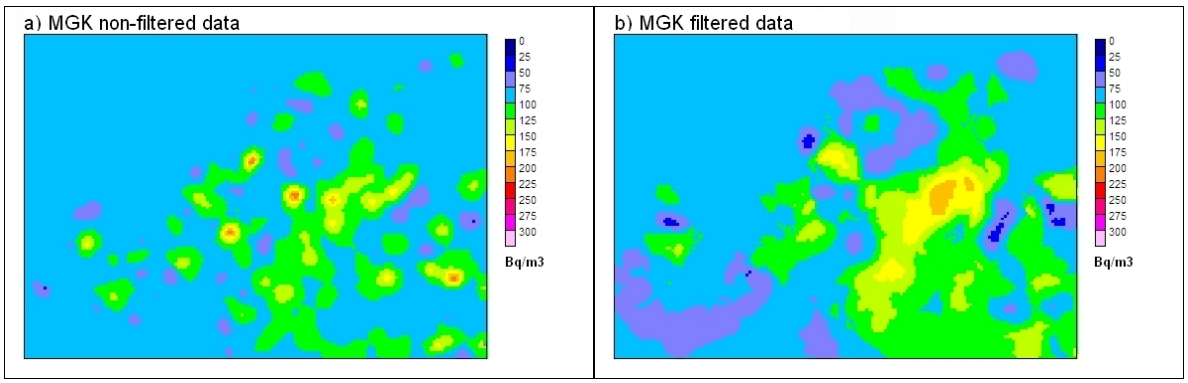



Figure 5.32: *MGK mapping of a) nscores of 3B data and b) nscores of CV filtered 3B data*

### SGS variogram reproduction

The influence of the number of neighbors was tested for set3B. For non-filtered data, variogram reproduction was tested for the ranges of 5 to 32 $n$ and 9 to 32 $n$ (Figure 5.33). Neighborhood ranges were also tested for filtered data: 1 to 8 $n$ and 5 to 12 $n$ (Figure 5.34).

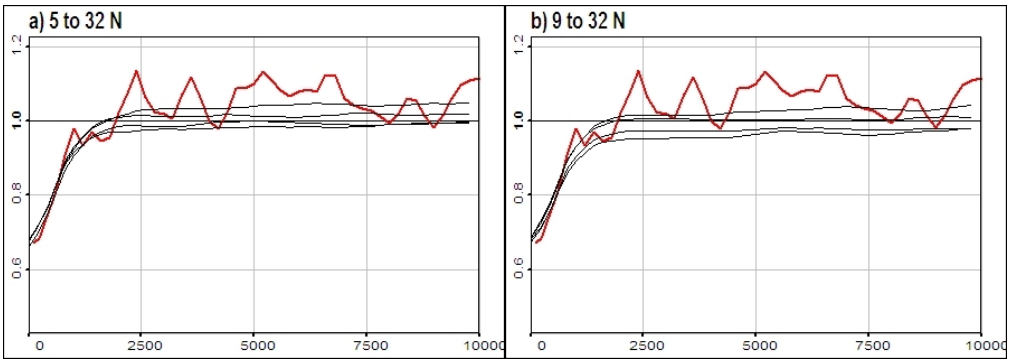



Figure 5.33: *SGS variogram reproduction of non-filtered data using a) 5 to 32 neighbors and b) 9 to 32 neighbors*

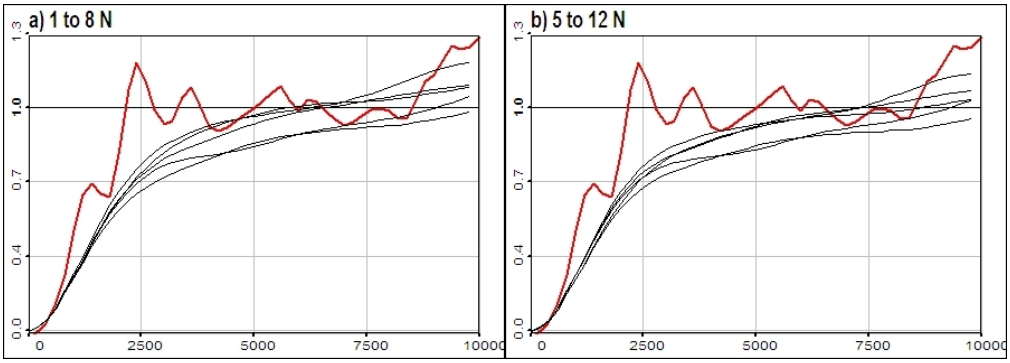



Figure 5.34: *SGS variogram reproduction of filtered data using a) 1 to 8 neighbors and b) 5 to 12 neighbors*

In set3, the variogram presented a pick of variability at a short distance: In this case, the use of a minimum number of neighbors $n$ helped with variogram reproduction. The option of minimum of 5 and a maximum of 32 $n$ has a lower range for fluctuations and seems to better reproduce the nscore non-filtered variogram. For the filtered data, the 5 to 12 $n$ option shows less fluctuations. These results are congruent with the MGK neighborhood optimization curves and the validation errors previously obtained.

### 5.4.2 Simulation scenario with proposed histogram

The complexity of stochastic simulations also gives the opportunity to build alternative scenarios, in other words, to propose alternative histograms and variograms for the estimations. To elaborate such a scenario, a series of arguments should be postulated based on the collected evidence. These arguments will help define the parameters and hyperparameters. The methods for defining spatial parameters have already been exposed in previous analysis. The first element to be introduced is the simulation net constrained to the urban area. The urban area is the natural spatial domain for indoor radon sampling, and it is assumed that almost all measurements should be contained within this domain. It is expected that this domain will rule all neighborhood parameters. Eventually, the evolution of sampling statistics can help delineate a tendency towards global statistics. That was the case in the cantons Neuchatel and Ticino, as exposed in chapters 2 and 3.

#### A spatial-statistical approach to the global mean

We arrive to a central question that has somehow been postponed in favor of spatial analysis: What is the best way to calculate the global statistical distribution? The available indoor radon samples indicate the existence of a positive skewness which give extreme values a heavy weight for all calculations. So, it is important to know the spatial distribution of high-extreme values and its relation with the sampling domain. Unfortunately, there is no recorded evidence that preferential sampling was committed for set3B towards high or low indoor radon values collection during sampling campaigns. Because this cannot be affirmed, it is difficult to decide whether a declustering procedure should end up reducing or increasing the mean. The declustering techniques proposed in the literature (31) (59) are based on the assumption that the preferential sampling design is known.

As stated above, if a tendency of actual sampling toward the full sampling domain is found, it can be used as evidence to approach a global histogram. With the QMI index, spatial clustering as a function of quantile indoor radon values was measured. Figure 3.40, in chapter 3, showed that higher values (third and fourth quartiles) were less clustered than lower values. This is the actual sampling state. So, what would be the state of the population statistics?

### Using the global sampling domain

As exposed in the chapter dedicated to exploratory spatial analysis, the global sampling domain is well defined for indoor radon measurements. This domain is constituted by all the existing buildings and is also the desired target area for making estimations. Because measurements are taken inside buildings, this space also constitutes the sampling support. It must be admitted that this spatial support is, in fact, small and that a change of support will be challenging.

In the work of Borgoni et al. (5), a realignment of the spatial data from the point support level to an area level was proposed. In this case, a good approximation to the municipality level was achieved due to the sampling design used in this study. With good criteria the sampling design, in this research, had a homogenous distribution between the municipality units. Unfortunately, for the case of Switzerland, the sampling schema was not planned in such a way because of the superposition of successive campaigns.

Nevertheless, the fact that buildings constitute the underlying sampling, can help on estimations. If a net of points formed with building locations describes the domain, we end up with a limited space, which is the real global sampling domain. This building's location net is shown in Figure 5.35 together with the actual sampling locations.

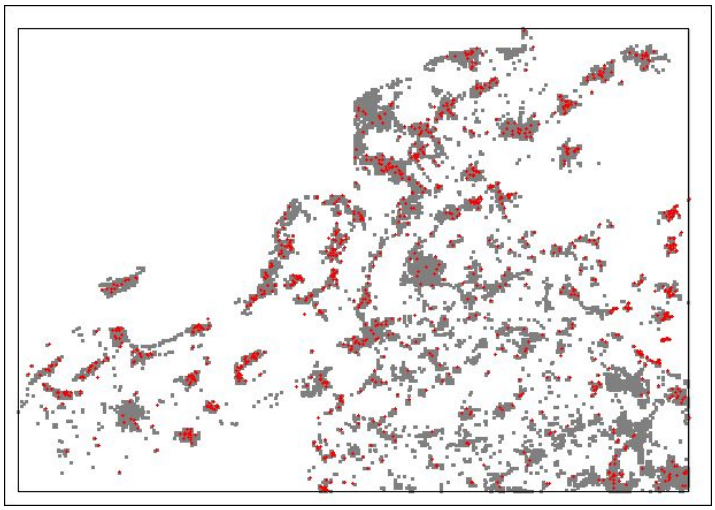


Figure 5.35: *Net of building locations (in grey) and indoor radon samples (in red)*

The question is: What would the global mean be if we had a measurement for each of these points? A possible answer is that we do not have a measurement but we can make an estimate. Estimates are of course not necessarily accurate but they can show a tendency in regards to the global domain. Map statistics, after making point estimations, can give a hint about this tendency.

A nearest neighbor and an IDW point interpolation were done using the set 3B data over the buildings net. The resulting maps are shown in Figure 5.36.

When using the NN method, the range of values (7 to 803 Bq/m3) is preserved because

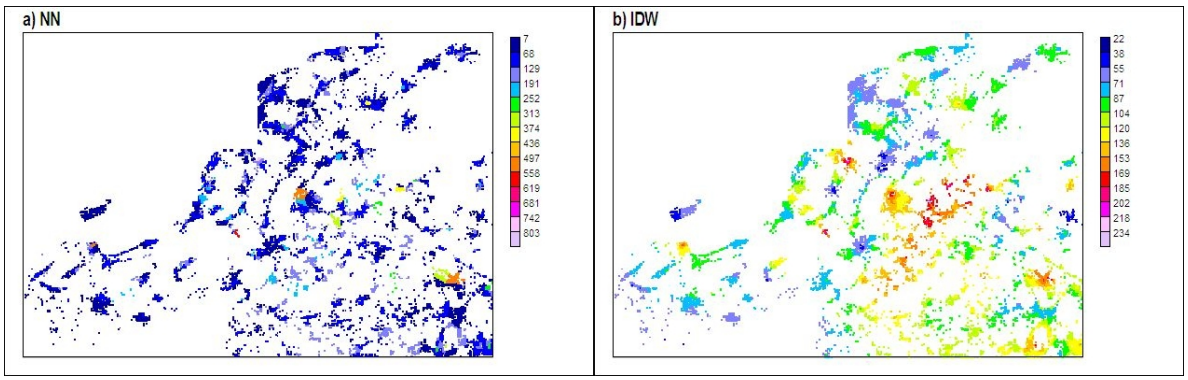



Figure 5.36: *Indoor radon map for set 3B over the building locations net domain using a) NN and b) IDW*

neighbors adopt existing values. What is interesting to observe, is that the mean of the map increased to 102.9 Bq/m3, while the mean of samples is 96.5 Bq/m3. The variance also increased from 4377 to 6714. With the IDW method, which produces smoothing, the mean also increases to 101 Bq/m3, while the variance is reduced to 895. In summary, an increment of the mean was observed by making estimations within the whole global point domain. According to these results, a tendency to a higher mean is proposed for the global histogram.

### Global mean by cell-declustering

Extreme values in indoor radon data are decisive in modifying statistics because they can reach such high magnitudes. They influence, of course, the declustering process and the modeling of the global histogram. In order to predict this influence, it is important to evaluate their clustering status before declustering. In this sense, a detailed calculation of functional clustering is necessary. In Figure 5.37b, a diagram of Quantile MI at an average distance for deciles of set3B is presented. The corresponding diagram for set3A is also presented for comparison (Figure 5.37a).

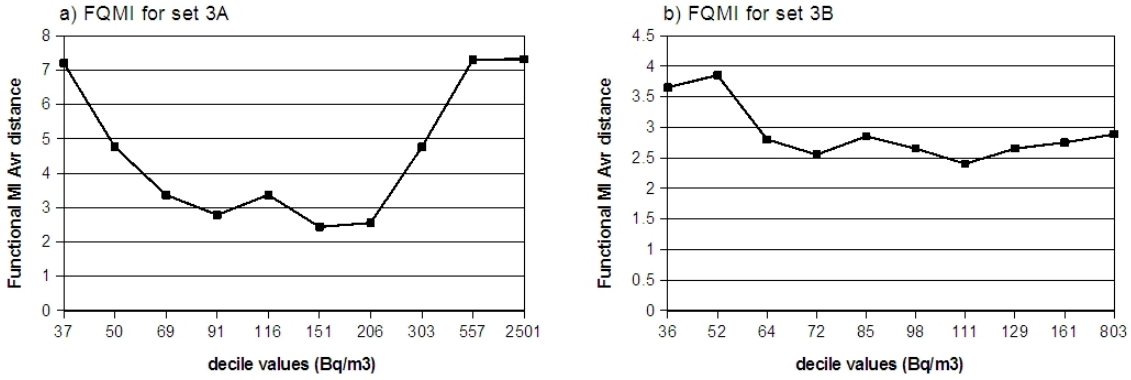



Figure 5.37: *Quantile decile MI at average distance for a) set 3A and b) set 3B*

In this detailed view of the QMI diagrams, sets 3A and 3B present the same clustering behavior as in quartiles. There is a clear tendency of spatial clustering for lower values and some clustering for very high values (the ninth and tenth deciles over 161 Bq/m3) in set3B.

