GIS-Based Advanced Spatial Analysis of Total Organic Carbon and Potentially Toxic Elements in European Agricultural and Irish Soils

Haofan Xu

Supervisor: Dr. Chaosheng Zhang

The thesis is submitted to National University of Ireland, Galway

in fulfilment of the requirement for the degree of Doctor of Philosophy,

in the School of Geography, Archaeology and Irish Studies

Submission date: Oct 2021

Abstract

With the increasing availability of data in environmental geochemistry, one of the biggest challenges is to extract useful knowledge and interpretable information from large and diverse data sources. The unprecedented volume and complexity of datasets make it difficult to rely on traditional tools for data analysis, which requires the applications and development of GIS-based spatial techniques. In this thesis, four advanced spatial analysis and machine learning (ML) techniques: (1) hot spot analysis (Get-is Ord G_i^*); (2) Geographically weighted regression (GWR); (3) K-means clustering analysis; and (4) Geographically Weighted Pearson Correlation Coefficient (GWPCC) were deployed to investigate the spatial patterns and to extract hidden information in large-scale datasets.

The total organic carbon (TOC) and potentially toxic elements (PTEs) were studied based on datasets of GEochemical Mapping of Agricultural Soil (GEMAS) of EuroGeoSurveys and Tellus of Geological Survey of Ireland. On the one hand, the TOC contents are receiving increasing attention in agricultural soils as an important indicator of soil nutrient, not only due to their close relationship with soil fertility, but also with carbon dioxide (CO_2) in the atmosphere. On the other hand, the advanced spatial techniques played important roles to evaluate concentrations and spatial variation of PTEs affected by multiple influencing factors from natural and anthropogenic sources. These studies provide demonstrations of applications of these advanced analytical techniques as possible solutions to the challenges of data analytics in the big data era.

(1) The hot spot analysis was performed on a total of 2,108 agricultural soil samples based on GEMAS data and revealed an overall negative correlation between TOC and pH, which was in line with the general relationship between these two variables. However, a ‘special’ feature of co-existence of comparatively low TOC and pH values was also identified in north-central Europe. It has been found that these ‘special’ patterns are strongly related to the high concentration of quartz (SiO_2) in the coarse-textured glacial sediments in north-central Europe.

(2) The GWR further explored the spatially varying relationships between TOC and pH based on the GEMAS data, with more than 50% original negative relationship changed to positive at the continental level. The significant positive correlations clustered in central-eastern Europe, while negative correlations were observed mainly in northern Europe. Mixed relationships occurred in southern Europe. Such results further highlighted the influences of the extensive occurrence of quartz-rich soils and climate factors on the 'special' positive correlations. In addition, anthropogenic inputs also interfered the relationships in the mixed southern European areas.

(3) The integration of hot spot analysis and K-means clustering analysis was applied to investigate the spatial patterns for 15 PTEs and associations with their controlling factors based on the Tellus data under the complicated geological background of Northern Ireland (NI). The spatial clustering patterns for the 15 PTEs from hot spot analysis and hidden patterns of 6,862 soil samples from K-means clustering were consistent with each other, highlighting the dominant control of peat and basalt in the topsoil of Northern Ireland.

(4) The GWPCC found that the relationships between lead (Pb) and aluminium (Al) are spatially varying, with both positive and negative correlations in the topsoil of northern half of Ireland based. The 'special' negative correlations were observed in more than 35% of the whole study area, mainly clustered in the north-eastern and western Ireland. The positive correlations were observed in the midlands. Mixed relationships of both negative and positive correlations occurred in the eastern coastal areas. The majority of negative correlation patterns showed clear association with blanket peat, which can be attribute to long-distance transportation of Pb from atmospheric deposition.

The main scientific contributions to the advancements in environmental geochemical studies of this research include the following:

(1) identified a 'special' feature of positive relationship of low TOC contents and low pH values in the north-central Europe;

(2) introduced the topic of ‘spatially varying relationships between TOC and pH’ which provide added value and clarification to the understanding of the controversy of their complicated relationship in the literature;

(3) provided latest understanding and classification of 15 PTEs in the topsoil of NI to enhance the current knowledge of their controlling factors under the complicated geological background;

(4) proved and observed the spatially varying relationships between Pb and Al which are associated with atmospheric deposition and anthropogenic activities.

Overall, these novel findings indicated that the spatial techniques have strong efficiency in processing large-scale datasets, providing demonstration and evidence for the application of GIS-based advanced spatial analysis on identification of the hidden spatial patterns for TOC and PTEs in the topsoil and to associate them with related influencing factors. These analytical results enhanced the current knowledge for soil management and risk assessment, and can be applied in environmental studies elsewhere.

Keywords: *Geographic information system (GIS); Total organic carbon (TOC); pH; Hot spot analysis; Geographically weighted regression (GWR); Spatially varying relationships; Potentially toxic elements (PTEs); Lead (Pb)*

Table of contents

Abstract iii

Table of contents vi

Declarations xi

Acknowledgments xii

Dedication and Publication xiii

List of Abbreviations xiv

Chapter 1 Introduction 1

1.1 General introduction 2

1.1.1 Background of spatial analysis in environmental geochemistry 2

1.1.2 TOC in European agricultural soil 3

1.1.3 PTEs in soils of Ireland 4

1.2 Existing policies on soil contamination of European Union (EU) and Ireland 5

1.3 Research hypothesis 7

1.4 Research objectives 8

1.5 Structure of Thesis 10

Reference 12

Chapter 2 Literature review 15

2.1 Overview 16

2.2 Compositional data analysis 16

2.3 Development and applications of GIS-based spatial techniques in environmental geochemistry 17

2.4 Big data era and spatial machine learning 21

2.5 The applications of GIS-based spatial analysis on TOC 23
2.5.1 Spatial distribution and variation of TOC 23
2.5.2 Spatially varying relationships between TOC contents and pH values 25
2.6 The application of GIS-based spatial analysis on PTEs 28
2.6.1 Potentially toxic elements (PTEs) 28
2.6.2 Sources and background of PTEs in the soils 29
2.6.3 Spatial distribution patterns and source identification of soil PTEs 32
2.6.4 Spatially varying relationships between Pb and Al in the soils 33
2.7 Summary 35
Reference 36
Chapter 3 Materials and methodologies 54
3.1 Study area and scales 55
3.1.1 European continent 55
3.1.2 Northern Ireland 56
3.1.3 Northern half of Ireland 57
3.2 Soil sampling and analyses 58
3.2.1 GEMAS project data 58
3.2.2 Tellus survey data 60
3.3 Data analysis 62
3.3.1 Descriptive statistics 62
3.3.1.1 Representatively descriptive parameters 62
3.3.1.2 Probability distribution 63
3.3.2 Data treatment 64

3.3.2.1 Data transformation for GEMAS data in European agricultural soil 64
3.3.2.2 Data transformation for Tellus data in Northern Ireland 64
3.3.2.3 Data transformation for Tellus data in the northern half of Ireland..... 65
3.3.3 Spatial analysis..... 65
3.3.3.1 Inverse distance weighted interpolation..... 65
3.3.3.2 Hot spot analysis (Getis-Ord G_i^* statistic) 66
3.3.3.3 Geographically weighted regression (GWR)..... 68
3.3.3.4 Geographically Weighted Pearson Correlation Coefficient (GWPCC).... 70
3.3.4 Multivariate analysis..... 72
3.3.4.1 Correlation analysis 72
3.3.4.2 Principal component analysis (PCA)..... 73
3.3.4.3 K-means clustering analysis 73
3.3.5 Computer software..... 75
3.4 Summary..... 76
Reference 77
Chapter 4 Research paper 85
4.1 Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data..... 86
4.2 Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression 98
4.3 Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis..... 110

4.4 Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient 125

4.5 Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR) .. 146

Chapter 5 Discussion 158

5.1 Overview of the Research Process 159

5.2 Contributions of Research 160

5.3 Advancement 162

5.4 Research limitations in this thesis 163

Reference 164

Chapter 6 Conclusion 167

6.1 Overview 168

6.2 Main conclusions 168

6.2.1 Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data 168

6.2.2 Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression 169

6.2.3 Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis 169

6.2.4 Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient 170

6.2.5 Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR) 170

6.3 Recommendations 171

6.4 Future research 171

6.4.1 Predicting and mapping of soil organic carbon contents using machine learning algorithms in the topsoil of Ireland 172

6.4.2 Investigation of spatially varying relationships between soil total organic carbon content and climate factors in European agricultural soil 172

6.4.3 Discovering hidden spatial patterns of selected potentially toxic elements using hot spot analysis and K-means clustering analysis in stream sediments in Northern Ireland 172

6.4.4 Exploring spatially varying relationships between Cd, Ni, Zn and Al in the topsoil of Ireland 173

Reference 174

Declarations

This thesis or any part thereof, has not been, or is not currently being submitted for any degree at any other university.

_____

Haofan Xu

The work reported herein is as a result of my own investigations, except where acknowledged and referenced.

_____

Haofan Xu

Acknowledgments

Primarily, I would like to express my sincere gratitude to my supervisor Dr. Chaosheng Zhang, for giving me the opportunity to start my Ph.D. studies in the department of Geography, Archaeology and Irish studies, National University of Ireland, Galway. I would like to thank Dr. Zhang for his continuous support and useful guidance during my three-year research study and life. Meanwhile, special thanks to his wife Mrs. Dongxia Zhang and their children who help and care for me to make me have the warmth of home.

I would like to thank Prof. Alecos Demetriades, Prof. Clemens Reimann and Prof. Peter Croot, for their thorough reviewing my manuscript and providing insightful suggestions.

I would like to thank Dr. Aaron Potito, Dr. Terry Morley and Dr. Mary Ryan as the members of my graduate research committee for their valuable comments on the progress of my research project.

I would like to thank Dr. Siubhán Comer for her assistance of my teaching skills in GIS and caring in my daily life.

I would like to thank my loving family. Thanks to my parents for their financial help. As a self-financed student, it brought a lot of financial pressure to my family. At the same time, I also thank them for their help and care in my whole life.

At the end, I would like to thank all the friends and colleagues in the geography department, who made my last three years full of warmth and joy. I would like to thank Dr. Yumin Yuan and Dr. Yuting Meng for the help and convenience they gave me during my first year of study, and it was a pleasure to work with them in the same research group. Special thanks to Axel Leahy, who is always friendly and kind. Special thanks to my girlfriend Ms. Yini Zhao, with whom spent much time with me during the COVID-19 pandemic period.

Dedication and Publication

This thesis comprises 4 publications as the first author and 1 publication as co-author:

1. **Xu, H.F.**, Demetriades, A., Reimann, C., Jiménez., J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94-104.
2. **Xu, H.F.**, Zhang, C.S., 2021. Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression. Sci. Total Environ., 752, 141977.
3. **Xu, H.F.**, Croot, P., Zhang, C.S., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environ. Int., 151, 106456.
4. **Xu, H.F.**, Croot, P., Zhang, C.S., 2021. Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient (Under Review)
5. Yuan, Y.M., Cave, M., **Xu, H.F.**, Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater., 393, 122377.

List of Abbreviations

AIC	Akaike information criterion
Al	Aluminium
As	Arsenic
Ba	Barium
Bi	Bismuth
CA	Cluster analysis
CEC	Cation exchange capacity
clr	Centred log-ratio
Co	Cobalt
CO_2	Carbon dioxide
Cr	Chromium
Cu	Copper
CV	Coefficient of variation
DL	Detection limit
EC	European Commission
EGS	EuroGeoSurveys
ESRI	Environmental Systems Research Institute
EU	European Union
FA	Factor analysis
GEMAS	GEochemical Mapping of Agricultural Soil
GIS	Geographic information system
GSI	Geological Survey of Ireland

GSNI	Geological Survey of Northern Ireland
GWPCC	Geographically Weighted Pearson Correlation Coefficient
GWR	Geographically weighted regression
ICP	Inductively Coupled Plasma
IDW	Inverse distance weighted
ilr	Isometric log-ratio
LISA	Local indicators of spatial association
ML	Machine learning
MLAs	Machine learning algorithms
Mn	Manganese
Mo	Molybdenum
Ni	Nickel
NI	Northern Ireland
OC	Organic carbon
OLS	Ordinary Least Square
OM	Organic matter
Pb	Lead
PCA	Principal component analysis
PCC	Pearson correlation coefficient
PMs	Parent materials
PTEs	Potentially toxic elements
Q-Q	Quantile-Quantile
Sb	Antimony

Sc	Scandium
SD	Standard deviation
SML	Spatial machine learning
Sn	Tin
Ti	Titanium
TOC	Total organic carbon
U	Uranium
V	Vanadium
XRF	X-ray fluorescence
Zn	Zinc
Zr	Zirconium

Chapter 1
Introduction

1.1 General introduction

1.1.1 Background of spatial analysis in environmental geochemistry

Geographic Information System (GIS) is a conceptual framework for collecting, managing and analysing geographic and spatial data (Clarke, 1986). It is an emerging computer-based tool that provides a platform to integrate geographic information and attribute data. With the unique functions of editing, managing and visualising geographic information, GIS has become a popular technique for revealing patterns and deeper insights of spatial data. It is originated in geography and geoscience, which has now also been applied in other fields such as engineering, transportation, economics and telecommunications, etc. (Maliene et al., 2011).

Since the 1990s, with the development of GIS, the exploratory research on environmental geochemistry has gradually relied on geostatistical analysis and modelling of spatial data (Burrogh and McDonnell, 1998). The datasets sampled for specific environmental variables and regions have been widely applied for environmental planning, management and risk assessment (Zhang et al., 2008a; Sollitto et al., 2010; Burrough et al., 2015). However, as the data volume and diversity continue to increase in the big data era, classical statistics are proposed as lacking of efficiency in data analysis of large-scale and multivariate datasets (Zuo, 2017). In recent years, spatial machine learning (SML) has been adopted in environmental studies as a novel and efficient approach for data mining. The SML techniques are the combination of advanced spatial techniques and machine learning algorithms (MLAs) applied onto the spatial data of environmental geochemistry (e.g., Kanevski et al., 2009; Li et al., 2011), including classification, prediction, visualisation, identification of cluster patterns and spatial relationships, etc. (Zhang, 2020). These techniques have the potential prior to the conventional techniques in revealing hidden spatial patterns, which are helpful to extract useful geochemical knowledge and associations from these patterns (Xie et al., 2004; Meshkani et al., 2011; Sergeev et al., 2019; Rahmati et al., 2020). In this thesis, four spatial analysis and SML techniques of hot spot analysis (Getis-Ord G_i^* statistic), geographically weighted regression

(GWR), K-means clustering analysis and Geographically Weighted Pearson Correlation Coefficient (GWPCC) were employed to investigate the spatial distribution patterns and spatially varying relationships of soil TOC and selected PTEs based on the GEMAS and Tellus datasets, respectively. These studies provide demonstration of applications of these advanced analytical techniques in environmental studies from both local and regional scales, and these spatial patterns provide an effective way to associate with related influencing factors or pollution sources. Moreover, considering the rapid development of spatial techniques in the big data era, the existing problems and possible solutions of GIS-based spatial analysis in environmental geochemistry were also reviewed and summarised.

1.1.2 TOC in European agricultural soil

Soil TOC content is a measure of the carbon stored in organic matter (OM), which is an important indicator of soil quality and productivity. In addition, soil is regarded as the largest organic carbon (OC) sink in the terrestrial ecosystem, with total amounts of carbon two or three times higher than that in the atmosphere or terrestrial vegetation (Batjes, 1996; Jobággy and Jackson, 2000; Schmidt et al., 2011). Therefore, even minor changes of TOC contents in soils can influence the atmospheric CO_2 concentrations (Johnston et al., 2004). The sequestration, decomposition and release of OC in soils play important roles in global carbon cycle, which is of great significance for mitigating global climate change and maintaining ecosystem services and functions (Yang et al., 2007). Due to the importance of TOC, its content has been widely studied on national and regional scales, such as Ireland (Zhang and McGrath, 2004; Zhang et al., 2011), France (Martin et al., 2011), Belgium (Meersmans et al., 2011), Spain (Rodríguez Martín et al., 2016) and the United Kingdom (Bradley et al., 2005). However, previous studies mainly focused on the storage and ignored the spatial variation and dynamic of TOC by using traditional statistical methods. As an essential component of agricultural soil, TOC exhibits a high degree of spatial variability at both horizontal and vertical levels under the complicated influencing factors including both natural and anthropogenic ones, such as climate, topography, soil properties, soil parent materials, fertiliser inputs and agricultural management (Jenny,

1980; Jackson et al., 2002; Lal, 2005; Jandl et al., 2007). In this case, it requires an improved way to explore the spatial variation patterns and varying relationships for TOC contents in agricultural soil. For example, the negative relationship between TOC contents and pH values in soils has been widely reported in previous studies (McGrath and Zhang, 2003; Korkanç, 2014; Reisser et al., 2016; Gebrehiwot et al., 2018; Zhang et al., 2018), while the positive relationship was also proposed in few study areas (Wang et al., 2010; Luo et al., 2017). These studies are based on global statistics or models (e.g., ordinary linear regression, correlation analysis), which cannot capture the varying relationships at different sampling locations. However, considering the complicated influencing factors on these two variables, the contradictory relationships should be objectively evaluated at the local scales, which can be done by Getis-Ord G_i^* statistic and GWR model. These two techniques can reveal the clustering patterns of positive and negative relationships, providing effective ways to investigate the spatial relationships between the soil TOC contents and pH values. Since the spatially varying relationships between soil TOC contents and pH values have not been quantitatively researched yet, it is a worthwhile topic at the European continental level.

1.1.3 PTEs in soils of Ireland

Exposure to excessive accumulation of PTE concentrations is harmful to human health, especially the highly toxic elements such as arsenic, cadmium, cobalt and lead (Bellinger, 2004; Zahran et al., 2009). Therefore, understanding the sources of PTEs is important for environment management and sustainability (Shazili et al., 2006; Huang et al., 2007). Considering the complicated geological background of the topsoil of NI, the hidden spatial patterns and controlling factors for the selected 15 PTEs including arsenic (As), barium (Ba), bismuth (Bi), chromium (Cr), cobalt (Co), copper (Cu), nickel (Ni), manganese (Mn), molybdenum (Mo), lead (Pb), antimony (Sb), tin (Sn), uranium (U), vanadium (V) and zinc (Zn) were identified by the combination of two SML techniques of hot spot analysis and K-means clustering analysis in the topsoil of NI. These two spatial clustering techniques have the potential to discover the spatial clustering patterns of high and low values of 15 PTEs and soil

samples, which provides an efficient way to reveal the clear spatial association between these patterns and their controlling factors with different geological features.

Moreover, among all the PTEs, the concentration and variation of Pb are the worthiest of attention in the environmental studies (Nriagu, 1983), due to its fate is extremely susceptible to interference from anthropogenic factors in the environment (Saby et al., 2006; Cheng et al., 2015). As a promising way to associate with influencing factors, the spatially varying relationships between Pb and Al concentrations were investigated by the GWPCC technique in the topsoil of northern half of Ireland. The element Al is a basic constituent of silicate clays, and Pb can not only be adsorbed to clay but it is also present in primary silicates as K-feldspar and mica (Spark, 2010). It is a conservative lithogenic element and often used as reference element (Shotyk et al., 2002; Sezgin et al., 2003; Le Roux et al., 2004), which is chemically stable and its fate in the environment media is not easily affected by human activities. The original relationship between Pb and Al is positive under most natural conditions due to their similar chemical properties (Schropp and Windom, 1988; Spark, 2010), which is expected for soils derived from continental crust (Walsh and Barry, 1957). However, this general relationship may be interfered by external factors including both natural and anthropogenic influence at a certain extend. Thus, the spatially varying relationships that revealed by GWPCC are able to be associated with related potential pollution sources of Pb, which is an interesting topic and can be also applied to identify influencing factors for other PTEs in the environmental studies elsewhere.

1.2 Existing policies on soil contamination of European Union (EU) and Ireland

Soil has the ability to buffer, filter, retain and degrade pollutants, and is regarded as a necessary but non-renewable resource based on its nature, providing food, biomass and raw materials to

humans (Mongwe and Fey, 2004; Swartjes et al., 2008; Ceci et al., 2019). Due to the inseparable relationship between soil and terrestrial ecosystem, even some minor pollution or damage to its structure will also affect other environmental media. For example, soil is considered as the largest OC pool, and the sequestration of OC in the soil can reduce the emission of CO_2 into the air, and thus effectively reduce the greenhouse effect. Soil degradation and contamination is one of the main threats affecting global soil health (FAO and ITPS, 2015). However, soil pollution is unique and invisible, while its impact is only visible when the pollution level has serious impacts on the environment and human health (Rodríguez Eugenio et al., 2018).

At present, extensive legislation for soil protection has been proposed at the European continental level, taking the EU Soil Thematic Strategy as core (Römbke et al., 2004; EC, 2006). The Soil Thematic Strategy identified the main threats to EU soils, including soil erosion, sealing and landslides, soil contamination, the loss of soil OM and biodiversity, soil compaction and salinisation (EC, 2006; Montanarella and Panagos, 2015). It introduced soil degradation trends in Europe and the world, as well as the challenges of ensuring protection. Although the European Commission (EC) withdrew the proposal for the Soil Framework Directive in 2014, the EU and its member states pledged to work on soil protection and study how to best achieve this goal in the future. In addition, other existing laws are mainly focused on the environmental objectives that are not explicitly on soil, such as reducing pollution, offsetting greenhouse gas emissions and preventing other environmental threats, such as Environmental Liability Directive (2004/35/EC), Industrial Emissions Directive (2010/75/EU), Environmental Impact Assessment Directive (85/337/EEC) and Sewage Sludge Directive (86/278/EEC), etc. In May 2020, the EC adopted the 2030 Biodiversity Strategy based on the European Green Deal, which aims to improve the ecosystem and achieve the sustainability of human habitation through long-term efforts.

In Ireland, the Environmental Protection Agency (EPA) is an official agency dedicated to environmental protection and improvement, which encourages local researchers to conduct

environmental monitoring and assessment. According to the current EU regulations and framework, both the EU and Ireland have carried out geochemical mapping of regional atlas to achieve the goal of environmental and soil assessment, including the GEMAS project (Reimann et al., 2014) and Tellus survey (Knights and Glennon, 2013).

1.3 Research hypothesis

Based on the literature review, three main research gaps related to spatial analysis in soil geochemistry were identified, including large-scale datasets, spatial patterns and spatially varying relationships between environmental variables. The big data era of soil geochemistry brings computational and statistical challenges to traditional techniques, as these techniques are reported lack efficiency on capturing hidden patterns or special features in large-scale datasets. Moreover, previous studies considered that the spatial relationships between environmental variables was spatially constant, while often ignoring their heterogeneity and spatially varying relationships between studied variables.

According to this, the specific research hypotheses of this thesis include:

(1) The spatial relationships between TOC and pH values are not constant, but spatially varying in different sampling locations across the European continent and can be captured by GWR model;

(2) The controlling factors for the 15 PTEs under the complicated geological background can be better evaluated through the hidden spatial patterns revealed by the advanced spatial clustering techniques;

(3) The spatially varying relationships between Pb and Al concentrations exist in the topsoil of Ireland, and the potential pollution sources of Pb can be associated with the spatial relationships revealed by the local statistics in the GWPCC approach.

1.4 Research objectives

The overall aims and objectives of this thesis is to use of GIS-based advanced spatial analysis to identify spatial patterns for soil nutrients (i.e., TOC) and PTEs based on the large-scale datasets (Fig. 1.1). These spatial patterns can be used to extract hidden knowledge and geochemical associations with different influencing factors and pollution sources, which are vital to the research of environmental geochemistry, as they are difficult to capture by traditional techniques. To this end, this thesis applied four spatial analytical technologies to demonstrate the process of identifying spatial patterns and extracting geochemical knowledge, and provided a summary of the development and applications of GIS-based spatial analysis. The unified link of these four papers in this thesis is the applications of different advanced spatial analysis techniques on identification of hidden patterns of soil geochemical mapping in large-scale datasets. These spatial patterns and spatially varying relationships between environmental variables can provide examples and evidence of application of SML in the big data era, as well as provide substantially supportive guidance on the soil monitoring and pollution assessment for relevant stakeholders and local government.

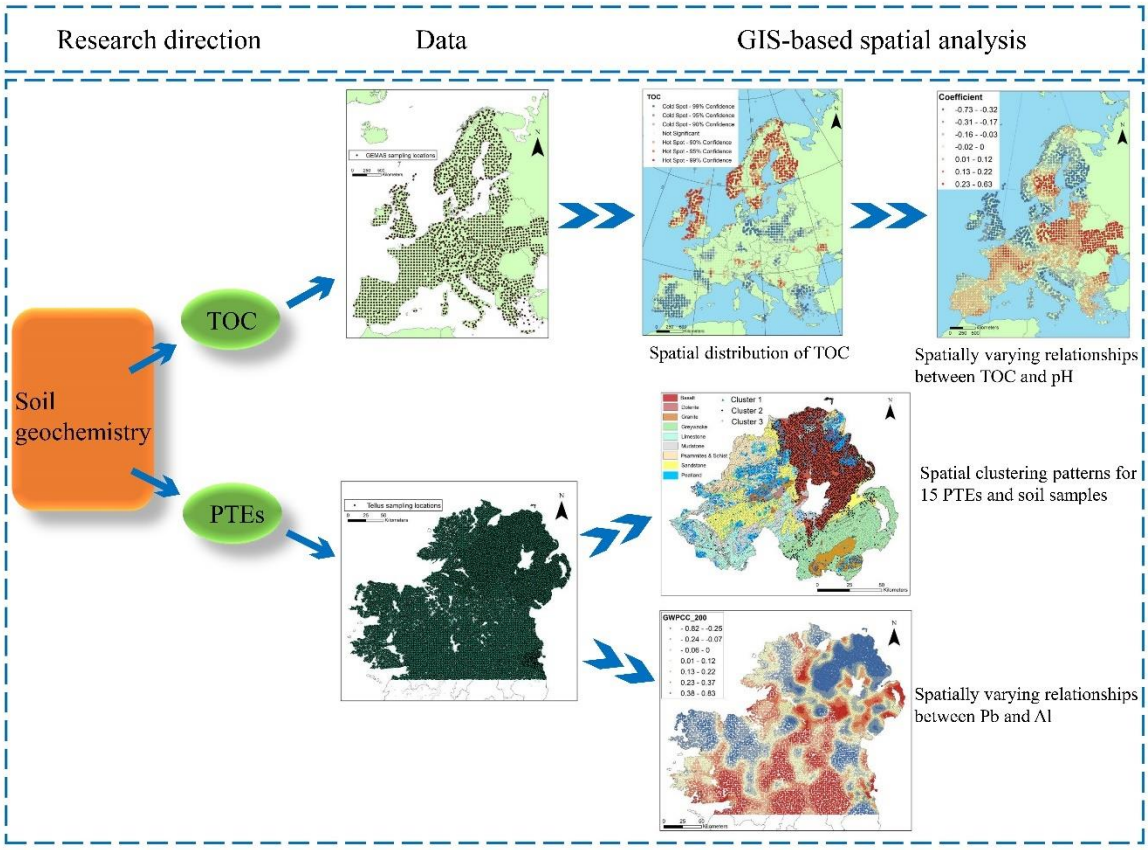


Figure 1.1 Overall research objectives of four studies in this thesis.

Specifically, the research objectives of identification of the co-existence of low TOC contents and low pH values in north-central Europe are:

(1) to reveal the spatial distribution patterns of TOC and pH;

(2) to identify the spatial relationship between soil TOC contents and pH values using hot spot analysis;

(3) to explore influencing factors of the special pattern of co-existence of both low TOC and pH values.

The research objectives of investigating spatially varying relationships between TOC contents and pH values in European agricultural soil are:

(1) to investigate the spatially varying relationships between TOC contents and pH values

using GWR model;

(2) to study the effects of bandwidths for identifying different patterns of the spatially varying relationships;

(3) to explore the influencing factors on the special positive relationship between TOC and pH values.

The research objectives of discovering hidden spatial patterns and their associations with controlling factors for PTEs in the topsoil of Northern Ireland are:

(1) to identify the spatial clustering patterns for 15 PTEs using hot spot analysis;

(2) to reveal the hidden patterns of soil samples using K-means clustering analysis;

(3) to explore the geochemical association between the spatial patterns and controlling factors on PTEs.

The research objectives of exploring main influencing factors of spatially varying relationships between Pb and Al concentrations in the topsoil of northern half of Ireland are:

(1) to investigate the spatial relationships between Pb and Al concentrations using GWPCC based on the currently available Tellus data set in the topsoil of northern half of Ireland;

(2) to identify the spatial associations with different influencing factors from the local correlation patterns;

(3) to further explore the underlying mechanisms between the 'special' negative correlation and potential pollution sources of Pb distribution.

1.5 Structure of Thesis

Chapter 2 reviews the development and applications of GIS-based spatial analysis, with the focus on the existing problems and possible solutions in the big data era of environmental geochemistry. It also summarises the applications of spatial analysis on the distribution and

variation of TOC and PTEs, including the associations with influencing factors from both natural and anthropogenic aspects.

Chapter 3 demonstrates the materials and methodologies used in this thesis, including soil sampling process, background of study area, data analysis and specific methodologies used in this study.

Chapter 4 comprises five published papers with their summaries and personal dedication descriptions.

Chapter 5 discusses the overview of the research process in this thesis, and highlights how the researches relate to each other, as well as the contributions and advancements to the current literature and wider research community.

Finally, Chapter 6 concludes the results of five papers, and recommends policy-relevant strategy and puts forward future research.

Reference

Ceci, A., Pinzari, F., Russo, F., Persiani, A.M. & Gadd, G.M. 2019. Roles of saprotrophic fungi in biodegradation or transformation of organic and inorganic pollutants in co-contaminated sites. Appl. Microbiol. Biotechnol., 103 (1): 53–68. https://doi.org/10.1007/s00253-018-9451-1

Coggins, A.M., Jennings, S.G., Ebinghaus, R., 2006. Accumulation rates of the heavy metals lead, mercury and cadmium in ombrotrophic peatlands in the west of Ireland. Atmos. Environ., 40 (2), 260-278.

de Moraes Sa, J.C., Cerri, C.C., Lal, R., Dick, W.A., de Cassia Piccolo, M., Feigl, B.E., 2009. Soil organic carbon and fertility interactions affected by a tillage chronosequence in a Brazilian Oxisol. Soil Till. Res., 104 (1), 56-64.

De Vleeschouwer, F., Gérard, L., Goormaghtigh, C., Mattielli, N., Le Roux, G., Fagel, N., 2007. Atmospheric lead and heavy metal pollution records from a Belgian peat bog spanning the last two Millenia: Human impact on a regional to global scale. Sci. Total Environ., 377, 282-295.

Eganhouse, R.P., 1997. Molecular markers in environmental geochemistry, ACS Symposium Series, American Chemical Society, Washington, DC.

FAO. 2015. Revised World Soil Charter. Rome, Italy, FAO. 10 pp. (also available at www.fao.org/3/i4965e/I4965E.pdf).

Fotheringham, S., Rogerson, P. (Eds.)., 2013. Spatial analysis and GIS. CRC Press.

Gebrehiwot, K., Desalegn, T., Woldu, Z., Demissew, S., Teferi, E., 2018. Soil organic carbon stock in Abune Yosef afroalpine and sub-afroalpine vegetation, northern Ethiopia. Ecol. Process., 7 (1), p. 6.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environ. Pollut., 231, 1188-1200.

Korkanç, S.Y., 2014. Effects of afforestation on soil organic carbon and other soil properties. Catena, 123, 62-69.

Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Chang. Biol., 23 (10), 4430-4439.

Macalady, D.L., (ed.), 1998. Perspectives in Environmental Chemistry. Oxford University Press, New York, pp. 138-166.

Manta, D. S., Angelone, M., Bellanca, A., Neri, R., & Sprovieri, M. (2002). Heavy metals in urban soils: a case study from the city of Palermo (Sicily), Italy. Science of the Total Environment, 300, 229–243.

McGrath, D., Zhang, C.S., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem. 18, 1629-1639.

Mongwe, H.G., Fey, M.V., 2004. The buffering capacity of soil materials for various contaminant types and the relationship between soil morphology, chemical properties and buffering capacity: a literature review. Pretoria, South Africa, University of Stellenbosch.

Nriagu, J.O., 1983. Lead and Lead Poisoning in Antiquity Wiley, New York.

Nriagu, J.O., 1996. A history of global metal pollution. Science, 272, 223-224.

Rodríguez Eugenio, N., McLaughlin, M.J., Pennock, D., 2018. Soil pollution: a hidden reality. Rome, Italy, Food and Agriculture Organization of the United Nations. 156 pp.

Schwarzenbach, R.P., Gschwent, P.M., Imboden, D.M., 1993. Environmental Organic Chemistry. Wiley, New York.

Wang, T., Kang, F., Cheng, X., Han, H., Ji, W., 2016. Soil organic carbon and total nitrogen stocks under different land uses in a hilly ecological restoration area of North China. Soil Tillage Res., 163, 176-184.

Wang, Z.M., Zhang, B., Song, K.S., Liu, D.W., Ren, C.Y., 2010. Spatial variability of soil organic carbon under maize monoculture in the Song-Nen Plain, Northeast China. Pedosphere, 20 (1), 80-89.

Wong, C. S., Li, X., & Thornton, I. (2006). Urban environmental geochemistry of trace metals. Environmental Pollution, 142, 1–16.

Xia, X., Chen, X., Liu, R., & Liu, H. (2011). Heavy metals in urban soils with various types of land use in Beijing, China. J. Hazard. Mater., 186, 2043–2050.

Yang, L., Pan, J., Shao, Y., Chen, J.M., Ju, W.M., Shi, X., Yuan, S., 2007. Soil organic carbon decomposition and carbon pools in temperate and sub-tropical forests in China. J. Environ. Manag., 85, 690-695.

Yuan, Y.M., Cave, M., Xu, H.F., Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater., 393, 122377.

Zhang, X., Liu, M., Zhao, X., Li, Y., Zhao, W., Li, A., Cheng, S. Cheng, S., Han, X., Huang, J., 2018. Topography and grazing effects on storage of soil organic carbon and nitrogen in the northern China grasslands. Ecol. Indic., 93, 45-53.

Chapter 2
Literature review

2.1 Overview

This chapter briefly reviewed the literatures on the development and applications of GIS-based spatial analysis in the big data era of environmental geochemistry, with focus on the existing problems and possible solutions from the spatial analysis perspective. As advanced and effective methods for environmental data mining, they are often applied to identify the spatial distribution and patterns of nutrients (e.g., TOC) and potentially toxic elements (PTEs) in the soil at the local or regional scale. Therefore, this chapter also summarised the applications of GIS-based spatial analysis techniques on the distribution and variation of TOC and PTEs, in order to provide a better understanding of geochemical knowledge and associations with related influencing factors from both natural and anthropogenic sources.

2.2 Compositional data analysis

Compositional data are non-negative data that carry relative (rather than absolute) information (Pawlowsky-Glahn and Buccianti, 2011). These data usually have a constant and constrained sample value. For example, the sum of proportions or percentages is 1 or 100%. These percentage data are considered as closed data (Aitchison, 1986, Buccianti et al., 2006), and it requires pre-processing (e.g., data transformation) to open these data to destroy the closure effects before the statistical or multivariate analysis of these data. To deal with this problem, log-ratio data transformation is recommended in current literature (Aitchison, 1986, Egozcue et al., 2003). However, in spatial analysis or geostatistics, many parametric techniques still require normal distribution of input datasets. Therefore, there is still controversy about the topic of compositional data, and the choice of data transformation method is under debate in the current literature.

Regarding these issues, the research in this thesis decided to follow the purpose of parametrical statistical analysis and spatial analysis, while data transformation is needed for the data which do not follow a normal distribution. Otherwise, non-parametric methods should be considered. When the data follow a lognormal distribution, the logarithm transformation is sufficient (Limpert et al., 2001). For the more general positively skewed distribution, a normal score transformation or Box-Cox transformation is recommended (Zhang et al., 2008). For compositional data, log-ratio transformations (e.g., centred log-ratio, isometric log-ratio) are recommended (e.g., Aitchison, 1986; Filzmoser et al., 2009).

2.3 Development and applications of GIS-based spatial techniques in environmental geochemistry

Data in environmental geochemistry are typical spatial data, containing geographic coordinates (i.e., longitude and latitude) and geochemical attributes (e.g., element concentrations), which can be stored in a geographical information system (Goodchild et al., 1992). The research on environmental geochemistry mainly relies on spatial data analysis in the GIS. This is different from traditional techniques because it requires considering of both geographic locations and attributes (Goodchild, 1987). Historically, environmental geochemical data were processed by using classic univariate statistics in most studies. Since the 2,000s, the combination of multivariate statistical analysis and GIS-based spatial analysis has become the mainstream tool in environmental geochemical studies (Hou et al., 2017). This is not only due to the growth of available data sets from geochemical survey, but also benefits from the advancement of the ability to manage large spatial data sets in the GIS platform.

With the improvement of computer hardware and software, a growing number of spatial analysis techniques have been integrated into GIS (Bailey, 1994; Fotheringham and

Rogerson, 2013), which has made great contributions to the environmental monitoring and assessment (Zhang and Selinus, 1998). As the amount of geoscience data continues to increase, the field of environmental geochemistry has also entered the era of big data. The development and applications of multivariate statistical analysis and spatial analysis on the spatial data bring new insights and opportunities to geochemical mapping, exploration as well as environmental and health assessment (Overpeck et al., 2011, Reichman et al., 2011). However, although various statistical and spatial analysis techniques existing, it is important to understand the advantages and disadvantages of these techniques and the practical problems that can be solved. Review on the past literature, several major problems in environmental geochemistry include probability distribution, spatial structures and patterns, correlation and spatial relationships, background and thresholds, visualisation, prediction, outlier detection and distinction of natural and anthropogenic factors (Darnley, 1990; Zhang and Selinus, 1998; Reimann and de Caritat, 2005; Zhang et al., 2008a).

Probability distribution is always the first step of spatial analysis in environmental geochemistry, as many multivariate statistical analysis and spatial analysis of geochemical data are based on the assumption that the data under study follows a normal distribution. In addition, the spatial structures and distribution patterns of the concentration for geochemical elements can reflect various geochemical phenomena and processes, and thus quantifying the spatial patterns can provide a deeper understanding of geochemical knowledge. Similarly, quantitatively evaluate the strength of the correlation or spatial relationships between geochemical variables can reveal more information in the datasets (Franzese and Iuliano, 2019). The background in environmental geochemistry can be defined as the natural concentration of harmful substances that are not disturbed by local human activities (Porteous, 1996), while technologies such as visualisation and prediction can provide better understanding on the concentration and distribution of nutrients or PTEs. Moreover, outlier detection and differentiate natural and anthropogenic factors are regarded as the ultimate issues due to their important relationships between environment and health (Wong et al., 2006). However, when dealing with multivariate datasets, these

problems cannot be solved in a simple or efficient way by using traditional techniques. Recently, with the increase in the amount of environmental geochemical data, a variety of advanced spatial analysis and statistical techniques have become useful and effective methods for outlier detection and potential sources identification at the local and regional level (Hou et al., 2017; Pan et al., 2017; Yadav et al., 2019), with spatial autocorrelation analysis (Ord and Getis, 1995), hot spot analysis (Getis and Ord, 1992; Anselin, 1995) and GWR models (Brunsdon et al., 1996; Fotheringham et al., 2002) being the most popular (e.g. Liang et al., 2017; Wu et al., 2019; Reyes et al., 2020). In response to these existing problems, some possible solutions using advanced spatial analysis are summarised in Table 2.1.

Table 2.1: Existing problems and possible solutions of spatial analysis in environmental geochemistry

Existing problem	Possible solution	Example	Reference
Compositional data/closure effect	Log-ratio transformation	Centred log-ratio transformation was applied before multivariate analysis of PTEs	Xu et al., 2021
Probability distribution	Data transformation (e.g., box-cox, normal score), K-S test	The influence of data transformation on identification of pollution hotspots	Zhang et al., 2008a
Spatial structures and patterns	spatial autocorrelation, variogram, fractal/multifractal	Spatial distribution of total organic carbon (TOC)	Xu et al., 2019
Correlation and spatial relationships	PCC, PCA, GWR	Spatially varying relationships between TOC and pH	Xu and Zhang, 2021
Background and thresholds	Plot (e.g., histogram, boxplot), regression analysis, fractal/multifractal	Establishing geochemical background and threshold for 53 chemical elements	Reimann et al., 2018
Visualisation	IDW, Kriging	Visualisation of point-to-area data transformation for environmental health research	Meng et al., 2019
Prediction	Kriging, multivariate regression, GWR	Spatial modelling and mapping of soil organic carbon	Zhang et al., 2011
Outlier detection	Univariate statistics, PCA, LISA	Identification of contamination hotspots of rare earth elements	Yuan et al., 2018
Distinction between natural and anthropogenic factors	EF, hot spot analysis, GWR	Spatially varying relationships between Pb and Al	Yuan et al., 2020

2.4 Big data era and spatial machine learning

The ‘four Vs’ characteristics of big data are exemplary in the field of environmental geochemistry: volume, variety, velocity and veracity (see Fig. 2.1; Reichstein et al., 2019). In the big data era of environmental geochemistry, one of the striking challenges is to extract useful knowledge and interpretable information from large and diverse data sources. The unprecedented volume and complexity of these data makes it difficult to rely on traditional tools for data management and processing (Vitolo et al., 2015), which requires the improvement of spatial techniques and statistical models. In recent years, the development of machine learning algorithms (MLA) become useful tools for processing problems on prediction, classification and regression, which have been widely applied in data mining and problem-solving in environmental geochemistry. In addition to traditional MLA, the commercial ArcGIS software developed by Environmental Systems Research Institute (ESRI) also supports SML technology by integrated advanced spatial analysis techniques. GIS-based SML technology can be also used to process prediction, classification and identification of hidden patterns of clusters. At present, the combination of classical MLA and SML technique has played a key role on spatial problem-solving in environmental geochemistry (Reimann et al., 2011; Fotheringham and Rogerson, 2013; Povak et al., 2014; Tarasov et al., 2018; Ghezelbash et al., 2019; Du et al., 2020; Xu et al., 2021). Moreover, incorporating geographical concepts directly into the spatial methods of calculation can lead to a deeper understanding of spatial data (Bennett, 2018). In the future, as the intersection of GIS and ML continues to expand, spatial analysis is expected to play a more important role in environmental geochemical studies.