Detailed mean diagrams of cell declustering for sets 3A and 3B are shown in Figure 5.38, in order to identify relations between functional clustering and declustering. Surprisingly, clustering and declustering diagrams show some similarities in shape but their lecture is not straightforward. For both sets the mean increases after declustering on short scales. Small cells create clustered domains, and then an increment in the mean is produced when high values are spread all over; in other words, when low values are clustered.

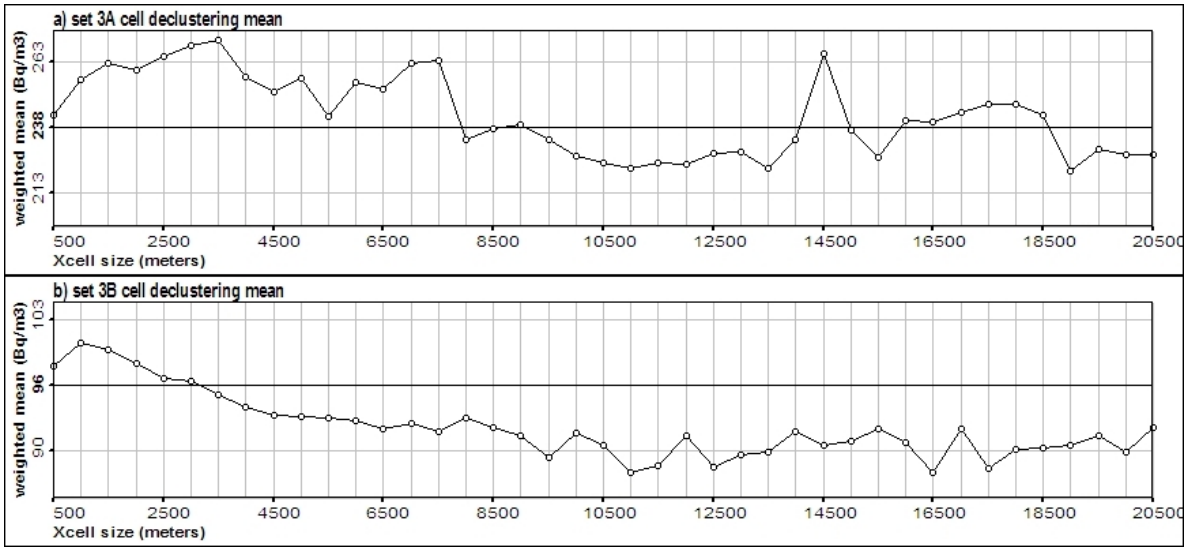



Figure 5.38: *Diagram of weighted mean after cell declustering for the a) set 3A and b) set 3B*

For set 3A, the effect of having high values clustered is an increment of the mean by declustering at shorter distances. In Figure 5.38a, the transformed mean happens to be above the actual sample mean reference line (at 238 Bq/m3) for distances between 500 and 7000 m. For set 3B (Figure 5.38b), the increment of the mean is lower and occurs within shorter distances (up to 3000 m). A detailed cell declustering mean diagram is presented for this range in Figure 5.39.

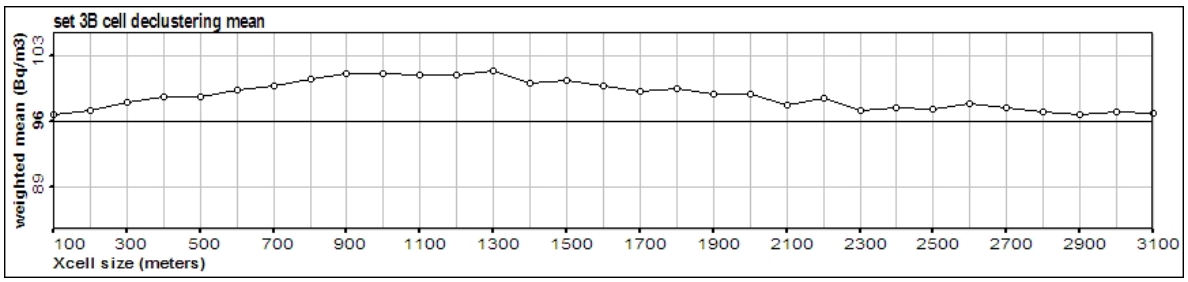



Figure 5.39: *Detail of the diagram of weighted mean after cell declustering for the set 3B*

At this point, it must be recalled that in the previous section the proposed tendency for set3B to reach the global mean was rather an increment. With the declustering diagram, it is possible to decide on the scale, producing an increment of the mean. Consequently, the recommended cell size for declustering is approximately 900 m.

**Variography and neighborhood parameters**

The urban spatial domain has an inherent spatial discontinuity and a multiscale behavior that is reflected in the variogram reproduction. Hence, the variogram model should consider some of this discontinuity. The variogram model in Figure 5.40a has up to 3 Gaussian structures fitted to an experimental model built with a tolerance of half the lag distance. This model pretends, on one side, to fit better with real data and on the other side, to express the multiscale behavior or discontinuity of the sampling domain. The structures have ranges going up to 1100, 1500 and 3300 m; with the main explained variance given to the middle range portion (till 1500 m). The corresponding coding is nug0.65+gaus0.05R1100+gaus0.2R1500+gaus0.1R3300. Using this variogram, a neighborhood optimization curve was launched with MGK (Figure 5.40b).

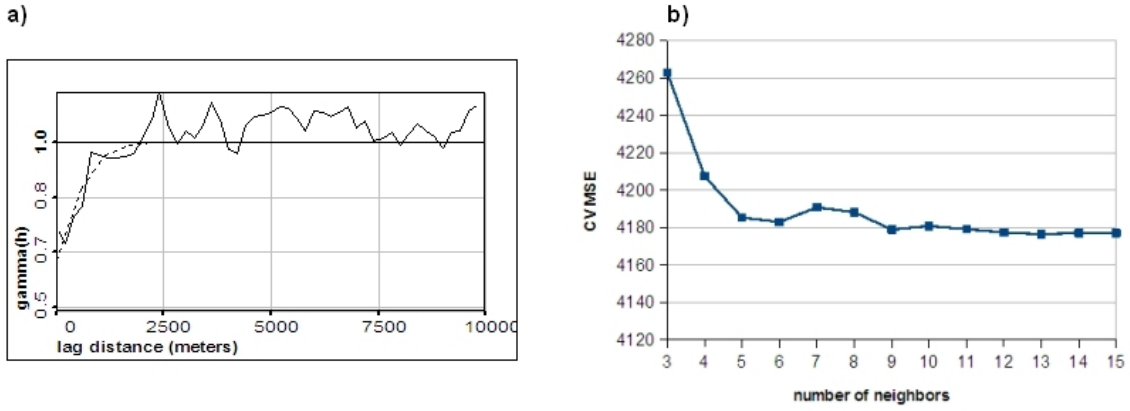



Figure 5.40: *a) Multiscale and multistructured variogram model for the set 3B b) MGK neigbors optimization curve using multistructured variogram*

The neighbors' number optimization curve is a bit smoother than the curve in Figure 5.30b, but has the same tendency. The same neighborhood range going from 5 to 32 n was, therefore, used for the SGS procedure. The variogram range, the variogram shape, the number of neighbors and the simulation net were linked together as common expressions of neighborhood.

**SGS results and mapping for an alternative scenario**

Using the defined statistical and spatial parameters, a series of 100 simulations were launched to evaluate the results and to produce some estimation maps. The variogram reproduction is depicted in Figure 5.41.

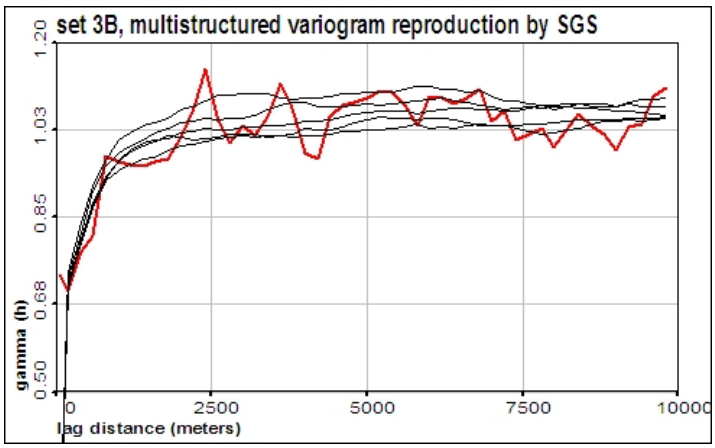



Figure 5.41: *Variogram reproduction of the set3B using a multistructured Gaussian variogram and the urban domain constrained net*

The influence of the constrained domain can be particularly perceived at a middle range distance (around 1500 m); most of the variance model is concentrated in this section. A probability map can be produced using the 100 simulations. In Figure 5.42, maps showing the probability to exceed four thresholds (50,100,200 and 400 Bq/m3) are presented.

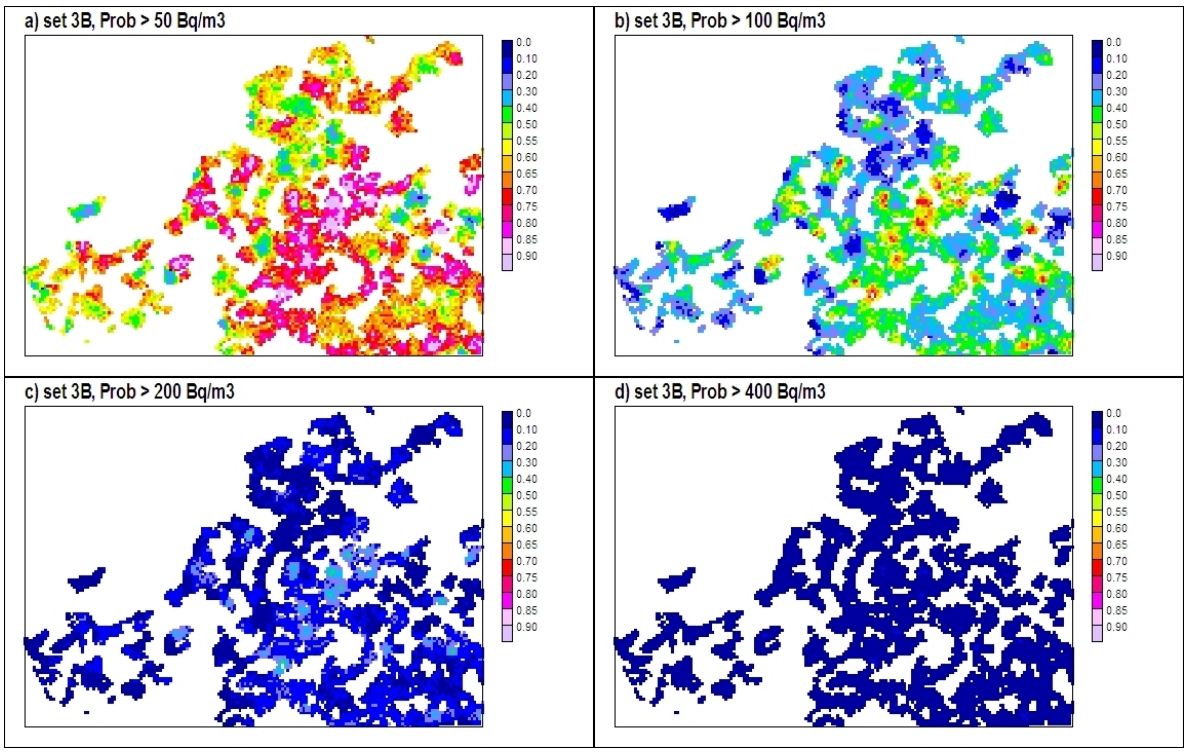



Figure 5.42: *Maps of the probability to exceed threshold values of a) 50 b) 100 c) 200 and d) 400 Bq/m3*

The simulated values obtained with the so-called alternative scenario are limited to the

imposed statistical distribution; with a mean of 101 and a maximum of 955 Bq/m3, they are a bit higher than training values. More than values itself, the desired information from simulations are the local distribution and the cumulate probability. In this example, the probability of exceeding critical values such as 400 Bq/m3 is very low. On the map, for the 100 Bq/m3 threshold, it is possible to identify some hotspots, which should be acknowledged.

Other simulation methodologies, like direct sequential simulation (DSSIM), aim to tackle the problem of histogram reproduction (60) (64). In the case of DSSIM the central idea is to avoid transformation into a Gaussian distribution and a corresponding loose of information during back-transform. The idea sounds appealing, in particular concerning extreme values. The major drawback of DSSIM is that it requires a structured variogram for raw data, because simulations are made directly from data without transformation. The option of obtaining a structured variogram model is not always available when it comes to indoor radon data. As with SGS, there is an option to propose a histogram, which should be reproduced.

## 5.5 Probability mapping with indicator kriging

Another proposed method to deal with highly skewed data is to score them into indicator values above or below a defined threshold. If a number of thresholds are set and the corresponding probability of being above (or below) is obtained with a kriging method; then, the local probability can also be constructed at each point (as was done for SGS). This method is called an indicator kriging (IK) (14).