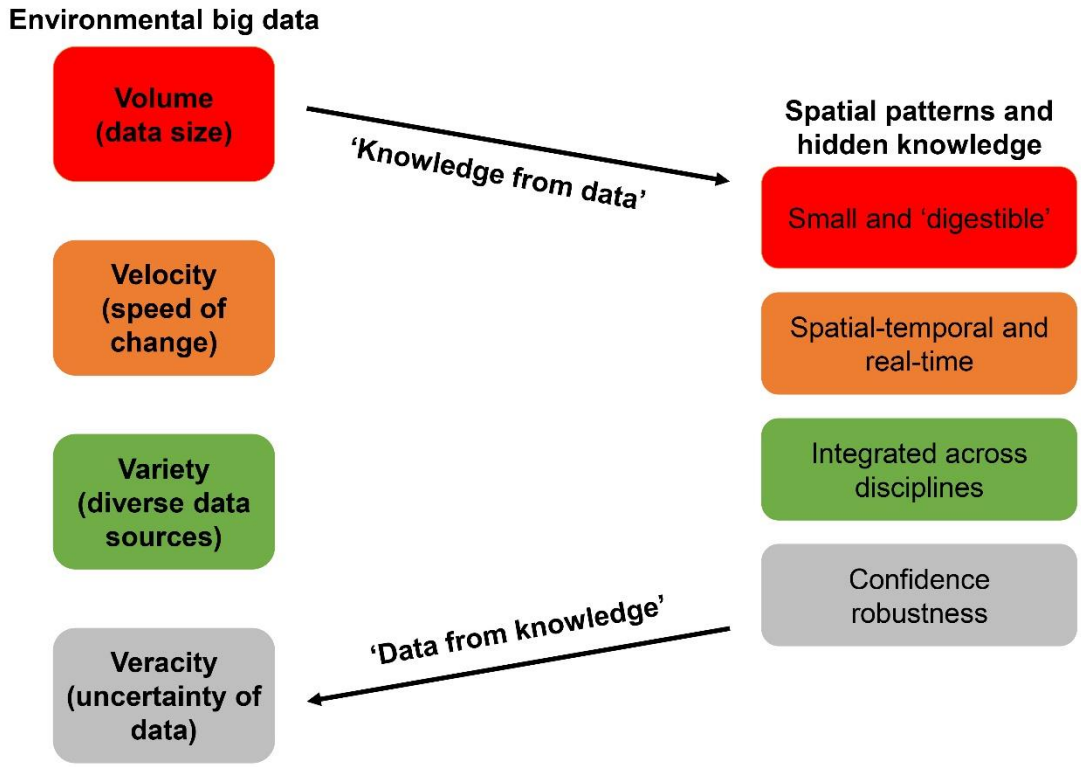


Figure 2.1: Characteristics and challenges of environmental geochemistry in the big data era (reproduced from Reichstein et al., 2019).

Another challenge is to explore the spatio-temporal trends of environmental geochemical data. Regional geochemical surveys are usually large-scale and time-consuming, and thus it is difficult to regularly conduct field works and collect samples (e.g., soil samples) over a long period of time. The spatial analysis of GIS is mature in identifying spatial distribution patterns, while it is difficult to deal with temporal variation due to the update and monitoring at the sampling locations. Therefore, it requires the advancement of spatio-temporal analysis models and combination with real-time data on regular environment monitoring in the future.

2.5 The applications of GIS-based spatial analysis on TOC

2.5.1 Spatial distribution and variation of TOC

As an important indicator of soil fertility and atmospheric environment, studies on the spatial distribution and variation of TOC contents are able to contribute to the improvement of agricultural productivity as well as mitigating global warming. Therefore, GIS-based spatial analysis techniques have been widely applied on quantification and visualisation of the spatial distribution patterns for TOC contents (Kumar et al., 2013). For example, McGrath and Zhang (2003) used spatial interpolation technique and local Moran's I index to investigate the spatial distribution and outliers of soil TOC contents in the grassland of Ireland. Subsequently, they performed spatial statistical techniques to explore the spatial-temporal changes of soil TOC contents during two period between 1964 and 1996, and successfully identified a significant increase of TOC storage in the eastern coastal areas (Zhang and McGrath, 2004). With the improvement of mapping and prediction requirements, an increasing number of advanced statistical techniques and regression models have been used to identify the spatial distribution patterns of TOC contents at the regional level. Meersmans et al. (2008) performed multiple regression model to assess the spatial distribution of TOC in Belgium. Zhang et al. (2011) introduced various environmental covariates into the GWR model and greatly improved the accuracy for prediction and mapping of soil TOC in Ireland. Moreover, in recent years, the development of SML has significantly improved the accuracy and efficiency of regional surveying and mapping on TOC contents (e.g., Were et al., 2015; Chen et al., 2019). These advanced spatial techniques have been proved to have greater potential over traditional spatial interpolation methods in predicting and mapping the spatial patterns of TOC content (Bhunia et al., 2018).

However, the previous studies mainly focused on mapping and prediction on TOC content in the soils, while they often ignored its spatial variation and dynamic. Total organic carbon

is a dynamic component of the terrestrial system, with not only internal changes at horizontal and vertical level, but also external exchanges with the atmosphere and the biosphere (see Fig. 2.2) (Zhang and McGrath, 2004). The influences on the spatial variation of TOC content are extremely complicated which include both natural and anthropogenic factors. The natural factors include topography, climate (i.e., temperature, precipitation), soil parent materials (PMs) and soil properties (e.g., pH, soil texture), while anthropogenic factors are related to human activities, such as land use, cultivation method, site management and fertilisers etc. (Jenny, 1980; Jackson et al., 2002; Lal, 2005; Jandl et al., 2007). Although all of these factors play some roles on the soil TOC content in different regions, most of them generally follow similar spatial patterns (Wiesmeier et al., 2019). For example, a climate of low temperature and high rainfall is conducive to the accumulation of OM (Rustad and Fernandez, 1998). Also, the TOC content in clay and silt particles is significantly higher than that of coarse-grained soil (Schimel and Parton, 1986). In addition, it is considered that land use and cultivation method have a great influence on TOC content, especially in poor quality soils, such as arid or semi-arid regions (West et al., 1994; Su et al., 2009). Therefore, the investigation on the spatial relationships between TOC and various environmental variables as well as related human factors is extremely important, which is able to provide a better understanding for its dynamic and variation.

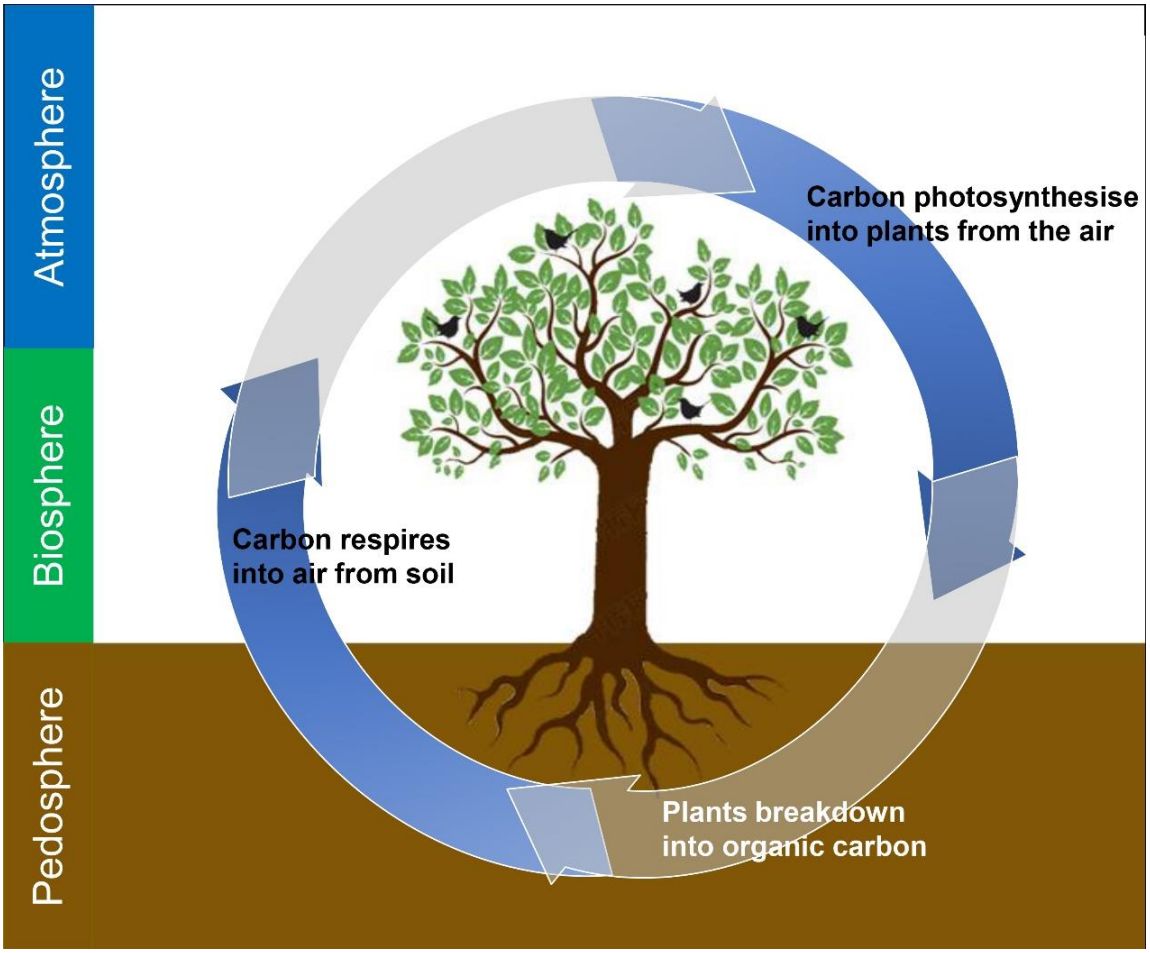


Figure 2.2: The simplified carbon cycle biosphere map (reproduced from www.biochar.org).

2.5.2 Spatially varying relationships between TOC contents and pH values

Soil pH value is considered as the most important parameter among all the environmental variables and influencing factors that related to TOC content. This is because pH is one of the key variables that determine the availability of almost all essential plant nutrients (Fabian et al., 2014). The ability of soil to maintain and supply nutrients is closely related to its cation exchange capacity (CEC), which is affected by soil pH values. Moreover, the natural inherent relationship between TOC and pH is well known (McGrath and Zhang,

2003; Fabian et al., 2014). Generally, the TOC contents and pH values have been found to maintain a negative correlation under natural conditions at various scales, a feature attributed to their innate internal relationship (Andersson and Nilsson, 2001; Reisser et al., 2016). The overall negative correlation between TOC contents and pH values could be related to multiple processes. Organic Carbon is the progenitor of carbonic acid, and its decomposition releases organic acids, leading to lower soil pH values (McGrath and Zhang, 2003). On the contrary, relatively high pH values accelerate the decomposition of soil organic carbon, and thus resulting in a decrease in TOC storage capacity (Andersson and Nilsson, 2001). However, due to complex influencing factors, this original negative relationship may be disturbed or masked at the local scale.

Previous studies have only briefly discussed the inherent negative correlation between TOC and pH, and only a few papers have conducted quantitative analysis on their negative correlation. The reviewed relevant literatures are listed in Table 2.2, as examples, four references with quantitative statistical analysis on the relationship between these two variables were found (McGrath and Zhang, 2003; Korkanç, 2014; Wang et al., 2016; Gebrehiwot et al., 2018). However, contradictory results of positive correlation have been reported in a limited number of two papers (Wang et al., 2010; Luo et al., 2017). The positive correlation is due to the combination of complex influencing factors and is worthy of further study. Moreover, their relationships can be varying at different locations of sub-regions (de Moraes Sa et al., 2009) or different soil layers (Zhang et al., 2018). When considering the ‘location’, the concept of ‘spatially varying relationship’ has not been well understood, which is the focus of this research. The concept of ‘spatially varying relationship’ refers to that the relationships between independent variable and dependent variable(s) are not constant over the space (Fotheringham et al., 2002), but constantly varying with the change of spatial locations. Use of the local statistical methods such as hot spot analysis (Getis-Ord Gi* statistic) and GWR is helpful to quantify the varying relationship between input environmental variables over space in an objective way.

Table 2.2: Overview of relationships between TOC and pH values in the past literatures.

Relationship	Author(s)	Method	Descriptions	Reasons
Negative correlation	McGrath and Zhang, 2003	Global Pearson (linear) correlation coefficient	Negative correlation between TOC and pH value in Ireland (r =0.17)	Innate relationship: organic acids lead to low pH
	Korkanç, S. Y., 2014	Global Pearson (linear) correlation coefficient	Negative correlation between TOC and pH value (r =-0.274)	Innate relationship: organic acids lead to low pH
	Wang et al., 2016	Global Redundancy analysis (RDA)	pH value negatively correlated with impact on soil organic carbon	Innate relationship: organic acids lead to low pH
	Gebrehiwot et al., 2018	Global Pearson (linear) correlation coefficient	(r=-0.126) Weak negative correlation was observed between TOC and pH value	Innate relationship: organic acids lead to low pH
Positive correlation	Wang et al., 2010	Global Pearson (linear) correlation coefficient	Positive relationship between TOC and pH value in the study area (r = 0.549, p =0.000)	Complex influences of soil bulk density and landscape, while the causative relationship between TOC and pH is complicated
	Luo et al., 2017	Path model (i.e., structural equation model)	pH significantly and positively associated with TOC	Combination of soil properties (e.g., particle size, CEC, clay and silt) and climate factors
Contradictory relationships	de Moraes Sa et al., 2009	Global Pearson (linear) correlation coefficients	Changes in the relationships between TOC concentration and pH values in the tillage chronosequence	Tillage chronosequence is the key factor to influence the soil pH values, and thus the original negative relationship changes
	Zhang et al., 2018	Liner regression model	Significant and positive correlation of soil pH value with storage of TOC was found for all soil layers except for 10–20 cm (negative)	Slope, climate, grazing intensity are main reasons for the contradictory relationships

2.6 The application of GIS-based spatial analysis on PTEs

2.6.1 Potentially toxic elements (PTEs)

Reimann et al. (2018) defined the background knowledge and thresholds of PTEs in detail based on GEMAS data in European agricultural soil, including silver (Ag), boron (B), arsenic (As), barium (Ba), bismuth (Bi), cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), mercury (Hg), manganese (Mn), molybdenum (Mo), nickel (Ni), lead (Pb), antimony (Sb), selenium (Se), tin (Sn), uranium (U), vanadium (V) and zinc (Zn). These PTEs are inherently non-biodegradable, and excessive emissions will cause abnormal enrichment of PTE concentrations in the environment (Wong et al., 2006). Although some PTEs (e.g., Cu and Zn) are regarded as essential elements for the growth of animals and plants, the presence of other PTEs can be highly toxic to plants and organisms, especially As, Cd, Hg and Pb (Kabata-Pendias, 2004; Hooda, 2010; Zeng et al., 2011). In urban environments, especially in urban soils, humans may be exposed to long-term input and accumulated PTE through inhalation, ingestion and dermal contact (Boyd et al., 1999, Mielke et al., 1999). Inorganic arsenic is highly toxic (Järup, 2003). Long-term As exposure can cause gastrointestinal symptoms, high blood pressure and chronic cardiovascular disease, etc. It may seriously damage the cardiovascular and central nervous system, and even lead to death (WHO, 2001). In addition, long-term exposure of the human body to Cd may cause kidney damage, and inhalation of cadmium dust or particles may even be life-threatening (Seidal et al., 1993; Barbee and Prince, 1999). Mercury is a global pollutant that affects the health of humans and ecosystems. The mercury pollution in the environment is mainly driven by anthropogenic emissions, and it greatly exceeds natural geogenic sources (FAO and UNEP, 2021). Lead is the most frequently reported soil pollutant over the world, and the exposure to Pb pollution can cause irreversible damage to the nervous system, especially for children, which are more susceptible to Pb contamination from urban soil (Li et al., 2011). For children, severe Pb damage includes symptoms such as mental decline, lack of attention, and anaemia (Bellinger, 2004; Counter

et al., 2008; Zahran et al., 2009). Moreover, the increase of Pb concentration in urban environment is closely related to the increase in the crime rate (Mielke and Zahran, 2012). Long-term exposure to Zn can affect cholesterol balance and fertility (Zhang et al., 2012), and Cu, Ni and Cr above the background values can also have adverse effects on the human body (USEPA, 2000). Therefore, identification of the spatial patterns of PTEs and assessment on their ecological and health risks has become one of the most important tasks in the current environmental research in many developed and developing countries (Rodriguez et al., 2007).

2.6.2 Sources and background of PTEs in the soils

Soil is regarded as the most important sink of PTEs (Wong et al., 2006), receiving contamination from both disposals from the ground and atmospheric deposition. Potentially toxic elements are usually found at trace levels in soils and plants, while the concentrations of PTEs in the soil have been increasing since the industrial revolution. To date, there are some special challenges in solving the soil PTE contamination (Hou et al., 2017):

(1) Potentially toxic elements are not biodegradable, and their concentrations are naturally accumulated in the soil rather than reduced (Maas et al., 2010);

(2) They have a wide range of health effects on humans, and the difference in bioavailability makes health risk assessment more complicated (Walker et al., 2003);

(3) There are many sources of diffusion for PTE pollution (Nriagu and Pacyna, 1988).

Therefore, the quantitative assessment of elevated PTE concentrations on the environment and health has always been the key focus of research in environmental studies, which includes not only the investigation on the increase in PTE concentrations caused by natural factors (Spijker, 2005; Jordan et al., 2007; Zhang et al., 2008b; Argyraki and Kelepertzis, 2014), but also focusing more on the evaluation of elevated concentrations associated with

anthropogenic factors (Hursthouse, 2001; Morillo et al., 2007; Meunier et al., 2010; Okorie et al., 2011). When there is no human interference, PTEs usually exist in trace levels and will not cause negative effects on the environment or human health (Alloway, 2013a). Natural factors are mainly related to geogenic occurrences, as well as soil formation and parent materials (Tipping et al., 2006; Reimann et al., 2014; Birke et al., 2017). Generally, each lithology has a relatively fixed control on a single element, and the overall associations between the 15 selected PTEs and different lithologies are summarised in Table 2.3. On the other hand, anthropogenic sources are related to human activities, including industrial, waste, traffic (vehicle emissions, fuel) and agricultural inputs, etc (e.g., Cloquet et al., 2006, Ettler et al., 2008, Aelion et al., 2009, Davis et al., 2009, Dao et al., 2014). The elevated PTE concentrations caused by geogenic sources are predominantly reflected in large and continuous spatial patterns, which are usually observed at a relatively large-scale area or with less human activities (Gloaguen and Passe, 2017; Jia et al., 2020; Xu et al., 2021). On the contrary, anthropogenic pollution is mainly characterised by points and scattered patterns on the spatial distribution maps of PTE concentrations, which is usually observed around urban areas with intensive industrial activities, mining and traffic emission (Zhang, 2006; Marchant et al., 2011; Delbecque and Verdoodt, 2016).

Table 2.3: Summary of the existing literatures on concentrations for selected PTEs in different bedrocks and geological features

Element	High concentration	Low concentration	Reference
As	Greywacke (shale), mudstone, schist	Basalt, quartzite, sandstone	Smedley and Kinniburgh, 2002; Tarvainen et al., 2013; McIlwaine et al., 2017
Ba	Carbonate, granite	Limestone, mafic (basalt)	Reimann et al., 2007; Reimann et al., 2014
Bi	Granite, shale	Sandstone	Reimann et al., 2014
Co	Greenstone, basalt	Limestone, sandstone	Farmer, 2014; McIlwaine et al., 2014; Albanese et al., 2015
Cr	Basalt	limestone, granite	Farmer, 2014; McIlwaine et al., 2014; Albanese et al., 2015
Cu	Basalt, shale	Granite, organic matter	Wedepohl, 1978; Reimann et al., 2014; Albanese et al., 2015
Mn	Basalt	Granite, quartzite, schist	Reimann et al., 2007; Reimann et al., 2014
Mo	Granite, greywacke (shale), schist	Basalt	McIlwaine et al., 2017; Reimann et al., 2018
Ni	Basalt, shale	Granite, limestone, quartzite, sandstone	Farmer, 2014; Reimann et al., 2014; Albanese et al., 2015; Jordan et al., 2018
Pb	Granite, peat, shale	Basalt, limestone	McIlwaine et al., 2014; Reimann et al., 2014; Palmer et al., 2015
Sb	Coal, peat	Basalt, sandstone	Reimann et al., 2014; McIlwaine et al., 2015
Sn	Granite, peat, shale	Basalt, limestone	Reimann et al., 2014; McIlwaine et al., 2015
U	Granite, shale	Basalt, sandstone	Alloway, 2013b; McKinley et al., 2013; Négrel et al., 2018
V	Basalt, shale	Limestone	Barsby et al., 2012; Reimann et al., 2014
Zn	Alluvium, basalt, shale	Granite	Reimann et al., 2014; McIlwaine et al., 2017

At present, based on large-scale regional survey, the background values and thresholds of soil PTEs have been established in different countries and regions, such as Finland (MEF, 2007), China (Wei and Yang, 2010), Australia (Reimann and de Caritat, 2017) and Europe (Carlon et al., 2007; Reimann et al., 2018), etc. The threshold (or baseline) is defined as the conservative concentration of PTE in the soil (Reimann et al., 2018), which indicates that no adverse effects on the environment and humans are expected to occur below this concentration. In contrast, the PTE concentration exceeds the soil background and threshold values are likely to cause adverse impacts on the environment and human body, especially through plants and crops in the agricultural soil (Reimann et al., 2014b).