### 5.5.1 Variography with indicator transforms

The a priori variance for indicator transforms will vary depending on the cutoff value. For indicators far from the median, one of the categories (1 or 0) will have a lower probability of occurrence; then, the a priori variance will be reduced. This could easily be calculated considering a Bernoulli type binary distribution, where the success probability is $p$ and the failure probability is $1 - p$. Then, its variance would be calculated as:

$$\sigma^2 = p(1 - p) \tag{5.5}$$

where $i$ is the indicator value and $p_i$ is its probability. Thus, for a median indicator the a priori variance will be 0.25, while for the sixth deciles it decreases to 0.24, and at the ninth deciles it is only 0.09. To illustrate this, the experimental indicator variograms for a series of cutoff values from 50 to 400 Bq/m3 for set3 are presented in Figure 5.43. Set3 is the optimal example to test improvements with variogram modeling after transformations, since its variogram of raw data is unstructured.

As seen in Figure 5.43, the variograms are quite different between indicator values. The best-structured variogram corresponds to the 100 Bq/m3 indicator, which is close to the median value (92 Bq/m3). The median indicator has the property of creating a balanced distribution between the two categories of indicators. For low and extremely high values,

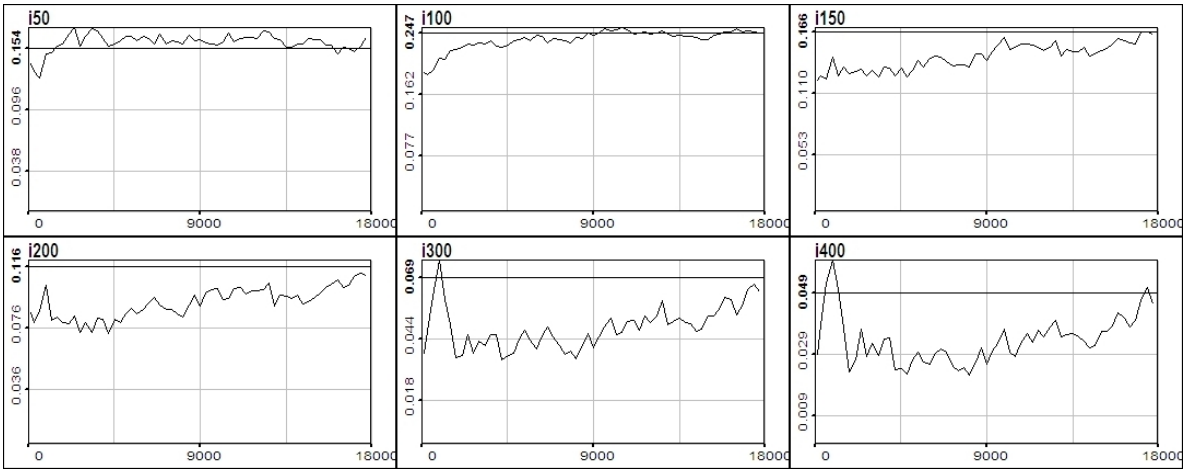



Figure 5.43: *Experimental indicator variograms for cutoff values from 50 to 400 Bq/m3)*

the spatial distribution is different. This is observed when plotting categories for the cutoff indicators 50 and 400 Bq/m3, as shown in Figure 5.44. The relatively high clustering of the category above 400 Bq/m3 (coded 1 in Figure 5.45b) can also be seen by comparing its trend diagram with the category below 50 Bq/m3 (coded 0 in Figure 5.45a). Spatial distribution and clustering can be also visualized together using Voronoi maps for indicators (Figures 5.46a and 5.46b).

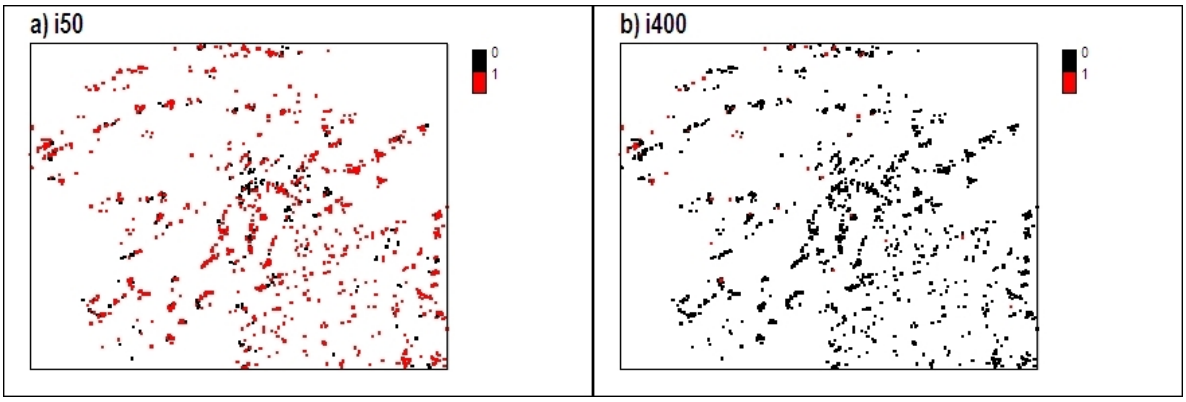



Figure 5.44: *Plot maps of indicator a) 50 and b) 400 Bq/m3)*

With the indicator 50 Bq/m3, the distribution between the two categories is more balanced than for the 400 Bq/m3 indicator. With an indicator of 50, the variance is distributed among the lag distances in the corresponding variogram (Figure 5.43) and even exceeds the a priori variance. With the 400 Bq/m3 indicator, high values are more clustered, and a large semivariance is present around the lag distance of 700 meters and reflected as a pick value in the variogram.

For the IK method, the main limitation during variography modeling is that the indicator variogram becomes unstructured, as the indicator value is far from the median value (either low or high), as can be seen in Figure 5.43. An alternative to modeling in such con-

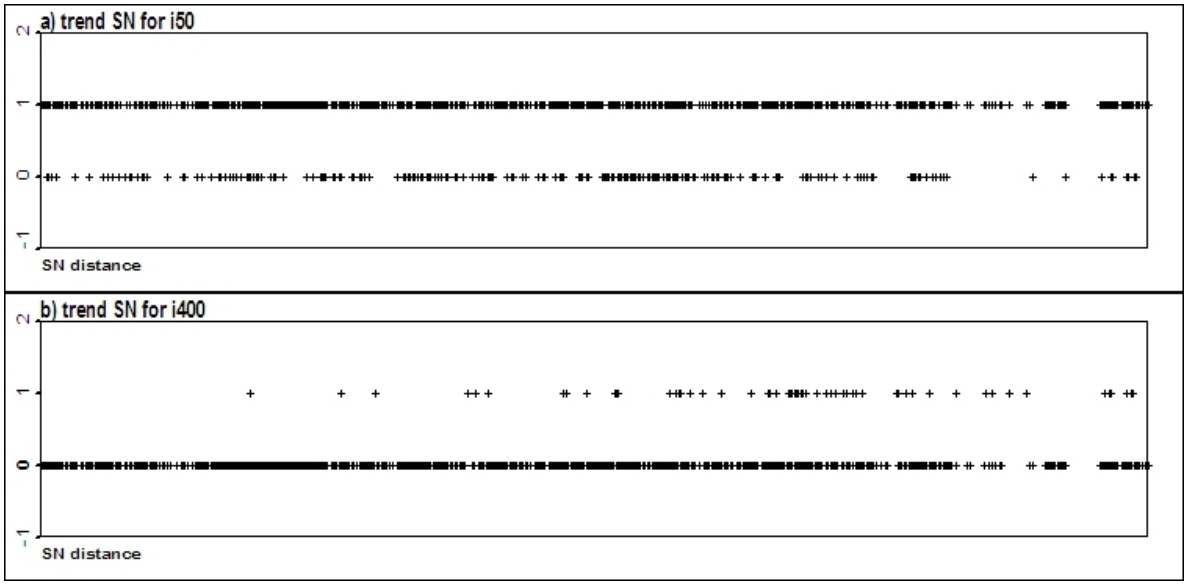



Figure 5.45: *North-South direction trend diagram of indicator a) 50 and b) 400 Bq/m3)*

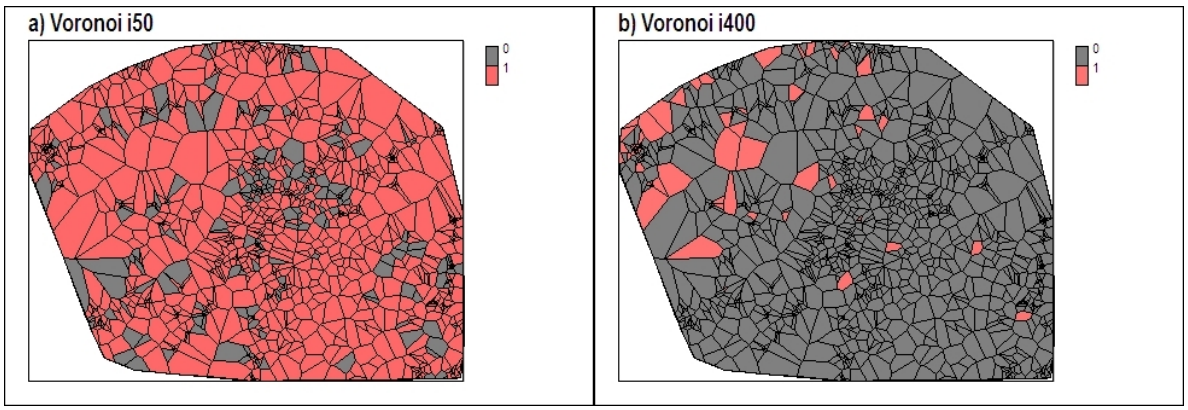



Figure 5.46: *Voronoi maps for categories of indicator a) 50 and b) 400 Bq/m3)*

ditions is to use a single well structured indicator variogram by assuming that all indicators eventually have similar spatial distributions. As stated, the median value is the one that creates a structured variogram and is evenly distributed between categories. A Voronoi polygon map can show this even distribution (Figure 5.47a), while the resulting variogram for median indicators seem structured (Figure 5.47b). It was possible to fit a model composed of nug0.18+exp0.07R4000 on top of the median variogram.

### 5.5.2 Kriging for individual indicators

The median indicator kriging approach proposes the assumption that spatial distribution has a similar configuration for all threshold values, and that the median indicator variogram may be used for all indicators. Although it seems to be an acceptable theoretical solution to the problem of the lack of structure for experimental variograms, in practice it is often impossible

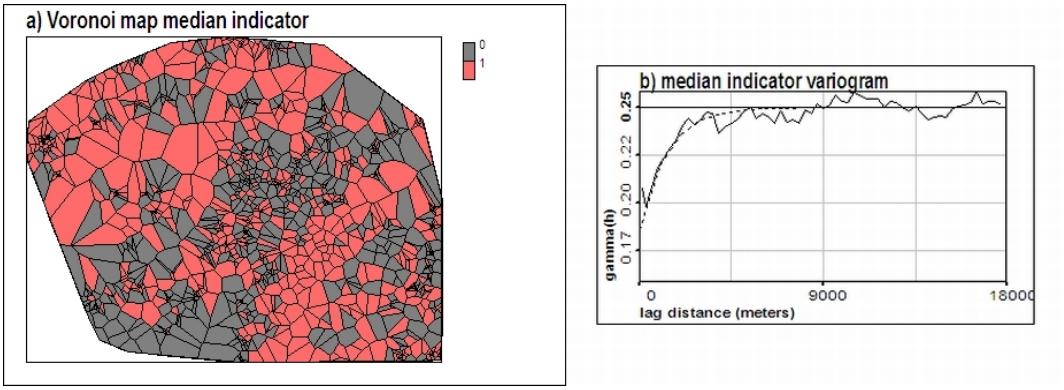



Figure 5.47: *a) Voronoi map for categories of the median indicator b) Median indicator variogram*

to preserve the ranking order to construct a local distribution. This is particularly difficult with highly skewed distributions, where variograms for higher values are very dissimilar.

In fact, the spatial structure between indicator experimental variograms would be more similar if this significant clustering were not present. The local variances of the values at certain lag distances are comparatively higher, and they produce picks in the variogram. IIf the semivariance value for a lag distance is standardized by the variance of the samples within the corresponding lag, then the influence of the local variance is filtered out. This results in standardized variograms, as presented in Figure 5.48, for the 200 and 400 Bq/m3 indicators.

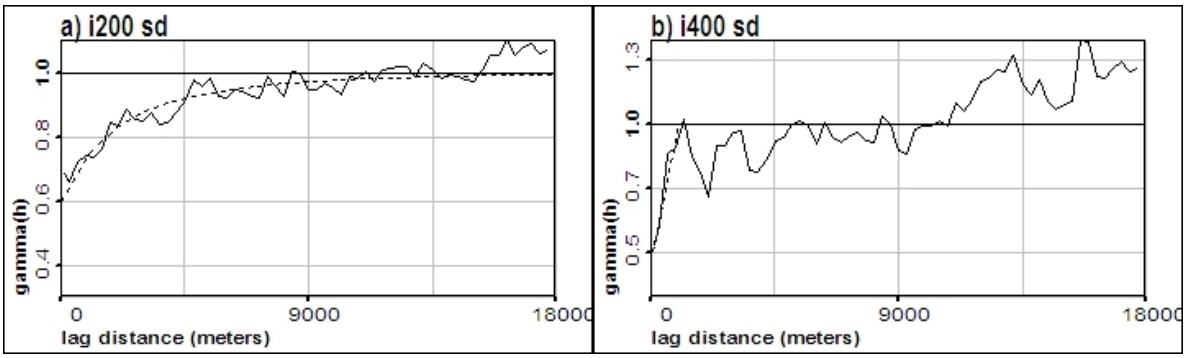



Figure 5.48: *Standardized variograms for indicators a) 200 and b) 400 Bq/m3*

This standardization also reveals a structure for high thresholds but unfortunately, cannot be used as a model for calculations. A simple alternative is to perform kriging only for the selected indicators of interest. Then, estimations would be focused on critical thresholds. The ultimate idea is not to construct the local distribution but to have an indicator probability as the estimation. In this case, the indicator is not set to 1 and 0 for indicators below and above the threshold, as is the case when a cumulative distribution ($P < z$) is constructed but to 0 and 1, as drawn in Figures 5.44. This can be interpreted as the probability of occurrence of being either below or above the selected indicator. The exceeding value has the indicator value 1 or the maximum probability to exceed the threshold.

For indoor radon set3, the threshold analyzed is the median threshold, since it is the only one that presents a structured variogram. The median of set3 is 92 Bq/m3, which is close to the actual critical threshold of 100 Bq/m3. The mapping results are presented in Figure 5.49, and below these maps is a map for the median threshold using SGS (Figure 5.50).

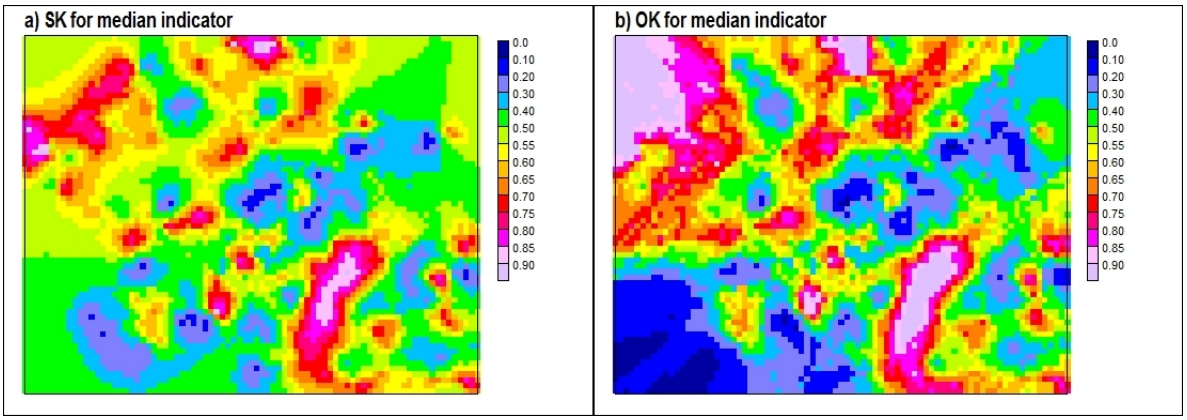



Figure 5.49: *Mapping of set3 for the median cutoff for a) simple and b) ordinary indicator threshold kriging*

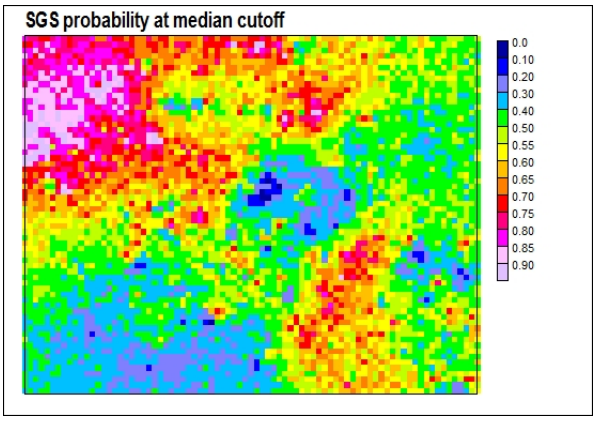



Figure 5.50: *Mapping of set3 for the median cutoff using SGS*

The probability map with SGS seems to have more similitude to ordinary indicator kriging than to simple indicator kriging. For higher thresholds of interest, such as 200 and 400 Bq/m3, the indicator method is probably more applicable when using stationary data like set3B.

### 5.5.3 Indicator probability mapping for set3B

For the indoor radon set3B, the thresholds of interest will be 100, 200 and 400 Bq/m3. The 1000 Bq/m3 cutoff seems to be more difficult to model, because of the few measurements exceeding this limit. As a first step, the corresponding variograms have been calculated for a series of thresholds (Figure 5.52).

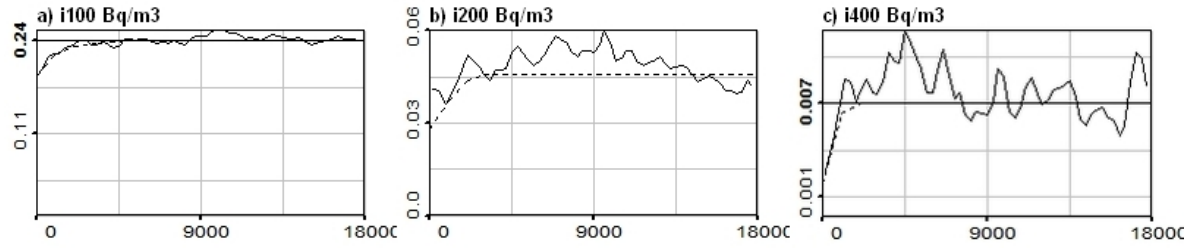



Figure 5.51: *Variograms of set 3B for indicator thresholds a) 100 and b) 200 and c) 400 Bq/m3*

The variogram model for indicator 100 Bq/m3 was nug0.19+exp0.0473R3250; for indicator 200 Bq/m3 was nug0.029+sph0.018R3000, and for indicator 400 Bq/m3 was nug0.002+sph0.047R1500. Two maps were produced using indicator OK for 100 and 200 Bq/m3 (Figure 5.52).

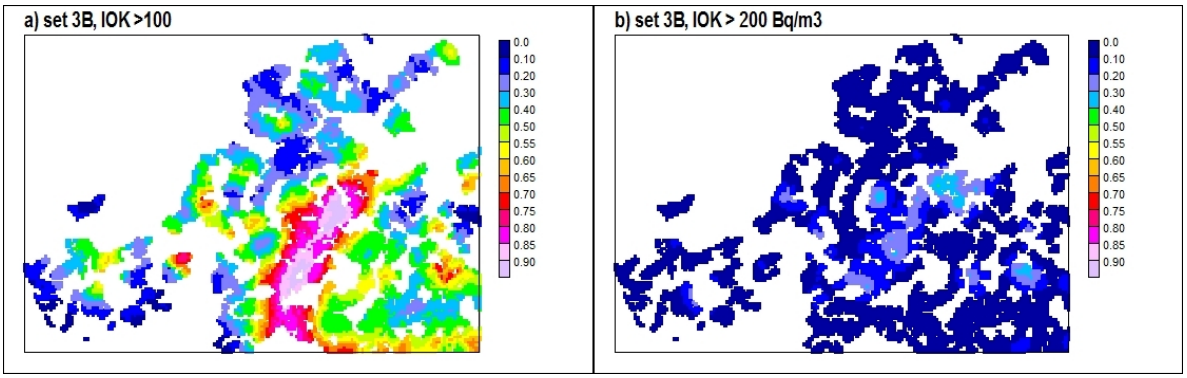



Figure 5.52: *Mapping of set3B using OK for a) 100 and b) 200 Bq/m3*

Once the IK interpolation has been performed for unsampled points, the resulting prediction will be made on a scale from 0 to 1, and p decision levels will be chosen to decide whether areas are considered to be below or above the threshold. Results have been compared to SGS probability maps because they are based on histogram reproduction and deemed more realistic. When looking at the corresponding maps calculated with the SGS method (Figure 5.42c), it is possible to see more similitude between the two methods for the 200 Bq/m3 threshold. For the 100 Bq/m3, it is clear that a generalization is produced by the indicator method and that probability values are higher than those on the SGS map (Figure 5.42b).

For the 100 Bq/m3 threshold, the OK for indicators has a pessimistic approach to probability mapping. If a pvalue of 0.7 is taken as critical limit, the OK indicator map has large areas that would be declared under exposure in comparison to SGS. Indicator kriging is a simpler method than SGS and offers a solution to the modeling of extreme values, but results are less accurate.

## 5.6 Classification methods for probability mapping

A tough feature to deal with regarding regression and simulation estimations of continuous values for indoor radon data was the high local variability and spatial clustering. Categorization of data and the use of robust methods may simplify this task, as proposed by Kanevski and Chaouch (37) (7).

The indicator kriging method exposed below is a first approach to making classifications by creating two categories based on thresholds of interest. Given a continuous variable and a set of decision thresholds, the problem of decision-oriented risk mapping was considered to be a classification one. Then, only the areas where the concentration of indoor radon exceeds a predefined decision threshold have to be determined.

Although IK is a sound geostatistical solution for making estimations over ranked data, it has some limitations. An important drawback is that variogram modeling may fail if the number of instances of some class is not large enough; this is a common situation for high value thresholds and hot spots. In such circumstances, the use of a structured model, such as the median indicator variogram, will not perform well in all cases. In (7), similar validation errors were obtained with indicator kriging, sequential indicator kriging (SIS) and Support Vector Machines (SVM). The threshold used in that research was close to the median value, which simplified the modeling task. The goal of modeling unbalanced categorizations resulting from high indoor radon thresholds was addressed in (71) (72).

Another limitation of geostatistical methods is that the calibration and optimization of parameters involves lot of time investment and a certain level of knowledge about the phenomena. The typical case involves the modeling of variograms for kriging and simulation procedures. Based on the previous research presented in this chapter, it should be acknowledged that simulations provide the most complete information and that some proposals were made regarding automation of kriging procedures. Nevertheless, the use of geostatistical methods is not always applicable because intrinsic conditions, such as stationarity or multi-Gaussianity, do not hold for all data. In mapping practice, it would be beneficial to have methodologies that work for all types of data configuration and that are not so exigent in terms of parameterization.

There is also a need for methods that handle data in a semi-automatic mode and produce a series of maps on a national level. Additionally, it must be considered that maps are required to be updated when new data are collected. Regional classification methods can be helpful tools for automatic modeling and prediction because they do not require expert knowledge injection.

Machine learning classification methods are more robust and simple to apply than kriging in the sense that they work with data categories. Moreover, they are more autonomous because they do not require expert guided variogram modeling, and they could provide an easy way to tune their parameters using cross-validation and other error minimization methods.

For this research, the use of a simple classification method, such as K Nearest Neighbors (KNN) has been proposed, as well as the use of some more complex method originating

from statistical learning, such as probabilistic neural networks (PNN) and support vector machines (SVM). KNN has been used as a benchmark model for the comparison of results. The dataset used will be set 3 split into training and validation sets.

### 5.6.1 K Nearest Neighbors (KNN)

KNN is the equivalent method of KNNR for classification as exposed in chapter 3. It is a simple method of data classification, where a category is assigned to every unsampled location based on the category of K-neighbors. The unique parameter is then the number of neighbors, which makes the method more data-driven. The most effective way of finding the optimal number of neighbors to be used is by performing a cross-validation. Prediction error is computed for different number of neighbors by leaving out one sample at a time, and the optimal number of neighbors is reached when the minimum error is attained. The class in the unsampled location is obtained by simply using a majority criteria; the label of the most frequent class among the neighbors is assigned. Calculating the mean value of the K neighbors is also an option. Subsequently, a classification is executed considering a decision level for the hardening of data.

### 5.6.2 Probabilistic neural networks (PNN)

The PNN method is closely related to GRNN (presented in chapter 4) (42). It also uses the Parzen-Rosenblatt density estimator and Bayesian formulations, but in a different way. In this case, we are only interested in the marginal distribution of the conditional data $x$ or the $f_X(x)$ component of formula 4.25:

$$f_Z(z \mid x) = \frac{f_{X,Z}(x,z)}{f_X(x)}$$

Where $X$ will be noted as $C$ because they are in fact subsets or categories of data from the whole set. So, $c_i$ will be a series of subsets labeled as a class $i = 1$ to $K$ classes. In other words, by PNN we are interested in the probability density function of $x$ given a certain class $c_i$ or, in other words, the probability of $x$ belonging to a class $c_i$: $p(x \mid c_i)$.

A Parzen-Rozenblatt density estimator of a class set ($f_C(x)$) is defined as:

$$\hat{f}_C(x) = \frac{1}{\sigma^{p+1} N} \sum_{i=1}^{N} K\left(\frac{x - x_i}{\sigma_x}\right) \tag{5.6}$$

Then, using a Gaussian kernel, the probability of $x$ belonging to a certain class $c_i$ will be expressed as:

$$p(x \mid c_i) = \frac{1}{(2\pi\sigma^2)^{p/2} N_i} \sum_{i=1}^{N_i} \exp\left(-\frac{\|x - x_i^{(n)}\|^2}{2\sigma^2}\right) \tag{5.7}$$

where $N_i$ is the size of the class, i.e. the number of samples belonging to class $c_i$, and $x_i^{(n)}$ is the $n^{th}$ sample of class $c_i$. Equation 5.7 is in fact equivalent to the input density estimator of GRNN (4.31) but for one class at a time.