Table 2.4: Thresholds and guideline values for selected PTEs in Europe (extract from MEF, 2007)

Element	Threshold value (mg/kg)	Lower guideline value (mg/kg)	Higher guideline value (mg/kg)
As	5	50	100
Cd	1	10	20
Co	20	100	250
Cr	100	200	300
Cu	100	150	200
Hg	0.5	2	5
Ni	50	100	150
Pb	60	200	750
Sb	2	10	50
V	100	150	250
Zn	200	250	400

2.6.3 Spatial distribution patterns and source identification of soil PTEs

Due to the complexity of the geochemical background, the concentration of PTE is usually influenced by multiple controls that combined both natural and anthropogenic factors, and

thus it is difficult to distinguish the potential sources of PTEs in the soil. The spatial distribution patterns of PTEs can usually be related to the source of enrichment, thereby revealing the spatial association with pollution sources from the local perspective. Looking back in history, it is a common way to apply classical statistics to the spatial distribution of PTE in soil (Webster et al., 1994; Einax and Soldt, 1995; Markus and McBratney, 1996). Geostatistical methods are useful tools for analysing and predicting the values of soil geochemical variables that associated with spatial or spatiotemporal phenomena (Goovaerts, 2005). However, it is still difficult to distinguish the source of element concentration only based on their spatial distribution patterns, while the combination of multivariate analysis and geostatistics is regarded as a more effective method. This method usually involves performing principal component analysis (PCA) or factor analysis (FA) and using geostatistical tools to map the derived components (factors) scores (Rodríguez et al., 2006, López et al., 2008, Lado et al., 2008). The combination of multivariate analysis and geostatistics provides a way to analyse multiple elements of the entire data set instead of an individual element, and describe the complex relationship and influencing factors of PTE in soil more precisely than univariate statistics (Borůvka et al., 2005). In recent years, with the development of GIS platforms, advanced spatial analysis and SML technologies can be easily performed on identification of the hidden patterns of PTEs and reveal the spatial association with pollution sources (Fotheringham and Rogerson, 2013; Hou et al., 2017). The GIS and GIS-based spatial techniques have been proven as promising tools for understanding the background of PTEs and studying the soil contamination (Zhou and Xia, 2010; Yuan et al., 2018; Meng et al., 2020).

2.6.4 Spatially varying relationships between Pb and Al in the soils

The spatial relationships and correlations between geochemical variables contain a large amount of information (Reimann and de Caritat, 2005), while many traditional statistical techniques do not have the potential to capture the role of spatial correlations (Hou et al., 2017). The development of the GWR model solved the limitation of exploring the spatially

varying relationships between geochemical elements and environmental variables (Brunsdon et al., 1996; Fotheringham et al., 2002), and has been widely applied on the pollution assessment and source appointment of PTEs (e.g., Fei et al., 2019; Liu et al., 2020).

Lead is considered to be one of the most common human-controlled PTEs in the soils. Previous studies have widely reported the abnormally elevated concentrations of Pb in soil caused by human activities in urban and industrial areas, including mining, leaded gasoline, coal combustion, industrial waste and construction, etc (e.g., Li et al., 2014; Marrugo-Negrete et al., 2017; Wu et al., 2019). In addition, the use of leaded gasoline and traffic emission in urban areas can cause Pb pollution in the air, which in turn pollutes rural soil through atmospheric deposition (Shotyk, 2002; Novák et al., 2003). The relationship between Pb and other reference elements (e.g., Al, Ti, Zr) generally maintains a positive correlation under most natural conditions (Schropp and Windom, 1988; Spark, 2010), which could be expected for soils derived from continental crust (Walsh and Barry, 1957). This relationship has been used to distinguish the natural and anthropogenic sources (Shotyk et al., 2002; Sezgin et al., 2003; Le Roux et al., 2004). However, the original positive correlation may be interfered or masked by external influences. Therefore, the varying relationships (i.e., negative correlation) or weakened relationships that explored by GWR and GWPCC can provide an effective way to indicate the spatial association with potential pollution sources. For example, Yuan et al. (2020) applied GWR to identify the spatially varying relationships between Pb and Al concentrations in the urban soil of London that associated by natural and human influence, and highlighted the effects of industry and green space in urban environment. In the big data era, the patterns of spatially varying relationships have great prospects and are worthy of further exploration.

2.7 Summary

The literature review summarised the development and applications of GIS-based spatial analysis in the big data era of environmental geochemistry, as well as the opportunities and challenges for environmental data mining. In addition, this chapter also specifically discussed the applications of GIS-based spatial analysis on the distribution patterns and spatially varying relationships of TOC and PTEs, which aimed to identify the potential influences from natural and anthropogenic factors.

Overall, the past literature indicated that GIS and GIS-based spatial techniques: (1) provide a promising and efficient way for processing environmental geochemical data sets; (2) can be used to reveal the spatial relationships and hidden spatial associations between geochemical variables; (3) can be used to identify and visualise the spatial distribution patterns and variation of TOC; (4) can be used to distinguish the natural and anthropogenic sources and controlling factors of PTE.

Reference

Aelion, C.M., Davis, H.T., McDermott, S., Lawson, A.B., 2009. Soil metal concentrations and toxicity: associations with distances to industrial facilities and implications for human health. Sci. Total Environ., 407, 2216-2223.

Aitchison, J., 1986. The statistical analysis of compositional data, Chapman and Hall, London, UK, p. 416.

Albanese, S., De Vivo, B., Lima, A., Cicchella, D., 2007. Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). Journal of Geochemical Exploration, 93(1), 21-34.

Albanese, S., Sadeghi, M., Lima, A., Cicchella, D., Dinelli, E., Valera, P., Falconi, M., Demetriades, A., De Vivo, B., The GEMAS Project Team, 2015. GEMAS: cobalt, Cr, Cu and Ni distribution in agricultural and grazing land soil of Europe. J. Geochem. Explor., 154, 81-93.

Ali, M.H., Mustafa, A.-R.A., El-Sheikh, A.A., 2016. Geochemistry and spatial distribution of selected heavy metals in surface soil of Sohag, Egypt: a multivariate statistical and GIS approach. Environ. Earth Sci., 75, 1257. https://doi.org/10.1007/s12665-016-6047-x.

Alloway, B.J., ed. 2013a. Heavy Metals in Soils. Environmental Pollution. Dordrecht, Springer Netherlands.

Alloway, B.J., 2013b. Bioavailability of Elements in Soil. O. Selinus (Ed.), Essentials of Medical Geology, Springer, Dordrecht, 10.1007/978-94-007-4375-5_15.

Alpaydin, E., 2010. Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0.

Andersson, S., Nilsson, S.I., 2001. Influence of pH and temperature on microbial activity, substrate availability of soil-solution bacteria and leaching of dissolved organic carbon in a mor humus. Soil Biol. Biochem., 33, 1181-1191.

Anselin, L., 1995. Local indicators of spatial association — LISA. Geogr. Anal., 27, 93-115.

Argyraki, A., Kelepertzis, E., 2014. Urban soil geochemistry in Athens, Greece: the importance of local geology in controlling the distribution of potentially harmful trace elements. Sci. Total Environ., 482–483, 366-377, 10.1016/j.scitotenv.2014.02.133.

Bailey, T.C., 1994. A review of statistical spatial analysis in geographical information systems. Spatial analysis and GIS, 13-44.

Barbee Jr, J.Y., Prince, T.S., 1999. Acute respiratory distress syndrome in a welder exposed to metal fumes. South. Med. J., 92 (5), 510-512.

Barsby, A., McKinleya, J.M., Ofterdinger, U., Young, M., Cave, M., Wraggd, J., 2012. Bioaccessibility of trace elements in soils in Northern Ireland. Sci. Total Environ., 433, 398-417.

Bellinger, D.C., 2004. Lead. Pediatrics, 113, 1016-1022.

Bennett, L., 2018. Machine learning in ArcGIS. Available at: https://www.esri.com/about/newsroom/arcuser/machine-learning-in-arcgis/?rmedium=arcuser&rsource=https://www.esri.com/esri-news/arcuser/spring-2018/machine-learning-in-arcgis.

Bhunia, G. S., Shit, P. K., Maiti, R., 2018. Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC). Journal of the Saudi Society of Agricultural Sciences, 17 (2), 114-126.

Birke, M., Reimann, C., Oorts, K., Rauch, U., Demetriades, A., Dinelli, E., Ladenberger, A., Halamić, J., Gosar, M., Jähne-Klingberg, F., the GEMAS Project Team, 2016. Use of GEMAS data for risk assessment of cadmium in European agricultural and grazing land soil under the REACH Regulation. Appl. Geochem., 74, 109-121.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer. ISBN 978-0-387-31073-2.

Borůvka, L., Vacek, O., Jehlička, J., 2005. Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. Geoderma, 128 (3–4), 289-300.

Boyd, H.B., Pedersen, F., Cohr, K.H., Damborg, A., Jakobsen, B.M., Kristensen, P., Samsøe-Petersen, L., 1999. Exposure scenarios and guidance values for urban soil pollutants. Regul. Toxicol. Pharmacol., 30, 197-208.

Brunsdon, C., Fotheringham, A.S., Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geogr. Anal. 28, 281-298.

Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., (Eds.) 2006. Compositional data analysis in the geosciences – from theory to practice. Geological Society of London, Special Publication 264.

Carlon, C., D'Alessandro, M., Swartjes, F., 2007. Derivation Methods of Soil Screening Values in Europe. A Review and Evaluation of National Procedures towards Harmonization. EUR 22805 EN. European Communities, Luxembourg.

Chen, D., Chang, N., Xiao, J., Zhou, Q., Wu, W., 2019. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. Sci. Total Environ., 669, 844-855.

Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev., 32, 314-324.

Cheng, Q., Xu, Y., Grunsky, E., 2000. Integrated spatial and spectrum method for geochemical anomaly separation. Nat. Resour. Res., 9, 43-52.

Clarke, K.C., 1986. Advances in geographic information systems, computers, environment and urban systems, 10, 175–184.

Cloquet, C., Carignan, J., Libourel, G., 2006. Isotopic composition of Zn and Pb atmospheric depositions in an urban/periurban area of northeastern France. Environ. Sci. Technol., 40, 6594-6600.

Counter, S.A., Buchanan, L.H., Ortega, F., 2008. Zinc protoporphyrin levels, blood lead levels and neurocognitive deficits in Andean children with chronic lead exposure. Ann. Clin. Biochem., 41, 41-47.

Dao, L.G., Morrison, L., Zhang, H., Zhang, C., 2014. Influences of traffic on Pb, Cu and Zn concentrations in roadside soils of an urban park in Dublin, Ireland. Environ. Geochem. Health, 36, 333-343.

Darnley, A.G., 1990. International geochemical mapping: a new global project. J. Geochem. Explor., 39 (1-2), 1-13.

Davis, H.T., Aelion, C.M., McDermott, S., Lawson, A.B., 2009. Identifying natural and

anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. Environ. Pollut., 157, 2378-2385.

de Moraes Sa, J.C., Cerri, C.C., Lal, R., Dick, W.A., de Cassia Piccolo, M., Feigl, B.E., 2009. Soil organic carbon and fertility interactions affected by a tillage chronosequence in a Brazilian Oxisol. Soil Till. Res., 104 (1), 56-64.

Delbecque, N., Verdoodt, A., 2016. Spatial patterns of heavy metal contamination by urbanization. J. Environ. Qual., 45, 9-17.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li., E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: A review. Journal of Geovisualization and Spatial Analysis, 4, 1-25.

Dung, T.T.T., Cappuyns, V., Swennen, R., Phung, N.K., 2013. From geochemical background determination to pollution assessment of heavy metals in sediments and soils. Rev. Environ. Sci. Biotechnol., 12 (4), 335-353.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol., 35, 279-300.

Einax, J., Soldt, U., 1995. Geostatistical investigations of polluted soils. Fresenius' J. Anal. Chem., 351, 48-53.

Ettler, V., Sebek, O., Grygar, T., Klementova, M., Bezdicka, P., Slavikova, H., 2008. Controls on metal leaching from secondary Pb smelter air-pollution-control residues. Environ. Sci. Technol., 42, 7878-7884.

Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E. 2014. GEMAS: spatial distribution of the pH of European agricultural and grazing land soil. Appl. Geochem., 48, 207-216.

Farmer, G.L., 2014. Continental basaltic rocks. In: Chapter 4.3 in R.L. Rudnick, H. Holland, K. Turekian (Eds.), The Crust, 2nd ed., Treatise on Geochemistry, no. 4, pp. 75–100.

FAO and UNEP., 2021. Global Assessment of Soil Pollution: Report. Rome. https://doi.org/10.4060/cb4894en.

Fei, X., Christakos, G., Xiao, R., Ren, Z., Liu, Y., Lv, X., 2019. Improved heavy metal mapping and pollution source apportionment in Shanghai City soils using auxiliary

information. Sci. Total Environ., 661, 168-177.

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci. Total Environ., 407, 6100-6108.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons Ltd., Chichester, UK, 284 pp.

Fotheringham, S., Rogerson, P. (Eds.)., 2013. Spatial analysis and GIS. CRC Press.

Gałuszka, A., 2007. A review of geochemical background concepts and an example using data from Poland. Environ. Geol. 52, 861–870.

Gebrehiwot, K., Desalegn, T., Woldu, Z., Demissew, S., Teferi, E., 2018. Soil organic carbon stock in Abune Yosef afroalpine and sub-afroalpine vegetation, northern Ethiopia. Ecol. Process., 7 (1), p. 6.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal., 24 (3), 189-206.

Ghezelbash, R., Maghsoudi, A., Carranza. E.J.M., 2019. Mapping of single- and multi-element geochemical indicators based on catchment basin analysis: application of fractal method and unsupervised clustering models. J. Geochem. Explor., 199, 90-104.

Gloaguen, T.V., Passe, J.J., 2017. Importance of lithology in defining natural background concentrations of Cr, Cu, Ni, Pb and Zn in sedimentary soils, northeastern Brazil. Chemosphere, 186, 31-42. 10.1016/j.chemosphere.2017.07.134.

Goovaerts, P., 2005. Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. Int. J. Health Geogr. 4, 31. https://doi.org/10.1186/1476-072X-4-31.

Hawkes, H.E., Bloom, H., 1955. Heavy metals in stream sediment used as exploration guides. Min. Eng., 8, 1121–1126.

Hooda, P.S., 2010. Elements in Soils. John Wiley & Sons, Chichester, U.K.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environ. Pollut., 231, 1188-1200.

Hursthouse, A., 2001. The relevance of speciation in the remediation of soils and sediments

contaminated by metallic elements— an overview and examples from Central Scotland, UK RID A-9005-2010. J. Environ. Monit., 3, 49-60.

Jackson, R.B., Banner, J.L., Jobbagy, E.G., Pockman, W.T., Wall, D.H., 2002. Ecosystem carbon loss with woody plant invasion of grasslands. Nature 418, 623-626.

Jandl, R., Lindner, M., Vesterdal, L., Bauwens, B., Baritz, R., Hagedorn, F., Johnson, D.W., Minkkinen, K., Byrne, K.A., 2007. How strongly can forest management influence soil carbon sequestration? A review. Geoderma 137, 253-268.

Jenny, H. 1980. The Soil Resource, Origin and Behavior. Springer-Verlag, New York, 392 pp.

Jia, Z., Wang, J., Zhou, X., Zhou, Y., Li, Y., Li, B., Zhou, S., 2020. Identification of the sources and influencing factors of potentially toxic elements accumulation in the soil from a typical karst region in Guangxi, Southwest China. Environ. Pollut., 256, 113505.

Johnson, C.C., Ander, E.L., 2008. Urban geochemical mapping studies: how and why we do them. Environ. Geochem. Health, 30 (6), 511. https://doi.org/10.1007/s10653-008-9189-2.

Jordan, C., Zhang, C.S., Higgins, A., 2007. Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland. J. Geochem. Explor., 93,135-152.

Jordan, G., Petrik, A., De Vivo, B., Albanese, S., Demetriades, A., Sadeghi, M., T.G.P. Team, 2018. GEMAS: spatial analysis of the Ni distribution on a continental-scale using digital image processing techniques on European agricultural soil data. J. Geochem. Explor., 186, 143-157, 10.1016/j.gexplo.2017.11.011.

Kabata-Pendias, A., 2004. Soil–plant transfer of trace elements—an environmental issue. Geoderma, 122, 143-149.

Korkanç, S.Y., 2014. Effects of afforestation on soil organic carbon and other soil properties. Catena, 123, 62-69.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

Kumar, S., Lal, R., Liu, D., Rafiq, R., 2013. Estimating the spatial distribution of organic carbon density for the soils of Ohio, USA. J. Geogr. Sci., 23 (2),280-296.

Lado, L.R., Hengl, T., Reuter, H., 2008. Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. Geoderma, 148, 189-199.

Lado, L.R., Hengl, T., Reuter, H.I., 2008. Heavy metals in European soils: a geostatistical analysis of the FOREGS geochemical database. Geoderma, 148, 189-199

Lal, R., 2005. Forest soils and carbon sequestration. Forest Ecol. Manag. 220, 242-258.

Lars, J., 2003. Hazards of heavy metal contamination, Br. Med. Bull., 68, 167–182, https://doi.org/10.1093/bmb/ldg032.

Le Roux, G., Weiss, D., Grattan, J., Givelet, N., Krachler, M., Cheburkin, A., Rausch, N., Kober, B., Shotyk, W., 2004. Identifying the sources and timing of ancient and medieval atmospheric lead pollution in England using a peat profile from Lindow bog, Manchester. J. Environ. Monit., 6, 502-510.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444.

Lee, C.S.-l., Li, X., Shi, W., Cheung, S.C.-n., Thornton, I., 2006. Metal contamination in urban, suburban, and country park soils of Hong Kong: a study based on GIS and multivariate statistics. Sci. Total Environ., 356, 45-61.

Li, H.B., Yu, S., Li, G.L., Deng, H., Luo, X.S., 2011. Contamination and source differentiation of Pb in park soils along an urban-rural gradient in Shanghai. Environ. Pollut., 159, 3536-3544.

Li, Z.Y., Ma, Z.W., van der Kuijp, T.J., Yuan, Z.W., Huang, L., 2014. A review of soil heavy metal pollution from mines in China: Pollution and health risk assessment. Sci. Total Environ., 468-469, 843-853.

Liang, J., Feng, C., Zeng, G., Gao, X., Zhong, M., Li, X., He, X.Y., Fang, Y., 2017. Spatial distribution and source identification of heavy metals in surface soils in a typical coal mine city, Lianyuan, China. Environ. Pollut., 225, 681-690.

Limpert, E., Stahel, W. A., Abbt, M., 2001. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—

normal or log-normal: that is the question. BioScience, 51 (5), 341-352.

Liu, Y., Fei, X., Zhang, Z., Li, Y., Tang, J., Xiao, R., 2020. Identifying the sources and spatial patterns of potentially toxic trace elements (PTEs) in Shanghai suburb soils using global and local regression models. Environmental Pollution, 264, 114171.

López, J.M., Borrajo, J.L., García, E.D.M., Arrans, J.R., Estévez, M.C.H., Castillo, A.J.S. 2008. Multivariate analysis of contamination in the mining district of Linares (Jaén, Spain). Appl. Geochem., 23, 2324-2336.

Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Chang. Biol., 23 (10), 4430-4439.

M.I. Jordan, T.M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349 (6245) (2015), pp. 255-260.

Maas, S., Scheifler, R., Benslama, M., Crini, N., Lucot, E., Brahmia, Z., Benyacoub, S., Giraudoux, P., 2010. Spatial distribution of heavy metal concentrations in urban, suburban and agricultural soils in a Mediterranean city of Algeria. Environ. Pollut., 158 2294-2301.

Marchant, B.P., Tye, A.M., Rawlins, B.G., 2011. The assessment of point-source and diffuse soil metal pollution using robust geostatistical methods: a case study in Swansea (Wales, UK). Eur. J. Soil Sci., 62, 346-358.

Markus, J.A., McBratney, A.B., 1996. An urban soil study: heavy metals in Glebe, Australia. Aust. J. Soil Res., 34, 453-465.

Marrugo-Negrete, J., Pinedo-Hernández, J., Díez, S., 2017. Assessment of heavy metal pollution, spatial distribution and origin in agricultural soils along the Sinú River Basin, Colombia. Environ. Res., 154, 380-388.

Matschullat, J., Ottenstein, R., Reimann, C., 2000. Geochemical background–can we calculate it?. Environ. Geol., 39 (9), 990-1000.

McGrath, D., Zhang, C., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem., 18 (10), 1629-1639.

McGrath, D., Zhang, C., Carton, O.T., 2004. Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. Environ. Pollut., 127 (2), 239-248.

McIlwaine, R., Cox, S.F., Doherty, R., 2015. When are total concentrations not total? Factors affecting geochemical analytical techniques for measuring element concentrations in soil Environ. Sci. Pollut. Res., 22, 6364-6371, 10.1007/s11356-015-4204-5.

McIlwaine, R., Doherty, R., Cox, S.F., Cave, M., 2017. The relationship between historical development and potentially toxic element concentrations in urban soils. Environ. Pollut., 220, 1036-1049.

McKinley, J.M., Ofterdinger, U., Young, M., Barsby, A. Gavin, A., 2013. Investigating local relationships between trace elements in soils and cancer data. Spat. Stat., 5, 25-41.

MEF (Ministry of the Environment, Finland), 2007. Government Decree on the Assessment of Soil Contamination and Remediation Needs. 214/2007 – in Finnish and Swedish.

Meng, Y.T., Cave, M., Zhang, C.S., 2018. Spatial distribution patterns of phosphorus in top-soils of Greater London Authority area and their natural and anthropogenic factors. Appl. Geochem., 88, 213-220.

Meng, Y., Cave, M., Zhang, C., 2019. Comparison of methods for addressing the point-to-area data transformation to make data suitable for environmental, health and socio-economic studies. Sci. Total Environ., 689, 797-807.

Meng, Y., Cave, M., Zhang, C., 2020. Identifying geogenic and anthropogenic controls on different spatial distribution patterns of aluminium, calcium and lead in urban topsoil of Greater London Authority area. Chemosphere 238, 124541.

Meunier, L., Walker, S.R., Wragg, J., Parsons, M.B., Koch, I., Jamieson, H.E., Reimer, K.J., 2010. Effects of soil composition and mineralogy on the bioaccessibility of arsenic from tailings and soil in gold mine districts of Nova Scotia. Environ. Sci. Technol., 44, 2667-2674.

Mielke, H.W., Gonzales, C.R., Smith, M.K., Mielke, P.W., 1999. The urban environment and children's health: soils as an integrator of lead, zinc, and cadmium in New Orleans, Louisiana, U.S.A. Environ. Res. (Section A), 81, 117-129.

Mielke, H.W., Zahran, S., 2012. The urban rise and fall of air lead (Pb) and the latent surge

and retreat of societal violence. Environ. Int., 43, 48-55.

Mishra, U., Lal, R., Liu, D., Van Meirvenne, M., 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. Soil. Sci. Soc. Am. J., 74 (3), 906-914.

Morillo, E., Romero, A.S., Maqueda, C., Madrid, L., Ajmone-Marsan, F., Grcman, H., Davidson, C.M., Hursthouse, A.S., Villaverde, J., 2007. Soil pollution by PAHs in urban soils: a comparison of three European cities RID A-9005-2010. J. Environ. Monit., 9, 1001-1008.

Négrel, P., De Vivo, B., Reimann, C., Ladenberger, A., Cicchella, D., Albanese, S., Birke, M., De Vos, W., Dinelli, E., Lima, A., O'Connor, P.J., Salpeteur, I., Tarvainen, T., the GEMAS Project Team, 2018. U-Th signatures of agricultural soil at the European continental scale (GEMAS): distribution, weathering patterns and processes controlling their concentrations. Sci. Total Environ., 622–623, 1277-1293.

Nezhad, M.T.K., Tabatabaii, S.M., Gholami, A., 2015. Geochemical assessment of steel smelter-impacted urban soils, Ahvaz, Iran. J. Geochem. Explor., 152, 91-109.

Novák, M., Emmanuel, S., Vile, M., Erel, Y., Véron, A., Pačes, T., Kelman Wieder, R., Vaněček, M., Štěpánová, M., Břízová, E., Hovorka, J., 2003. Origin of lead in eight European peat bogs determined from isotope ratios, strengths, and operation times of regional pollution sources. Environ. Sci. Technol., 37, 437-445.

Nriagu, J.O., Pacyna, J.M., 1988. Quantitative assessment of worldwide contamination of air, water and soils by trace metals. Nature, 333, 134-139.

Okorie, A., Entwistle, J., Dean, J.R., 2011. The application of in vitro gastrointestinal extraction to assess oral bioaccessibility of potentially toxic elements from an urban recreational site. Appl. Geochem., 26, 789-796.

Ord, J. K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal., 27 (4), 286-306.

Overpeck, J.T., Meehl, G.A., Bony, S., Easterling, D.R., 2011. Climate data challenges in the 21st century. Science, 331 (6018), 700-702.

Palmer, S., Mcllwaine, R., Ofterdinger, U., Cox, S.F., Mckinley, J.M., Doherty, R., Wragg, J., Cave, M., 2015. The effects of lead sources on oral bioaccessibility in soil and implications for contaminated land risk management. Environ. Pollut., 198, 161-171.

Pan, H., Lu, X., Lei, K., 2017. A comprehensive analysis of heavy metals in urban road dust of Xi'an, China: contamination, source apportionment and spatial distribution. Sci. Total Environ., 609, 1361-1369.

Pawlowsky-Glahn, V., and Buccianti, A. (Eds.). (2011). Compositional data analysis: Theory and applications. John Wiley & Sons.

Povak, N.A., Hessurg, P.H., McDonnell, T.C., Reynolds, K.M., Sullivn, T.J., Salter, R.B., Crosby. B.J., 2014. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Water Resour. Res., 50, 2798-2814.

Rao, C.R.M., Sahuquillo, A., Sanchez, J.L., 2008. A review of the different methods applied in environmental geochemistry for single and sequential extraction of trace elements in soils and related materials. Water Air Soil Pollut., 189 (1), 291-333.

Reichman, O.J., Jones, M.B., Schildhauer, M.P., 2018. Challenges and opportunities of open data in ecology. Science, 331 (6018), 703-705.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, K., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Reimann, C., Siewers, U., Tarvainen, T., Bityukova, L., Eriksson, J., Gilucis, A., Gregorauskiene, V., Lukashev, V.K., Matinian, N.N., Pasieczna, A., 2003. Agricultural Soils in Northern Europe: A Geochemical Atlas. Geologisches Jahrbuch, Sonderhefte, Reihe D, Heft SD 5, Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany.

Reimann, C., de Caritat, P., 2005. Distinguishing between natural and anthropogenic sources for elements in the environment: regional geochemical surveys versus enrichment factors. Sci. Total Environ., 337 (1-3), 91-107.

Reimann, C., Garrett, R.G., 2005. Geochemical background—concept and reality. Sci. Total Environ., 350 (1-3), 12-27.

Reimann, C., Arnoldussen, A., Englmaier, P., Filzmoser, P., Finne, T.E., Garrett, R.G., Koller, F., Nordgulen, O., 2007. Element concentrations and variations along a 120-km transect in southern Norway – anthropogenic vs. geogenic vs. biogenic element sources and cycles. Appl. Geochem., 22, 851-871.

Reimann, C., Matschullat, J., Birke, M., Salminen, R., 2010. Antimony in the environment: lessons from geochemical mapping. Appl. Geochem., 25 (2), 175-198.

Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2011. Statistical data analysis explained: applied environmental statistics with R. John Wiley & Sons.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O' Connor, P., 2014a. Chemistry of Europe's Agricultural Soils, Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, 523 pp.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014b. Chemistry of Europe's Agricultural Soils, Part B: General Background Information and Further Analysis of the GEMAS Data Set, Geologisches Jahrbuch Reihe B, Band B 103. Schweizerbart Science Publishers, Stuttgart, p. 352.

Reimann, C., de Caritat, P., 2017. Establishing geochemical background variation and threshold values for 59 elements in Australian surface soil. Sci. Total Environ., 578, 633-648.

Reimann, C., Fabian, K., Birke, M., Filzmoser, P., Demetriades, A., Négrel, P., Oorts, K., Matschullat, J., 2018. GEMAS: Establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Appl. Geochem., 88, 302-318.

Reisser, M., Purves, R.S., Schmidt, M.W.I., Abiven, S., 2016. Pyrogenic carbon in soils: a literature-based inventory and a global estimation of its content in soil organic carbon and stocks. Front. Earth Sci., 4, p. 80, 10.3389/feart.2016.00080.

Reyes, A., Thiombane, M., Panico, A., Daniele, L., Lima, A., Di Bonito, M., De Vivo, B., 2020. Source patterns of potentially toxic elements (PTEs) and mining activity contamination level in soils of Taltal city (northern Chile). Environ. Geochem. Health, 42 (8), 2573-2594.

Robertson, D.J., Taylor, K.G., Hoon, S.R., 2003. Geochemical and mineral magnetic characterisation of urban sediment particulates, Manchester, UK. Appl. Geochem., 18 (2), 269-282.

Rodríguez, M.J.A., López, A.M., Grau, C.J.M., 2006. Heavy metal contents in agricultural topsoils in the Ebro basin (Spain). Application of the multivariate geostatistical methods to study spatial variations. Environ. Pollut., 144, 1001-1012.

Rodriguez, L., Rincón, J., Asencio, I., Rodríguez-Castellanos, L., 2007. Capability of selected crop plants for shoot mercury accumulation from polluted soils: phytoremediation perspectives. Int. J. Phytoremediation, 9, 1-13.

Rustad, L.E., Fernandez, I.J., 1998. Experimental soil warming effects on CO2 and CH4 flux from a low elevation spruce–fir forest soil in Maine, USA. Glob. Change Biol., 4, 597-605.

Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamić, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P.J., Olsson, S.A., Ottesen, R.-T., Petersell, V., Plant, J.A., Reeder, S., Salpeteur, I., Sandstrom, H., Siewers, U., Steenfelt, A., Tarvainen, T., 2005. Geochemical atlas of Europe, part 1, background information, methodology and maps. Espoo: Geological Survey of Finland.

Schropp, S.J., Windom, H.L., 1988. 'A Guide to the Interpretation of Metal Concentrations in Estuarine Sediments', in Schropp S.J., and Windom, H.L., (eds). Savannah, Georgia.

Seidal, K., Jörgensen, N., Elinder, C.G., Sjögren, B., Vahter, M., 1993. Fatal cadmium-induced pneumonitis. Scandinavian journal of work, environment & health, 429-431.

Selinus, O.S., Esbensen, K., 1995. Separating anthropogenic from natural anomalies in environmental geochemistry. J. Geochem. Explor., 55, 55–66.

Sezgin, N., Ozcan, H.K., Demir, G., Nemlioglu, S., Bayat, C., 2003. Determination of heavy metal concentrations in street dusts in Istanbul E-5 highway. Environ. Int., 29, 979-985.

Shotyk, W., 2002. The chronology of anthropogenic, atmospheric Pb deposition recorded by peat cores in three minerogenic peat deposits from Switzerland. Sci. Total Environ., 292, 19-31.

Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs. J. Geochem. Explor., 3, 129-149.

Singh, A.K., Hasnain, S.I., Banerjee, D.K., 1999. Grain size and geochemical partitioning of heavy metals in sediments of the Damodar River–a tributary of the lower Ganga, India. Environ. Geol., 39 (1), 90-98.

Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behaviour and distribution of arsenic in natural waters. Appl. Geochem., 17, 517-568.

Spark, D.L., 2010. Environmental surfaces and interfaces from the nanoscale to the global scale. J. Environ. Qual., 39 (4), 1535. doi:10.2134/jeq2010.0007br.

Spijker, J., 2005. Geochemical patterns in the soils of Zeeland: natural variability versus anthropogenic impact. Utrecht University.

Su, Y.Z., Liu, W.J., Yang, R., Chang, X.X., 2009. Changes in soil aggregate, carbon, and nitrogen storages following the conversion of cropland to alfalfa forage land in the marginal oasis of northwest China. Environ. Manag., 43,1061-1070.

Tarasov, D.A., Buevich, A.G., Sergeev, A.P., Shichkin, A.V., 2018. High variation topsoil pollution forecasting in the Russian Subarctic: using artificial neural networks combined with residual kriging. Appl. Geochem., 88, 188-197.

Tarvainen, T., Albanese, S., Birke, M., Poňavič, M., Reimann, C., 2013. Arsenic in agricultural and grazing land soils of Europe. Appl. Geochem., 28, 2-10.

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem., 23 (8), 2198-2213.

Teng, Y., Ni, S., Wang, J., Zuo, R., Yang, J., 2010. A geochemical survey of trace elements in agricultural and non-agricultural topsoil in Dexing area, China. J. Geochem. Explor., 104 (3), 118-127.

Thornton, I., Webb, J.S., 1979. Geochemistry and health in the United Kingdom. Philosophical Transactions of the Royal Society, B288, 151–168.

Tipping, E., Lawlor, A., Lofts, S., Shotbolt, L., 2006. Simulating the long-term chemistry of an upland UK catchment: heavy metals. Environ. Pollut., 141, 139-150.

Tobiszewski, M., Namieśnik, J., 2012. PAH diagnostic ratios for the identification of pollution emission sources. Environ. Pollut., 162, 110-119.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, USA.

USEPA (United States Environmental Protection Agency), 2000. Risk Based Concentration Table. United States Environmental Protection Agency, Washington (DC).

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C.J.A., Buytaert, W., 2015. Web technologies for environmental Big Data. Environ. Modell. Software, 63, 185-198.

Walker, D.J., Clemente, R., Roig, A., Bernal, M.P., 2003. The effects of soil amendments on heavy metal bioavailability in two contaminated Mediterranean soils. Environ. Pollut., 122, 303-312.

Walsh, T., Barry, T., 1957. The Chemical Composition of Some Irish Peats. Proceedings of the Royal Irish Academy. Section B: Biological, Geological, and Chemical Science, 59, 305-328.

Wang, T., Kang, F., Cheng, X., Han, H., Ji, W., 2016. Soil organic carbon and total nitrogen stocks under different land uses in a hilly ecological restoration area of North China. Soil Tillage Res., 163, 176-184.

Wang, Z.M., Zhang, B., Song, K.S., Liu, D.W., Ren, C.Y., 2010. Spatial variability of soil organic carbon under maize monoculture in the Song-Nen Plain, Northeast China. Pedosphere, 20 (1), 80-89.

Webster, R., Atteia, O., Dubois, J.-P., 1994. Coregionalization of trace metals in the soil in the Swiss Jura. Eur. J. Soil Sci., 45, 205-218.

Wedepohl, K.H., 1978. Handbook of Geochemistry Springer-Verlag, Berlin-Heidelberg.

Wen, Y., Li, W., Yang, Z., Zhang, Q., Ji, J., 2020. Enrichment and source identification of Cd and other heavy metals in soils with high geochemical background in the karst region, Southwestern China. Chemosphere, 245, 125620.

Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic., 52, 394-403.

West, N.E., Stark, J.M., Johnson, D.W., Abrams, M.M., Wight, J.R., Heggem, D., Peck, S., 1994. Effects of climatic change on the edaphic features of arid and semiarid lands of western North America Arid Soil Res. Rehabil., 8, 307-351.

WHO, 2001. Arsenic and Arsenic Compounds. Environmental Health Criteria, vol. 224. Geneva: World Health Organization, 2001.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., Von Lützow, M., Marin-Spiotta, E., van

Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma, 333, 149-162.

Wong, C.S., Li, X., Thornton, I., 2006. Urban environmental geochemistry of trace metals. Environ. Pollut., 142, 1-16.

Wu, S., Zhou, S., Bao, H., Chen, D., Wang, C., Li, B., Tong, G.J., Yuan, Y.J., Xu, B., 2019. Improving risk management by using the spatial interaction relationship of heavy metals and PAHs in urban soil. J. Hazard. Mater., 364, 108-116.

Xie, X.J., Mu, X.Z., Ren, T.X., 1997. Geochemical mapping in China. J. Geochem. Explor., 60 (1), 99-113.

Xu, H.F., Demetriades, A., Reimann, C., Jiménez., J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94-104.

Xu, H.F., Zhang, C.S., 2021. Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression. Sci. Total Environ., 752, 141977.

Xu, H.F., Croot, P., Zhang, C.S., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environ. Int. 151, 106456.

Yadav, I. C., Devi, N. L., Singh, V. K., Li, J., Zhang, G., 2019. Spatial distribution, source analysis, and health risk assessment of heavy metals contamination in house dust and surface soil from four major cities of Nepal. Chemosphere, 218, 1100-1113.

Young, M.E., Donald, A.W., (eds.), 2013. A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast.

Yuan, Y., Cave, M., Zhang, C. 2018. Using local Moran's I to identify contamination hotspots of rare earth elements in urban soils of London. Appl. Geochem., 88 (2018), 167-178, 10.1016/j.apgeochem.2017.07.011.

Yuan, Y.M., Cave, M., Xu, H.F., Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using

geographically weighted regression (GWR). J. Hazard. Mater., 393, 122377. https://doi.org/10.1016/j.jhazmat.2020.122377.

Zahran, S., Mielke, H.W., Weiler, S., Berry, K.J., Gonzales, C., 2009. Children's blood lead and standardized test performance response as indicators of neurotoxicity in metropolitan New Orleans elementary schools. Neurotoxicology, 30, 888-897.

Zeng, F.R., Ali, S., Zhang, H.T., Ouyang, Y.N., Qiu, B.Y., Wu, F.B., Zhang, G.P., 2011. The influence of pH and organic matter content in paddy soil on heavy metal availability and their uptake by rice plants. Environ. Pollut., 159 (1), 84-91.

Zhang, C.S., Selinus, O., 1998. Statistics and GIS in environmental geochemistry—some problems and solutions. J. Geochem. Explor., 64 (1-3), 339-354.

Zhang, C.S., McGrath, D., 2004. Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of southeastern Ireland from two different periods. Geoderma 119, 261-275.

Zhang, C.S., 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ. Pollut., 142 (3), 501-511.

Zhang, C.S., Luo, L., Xu, W., Ledwith, V., 2008a. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci. Total Environ., 398 (1-3), 212-221.

Zhang, C.S., Fay, D. McGrath, D., Grennan, R., Carton, O.T., 2008b. Statistical analyses of geochemical variables in soils of Ireland. Geoderma, 146, 378-390.

Zhang, C.S., 2020. Towards spatial machine learning to reveal hidden patterns and relationships in national and international geochemical databases. In EGU General Assembly Conference Abstracts (p. 2179).

Zhang, H., Huang, B., Dong, L., Hu, W., Akhtar, M. S., Qu, M., 2017. Accumulation, sources and health risks of trace metals in elevated geochemical background soils used for greenhouse vegetable production in southwestern China. Ecotoxicol. Environ. Saf., 137, 233-239.

Zhang, X.W., Yang, L.S., Li, Y.H., Li, H.R., Wang, W.Y., Ye, B.X., 2012. Impacts of lead/zinc mining and smelting on the environment and human health in China. Environ. Monit. Assess., 184, 2261-2273.

Zhang, X., Liu, M., Zhao, X., Li, Y., Zhao, W., Li, A., Cheng, S. Cheng, S., Han, X., Huang, J., 2018. Topography and grazing effects on storage of soil organic carbon and nitrogen in the northern China grasslands. Ecol. Indic., 93, 45-53.

Zhou, X., Xia, B., 2010. Defining and modelling the soil geochemical background of heavy metals from the Hengshi River watershed (southern China): integrating EDA, stochastic simulation and magnetic parameters. J. Hazard. Mater., 180, 542-551.

Zuo, R., 2017. Machine Learning of Mineralization-Related Geochemical Anomalies: A Review of Potential Methods. Nat. Resour. Res. 26, 457–464.

Zuo, R., Carranza, E.J.M., Wang, J., 2016. Spatial analysis and visualization of exploration geochemical data. Earth Sci. Rev., 158, 9-18, 10.1016/j.earscirev.2016.04.006.

Zuo, R., Carranza, E.J.M., Wang, J., 2016. Spatial analysis and visualization of exploration geochemical data. Earth Sci. Rev., 158, 9-18, 10.1016/j.earscirev.2016.04.006.

Zuo, R., Xiong, Y., 2020. Geodata science and geochemical mapping. J. Geochem. Explor., 209, 106431.

Chapter 3
Materials and methodologies

3.1 Study area and scales

In order to demonstrate the exploration of GIS-based spatial analysis for different environmental geochemical data sets, three large-scale regional data sets were studied in different research areas, including GEMAS project data in European agricultural soil, and Tellus survey data in Northern Ireland and the republic of Ireland, respectively.

Based on different study areas and datasets, there are different research scales of local, regional and national level in this study, which need to be elaborate here to make a clear statement. The local scale refers to a localised research area (i.e., county level in this study), using local statistics in these advanced analytical techniques to study the hidden spatial patterns of TOC and PTEs. The regional and national scale refers to a wider research area (i.e., European continent, Northern Ireland and Republic of Ireland) which requires large amount of sampling works from geochemical surveys. It is worth noting that the choice of research scale needs to be carefully considered based on the availability of datasets, and the scales used for different spatial analysis techniques are not the same. There is no actual limitation on the big datasets in the advanced spatial analysis techniques. Generally, for low-density and large-scale sampling survey (e.g., GEMAS), larger scale is applicable. While smaller scale is more suitable for discovering interesting spatial patterns within a low-density dataset.

3.1.1 European continent

The European continent is completely located in the northern hemisphere, mostly in the eastern hemisphere, located in the western part of Eurasia and occupies one-fifth of its total landmass. The European continent is surrounded by the sea on three sides, its maritime boundaries include the Arctic Ocean to the north, the Atlantic Ocean to the west and the Mediterranean Sea, the Black Sea and the Caspian Sea to the south (Berentsen, 1998). Europe is mainly located in a temperate climate zone with a mild climate. The climate in

the west is more oceanic, while the climate in the east is even less. The geological conditions of European continent are very different, complex and diverse, and have created wide variety of landscapes across the whole continent.

3.1.2 Northern Ireland

Although the total area of Northern Ireland (NI) is only 14,120 square km^2, with 13,480 square km^2 land area and 640 square km^2 inland water area, it is a microcosm of geology of the earth (Zhang et al., 2007). The history of bedrock in NI covers almost every period from Mesoproterozoic to Paleogene, and almost all known types of rocks can be found. A simplified bedrock geology map is displayed in Fig. 3.1, with the locations of the peatland overlaid. The history of NI involves the development of ice sheets and meltwater from the last 100,000 years, which resulted in more than 80% of the bedrock being covered by various superficial deposits (e.g., alluvium, peat). According to reports, peatlands account for more than 12% of the total land area (Davies and Walker, 2013), which is a major soil subtype in NI. The north-eastern part is composed of a large area of extrusive basalt, and the north-western area is dominated by psammites (schist). The south-western terrain is a mixture of sandstone, mudstone and limestone, while south-eastern is controlled by greywacke shales, as well as significant granite intrusions were found in this area.

Northern Ireland is rich in minerals, includes iron ore, lead, coal and salt. Nowadays, there are more than 2,000 abandoned mines, most of them worked during the 18^{th} and early 20^{th} centuries. In recent years, gold, lignite and industrial minerals have dominated in commercial mining exploration activities in NI. For example, the county Tyrone is reported to hold “one of the most promising undeveloped gold deposits over the world” (Dalradian, 2019). In addition, there are two main urban areas in Northern Ireland: the Belfast Metropolitan Area and Londonderry.