Having defined the Parzen-Rosenblatt *pdf* model for each class, it is possible to decide to which one of the $K$ classes an unknown sample $x$ belongs. For this purpose, the Bayesian optimal or maximum a posteriori (MAP) decision rule is applied:

$$C(x) = \{c_1, c_2, \ldots, c_k\} = \quad \text{argmax}_{c_i} \quad P(c_i)p(x \mid c_i) \quad i = 1, 2 \ldots, K \tag{5.8}$$

The decision regarding class membership, for an unknown $x$, can be made by selecting the higher-class probability. Due to the assumption that any sample $x$ can have a degree of membership in each of the $K$ classes, the Bayesian confidence $P(x \mid c_i)$ (a posterior probability of $x$ belonging to class $c_i$) can be estimated as:

$$P(x \mid c_i) = \frac{p(x \mid c_i)}{\sum_{k=1}^{K} p(x \mid c_k)} \tag{5.9}$$

Additionally, if it is believed that one class is more probable to occur (not because of preferential sampling but because of the phenomenon), then it is possible to introduce the prior probability of the class $P(c_i)$. This prior probability class membership can be considered during the decision rule (as noted in equation 5.8) as well as for the posterior probability as:

$$P(x \mid c_i) = \frac{P(c_i)p(x \mid c_i)}{\sum_{k=1}^{K} P(c_i)p(x \mid c_k)} \tag{5.10}$$

As seen, two products can be obtained with PNN mapping: a classification map and the maps of posterior probability memberships to the defined classes.

### 5.6.3 Support Vector Machines (SVM)

SVM is a contemporary technique introduced and specially tailored for approaching binary classification problems. It is theoretically founded by Statistical Learning Theory (77), which states that a predictive model should have an optimal complexity for a particular dataset in order to generalize it. When proposing a model, a level of complexity must be considered until the limit where training and testing errors will not compromise predictions. The structure of the model can be so complex that it over fits data or so simple that it over smoothes the results. An optimal model should consider both complexity and fit to the available training data. This approach relates to the principle that constitutes the Tikhonov theory of regularization in the context of linear models (2).

The structure of a model provided by SVM is defined by the most discriminative samples of the dataset (support vectors). They delineate a decision function to divide data into different categories by maximizing the margin between them. Given a training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$ of $L$ samples, a linear classifier decision function $f(x)$ can be defined as:

$$f(x) = \sum_{i=1}^{L} y_i \alpha_i K(x_i, x) + b \tag{5.11}$$

where $K(x_i, x)$, is a kernel function as the one used for GRNN and PNN and is based in the best structured points or support vectors $x_i$. $y_i$ are class labels that can be defined as $(y_i \in \{+1, -1\})$ for a binary classification task. $b$ is a constant and $\alpha_i$ are the weights.

The kernel function is often a Gaussian RBF of some sigma $\sigma$ width, which needs to be tuned. The most important parameter to tune is the regularization factor $C$, which penalizes miss-classification of training samples. Practically, this is the upper bound of the weights, that is $0 < \alpha_i < C$. A high $C$ value will give more penalization when data appears to overfit, and the value of $C$ is usually lower for noisy and uncertain data. Optimal values for both kernel parameters and $C$ can be chosen based on the number of support vectors and the testing or cross-validation error, which estimate the generalization ability of the model. In general, best results are obtained with values that produce low testing error by using the simplest solution, i.e. the low number of support vectors.

### 5.6.4 Evaluation error and mapping

For the present classification task only two categories were defined: below or above a critical threshold. In Figure 5.53, there is a table for the error matrix for threshold classification. For indoor radon, the positive category was defined as the samples exceeding a critical threshold. Viewed in an error matrix the positive events will correspond to predicting the class over the threshold (Figure 5.53).

| | Actual value | |
|---|---|---|
| Outcome | Class > Threshold (P) | Class < Threshold (N) |
| Class > Threshold (P) | TP | FP |
| Class < Threshold (N) | FN | TN |

Figure 5.53: *Error matrix validation table defined for radon regional classification task*

A common measure of error derived from this table is the total error classification TE, which is the sum of false negatives (FN) and (FP) divided by the total number of validation samples. A problem of interpretation can arise with this error measure for critical thresholds that are well above the median value. When categories are unbalanced (one has more data than the other) the total error could be masked by the class size. It could have a very low total classification error simply because the class over the critical value is represented by very few samples and the number of FN is low.

If we choose a pessimistic approach, the task of the classifier would be to identify as many of the places as possible where the threshold is being exceeded. That would be equivalent to increasing the true positives TP and decreasing the FN. These measures solely involve the critical class, and the error will be expressed as the false negative ratio (FNR), which is calculated as:

$$FNR = \frac{\text{FN}}{\text{TP+FN}} \tag{5.12}$$

Other tasks could be chosen as well, such as the correct detection of the lower threshold class. In this case the goal of the classifier would be to reduce the FP ratio. This would be

an optimistic approach related to safety decisions. As aforementioned, the pessimistic interpretation was chosen since it is related to the prediction of critical situations, remediation measures and resampling campaigns.

### 5.6.5 Pessimistic decision levels

An additional difficulty in the classification of unbalanced categories is that the resulting probability scores at unknown locations are very low for the underrepresented class. If the maximum a posteriori (MAP) decision rule 5.8 is used to make classifications, the category often simply does not appear in the results. This is particularly true for methods with high generalization capabilities like PNN. An alternative proposed here, in order to force an area to be classified as the above threshold category, is to use a concept for hardening based on mapping areas.

As seen with the SGS and IK methods, the resulting probability maps can have dissimilar scales, and the decision levels are subject to interpretation. Depending on the wish to create a more optimistic or pessimistic scenario, a low or high pvalue will be chosen. Consequently, the extension of a class area in a map will vary. In any case, what is relevant is to obtain a shape and a distribution of the declared area that corresponds most with reality. Additionally, the extension of the class area will also vary depending on the method used. For the purpose of the present study, which aims at making an intercomparison of methods, a common evaluation task has been defined.

The decision functions maps resulting from KNN, PNN and SVM can be hardened to force a certain percentage of the map to be declared above the threshold. In this way, methods can be compared based on equal areas per category. For instance, to obtain half of the area for a class, the decision level will correspond to the median value of probability in order to belong to that class. Since we only work with two categories, the other half of the area will correspond to the second class. Half the area will be a realistic selection for thresholds that are close to the median value. For thresholds far from the median it is logical to expect a smaller area to be declared above the limit; in this case, a quarter of the area could be declared to be above the critical value by using a quartile decision level.

### 5.6.6 Indoor radon classification results

As mentioned, set3 consists of 1710 samples subdivided into a training set with 1310 values and a validation set with 400 values. For the SVM method, the training test was subdivided into a proper training set of 1110 and a test set with 200 samples.

In order to define categories, the datasets were classified into below and above threshold category. The thresholds considered are legal limit values for indoor radon intervention: 100, 200, 400 and 1000 Bq/m3. Whether a certain level of indoor radon constitute a health risk, is a subject of constant debate; it is more suitable to talk about action levels, as they are legally defined. In Switzerland, remediation actions are defined for dwellings that exceed 400 and 1000 Bq/m3, while prevention actions are considered for 100 and 200 Bq/m3.

The KNN and PNN methods require having categories coded as positive integers; thus, categories have been assigned labels 1 and 2. For the SVM method, coding of categories are -1 and 1. In Figure 5.54, it can be observed that the spatial distribution of categories changes according to the selected cutoff. The over 200 Bq/m3 category is distributed more homogenously throughout the zone while the over 1000 Bq/m3 class is concentrated to the northwest, creating 'hotspots'.

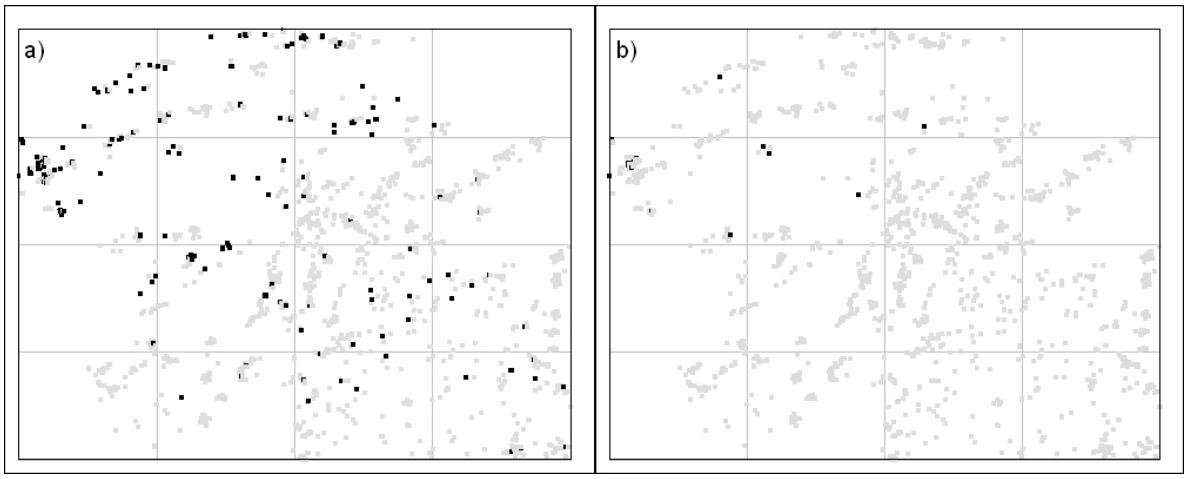



Figure 5.54: *Distribution of indoor radon samples after classification based on a) the 200 Bq/m3 and b) 1000 Bq/m3 thresholds. Dots in bold indicate the above limit category.*

After categorization of the training data we have a different realization of the spatial distribution depending on the selected cutoff. The results of interpolation using these different sets will each produce a classification map representing zones where there is a high probability of exceeding the corresponding critical value. Such probability maps, or decision function maps, can be subsequently used to decide which areas can be declared to be above a certain level. These decision function maps can be also useful in incorporating the uncertainty of predictions, as was done in the simulations.

### 5.6.7 Classification using KNN

The KNN method was run for each threshold set using an optimal number of neighbors obtained by cross-validation (CV) with the leave-one-out method (Figure 5.55). The prediction was done for each point in a grid using regression. Regression was used because it provides continuous values, which makes it possible to decide upon a convenient level for classification. The purpose of doing this is to make a comparison with other methods producing decision functions.

The classification of results was made using different decision values. For the 100 and 200 Bq/m3 thresholds, the median was used as the decision level, to produce a critical area covering 50% of the total area. For 400 and 1000 Bq/m3, the decision level used was the one that covered a quarter of the area. A map representing the result with KNN regression and a

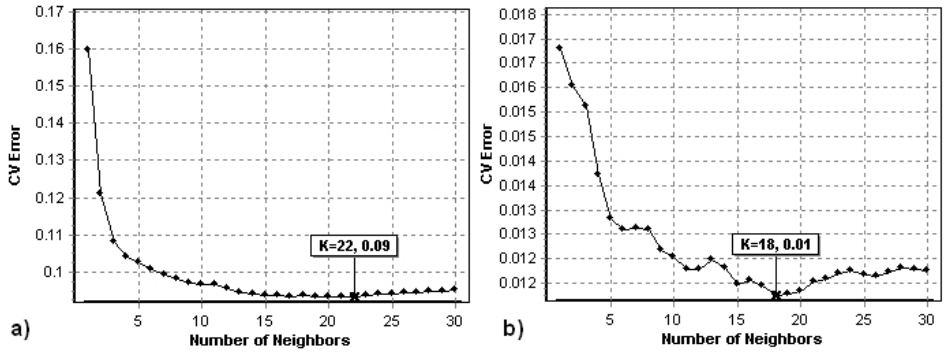



Figure 5.55: *Graphs of cross validation for KNN method to find optimal neighbor numbers for a) the 200 bq/m3 cutoff and b) the 1000 bq/m3 cutoff.*

subsequent map of exposure for the 200 Bq/m3 threshold are presented in Figure 5.56. The validation samples with values exceeding the threshold are superimposed to the classified map to perceive the level of false negatives or omission error (crosses that fall within the below threshold category).

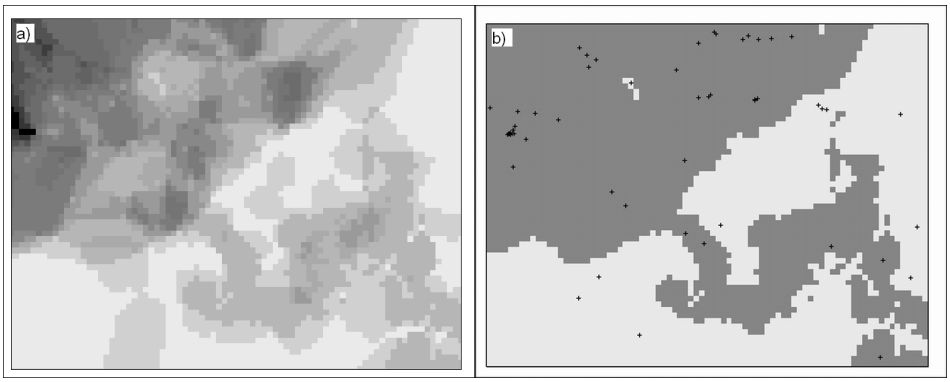



Figure 5.56: *a) decision function map using KNN regression for the 200 Bq/m3 threshold and b) classified map produced with a median decision level. Bold areas indicate a higher probability of exceeding the threshold.*

Figure 5.57, shows the classified maps for the 400 and 1000 Bq/m3 limits. The 400 Bq/m3 decision class map has an area declared to be above limit that covers 34% of the area. In this case, it was not possible to produce a map with exactly a quarter of its area having more probability of indoor radon exposure, due to the limited detail of predictions. The 1000 Bq/m3 critical classified map has an area above the limit covering 19% of the total area. The resulting omission errors (the FNR) are presented in a comparison table together with the results for other methods (Table 5.1).