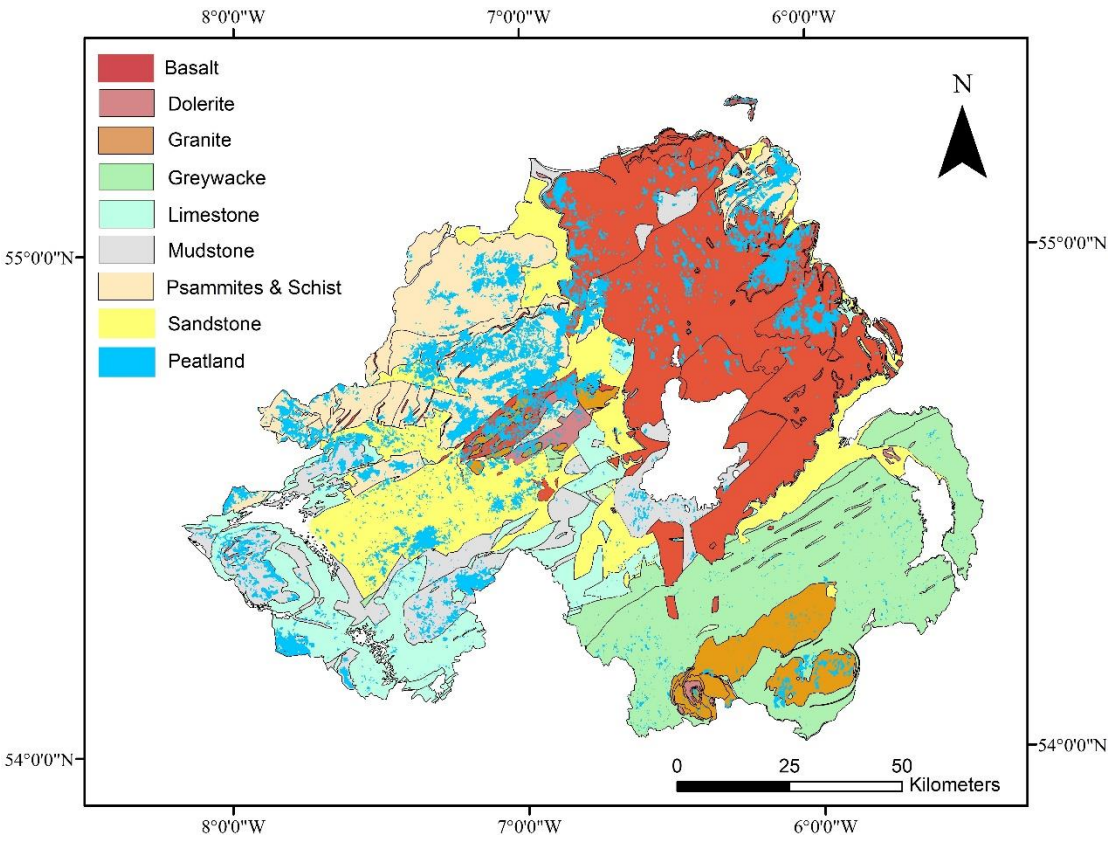


Figure 3. 1: Simplified bedrock geology maps of Northern Ireland and areas of peatland (original GIS shapefiles from GSNI, 1998).

3.1.3 Northern half of Ireland

The island of Ireland is in the north-western Europe, with a total area of 84,421 km^2, containing a diverse geology of two major domains including the continent of Laurentia in the north and Gondwana in the south. Due to the availability of only 50% on the geochemical surveys have been completed by the current Tellus dataset in Ireland, the study area is the northern half of Ireland. Based on the bedrock unit map from Geological Survey of Ireland (GSI), a simplified bedrock map of the study area is classified and shown in Fig. 3.2a (McConnell and Gately, 2006), mainly comprises basalt, clay, granite, greywacke shale, limestone, sandstone and schist. There are two main types of peat in the island (Fig. 3.2b), including blanket peat and basin peat. The blanket peat is mostly

concentrated in the mountains of the north-eastern and western coastal of Ireland, while basin peat is mainly distributed in the central part of midland areas. It is reported that mineral deposits are enriched in Ireland, especially in the western and north-eastern part (counties Mayo, Galway, Tyrone and Down) (EPA, 2009; Lusty et al., 2012). In addition, there are three major urban areas, including Galway in the west, Dublin in the east and Belfast in the north-eastern areas.

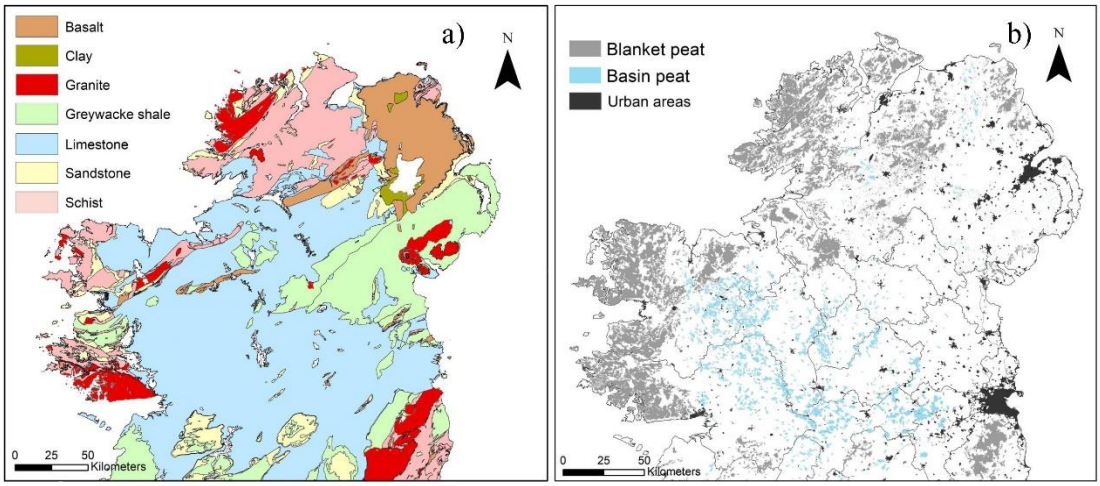


Figure 3.2: Maps showing background of study area: a) simplified bedrock map (original 1:500,000 shapefile from GSI, 2006); b) spatial distribution of locations for peatland and urban areas.

3.2 Soil sampling and analyses

3.2.1 GEMAS project data

GEochemical Mapping of Agricultural Soil (GEMAS) is a collaborative project between the Geochemistry Expert Group of EuroGeoSurveys (EGS) and Eurometaux (Reimann et al., 2014a, Reimann et al., 2014b). The GEMAS project mainly targets on agricultural and grazing land soil. During 2008 and early 2009, a total of 2,108 agricultural and 2,024 grazing land soil samples were collected, covering 33 European countries and

5.6 million km^2 (Reimann et al., 2014a). The sampling locations are presented in Fig. 3.3. Soil samples from agricultural and grazing land were taken at depths of 0–20 and 0–10 cm, respectively (ECHA, 2012). The sample density was 1 site per 2,500 square km^2. Each sample was taken as composite samples from five sub-sites, with an average weight of approximately 3.0 kg. All soil sampling materials and equipment, especially the bags used for packing samples were centrally provided to the field sampling teams (EGS, 2008). After collection, soil samples were prepared in the central laboratory of the Geological Survey of Slovakia and completed by May 2009. The soil samples were air-dried and sieved through a nylon sieve of 2 mm pore size, and subsequently homogenised and split into 10 aliquots for further study and analysis (Mackových and Lučivjanský, 2014). Specifically for the study of spatial relationship between TOC and pH, the soil pH value was determined at NGU laboratory by measurement in 0.01 M $CaCl_2$-solution extraction using pH meters (Fabian et al., 2014), and TOC content was determined at FUGRO Consult GmbH in Germany (now KIWA Control GmbH) (Reimann et al., 2011). In order to remove any inorganic carbon, 1 g sample was treated with hydrochloric acid (4 mol^{-1}) and left to stand at room temperature for 4 hours. Then, the sample was dried in an oven at 70°C for 16 hours. Then, 100–200 mg per sample was placed into the furnace and TOC determined by IR spectroscopy (Reimann et al., 2014a; Matschullat et al., 2018).

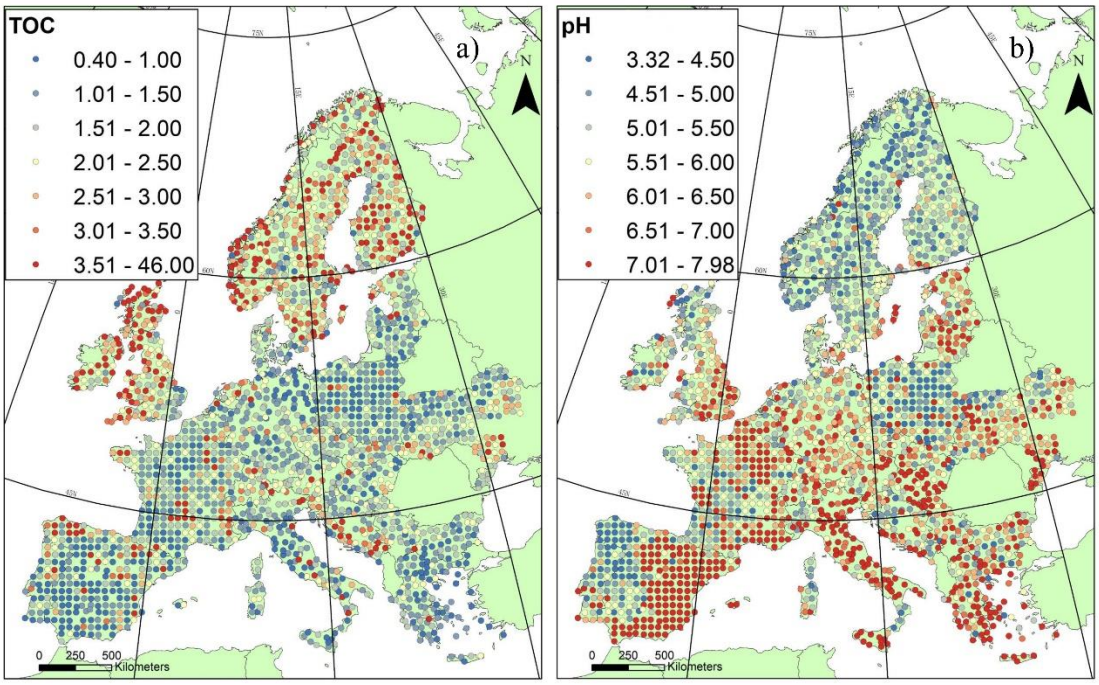


Figure 3.3: Spatial distribution maps showing sampling locations of GEMAS project data and study area in European agricultural soil: a) TOC; b) pH.

3.2.2 Tellus survey data

The Tellus project is a national-level collaborative project aimed at collecting geophysical and geochemical data across the entire island of Ireland. It is managed and undertaken by Geological Survey Ireland (GSI) and Geological Survey of Northern Ireland (GSNI) in Republic of Ireland and NI, respectively. During 2004 and 2019, a total of 17,867 regional topsoil samples (surface to 20 cm depth) were collected in the northern half part, marking more than 50% completion in geochemical survey of Ireland. Each sample was taken as composite samples from five sub-sites (approx. 750 g), with sampling density is on an average of one sample per 4 km^2 and 2 km^2 in Ireland and NI, respectively. In the republic of Ireland, the sample density was increased to one site per 2 km^2 in the urban areas of Galway and Dublin. All samples were collected in paper bags and air-dried initially before further preparation process. Then, the samples were sieved through a 2 mm pore size nylon mesh, while repetition was prepared by shallow-splitting of each duplicate sample. After

sample preparation, the geochemical composition was analysed in the laboratory by Inductively Coupled Plasma (ICP-OES/-MS) method following aqua regia digestion and X-ray fluorescence (XRF) analysis under a series of strict quality controls. Specific details on the sampling program, including protocols, and all data are publicly available from the Geologic Survey of Ireland (https://www.gsi.ie/tellus).

Specifically, the data used for the study of discovering hidden spatial patterns for 15 PTEs in the topsoil of NI was 6,862 regional topsoil samples which collected between 2004 and 2006 (Fig. 3.4). For the study of the spatially varying relationships between Pb and Al in northern half of Ireland, the total number of samples was 17,798 (69 samples with missing values were excluded) (Fig. 3.5).

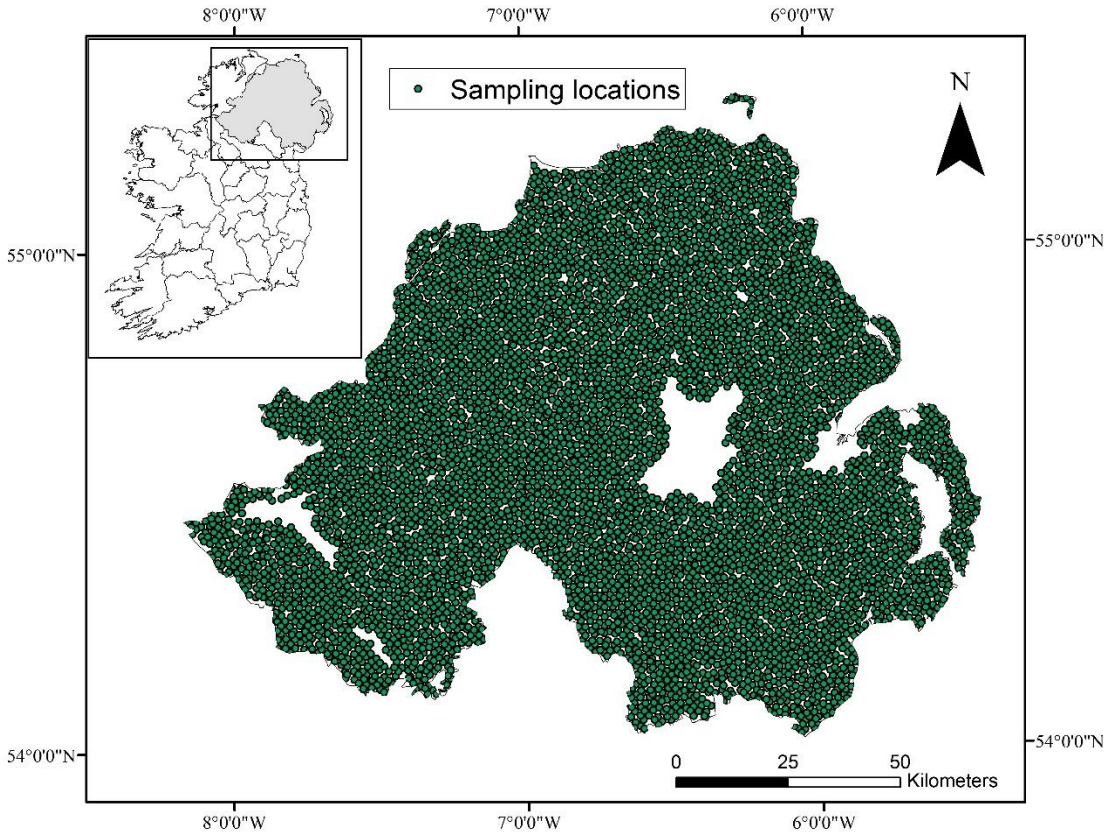


Figure 3.4: Spatial distribution map with locations of 6,862 topsoil samples in Northern Ireland.

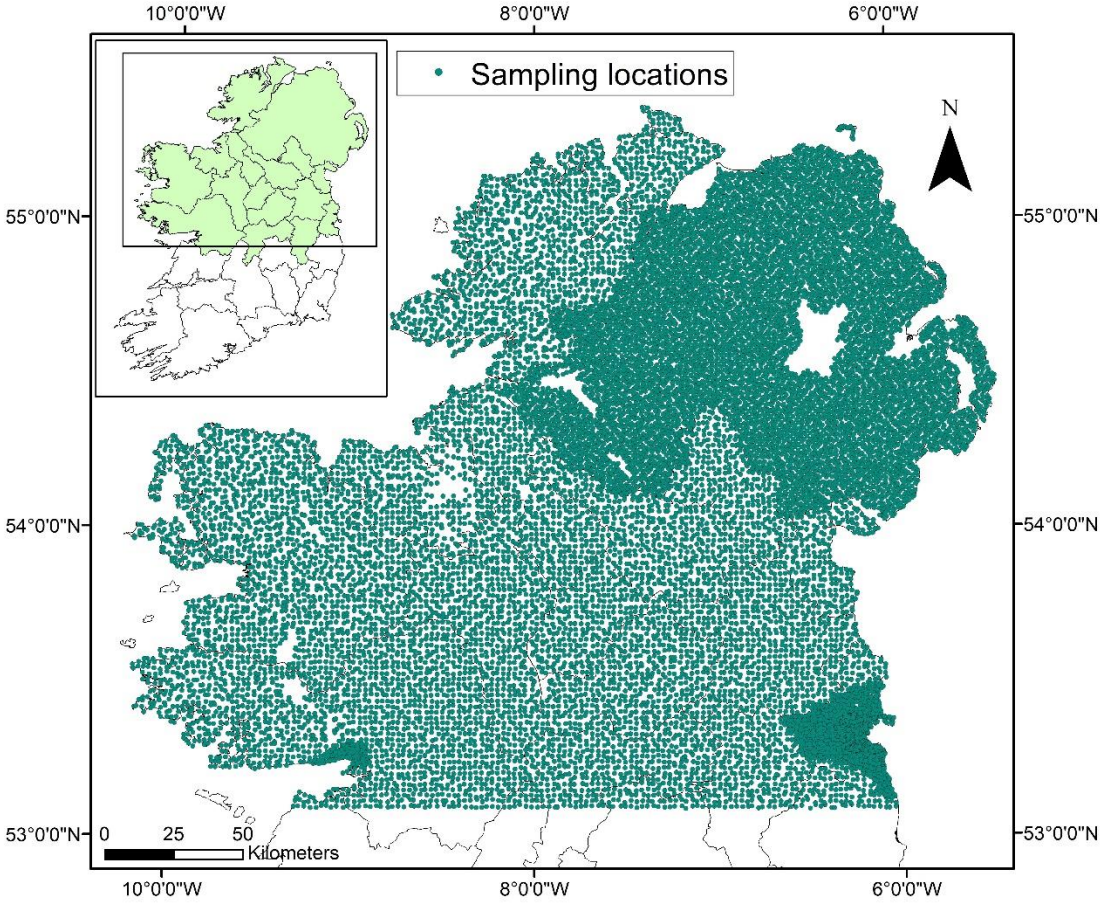


Figure 3.5: Spatial distribution map with locations of 17,798 topsoil samples in the northern part of the island of Ireland.

3.3 Data analysis

3.3.1 Descriptive statistics

3.3.1.1 Representatively descriptive parameters

The first step in processing environmental geochemical data is usually to explore the descriptive parameters and examine the probability distribution of the variables in the data set. The descriptive parameters used in this study include the total number of samples, minimum, maximum, percentile (i.e., 25%, 75%, 95%), mean and median value, standard

deviation (SD), coefficient of variation (CV) and detection limit (DL), etc. Summarising the descriptive parameters of the studied variables provides a way to understand the background knowledge of the concentration of soil elements in the study area. The measure of SD and CV can reflect the degree of dispersion of the data set, with usually larger values indicating the existence of potential outliers (extreme high values). These outliers will interfere the results of spatial analysis and statistics, and should be carefully treated prior to the further analysis.

3.3.1.2 Probability distribution

Many multivariate analysis and spatial analysis of geochemical data are based on the assumption that the data under study follows a normal or lognormal distribution. However, previous research has proposed that most elements are not normally or lognormally distributed under natural conditions (e.g., Zhang and Selinus, 1998; Reimann and Filzmoser, 2000), instead, they usually display a right skewed distribution due to the presence of outliers in the data set. Therefore, it is necessary to test the probability distribution of variables through probability plots before further analysis, such as histograms and Quantile-Quantile (Q-Q) plots. The histogram shows the frequency distribution and aggregate the data into different groups. In the case of a large amount of data, superimposing the histogram and the normal probability curve can simply and effectively examine the normality of the input variable (Lin and Mudholkar, 1980). On the other hand, the Q-Q plot displays the expected values of normal distribution against the actual values for studied variables (Wilk and Gnanadesikan, 1968). If the values are normally distributed, the points should cluster near to a straight line on the plot.

3.3.2 Data treatment

3.3.2.1 Data transformation for GEMAS data in European agricultural soil

As mentioned earlier, the spatial analysis and statistics (i.e., Getis-Ord G_i^*statistic; GWR model) is required the normality of input variables. However, the TOC contents and pH values in the GEMAS data set do not follow a normal distribution. The significance of Kolmogorov-Smirnov normality test (K-S test p value < 0.05) also suggested the non-normality of raw data. Therefore, in order to limit the impact of outliers and deal with 'non-normality' of the raw data prior to the spatial analysis (Zhang et al., 2008a), a normal score transformation was applied to the raw data set of TOC and pH. The normal score transformation is regarded as an efficient tool to transform the original distribution of a data set to a near symmetrical distribution.

3.3.2.2 Data transformation for Tellus data in Northern Ireland

The spatial clustering patterns of 15 PTEs and soil samples were investigated by the hot spot analysis and K-means clustering analysis in the topsoil of NI. Data without transformation can lead to relatively unreliable results of spatial clustering analysis. For hot spot analysis, data transformation is a standard process as it belongs to a parametric statistic. The effects on the raw data and results of different transformation methods have been discussed in previous studies (e.g., Zhang et al., 2008a, Xu et al., 2019). For K-means clustering analysis, centred log-ratio (clr) transformation and isometric log-ratio (ilr) transformation have been reported as better methods to capture spatial hidden patterns in the geochemical datasets (Templ et al., 2008). Therefore, for consistency, a clr-transformation was subjected to the raw data based on 15 variables of PTEs.

3.3.2.3 Data transformation for Tellus data in the northern half of Ireland

The spatial relationships between Pb and Al were studied by GWPCC in the topsoil of the northern half of Ireland. The GWPCC also belongs to parametric spatial statistic that depends on classic statistical parameters (i.e., mean value), which requires the normality of distribution for the data. However, the raw Tellus data set of Pb and Al do not follow a normal distribution, thus necessary data transformation process is required. In order to meet the normality requirement of GWPCC (Fotheringham et al., 2002), the normal score transformation was performed on the Pb and Al concentrations.

3.3.3 Spatial analysis

3.3.3.1 Inverse distance weighted interpolation

Inverse distance weighted (IDW) interpolation is a deterministic interpolation method, which is widely used in environmental and geochemical mapping as the simplest spatial interpolation method (Shepard, 1964; Wackernagel, 1998). The IDW interpolation can predict values in unsampled locations by using weights on the measured values of surrounding sampled points within a defined distance (Robinson and Metternicht, 2006). It assumes that each estimated point has a local influence that diminishes with distance, and, thus, gives a higher weight to points that are closer to the prediction point, and the weights gradually decrease as a function of distance. The IDW interpolation has two main parameters, including the power value and the number of neighbours (Zhang et al., 2011). However, there is no standard criteria for determining the optimal parameters, which depends on the actual objectives of study. The power was chosen as 2 and the searched neighbours were between 10 to 15, which is able to create smooth surface at the regional level.

In this study, the IDW interpolation was used to produce the continuous colourful surface maps for the spatial distribution of 15 PTEs concentration in NI, and the concentration of Pb and Al in the island of Ireland.

3.3.3.2 Hot spot analysis (Getis-Ord G_i^* statistic)

Hot spot analysis is a mapping technique that can reveal spatial clusters based on the distance between samples, and can identify locations with statistically significant high and low values in a certain geographic area based on a calculated distance. This particular analysis groups samples based on the similar high or low values which are found in a cluster. In fact, hotspot analysis requires the presence of clustering within the spatial data set. The hot spot analysis is based on Tobler's First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This first law is the foundation of the fundamental concepts of spatial dependence and spatial autocorrelation.

Getis-Ord G_i^* statistic (local G_i^* statistic) is a measure of spatial autocorrelation from a local perspective (Ord and Getis, 1995), which belongs to one of the methods of the hot spot analysis. The local G_i^* statistic returns the z-scores and *p*-values by calculating the local sum for the values of each feature and its corresponding neighbours. A high z-score and a small *p*-value for a feature indicate a significant hotspot (high-value cluster). On the same premise, a low negative z-score and a small *p*-value indicate a significant cold spot (low-value cluster). A statistically significant hotspot is a location surrounded by other samples with high values (the reverse applies for a cold spot). Also, this tool can help identify hot and cold spots with different significant levels, so priorities can be set up based on practical situations and requirements. The equations for calculation of Getis-Ord Gi* statistic are given below (Getis and Ord, 1992):

$$G_i^* = \frac{\sum_{j=1}^{n} \omega_{i,j} x_j - \bar{X} \sum_{j=1}^{n} \omega_{i,j}}{S \sqrt{\frac{\left[n \sum_{j=1}^{n} \omega_{i,j}^2 - \left(\sum_{j=1}^{n} \omega_{i,j} \right)^2 \right]}{n-1}}} \quad (3.1)$$

where *i* is the centre of the local neighbourhood; x_j is the value of the variable in the sample at location *j*; $\omega_{i,j}$ is the spatial weight between sample locations i and *j*; *n* is the total number of samples.

The following equation calculates the mean value of the whole data set:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (3.2)$$

and the standard deviation of the whole data set is calculated by the following equation:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \quad (3.3)$$

In this study, Getis-Ord Gi* statistic was used to identify the spatial clustering patterns for TOC contents and pH values in European agricultural soil, and the spatial clustering patterns for the 15 PTEs concentration in the topsoil of NI. The distance bands for GEMAS data in European continent and Tellus data in Northern Ireland were 100,000 m and 3,000 m, respectively.

3.3.3.3 Geographically weighted regression (GWR)

Since the 1990s, the GWR model is known as a powerful method to explore spatial non-stationarity and capture spatially varying relationship (Brunsdon et al., 1996; Fotheringham et al., 2002). This technique is an extension of ordinary regression model, such as Ordinary Least Square (OLS), and is used to reveal spatial relationships between the dependent and independent variable(s) from the local perspective (Fotheringham et al., 2001). The GWR can generate a set of regression coefficients at the local level that reveal how the relationship between the input variables change over space (Fotheringham et al., 2002), while the spatial patterns of such local parameters cannot be identified by the traditional regression model (e.g., OLS). The traditional regression model assumes the studied relationship is linear and spatially constant over the space, and thus the estimated parameters (i.e., regression coefficients) remain the same in the whole study area (Tu and Xia, 2008). These conventional techniques should be regarded as global statistics. On the contrary, the GWR has the potential to estimate the local regression coefficients at each sample site by allowing the parameter estimation between the dependent and independent variable(s) to vary concurrently at each location (Fotheringham et al., 2002; Kumar et al., 2012). Therefore, GWR can explore the spatially varying relationships between input variables by including the spatial coordinates of each sample site, which are often ignored in the traditional linear regression modelling. The traditional OLS equation is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \quad (3.4)$$

where y_i is the value of the dependent variable (TOC) at the i^{th} location, x_{i1} is the value of independent variable (pH) at the i^{th} location, β_0 is the intercept on the y-axis, β_1 is the regression coefficient that is estimated for the independent variable (pH) at location i, and ε_i is the error term.

Based on Fotheringham et al. (2002), the GWR model can estimate local coefficients rather than global ones by adding the geographical location in the function, which is expressed as:

$$y_i = \beta_0(\mu_i, v_i) + \beta_1(\mu_i, v_i)x_{i1} + \varepsilon_i \quad (3.5)$$

where (μ_i, v_i) represent the coordinates for sample location *i*, $\beta_0(\mu_i, v_i)$ is the intercept for location *i*, and $\beta_1(\mu_i, v_i)$ is the local regression coefficient for the independent variable (pH) at location *i*.

In contrast with the ordinary regression model, the local regression coefficients in GWR can be estimated by using a weighted function (Fotheringham et al., 2002), as expressed by the following equation:

$$\hat{\beta}(\mu_i, v_i) = (X^T W(\mu_i, v_i)X)^{-1} X^T W(\mu_i, v_i)Y \quad (3.6)$$

where X is the matrix formed by the values of the independent variable x; Y is the corresponding matrix generated by the values of the dependent variable y; $W(\mu_i, v_i)$ represents the weight matrix chosen to ensure that observations closer to the specific location (μ_i, v_i) have greater influence on the final result.

There are two important parameters when implementation of GWR model, including the kernel function and bandwidth. The adaptive kernel type was selected as the weight function because it can reduce the 'border effect' when sample sites are located near to coastal or country border areas (Zhang et al., 2011), which is suitable for the study area (i.e., European continent). There are two types of bandwidths in the GWR model, one is

the spatial distance and the other is the number of nearest neighbours. By applying the adaptive kernel function, the latter bandwidth was selected. The bandwidth was chosen by using the AIC function, which is effective in finding the 'optimal' distance band in the GWR model (Fotheringham et al., 2002). Regarding technical details, there is no consensus on the choice of the 'best' bandwidth and this has been extensively debated in the literature (e.g., Farber and Pa´ez, 2007; Guo et al., 2008). The GWR results vary by selecting different bandwidths and spatial weights. With smaller bandwidth, it can reveal more spatial variation at the local level, and the spatial patterns of regression coefficients are scattered over the study area. With larger bandwidth, the GWR approach tends to reach a global regression, and the spatial patterns of the estimated parameters become larger and smoother. Therefore, the selection of bandwidth depends on the specific aims and objectives of the research. Considering the research objectives to reveal the spatially varying relationships between TOC and pH at different scales, eight different bandwidths (with the number of neighbours being 25; 50; 75; 100; 125; 150; 200; 250) were investigated in this study.

3.3.3.4 Geographically Weighted Pearson Correlation Coefficient (GWPCC)

Geographically Weighted Pearson Correlation Coefficient (GWPCC) is an extension of traditional Pearson Correlation Coefficient (PCC) which adopts the concept of geographical weights (GW) around observations for calculating local statistics (Fotheringham et al., 2002; Kalogirou, 2012). The traditional PCC is regarded as a global statistic that assumes the measured correlation between two variables are constant and remain the same across the study area (Tu and Xia, 2008), and thus cannot capture the correlation at the local level. Based on the same principle as GWR, the GWPCC estimates the local correlation coefficients at each sample point by measuring the parameters of relationship locally (Fotheringham et al., 2002). Therefore, it can capture the spatially varying relationships between input variables by including the information of spatial locations for each sample site, which are ignored by traditional PCC. Moreover, the local coefficients of GWPCC can represent strong or weak correlation between variables, rather

than the regression (slope) coefficients in the GWR (Xu and Zhang, 2021). A series of significance tests are provided by GWPCC, which can identify the local variations at different significance levels (Kalogirou, 2014). The formula of traditional PCC is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.7}$$

where x_i is the value of Al at the i^{th} location, y_i is the value of Pb at the i^{th} location, $\bar{x}$ is mean value of Al which is calculated by $\sum_{i=1}^{n} x_i/n$, $\bar{y}$ is the mean value of Pb which calculated by $\sum_{i=1}^{n} y_i/n$, n is the total number of samples.

The GWPCC can estimate local correlation coefficients (r_i) at a location i by adding geographical weighting w_{ij} in the equation, which is expressed as (Kalogirou, 2014):

$$gwpcc_i = \frac{\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} w_{ij}(y_i - \bar{y})^2}} \tag{3.8}$$

where $\bar{x}$ is the geographically weighted mean value of Al which calculated by $\sum_{i=1}^{n} w_{ij}x_i \Big/ \sum_{i=1}^{n} w_{ij}$, $\bar{y}$ is the geographically weighted mean value of Pb which calculated by $\sum_{i=1}^{n} w_{ij}y_i \Big/ \sum_{i=1}^{n} w_{ij}$.

The weights are calculated by a bi-square function expressed as:

$$w_{ij} = \begin{cases} [1 - \left(\frac{d_{ij}}{h_i}\right)^2]^2 & \text{if } d_{ij} < h_i \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$

where d_{ij} is the distance between location i and j, h_i is the selected bandwidth (nearest neighbours) using adaptive kernel type function of location i.

As same concept with GWR, bandwidth is an important parameter in the GWPCC and also other GW models, which has been extensively debated in the literature (e.g., Farber and Páez, 2007; Guo et al., 2008; Gao and Li, 2011). The details of the discussion of bandwidth selection can be found in *section 3.3.3.3*. Also, the adaptive kernel type was chosen to reduce the 'border effect' in the study area (i.e., island of Ireland), and the technically optimal bandwidth was chosen by AIC (n = 43). Considering the research objectives of revealing large and smooth patterns of local correlation, six bandwidths with large nearest neighbours (43; 100; 150; 200; 250; 300) were investigated. In this study, the GWPCC was applied to explore the spatially varying relationships and local correlations between Pb and Al in the topsoil of the northern half of Ireland.

3.3.4 Multivariate analysis

3.3.4.1 Correlation analysis

Correlation analysis is a bivariate method used to quantitatively evaluate the strength of the relationship between two variables (Franzese and Iuliano, 2019). A statistically significant high correlation indicates that there is a strong relationship between these two variables, while a statistically significant weak correlation indicates that these two variables are poorly related (Koch and Link, 2002). The most popular correlation analysis method is Pearson's correlation coefficient analysis. However, it has the prerequisite of 'normality' for environmental geochemical data sets. Another commonly used correlation

analysis technique is Spearman's correlation coefficient analysis, which belongs to the non-parametric statistic, and does not carry any assumptions on the distribution of the data set. Therefore, in this study, Spearman's correlation coefficient was used to investigate the correlation among TOC, pH and other environmental variables in the European agricultural soil.

3.3.4.2 Principal component analysis (PCA)

Principal component analysis (PCA) is one of the most popular methods in multivariate statistical analysis, which has become a standard approach and widely used to extract useful geochemical information. It combines multiple correlated variables into fewer principal components based on correlation or covariance matrix. These components are not correlated with each other, which can represent the interrelationships between the multi-variables in the original data set (Jolliffe, 2002). The advantage of adopting these extract components is that the input data sets can be replaced by fewer comprehensive indicators with as little loss of information as possible (Jolliffe, 2002), and this step is called dimension reduction. The appropriate number of components can be determined by a significant inflection point on the output scree plot (Cattell, 1966). In addition, PCA can enhance the interpretability of results among multiple variables by selecting appropriate rotation methods (Cheng et al., 2006), including Varimax, Promax, Oblimin and Quartimin (Carroll, 1953; Kaiser, 1958; Hendrickson and White, 1964; Harman, 1976). In this study, PCA was performed to reduce dimension for the 15 PTEs in the topsoil samples of NI prior to the K-means clustering analysis.

3.3.4.3 K-means clustering analysis

K-means clustering analysis is a partitioning clustering algorithm, which is adopted as the most widely used clustering method in ML and data mining due to its simplicity and efficiency (Han and Kamber, 2006). It is usually performed as the initial step of data

analysis, which has been proved to be powerful for capturing the hidden spatial patterns in environmental geochemistry (e.g., Bengio et al., 2013; LeCun et al., 2015; Zuo et al., 2017). The principle of K-means clustering is to partition the space into *k* non-overlapping clusters, and classify each observation to the nearest centre in order to minimise the within-cluster variance as well as maximise the between-cluster variance (Hartigan, 1975; Alizadeh et al., 2017). In other words, it aims to divide the samples with higher similarity into the same cluster, while the samples between each cluster are very dissimilar. The function of K-mean clustering is presented as follow (MacQueen, 1967; Hartigan and Wong, 1979):

$$J = \sum_{i=1}^{k} \sum_{j \in C_i}^{n_i} \left\| x_j - \mu_i \right\|^2 \tag{3.10}$$

Where J is the objective function, C_i is the i^{th} cluster, n_i is the number of samples in i^{th} cluster, distance function $d_{ji} = ||x_j - \mu_i||^2$ represents the calculation of the distance between each sample point x_j and centroid μ_i in the i^{th} cluster. The centroid μ_i can be calculated based on the function as below:

$$\mu_i = \frac{1}{|C_i|} \sum_{j \in C_i} x_j \tag{3.11}$$

The implementation of K-means clustering algorithm can be summarised in the following steps (Zagouras et al., 2013):

(1) Randomly initializing the cluster centroid μ_1, μ_2, …, μ_k;

(2) Calculating the distance function d_{ji} between each sample point x_j and centroid μ_i in the i^{th} cluster. The distance function d_{ji} was based on the Euclidean distance in this study.

(3) Moving each sample point x_j to the cluster of its nearest centroid $\mu_{nearest}$, and update cluster centroids from which sample points have been disjointed or reassigned.
(4) Computing the objective function J, as given above in formula (1). If function J converges, the centroids do not change from the previous iterations, and the K-means clustering algorithm derives the final centroids of cluster. Otherwise, the step 2 and 3 are repeated until the objective function J converges.

The number of clusters is an important parameter when using the partition clustering (Weatherill and Burton, 2008). The choice of optimal cluster numbers can be achieved by various methods and the prior knowledge, including Davies-Bouldin Index (Davies and Bouldin, 1979), Silhouette method (Rousseeuw, 1987), elbow method (Ketchen and Shook, 1996), information criterion approach (Goutte et al., 2001). In this study, Silhouette method was applied to choose the appropriate cluster number. It can provide succinct graphics to display the quality of classification, as well as silhouette values to interpret and validate the consistency of clusters within samples (Rousseeuw, 1987). The silhouette values can represent how similar an observation belongs to its cluster compared to others, where a high value implies the good cohesion of one object to its own cluster and poor match with adjacent clusters. This principle corresponds well to the classification criteria of cluster analysis.

In this study, K-means clustering analysis was performed to reveal the hidden spatial patterns of the topsoil samples based on the 15 PTEs in NI.

3.3.5 Computer software

All the data sets are stored in Microsoft Excel (ver. 2016), and data statistics were computed in SPSS (ver. 24). The normal score transformation was conducted in SPSS (ver. 24), and clr-transformation was conducted using R project (ver. 3.56). All the spatial

distribution maps were produced using IDW interpolation in ArcGIS (ver. 10.4). The Getis-Ord G_i^* statistic and GWR were also performed in ArcGIS (ver. 10.4). Principal component analysis was conducted in SPSS (ver. 24), while K-means clustering was compiled using '*cluster*' package (ver. 2.10) in R project (Maechler et al., 2019; https://cran.r-project.org/web/packages/cluster/cluster.pdf). The local correlation coefficients and their significance level were calculated using the GWPCC in the R package '*lctools*' (ver. 3.56, in http://cran.r-project.org/web/packages/lctools/index.html).

3.4 Summary

This chapter describes the background knowledge of the study area in European continent, Northern Ireland and the northern half of Ireland. In addition, the sampling locations, preparation and laboratory analysis for the GEMAS project and Tellus survey data were also discussed. Furthermore, the detailed information about methodologies and data analysis were provided in this chapter.

Reference

Alizadeh, M. J., Shahheydari, H., Kavianpour, M. R., Shamloo, H., Barati, R., 2017. Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network. Environ. Earth Sci., 76 (2), 86.

Bengio, Y., 2013. Deep learning of representations: looking forward. In: International Conference on Statistical Language and Speech Processing. Springer, Berlin, Heidelberg, 1–37.

Berentsen, W.H., 1998. Europe continent. Available at https://www.britannica.com/place/Europe. [Accessed: 21/03/2021].

Brunsdon, C., Fotheringham, A.S., Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geogr. Anal. 28, 281-298.

Carroll, J.B., 1953. An analytic solution for approximating simple structure in factor analysis. Psychometrika 18, 23-38.

Cattell, R.B., 1966. The scree test for the number of factors. Multivar. Behav. Res., 1 (2), 245-276.

Cheng, Q., Jing, L., Panahi, A., 2006. Principal component analysis with optimum order sample correlation coefficient for image enhancement. Int. J. Remote Sens, 27 (16), pp. 3387-3401.

Dalradian, 2019. Making the most of County Tyrone's gold deposits. Available at: https://www.newsletter.co.uk/business/making-the-most-of-county-tyrone-s-gold-deposits-1-9081043.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell., 1 (4), 224-227.

Davies, H., Walker, S., 2013. Strategic planning policy statement (SPPS) for Northern Ireland: Strategic Environmental Assessment (SEA) Scoping report. Leeds.

ECHA, 2012. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.16: Environmental Exposure Assessment. Version 2.1. European Chemicals Agency (147 pp).

EGS, 2008. EuroGeoSurveys geochemistry working group. EuroGeoSurveys geochemical mapping of agricultural and grazing land in Europe (GEMAS) - field manual. Norges Geologiske Undersøkelse Report, 2008.038, 46 pp.

Environmental protection agency (EPA), 2009. Historic Mine Sites - Inventory and Risk Classification Volume 1. ISBN: 1-84095-318-3.

Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E. 2014. GEMAS: spatial distribution of the pH of European agricultural and grazing land soil. Appl. Geochem., 48, 207-216.

Farber, S., Páez, A., 2007. A systematic investigation of crossvalidation in GWR model estimation: empirical analysis and Monte Carlo simulations. J. Geogr. Syst. 9, 371-396. http://doi:10.1007/s10109-007-0051-3.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons Ltd., Chichester, UK, 284 pp.

Franzese, M., Iuliano, A., 2019. Correlation analysis. Encyclop. Bioinformatics Comput. Biol., 1, 706-721.

Gao, J.B., Li, S.C., 2011. Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using Geographically Weighted Regression. Appl. Geogr., 31, 292-302.

Getis, A., Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. Geogr. Anal. 24 (3), 189-206.

Goutte, C., Hansen, L.K., Liptrot, M.G., Rostrup, E., 2001. Feature-space clustering for fMRI meta-analysis. Hum. Brain Mapp., 13 (3), 165-183.

Guo, L., Ma, Z., Zhang, L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can. J. For. Res. 38, 2526-2534.

Han, J., Kamber, M., 2006. Data Mining, Concepts and Techniques. Morgan Kaufman Publishers, San Francisco, USA.

Harman, H.H., 1976. Modern Factor Analysis, 3rd Edition. University of Chicago Press, Chicago.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100–108.

Hendrickson, A.E., White, P.O., 1964. PROMAX: a quick method for rotation to oblique simple structure. Brit. J. Stat. Psychology, 17, 65-70.

Jolliffe, I.T., 2002. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed. Springer, New York, 487 pp.

Kaiser, H.F., 1958. The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187-200.

Kalogirou, S., 2012. Testing local versions of correlation coefficients, Review of Regional Research-Jahrbuch für Regionalwissenschaft, 32 (1), 45-61. doi: 10.1007/s10037-011-0061-y.

Kalogirou, S., 2014. A spatially varying relationship between the proportion of foreign citizens and income at local authorities in Greece. Proceedings of the 10th International Congress of the Hellenic Geographical Society, 5, 1458-1466.

Ketchen Jr., D.J., Shook., C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manag. J., 17 (6), 441-458.

Knights, K.V., Glennon, M.M., 2013. Tellus Border Project Geochemistry Data User Guide Version 1. Geological Survey of Ireland and Geological Survey of Northern Ireland Joint Report.

Koch, G.S., Link, R.F. Statistical analysis of geological data. New York: Dover Publications. 2002.

Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189, 627-634.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444.

Lin, C. C., Mudholkar, G. S., 1980. A simple test for normality against asymmetric alternatives. Biometrika 67(2), 455-461.

Lusty, P.A.J., Scheib, C., Gunn, A.G., Walker, A.S.D., 2012. Reconnaissance-scale prospectivity analysis for gold mineralisation in the Southern uplands-down-longford terrane, Northern Ireland. Nat. Resour. Res., 21 (3), 359–382.

Mackových, D., Lučivjanský, P., 2014. Preparation of GEMAS project samples and standards. Chapter 4 in C. Reimann, M. Birke, A. Demetriades, P. Filzmoser, P. O'Connor (Eds.), Chemistry of Europe's Agricultural Soils. Part a: Methodology and Interpretation of the GEMAS Data Set, Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover (2014), pp. 37-40.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, 1, pp.281--297. Berkeley, CA: University of California Press.

Matschullat, J., Reimann, C., Birke, M., Dos Santos Carvalho, D., GEMAS Project Team. GEMAS: CNS concentrations and C/N ratios in European agricultural soil. Sci. Total Environ., 627, 975-984.