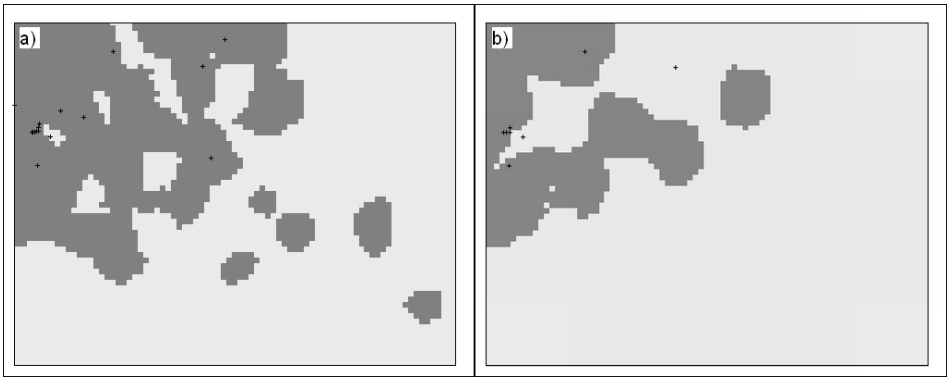



Figure 5.57: *KNN classification maps representing areas with a higher probability of exceeding a) the 400 B/m3 limit and b) the 1000 B/m3 limit.*

### 5.6.8 Classification using PNN

The PNN method has an advantage over KNN when defining a model of values' distribution taking into account density and anisotropy. A single parameter sigma is acquired and optimized through cross-validation for each threshold classification. The algorithm calculates the probabilities of belonging to either the below or the above class. For our application, which emphasizes the critical area, the probability of belonging to the above limit category was used as the decision function for classification. As for the previous method, decision levels were selected with the scope of obtaining a fixed area with higher probability to be above the limit.

Training of the model consisted on CV error minimization. Optimization of sigma was done considering two dimensions, to account for spatial anisotropy. Gradient descent was used to make the search more precise within the MLO software. For instance, an optimal sigma of 1535 by 1155 meters was obtained for the 200 Bq/m3 threshold. In order to calculate the class probability, the class size was used as prior probability. In Figure 5.58 the probability map and the classification decision map for this limit is presented. Figure 5.59 shows the classification maps for the 400 and 1000 Bq/m3 limits obtained with the PNN method.

### 5.6.9 Classification using SVM

Classification using SVM is more complex in the sense that it requires the optimization of two parameters: the kernel parameter and the regularization parameter $C$ (to allow for misclassification or uncertainty). Three levels of regularization (10, 100 and 10000) were tested to find the best results. The 10000 $C$ level gave less error and therefore was fixed as a constant parameter for all thresholds. Indeed, the high level of local variability of indoor radon values tends to over-fit with models; a higher level of generalization seems to improve modeling.

The sigma parameter for the kernel was optimized considering the structural and empirical risks as explained earlier. This implies the choice of a kernel size for which a good compromise between the minimum number of support vectors and the minimum testing

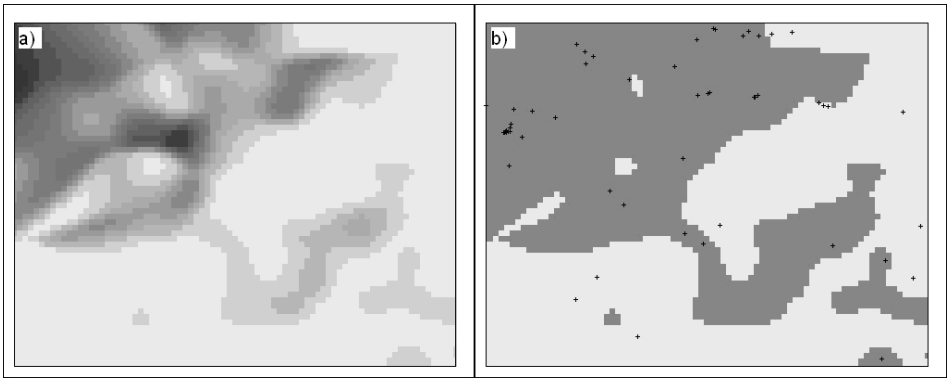



Figure 5.58: *a) Decision function map using PNN for the 200 Bq/m3 threshold and b) classification map produced with a median decision level. Bold areas indicate a higher probability of exceeding the threshold.*

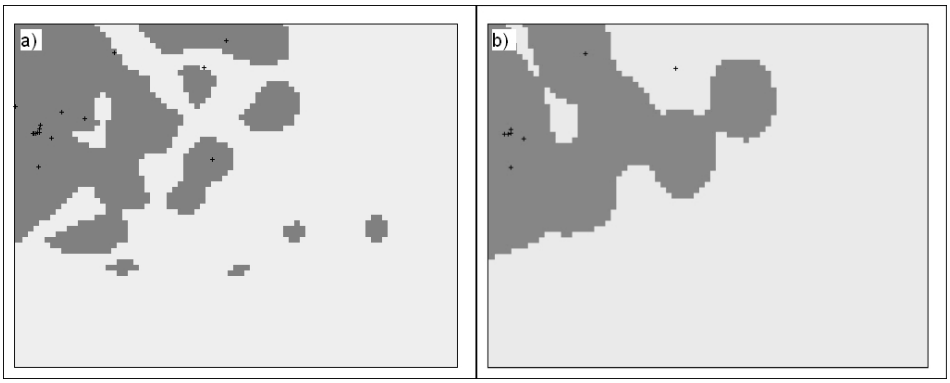



Figure 5.59: *a) PNN classification maps using the upper quartile decision level representing areas exceeding the 400 Bq/m3 limit and b) the 1000 B/m3 limit.*

error are attained. To find this compromise, the normalized support vectors (NSV) number was summed to the CV testing error. Figure 5.60 shows the testing error and NSV curves for tuning the kernel size when classifying at the 200 Bq/m3 cutoff.

In Figure 5.60, the training error simply increases as the kernel size enlarges. Using this criterion, different optimal kernel sizes were obtained for the threshold limits. For instance, the optimal kernel size for the 200 Bq/m3 threshold was 1600.

The SVM algorithm generates a decision function, which indicates the relative membership to one or another category. This function was used to classify data according to the research task of defining areas where there higher probability of exceeding the legal thresholds. Figure 5.61 shows a decision function map for 200 Bq/m3 along with its corresponding classification map.

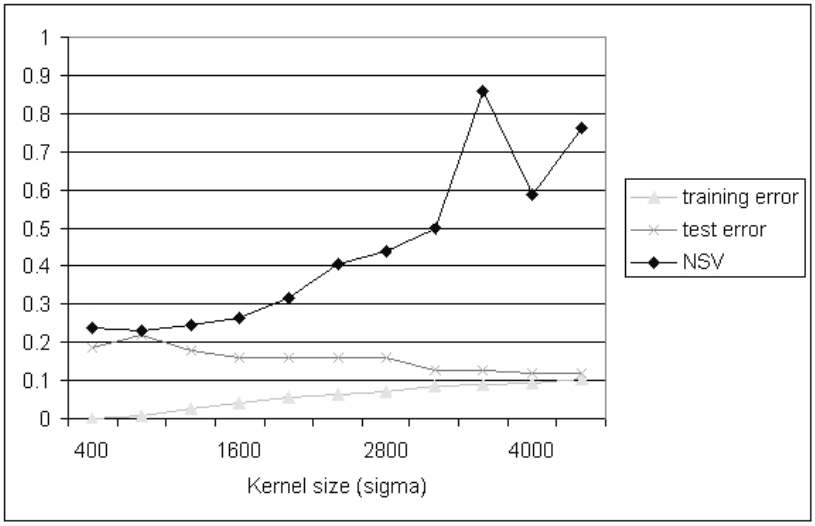



Figure 5.60: *Tuning of parameter sigma, considering NSV and test error, for the 200 Bq/m3 threshold.*

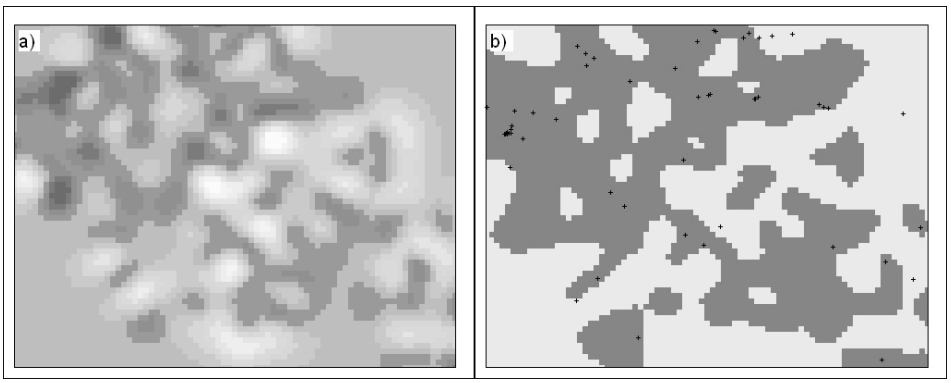



Figure 5.61: *a) Decision function map using SVM for the 200 Bq/m3 threshold and b) classification map produced with a median decision level. Bold areas indicate a higher probability of exceeding the threshold.*

### 5.6.10 Results comparison

The omission error for the category above limit was computed for every threshold and plotted together into a comparison table (Table 5.1). For the lower thresholds (100 and 200 Bq/m3), the SVM method performed a better classification, while for the higher cutoffs, the PNN method gave lower omission errors. Results with the KNN method were not far from the other methods.

As a way of comparison, is possible to produce a classification map out of the SGS probabilities. The z-cut 200 Bq/m3 SGS map for set3 was hard-classified according to the pessimistic approach criterion, and the result is presented in Figure 5.62.

The interpretation of the classification maps should be done carefully. One must be aware that **the error measures are all relative to a fixed mapping area**, and they have the

Table 5.1: *FNR error for classification using KNN, PNN and SVM at 4 levels of threshold categorization. Half of the area is declared above the limit of 100 and 200 Bq/m3 and a quarter of the area above 400 and 1000 Bq/m3*

| Method | Error (%) per Threshold value | | | |
|---|---|---|---|---|
| | 100 Bq/m3 | 200 Bq/m3 | 400 Bq/m3 | 1000 Bq/m3 |
| **KNN** | 25 | 21 | 6 | - |
| **PNN** | 35 | 21 | 6 | 13 |
| **SVM** | 18 | 15 | 12 | 38 |

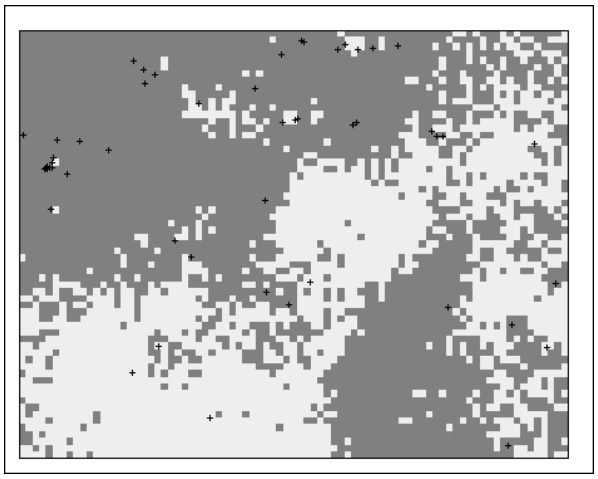


Figure 5.62: *Classification map of the SGS zcut 200 Bq/m3 for set3*

sole purpose of comparing results of methods. In addition, hardening of data has not exactly given the desired proportion of area due to a precision limitation in the decision function values.

## 5.7 Conclusions about SGS and IK

SGS realizations provide the option of deriving the local probability of exceeding critical thresholds, which constitutes a valuable indoor radon exposure assessment. The SGS basic assumption is the multi-Gaussian joint distribution of RVs. MGK can be seen as a simplification of the SGS method or a single realization of this multi-Gaussian distribution. MGK was used as a rapid-mapping tool, and as a referential method to optimize the neighborhood parameter for SGS.

As heavy spatial clustering is present in data, the neighborhood parameters appeared to play a relevant during modeling. High spatial clustering of indoor radon in Switzerland is a consequence of the urban area fragmentation. So, the use of a simulation net constrained to the urban area was also considered as a neighborhood parameter because it constitutes the global domain for indoor radon samples.

A first trial of the SGS method was performed on the national scale. In this case, a MW filter was applied in order to reduce local variability and to obtain a structured variogram

model. It also produced a reduction of data which helped speed-up simulations for a group of samples as large as the national set. The use of a constrained simulation net also reduced the calculation time. This test has shown, throughout variogram reproduction, that the long-range structure (80 km) explains most of the samples' spatial variability.

Set3 was used next to analyze the influence of clustering and neighborhood parameters. the number of maximum and minimum neighbors, as well as three simulation nets, were tested. Results showed that fluctuations of realizations adjust better to the variogram model when using a neighborhood between 40 and 60. For the neighborhood definition a minimum $n$ has the effect of smoothing the local variance, while the maximum limit has the effect of reducing time calculations.

In general, the reproduced variograms don't adjust at the middle range using a constrained net made out of buffer zone around samples. The use of a large number of neighbors and an enlarged simulation net gave a better reproduction of variances at long ranges, while the constrained nets and a low number of neighbors helped model the short range variograms. In general, a bounding box net appeared good enough to reproduce the variogram for set3.

In the sequential play of SGS, the level of conditioning increases with the number of nodes considered in the net. SGS is a process of gradual space filling and therefore a gradual change in neighborhood is produced. A bounding box simulation net can contribute to variogram reproduction at all ranges because of the continuity of points which works better with the sequential play. Alternative variogram models expressing some discontinuity were also proposed.

Regarding the uncertainty of variogram and histogram reproduction, the following objective was to compare two possible simulation scenarios that are either more conservative or more alternative. With a more conservative way of thinking, the information from samples and the derived models were assumed to be well enough to explain the process. With the hypothesis that the samples are not totally representative of the phenomenon, the simulation task was enlarged to the definition of an alternative target variogram and histogram. Set 3B was used for this case.

For the conservative scenario, the filtered and non-filtered data were compared. The objective was to orient simulations to the automatization of procedures. For instance, the optimal number of neighbors obtained with MGK CV optimization provided a better corresponding variogram reproduction.

For the alternative scenario a spatial distribution of population elements considering the natural domain (the urban area) was proposed. Also a multi-structured variogram was proposed to express a scale discontinuity of the phenomenon. Regarding the histogram reproduction, a global histogram scenario was proposed, considering the natural sampling domain, which has buildings as the support size. Interpolations with NN and IDW indicate a possible increment of the global mean for the case of wholeness sampling. Based on this tendency a weighting of the available samples was proposed using cell declustering. The weighting transform was used to define a target histogram for SGS.

Indicator kriging for thresholds of interest was proposed as another method to be used for probability mapping. This was a simplification alternative to the IK histogram reproduction to tackle the problem of unstructured variograms for higher cutoffs. Even though the IK method is an alternative for extreme values modeling, results are very coarse compared to SGS. The ordinary IK appeared to produce a more realistic mapping than simple IK. Probability maps with IK at lower thresholds are more pessimistic than SGS probability maps. Probability maps are valuable tools for decision-oriented indoor radon mapping because they provide the option to decide whether or not a critical value could be exceeded (decision maps).

## 5.8 Conclusions about classification methods

Categorization of data and the use of robust methods may simplify the task of prediction and classification of indoor radon data. Regional classification has the potential for mapping large datasets in automatic mode.

Classification methods, such as KNN, PNN and SVM obtain optimal parameters through cross validation. KNN is a simple method that requires only the optimal number of neighbors as the parameter. PNN only requires calculating the sigma density. SVM needs to provide a sigma value and an uncertainty parameter that is obtained from the testing error and the number of support vectors. The level of uncertainty in the SVM model relates to the local variability of data.

Decision maps were obtained for the three methods based on different decision functions. For KNN, the regression over neighbors produced continuous values from which a decision value could be chosen. For PNN, the probability of belonging to the above limit category was used for decision mapping. SVM provided a decision function to derive classification maps. PNN and SVM had the advantage of more efficiently incorporating uncertainty into predictions by calculating class probability in the former case, and by using an uncertainty parameter in the latter.

The level of omission error at different cutoffs for the three tested methods, showed an advantage using SVM for classification of lower thresholds, while there was a slight advantage with the use of PNN for larger limits. The use of the class size proportion as a prior probability has contributed to better results with PNN and deal with the unbalanced distribution of samples and the smoothing effect of PNN.

# Discussion and General Conclusions

## Decision mapping

Fundamental questions are: how should indoor radon mapping be done, and which information should be included in the maps? Regarding the first point, there are many interpolation techniques that have been proposed in the present research. An important feature to take into account is that the method must fit the particular conditions of indoor radon data: high local variability and significant important spatial clustering.

There is also a need for methods to be able to handle large volumes of data and to produce series of maps on a national level. Additionally, it must also be considered that maps are required to be updated when new data are collected. To illustrate this, indoor radon data for the Swiss territory accounted for around 140,000 by 2008 and campaigns are still ongoing at present time. Thereafter, proposed methods may be applied without the need for deep modeling or expert knowledge injection.

Categorization of data and the use of robust methods may simplify the task of extreme values modeling while regional classification could be a solution to automatize processes. Nevertheless, regression over data will tend, in many cases, to smooth maps. This neglects the fact that indoor radon can reach, in some cases, truly elevated concentrations.

Simulation has also showed to be a powerful approach to spatial modeling because it provides the most realistic and complete global information. However, the results should be interpreted in probability terms and not in estimation terms. With probability maps, a decision map was created, by considering a critical threshold value within either optimistic or pessimistic scenarios. A drawback of classical geostatistical methods, including SGS, is that calibration and optimization of parameters involves a lot of time investment and a certain level of knowledge of the phenomena. A Typical case is the modeling of variograms for kriging and simulation procedures. Intrinsic conditions, such as stationarity or multi-Gaussianity should also be taken into account.

Being that the indoor radon sampling schema in Switzerland is quite heterogeneous, the spatial estimation task should not rely on a single method but rather on various modeling capabilities, and results should be combined. Regarding the issue of which information should be included on the indoor radon maps, it would be advisable for authorities to create some sort of decision maps. For instance, rather than binding the map content to predictions (possibly inaccurate ones in some cases), it would be convenient to produce maps combining

the probability of having a certain indoor radon value with the liability of estimations based on the estimation variances.

These kinds of combined maps can be subsequently used to decide which areas are declared to be above a certain level. As indoor radon exposure is a subject of debate concerning which levels should be declared a threat for health, it is more convenient to have flexible and realistic maps. Incorporating the uncertainty of the phenomena and expressing the possible committed error can make maps more reliable.

## General Conclusions

General conclusions have been obtained out of the partial conclusions presented at the end of each chapter. From the research on the analysis and modeling of Swiss indoor radon data and following the stated objectives, the conclusions of this thesis are:

- Indoor radon concentrations in Switzerland cannot be explained based on the influence of a single factor. Significant but weak associations with geotechnical units and elevations have been found. Nevertheless, these factors have been unable to explain the high local variations of indoor radon.
- A spatially constrained domain based on the built-up area has been defined. This domain has provided a support-size for obtaining coherent results for declustering, neighborhood definition and simulation estimations.
- A non-linear behavior of the functional clustering for various indoor radon datasets has been found, using the proposed Quantile Morishita Index (QMI) profile method.
- Data spatial partition based on criteria of MW statistics, natural regions and sampling design were proposed as a way to improve modeling and estimations on a multiscale frame.
- In addition to nscores and categorization transforms, the use of MW averaging at an optimal distance and KNNR filtering have been proposed as methods for revealing and reducing the high local variability influence on variography modeling. Data transformation has helped reveal the underlying spatial continuity structures to different degrees.
- Methods used for interpolation, like KNNR, Kriging or GRNN can also be used as exploratory tools to optimize neighborhood parameters and perform data selection.
- In a comparative analysis between regression and interpolation methods, it has been found that the validation errors were less sensitive to the type of method used than to the data selected for the validation. The sole advantages of the GRNN and IDW methods are their modeling simplicity and their options for automation.
- Sequential Gaussian simulation provided a more complete and realistic representation of indoor radon than regression methods, including the natural uncertainty of the phenomena. Exploratory and transformation methods were used to provide an alternative simulation scenario with optimal spatial parameters and histogram reproduction.

- Space filling of the built-up area has been proposed as a procedure to identify the deviation between the sample mean and a hypothetical global mean for a case study, in order to approach global parameters by declustering, to be reproduced throughout simulations.
- Regional classification mapping has been presented as an option to create decision maps with pessimistic or optimistic scenarios for decision-making. The false negative ratio (FNR) classification error has been used to propose a pessimistic scenario.
- An operational diagram flow has been proposed to integrate the results of simulations and other estimation methods in order to obtain complementary results for different conditions of scale and neighborhood definitions in a more automated mode.

## Operational Diagram for indoor radon modeling and mapping

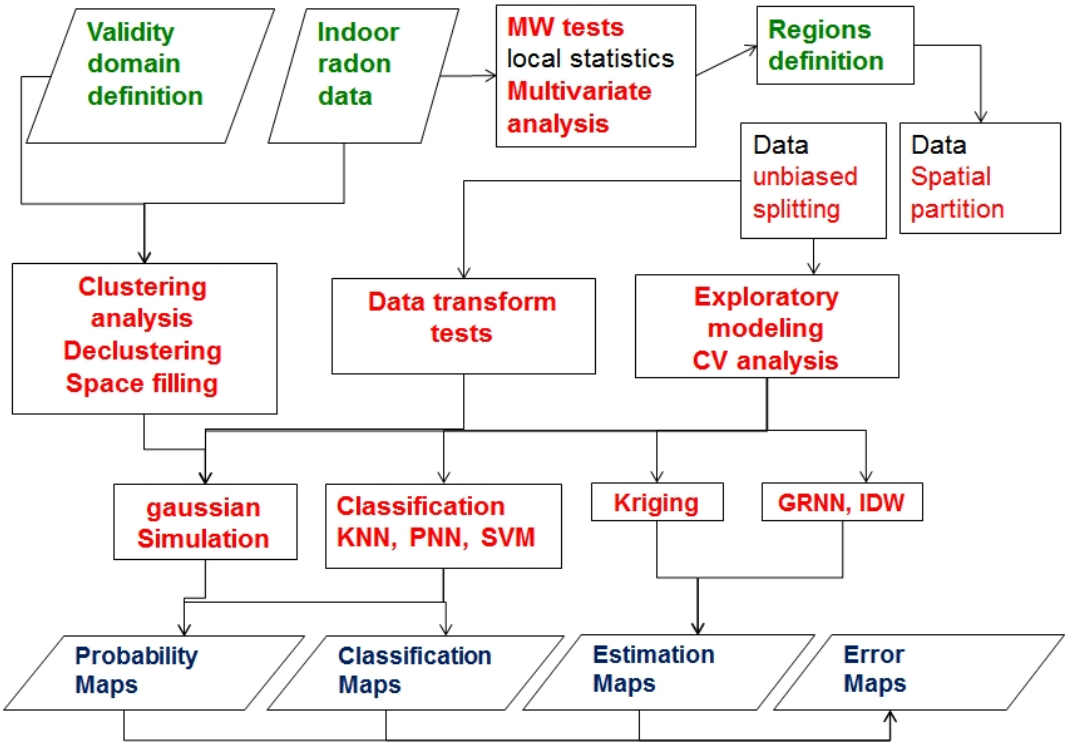



Figure 5.63: *Operational diagram for indoor radon modeling and mapping*

# Bibliography


[1] Andersen, C. E., K. Ulbak, A. Damkjæra, P. Kirkegard, and P. Gravesen: 2001, 'Mapping indoor radon-222 in Denmark: design and test of the statistical model used in the second nationwide survey'. *The Science of the Total Environment* **272**, 231–241.

[2] Bishop, C. et al.: 1995, 'Neural networks for pattern recognition'.

[3] Boehm, C.: 2004, 'Radon risk mapping by means of hydrogeological investigations - example of the Grisons, Switzerland'. In: I. Barnet, M. Neznal, and P. Pacherova (eds.): *Radon investigations in the Czech Republic X and 7th international workshop on the Geological Aspects of Radon Risk Mapping*.

[4] Böhm, C.: 2003, 'Einfluss des Untergrundes auf die Radonkonzentration in Gebäuden–dargestellt anhand einiger Gebäude'. *Bundesamt für Gesundheit, Bern* **20**.

[5] Borgoni, R., P. Quatto, G. Somà, and D. De Bartolo: 2010, 'A geostatistical approach to define guidelines for radon prone area identification'. *Statistical Methods & Applications* **19**(2), 255–276.

[6] Chaouch, A.: 2004, 'Indoor Radon data mining with Geostatistical Tools'. Master's thesis, Université de Lausanne.

[7] Chaouch, A., M. Kanevski, M. Maignan, A. Pozdnoukhov, J. Rodriguez, and G. Piller: 2009, 'Regional Classification of Indoor Radon Data with Support Vector Machines and Geostatistical Tools'. *Interfacing Geostatistics and GIS* p. 65.

[8] Chilès, J.-P. and P. Delfiner: 1999, *Geostatistics, modellind spatial uncertainty*, Wiley series in probability and statistics. Jhon Wiley and sons.

[9] Cohen, J.: 1988, *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

[10] Darby, S., D. Hill, H. Deo, A. Auvinen, J. Barros-Dios, H. Baysson, F. Bochicchio, R. Falk, S. Farchi, A. Figueiras, et al.: 2006, 'Residential radon and lung cancer–detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14,208 persons without lung cancer from 13 epidemiologic studies in Europe.'. *Scandinavian journal of work, environment & health* **32**, 1.

[11] DeLeeuw, J.: February 2011, 'History and Theory of Nonlinear Principal Component Analysis'. Presentation at UCLA Departement of Statistics.

[12] Demyanov, V., M. Kanevski, Maignan, E. Savelieva, V. Timonin, S. Chernov, and G. Piller: 2000, 'Indoor radon risk assessment with geostatistics and artificial neural networks'. In: *Proceedings of the Geostatistical Congress, Capetown*.

[13] Deutsch, C.: 1996, 'Correcting for negative weights in ordinary kriging'. *Computers & Geosciences* **22**(7), 765–773.

[14] Deutsch, C. V. and A. Journel: 1998, *GSLIB, Gesotatistical software library and user's guide*, Applied Geostatistics. Oxford University Press, 2 edition.

[15] Dubois, G.: 2005, 'An overview of radon surveys in Europe'. Technical report, Radioactivity Environmental Monitoring Emissions and Health Unit, Institute for Environment and Sustainability, JRC.

[16] Dubois, G., P. Bossew, T. Tollefsen, and M. De Cort: 2010, 'First steps towards a European atlas of natural radiation: status of the European indoor radon map'. *Journal of environmental radioactivity* **101**(10), 786–798.

[17] Durrani, S. A. and I. Badr: 1995, 'Geostatistically controlled field study of radon levels and the analysis of their spatial variation'. *Radiation Measurements* **25**(1-4), 565–572.

[18] Emery, X. and J. Ortiz: 2005, 'Histogram and variogram inference in the multigaussian model'. *Stochastic environmental research and risk assessment* **19**(1), 48–58.

[19] Euratom: 1990, 'Recommendation of 21 febraury 1990 on the protection of the public against indoor exposure to radon'. Technical report, Commission of the european communities.

[20] Finkelstein, M., L. V. Eppelbaum, and C. Price: 2006, 'Analysis of temperature influences on the amplitude-frequency characteristics of Rn gas concentration'. *Journal of Environmental Radioactivity* **86**(2), 251 – 270.

[21] Friedmann, H.: 2005, 'Final results of the Austrian radon project'. *Health physics* **89**(4), 339.

[22] Friedmann, H., W. Hofmann, H. Lettner, F. SteinhäuslerE, F. Maringer, L. MossbauerH, E. Nadschläger, S. SperkerP, C. Kralik, K. Pock, et al.: 1996, 'The Austrian radon project'. *Environment International* **22**, 677–686.

[23] Goovaerst, P.: 1997, *Geostatistics for Natural Resources Evaluation*. Oxford Univeristy Press.

[24] Goreaud, F. and R. Pélissier: 1999, 'On explicit formulas of edge effect correction for Ripley's K-function'. *Journal of Vegetation Science* **10**(3), 433–438.

[25] Gruson, M.: 2006, 'Radon mapping in Switzerland'. In: I. Barnet, M. Neznal, and P. Pacherova (eds.): *Radon investigations in the Czech Republic XI and 8th international workshop on the Geological Aspects of Radon Risk Mapping*.

[26] Gundersen, L. C. S. and R. R. Schumann: 1996a, 'Mapping the radon potential of the United States: Examples from the Appalachians'. *Environment International* **22**(Supplement 1), 829 – 837. The Natural Radiation Environment VI.

[27] Gundersen, L. C. S. and R. R. Schumann: 1996b, 'Mapping the radon potential of the United States: Examples from the Appalachians'. *Environment International* **22**(Supplement 1), 829 – 837. The Natural Radiation Environment VI.

[28] Huguenot, P.: 2008, 'Analyse de données spatiales complexes: application à la cartographie du radon indoor'. Master's thesis, Université de Lausanne.

[29] Iakovleva, V. S. and V. D. Karataev: 2001, 'Radon levels in Tomsk dwellings and correlation with factors of impact'. *Radiation Measurements* **34**(1-6), 501 – 504.

[30] Iakovleva, V. S. and N. K. Ryzhakova: 2003, 'Spatial and temporal variations of radon concentration in soil air'. *Radiation Measurements* **36**(1-6), 385 – 388. Proceedings of the 21st International Conference on Nuclear Tracks in Solids.

[31] Isaaks, E. and R. Srivastava: 1989, *Applied geostatistics*. Oxford University Press New York.

[32] Jönsson, G.: 2001, 'Soil radon depth dependence'. *Radiation measurements* **34**(1), 415–418.

[33] Journel, A.: 1989, *Fundamentals of geostatistics in five lessons*, Vol. 8. Amer Geophysical Union.

[34] Kanevski, M., S. Chernov, V. Demyanov, E. Savelieva, and V. Timonin: 2003, 'GEOSTAT OFFICE, Software Solution for Spatial Data Analysis. USER'S GUIDE Version 7.10'. IBRAE.

[35] Kanevski, M., S. Chernov, V. Demyanov, E. Savelieva, V. Timonin, and A. Pozdnukhov: 2004a, 'GEOSTAT OFFICE USERS GUIDE'. IBRAE, GEOSTAT OFFICE (GSO) GROUP, Moscow, Russia.

[36] Kanevski, M., V. Demyanov, S. Chernov, E. Savelieva, A. Serov, V. Timonin, and M. Maignan: 1999, 'Geostat Office for environmental and pollution spatial data analysis'. *Mathematische Geologie* **3**, 73–83.

[37] Kanevski, M. and M. Maignan: 2004, *Analysis and Modelling of Spatial Environmental Data*. EPFL Press.

[38] Kanevski, M. and M. Maignan: 2005, 'Analysis and modelling of indoor radon data in Switzerland: geostatistical approach and machine learning algorithms'. presented on International Workshop of Radon Data: valorisation, analysis and mapping. Lausanne 4-5 march 2005, Switzerland.

[39] Kanevski, M., M. Maignan, and G. Piller: 2004b, 'Advanced analysis and modelling tools for spatial environmental data. Case study: indoor radon data in Switzerland'. In: *Proceedings of the XVIII International Conference Enviroinfo 2004*.

[40] Kanevski, M., M. Maignan, and R. Tapia: 2006, 'Indoor radon risk mapping using geostatistical simulations'. In: E. Pirard, A. Dassargues, and H. Havenish (eds.): *XI International Congress for Mathematical Geology, Quantitative Geology from Multiple Sources*.

[41] Kanevski, M., R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V. Demyanov, and S. Canu: 2004c, 'Environmental data mining and modeling based on machine learning algorithms and geostatistics'. *Environmental Modelling & Software* **19**(9), 845–855.

[42] Kanevski, M., A. Pozdnoukhov, and V. Timonin: 2009, *Machine Learning for Spatial Environmental Data: theory, applications and software*. EPFL Press.

[43] Kanevski, M., A. Pozdnukhov, S. Canu, M. Maignan, P. Wong, and S. Shibli: 2002, 'Support vector machines for classification and mapping of reservoir data'. *Studies In Fuzziness And Soft Computing* **80**, 531–558.

[44] Kanevsky, M., R. Arutyunyan, L. Bolshov, S. Chernov, V. Demyanov, N. Koptelova, I. Linge, E. Savelieva, T. Haas, M. Maignan, et al.: 1997, 'Chernobyl fallout: review of advanced spatial data analysis.'. In: *GeoENV I-geostatistics for environmental applications. Proceedings, Lisbon, Portugal, 18-19 November 1996.* pp. 389–400.

[45] Kemski, J., R. Klingel, and A. Siehl: 1996, 'Classification and mapping of radon-affected areas in Germany'. *Environment International* **22**(Supplement 1), 789 – 798. The Natural Radiation Environment VI.

[46] Kemski, J., A. Siehl, R. Stegemann, and M. Valdivia-Manchego: 2001, 'Mapping the geogenic radon potential in Germany'. *The Science of The Total Environment* **272**(1-3), 217 – 230.

[47] Kruskal, J.: 1964, 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis'. *Psychometrika* **29**(1), 1–27.

[48] Marley, F.: 2001, 'Investigation of the influences of atmospheric conditions on the variability of radon and radon progeny in buildings'. *Atmospheric Environment* **35**(31), 5347 – 5360.

[49] Michailidis, G. and J. De Leeuw: 1998, 'The Gifi system of descriptive multivariate analysis'. *Statistical Science* pp. 307–336.

[50] Miles, J.: 1998, 'Development of maps of radon-prone areas using radon measurements in houses'. *Journal of Hazardous Materials* **61**(1-3), 53 – 58.

[51] Miles, J. and J. Appleton: 2005, 'Mapping variation in radon potential both between and within geological units'. *Journal of Radiological Protection* **25**, 257.

[52] Miles, J. and K. Ball: 1996, 'Mapping radon-prone areas using house radon data and geological boundaries'. *Environment International* **22**(Supplement 1), 779 – 782. The Natural Radiation Environment VI.

[53] Neznal, M., M. Neznal, M. Matolín, I. Barnet, and J. Mikšová: 2004, 'The New Method for Assessing the Radon Risk of Building Sites'. Technical report, Czech Geological Survey.

[54] OFS: 2007, 'Recensement fédéral de la population 2007, Bâtiments, logements et conditions dhabitation'. Technical report, Office Federal de la Statistique, Centrale dinformation sur le recensement de la population.

[55] OFSP: 2007, 'Manuel suisse du radon'. Technical report, Office fdral de la sant publique.

[56] OFSP: 2008, 'Radon, information sur un thème rayonnant'. Technical report, Office fédéral de la santé publique.

[57] OFSP: 2011, 'Plan d'action national radon 2012-2020'. Technical report, Office fédéral de la santé publique.

[58] OFSP: 2012, 'Office fédéral de la santé publique, http://www.bag.admin.ch/'.

[59] Olea, R.: 2007, 'Declustering of clustered preferential sampling for histogram and semivariogram inference'. *Mathematical Geology* **39**(5), 453–467.

[60] Oz, B., C. Deutsch, T. Tran, and Y. Xie: 2003, 'DSSIM-HR: a FORTRAN 90 program for direct sequential simulation with histogram reproduction'. *Computers & geosciences* **29**(1), 39–51.

[61] Panatto, D., P. Ferrari, P. Lai, and G. Gallelli: 2006, 'Relevance of air conditioning for 222Radon concentration in shops of the Savona Province, Italy'. *Science of The Total Environment* **355**(1-3), 25 – 30.

[62] SFC, S. F. C.: 1994, 'Ordonnance sur la Radioprotection (ORaP), du 22 juin 1994'.

[63] Shepard, R.: 1962, 'The analysis of proximities: Multidimensional scaling with an unknown distance function. I.'. *Psychometrika* **27**(2), 125–140.

[64] Soares, A.: 2001, 'Direct sequential simulation and cosimulation'. *Mathematical Geology* **33**(8), 911–926.

[65] 'SPSS': 2007, 'SPSS Version release 16.0. CATPCA algorithms'. Software documentation.

[66] Strand, T., C. Lunder Jensen, K. Anestad, L. Ruden, and G. B. Ramberg: 2005, 'High radon areas in Norway'. *International Congress Series* **1726**, 212–214.

[67] Sundal, A. V., H. Henriksen, O. Soldal, and T. Strand: 2004, 'The influence of geological factors on indoor radon concentrations in Norway'. *Science of The Total Environment* **328**(1-3), 41 – 53.

[68] Tapia, R., P. Huguenot, and M. Kanevski: 2008, 'Sequential gaussian simulations for indoor radon risk mapping using declustering and constrained nets'. In: *European Geosciences Union, Vienna, Austria, 14 - 18 April*.

[69] Tapia, R., M. Kanevski, M. Maignan, and M. Gruson: 2006, 'Comprehensive multivariate analysis of indoor radon data in Switzerland'. In: I. Barnet, M. Neznal, and P. Pacherova (eds.): *Radon investigations in the Czech Republic XI and the 8th international workshop on the Geological Aspects of Radon Risk Mapping*. pp. 228–238.

[70] Tapia, R., M. Kanevski, M. Maignan, M. Piller, and M. Gruson: 2005, 'Efficient spatial predictors for indoor radon mapping'. In: *3rd Swiss Geoscience Meeting, Zurich*.

[71] Tapia, R., M. Kanevski, and V. Timonin: 2007a, *Geocomputation, Geosimulation, Geovisualisation: metodi innovativi a supporto della pianificazione urbana e territoriale. Collana di ingegneria della citt e del territorio*, Chapt. Regional classification of indoor radon data. Alinea Editrice Firenze.

[72] Tapia, R., V. Timonin, A. Pozdnukhov, M. Kanevski, and M. Gruson: 2007b, 'Automatic Regional Classification of Environmental Data'. In: *Geophysical Research Abstracts*, Vol. 9. p. 01307.

[73] Thomas, J., J. Hulka, L. Tomsek, I. Fojtkov, and I. Barnet: 2002, 'Determination of radon prone areas by probabilistic analysis of indoor survey results and geological prognostic maps in the Czech Republic'. *International Congress Series* **1225**, 49 – 54.

[74] Tuia, D. and M. Kanevski: 2006, 'Indoor Radon Data Monitoring Networks: Topology'. In: *Fractality and Validity Domains, Congress of the International Association of Mathematical Geology (IAMG), Liège, Belgium*.

[75] Tuia, D. and M. Kanevski: 2008a, *Advanced Mapping of Environmental Data*, Chapt. Chapter 3, Environmental Monitoring Network Characterization and Clustering. ISTE.

[76] Tuia, D. and M. Kanevski: 2008b, 'Indoor radon distribution in Switzerland: lognormality and Extreme Value Theory'. *Journal of Environmental Radioactivity* **99**(4), 649–657.

[77] Vapnik, V.: 2000, *The nature of statistical learning theory*. Springer-Verlag New York Incorporated.

[78] Wackernagel, H.: 2003, *Multivariate geostatistics*. Springer.

[79] Webster, R. and M. A. Oliver: 2001, *Geostatistics for Environmental Scientists*. John Wiley & Sons, Inc.

[80] Weltner, A., I. Mkelinen, and H. Arvela: 2002, 'Radon mapping strategy in Finland'. *International Congress Series* **1225**, 63 – 69.

[81] Yamamoto, J.: 2000, 'An alternative measure of the reliability of ordinary kriging estimates'. *Mathematical Geology* **32**(4), 489–509.

[82] Zeeb, H., F. Shannoun, et al.: 2009, *WHO handbook on indoor radon: a public health perspective/edited by Hajo Zeeb, and Ferid Shannoun*.