McConnell, B., Gatley, S., 2006. Bedrock geological map of Ireland: 1:500,000 scale. Dublin: Geological Survey of Ireland. ISBN: 189970244X.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27 (4), 286-306.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O' Connor, P., 2014a. Chemistry of Europe's Agricultural Soils, Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, 523 pp.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014b. Chemistry of Europe's Agricultural Soils, Part B: General Background Information and Further Analysis of the GEMAS Data Set, Geologisches Jahrbuch Reihe B, Band B 103. Schweizerbart Science Publishers, Stuttgart, p. 352.

Reimann, C., Demetriades, A., Eggen, O.A., Peter, P., 2011. Filzmoser, the EuroGeoSurveys Geochemistry Expert Group. The EuroGeoSurveys GEochemical Mapping of Agricultural and Grazing Land Soils Project (GEMAS) – Evaluation of Quality Control Results of Total C and S, Total Organic Carbon (TOC), Cation Exchange Capacity (CEC), XRF, pH, and Particle Size Distribution (PSD) Analysis. Geological Survey of Norway, Trondheim, NGU Report 2011.043. 90 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2011/2011_043.pdf.

Robinson, T.P., Metternicht, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Comput. Electron. Agric., 50 (2), 97-108.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53-65.

Shepard, D., 1964. A Two Dimensional Interpolation Function for Irregularly Data Spaced. ACM Nat. Conf., 517-524.

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. Appl. Geochem., 23 (8), 2198-2213.

Tobler W., 1970. A computer movie simulating urban growth in the Detroit region. Economic Geography, 46 (Supplement), 234-240.

Tu, J., Xia, Z.G., 2008. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. Sci. Total Environ. 407 (1), 358-378.

Wackernagel, H., 1998. Multivariate Geostatistics: an Introduction with Applications. Springer-Verlag, Berlin.

Weatherill, G., Burton, P.W., 2008. Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. Geophys. J. Int., 176 (2), 565-588.

Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika, 55 (1), 1-17.

Xu, H.F., Demetriades, A., Reimann, C., Jiménez., J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94-104.

Xu, H.F., Zhang, C.S., 2021. Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression. Sci. Total Environ., 752, 141977.

Young, M.E., Donald, A.W., (eds.), 2013. A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast.

Zagouras, A., Kazantzidis, A., Nikitidou, E., Argiriou, A.A., 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. Sol. Energy, 97 (5), 1-11.

Zhang, C.S., Jordan, C., Higgins, A., 2007. Using neighbourhood statistics and GIS to quantify and visualize spatial variation in geochemical variables: An example using Ni concentrations in the topsoils of Northern Ireland. Geoderma, 137, 466-476.

Zhang, C.S., Luo, L., Xu, W., Ledwith, V., 2008a. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci. Total Environ., 398 (1-3), 212-221.

Zhang, C., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem., 26, 1239-1248.

Zuo, R.G., 2017. Machine Learning of Mineralization-Related Geochemical Anomalies: A Review of Potential Methods. Nat. Resour. Res. 26, 457-464.

Chapter 4
Research paper

4.1 Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data

Xu, H.F., Demetriades, A., Reimann, C., Jiménez., J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94-104.

Summary: This paper investigated the spatial patterns of TOC contents and its relationship with pH values using hot spot analysis (Getis-Ord G_i^* statistic) based on 2,108 topsoil samples that collected from GEMAS project in European agricultural soil. The overall patterns revealed by the hot spot maps showing a general negative relationship between these two variables at the European continent scale. High TOC contents accompanying low pH values in the north-eastern Europe, while low TOC with high pH values in the southern part. Moreover, a 'special' feature of co-existence of comparatively low TOC contents and low pH values in north-central Europe was also identified by hot spot analysis, and this hidden pattern showed clear association with high concentration of SiO_2 (quartz) in the coarse-textured glacial sediments in north-central Europe. The results demonstrated that hot spot analysis is effective in highlighting the spatial patterns of TOC in European agricultural soil and helpful to identify hidden relationships between environmental variables.

My contribution in this paper accounted for ~80% in reviewing literatures, exploring data and writing manuscript.

Science of the Total Environment 678 (2019) 94–104

Contents lists available at ScienceDirect

Science of the Total Environment

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

Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data

Haofan Xu [a], Alecos Demetriades [b,1], Clemens Reimann [c], Juan J. Jiménez [d], Juliane Filser [e], Chaosheng Zhang [a,*], GEMAS Project Team [2]

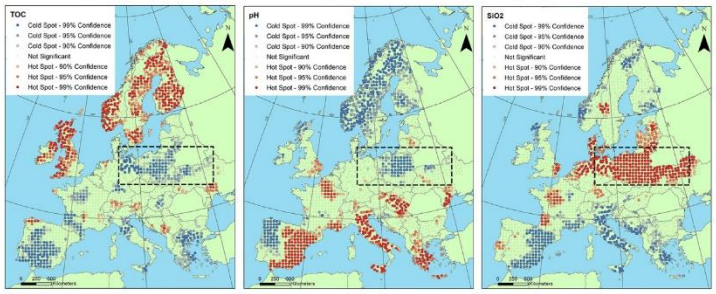
[a] International Network for Environment and Health (INEH), School of Geography and Archaeology & Ryan Institute, National University of Ireland, Galway, Ireland
[b] Institute of Geology and Mineral Exploration, Athens, Hellas
[c] Geological Survey of Norway, P.O. Box 6315, Torgarden, N-7491 Trondheim, Norway
[d] ARAID Researcher, Pyrenean Institute of Ecology-National Spanish Research Council, IPE-CSIC, Av. Nuestra Señora de la Victoria 16, 22700 Jaca, (Huesca), Spain
[e] University of Bremen, UFT, Department of General and Theoretical Ecology, Leobener Str. 6, D - 28359 Bremen, Germany

HIGHLIGHTS

- Hot spot analysis reveals spatial patterns of TOC in European agricultural soil.
- Soil TOC and pH values are negatively correlated at European scale.
- Low TOC and low pH values are related to coarse-textured glacial sediments.
- Co-existence of low TOC and low pH values occur in north-central Europe.

GRAPHICAL ABSTRACT

A special co-existence of both low soil total organic carbon contents and low pH values was identified in agricultural soil in north-central Europe, which is highly correlated with the high concentration of SiO_2 in those soil.



ARTICLE INFO

Article history:
Received 22 March 2019
Received in revised form 24 April 2019
Accepted 26 April 2019
Available online 26 April 2019

ABSTRACT

Total organic carbon (TOC) contents in agricultural soil are presently receiving increased attention, not only because of their relationship to soil fertility, but also due to the sequestration of organic carbon in soil to reduce carbon dioxide emissions. In this research, the spatial patterns of TOC and its relationship with pH at the European scale were studied using hot spot analysis based on the agricultural soil results of the Geochemical Mapping of Agricultural Soil (GEMAS) project. The hot and cold spot maps revealed the overall spatial patterns showing a

* Corresponding author.
E-mail addresses: H.XU2@nuigalway.ie (H. Xu), alecos.demetriades@gmail.com (A. Demetriades), clemens.reimann@ngu.no (C. Reimann), jjimenez@ipe.csic.es (J.J. Jiménez), filser@uni-bremen.de (J. Filser), Chaosheng.Zhang@nuigalway.ie (C. Zhang).
[1] (retired).
[2] GEMAS Project Team: S. Albanese, M. Andersson, R. Baritz, M.J. Batista, A. Bel-lan, M. Birke, D. Cicchella, B. De Vivo, W. De Vos, E. Dinelli, M. Ďuriš, A. Dusza-Dobek, M. Eklund, V. Ernstsen, P. Filzmoser, B. Flem, D.M.A. Flight, S. Forrester, M. Fuchs, U. Fügedi, A. Gilucis, M. Gosar, V. Gregorauskiene, W. De Groot, A. Gulan, J. Halamić, E. Haslinger, P. Hayoz, R. Hoffmann, J. Hoogewerff, H. Hrvatovic, S. Husnjak, L. Janik, G. Jordan, J. Kirby, V. Klos, F. Krone, P. Kwecko, L. Kuti, A. Ladenberger, A. Lima, J. Locutura, P. Lucivjansky, A. Mann, D. Mackovych, M. McLaughlin, B.I. Malyuk, R. Maquil, J. Matschullat, R.G. Meuli, G. Mol, P. Négrel, P. O'Connor, K. Oorts, A. Pasieczna, V. Petersell, S. Pfleiderer, M. Poňavič, C. Prazeres, U. Rauch, S. Radusinović, M. Sadeghi, I. Salpeteur, R. Scanlon, A. Schedl, A. Scheib, I. Schoeters, E. Sellersjö, I. Slaninka, J.M. Soriano-Disla, A. Šorša, R. Svrkota, T. Stafilov, T. Tarvainen, V. Trendavilov, P. Valera, V. Verougstraete, D. Vidojević, A. Zissimos, Z. Zomeni.

https://doi.org/10.1016/j.scitotenv.2019.04.382

Editor: Jay Gan

Keywords:
TOC
pH
Hot spot analysis
GEMAS
European agricultural soil

negative correlation between TOC contents and pH values in European agricultural soil. High TOC contents accompanying low pH values in the north-eastern part of Europe (e.g., Fennoscandia), and low TOC with high pH values in the southern part (e.g., Spain, Italy, Balkan countries). A special feature of co-existence of comparatively low TOC contents and low pH values in north-central Europe was also identified on hot and cold spot analysis maps. It has been found that these patterns are strongly related to the high concentration of SiO_2 (quartz) in the coarse-textured glacial sediments in north-central Europe. The hot spot analysis was effective, therefore, in highlighting the spatial patterns of TOC in European agricultural soil and helpful to identify hidden patterns.

1. Introduction

Global climate change currently is an important research topic (Mathieu et al., 2015; Fang et al., 2017), as the concentration of carbon dioxide in the atmosphere has increased continuously during the past decades (Yu et al., 2014; O'Rourke et al., 2015). Soil is regarded the largest pool of organic carbon (OC) in the terrestrial ecosystem, with total amounts of carbon two or three times higher than that in the atmosphere or terrestrial vegetation (Eswaran et al., 1993; Batjes, 1996; Jobággy and Jackson, 2000; Schmidt et al., 2011). Therefore, even small changes in soil organic carbon storage can influence the atmospheric CO_2 concentration (Johnston et al., 2004; Xu et al., 2011a). The preservation and release of this large OC pool has been considered as a vital factor in controlling atmospheric CO_2 concentrations (Pan et al., 2003). This has increased interest in soil organic carbon sequestration as a helpful way to offset carbon dioxide emissions (Lenka and Lal, 2013). Nowadays, due to the current low OC content in agricultural soil at the global level, they are likely to store about 5500–6000 Mt CO_2-eq·yr^{-1} by 2030 (Smith et al., 2008). Moreover, a high level of total organic carbon (TOC) in agricultural soil significantly raises the nutrient levels and improves soil structure conditions (Tiessen et al., 1994; Hati et al., 2007). Hence, studies on TOC in agricultural soil can contribute to the improvement of agricultural productivity as well as mitigating global warming.

Due to the importance of soil organic carbon pools in terrestrial ecosystems, much effort has focused on the study of TOC contents at national and regional scales. For instance, the organic carbon stock has been investigated at the national level in some European countries, including Belgium (Meersmans et al., 2011), Ireland (Zhang and McGrath, 2004; Xu et al., 2011b), the United Kingdom (Bradley et al., 2005), France (Martin et al., 2011) and Spain (Rodríguez Martín et al., 2016). In addition, the dynamics and influencing factors of TOC contents in soil have also been widely studied (McGrath and Zhang, 2003; Reisser et al., 2016; Zhang et al., 2018). The influencing factors of TOC contents include both natural and anthropogenic ones, such as land use, elevation, climate, parent materials, soil properties (e.g., pH, soil texture), cultivation method, human input (e.g., fertilisers) and site management (Jenny, 1980; Guo and Gifford, 2002; Jackson et al., 2002; Lal, 2005; Jandl et al., 2007). Deploying some of these factors can effectively increase the OC sequestration in soil, but this requires evaluating reliably changes based on statistically sound analysis.

Hot spot analysis can investigate where the spatial features under study are concentrated (Alessa et al., 2008). It is based on the methodologies of Local Moran's I and Getis-Ord Gi* (or referred to as Gi* statistic) (Braithwaite and Li, 2007). The Gi* statistic takes the values of all neighbouring features into consideration and reports the hot and cold spots at different statistically significance levels. Hot spot analysis has been widely used in crime rates analysis, traffic accidents, epidemiology, economic geography, species populations and demographics (e.g., Barro et al., 2015; ESRI, 2016; Lu et al., 2017; Ansong et al., 2018). In recent years, it has been often applied in studies of environmental science (Zhang et al., 2012; Tran et al., 2017; Kumar et al., 2018), spatial clusters of diseases (Wang et al., 2012; Wang et al., 2016) and biodiversity (Di Minin et al., 2013).

The relationship between soil TOC contents and pH values has been reported in numerous studies (McGrath and Zhang, 2003; Reisser et al., 2016; Zhang et al., 2018). This study attempts to employ hot spot analysis to directly reveal spatial patterns by identifying hot and cold spots of TOC and pH at the European scale based on the GEMAS data set. This approach should help to identify 'hidden' relationships and patterns. TOC contents and pH values in agricultural soil samples are presented in the GEMAS atlas (Reimann et al., 2014a). TOC is also discussed in Baritz et al. (2014) and Matschullat et al. (2018), and pH values are presented in Fabian et al. (2014).

The objectives of this study were: (1) to study the spatial distribution patterns of TOC and pH in European agricultural soil based on GEMAS data using hot spot analysis; (2) to identify the spatial relationship between soil TOC contents and pH values at the European scale using mapping techniques based on hot spot analysis, and (3) to explore influencing factors of the special pattern of co-existence of low TOC contents and low pH values in north-central Europe.

2. Data and methods

2.1. The GEMAS project

GEMAS (GEochemical Mapping of Agricultural Soil) was a collaborative project between the Geochemistry Expert Group of EuroGeoSurveys (EGS) and Eurometaux (Reimann et al., 2014b, 2014c). The GEMAS project aimed at generating consistent soil geochemistry data at the continental-scale based on REACH regulation requirements (EC, 2006). The GEMAS project mainly focused on agricultural and grazing land soil, with a total of 2108 agricultural (Ap) and 2023 grazing land (Gr) soil samples collected during 2008 and early 2009, covering 33 European countries and 5.6 million km^2 (Reimann et al., 2014b, p.24). Soil samples from agricultural and grazing land were taken at depths of 0–20 and 0–10 cm, respectively, according to the REACH regulation specifications (ECHA, 2012). The entire project area was covered by a 50 × 50 km grid and within each cell of 2500 km^2 one sampling site of each sample type (Ap and Gr) was chosen. It was in general decided to follow the sample density used in the Baltic Soil Survey project (Reimann et al., 2003). This was a practical design as it allowed the collection of Ap and Gr samples evenly all over Europe with a very different spatial distribution of agricultural and grazing land. Thus, the field teams were free to choose where they took the soil samples from Ap (agricultural) and Gr (grazing) land in each grid cell (2500 km^2 area) (Reimann et al., 2014b, p.33). All soil sampling materials and equipment, especially the bags used for packing samples were centrally provided to the field sampling teams (EGS, 2008).

2.2. Soil sampling

The agricultural land (arable land, Ap) soil samples were used in this study. The samples of an average weight of approximately 3.0 kg each, were taken as composite samples from five sub-sites from the corners and centre of a 10 × 10 m square. Each sampling site was carefully recorded and replicated on-site at every 20 sites (EGS, 2008).

There were 13 samples with missing TOC values in the GEMAS database, thus the corresponding 13 pH sample values were excluded. The total number of Ap soil samples used in this study was, thus, 2095 (Fig. 1).

2.3. Soil sample preparation and analysis

All soil samples in the GEMAS project were prepared in the central laboratory of the Geological Survey of Slovakia and completed by May 2009. The soil samples were air-dried and sieved through a nylon sieve of 2 mm pore size, and subsequently homogenised and split into 10 aliquots for further study and analysis (Mackových and Lučivjanský, 2014). The Geological Survey of Norway (NGU) prepared a list of random numbers for each sample set, allowing insertion of analytical replicates and project standards in batches of twenty samples, so that quality control samples could not be recognised by the analytical laboratories.

Soil pH was determined at NGU laboratory by measurement in 0.01 M $CaCl_2$-solution. Total organic carbon (TOC) was determined at FUGRO Consult GmbH in Germany (now KIWA Control GmbH) according to the ISO standard 10694 (ISO, 1995; Reimann et al., 2011).

In order to generate harmonised and comparable data sets across national borders, all elements and parameters were analysed in the same laboratory and under strict quality control procedures. Reimann et al. (2009, 2011), Birke et al. (2014) and Demetriades et al. (2014) have already clearly described the analytical methods and quality control procedures.

2.4. Hot spot analysis

Hot spot analysis can identify locations with statistically significant high and low values over a geographical area by aggregating sample values that are in proximity to one another based on a calculated distance. This particular analysis groups samples when similar high (hot) or low (cold) values are found in a cluster. In fact, hotspot analysis requires the presence of clustering within the spatial data set. Hot spot analysis is based on Tobler's First Law of Geography, which states that "*everything is related to everything else, but near things are more related than distant things*" (Tobler, 1970). This first law is the foundation of the fundamental concepts of spatial dependence and spatial autocorrelation.

In this study, the Getis-Ord Gi* statistic (Ord and Getis, 1995), which is a measure of spatial autocorrelation from a local perspective, was used to identify the high and low spatial clusters in the GEMAS Ap data set for soil total organic carbon (TOC) and pH using ESRI's ArcMap software version 10.4. This technique calculates the local sum for a feature and its neighbours, and is compared proportionally to the sum of all features, i.e., in this case the Ap sample values of TOC and pH and the corresponding neighbouring sample values. When the local sum is very different from the expected local sum, and this difference is too large to be the product of random choice, a statistically significant z-score results. A high z-score and a small p-value for a feature indicate a significant hotspot (high value cluster). On the same premise, a low negative z-score and a small p-value indicate a significant cold spot (low value cluster). A statistically significant hotspot is a location surrounded by other samples with high values (the reverse applies for a cold spot). Also, this tool can help identify hot and cold spots with different significant levels, so priorities can be set up based on practical situations and requirements. The equations for the calculation of Getis-Ord Gi* statistic are given below (Getis and Ord, 1992; ESRI, 2016):

$$G_i^* = \frac{\sum_{j=1}^{n} \omega_{i,j} x_j - \bar{X} \sum_{j=1}^{n} \omega_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} \omega_{i,j}^2 - \left(\sum_{j=1}^{n} \omega_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where i is the centre of the local neighbourhood; x_j is the value of the variable in the sample at location j; $\omega_{i,j}$ is the spatial weight between sample locations i and j; n is the total number of samples.

The following equation calculates the mean of the whole data set:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2)$$

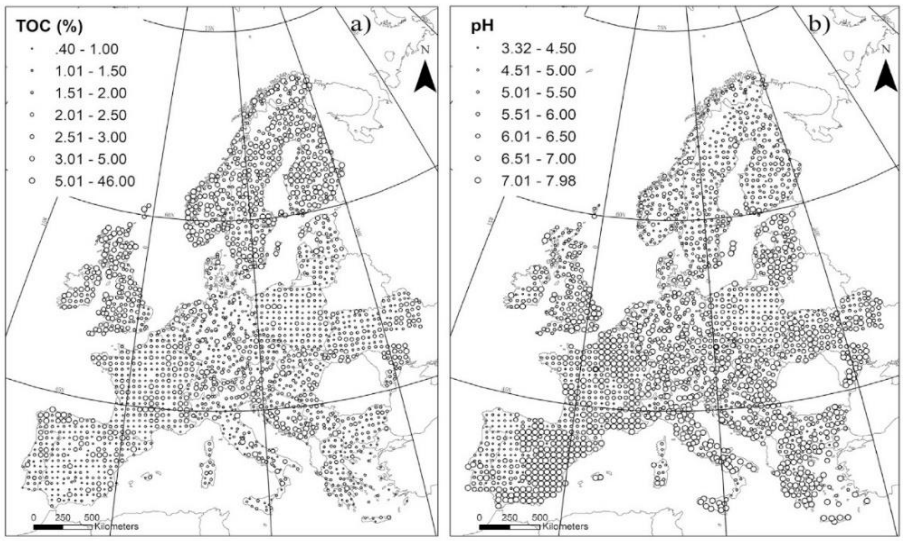


Fig. 1. Growing dot maps showing agricultural soil sampling locations and TOC contents and pH values in European agricultural soil: a) TOC contents; b) pH values ($n = 2095$).

and the standard deviation of the whole data set is calculated by the following equation:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\overline{X})^2} \quad (3)$$

The final Gi* statistic returned for each location is a z-score. For positive z-scores, with statistically significant positive values, the higher the z-score the more intense is the clustering of high-values (hot spots). Similarly, for statistically significant negative z-scores, the lower the z-score is, the more intense is the cluster of low-values (cold spots). The z-score indicates the statistical significance of the cluster at a specified level. For 99%, 95% and 90% significance levels, the z-score should be between ±2.58, ±1.98 and ±1.58, respectively. In this paper, the TOC values were divided into seven classes based on the statistical results.

2.5. Data transformation and computer software

The GEMAS TOC and pH data were treated using different computer software programs. In order to limit the impact of outliers and non-normality of the raw data on spatial analysis, a normal score transformation was applied to the raw data set. The normal score transformation is regarded as an efficient tool to transform the original distribution of a data set to a relatively standard normal distribution. This tool ranks the data from the lowest to the highest value and matches these ranks to equivalent ranks produced in the normal distribution. Zhang et al. (2008) have discussed the effects of normal score transformation in spatial analysis in detail. The data transformation and descriptive statistics were calculated using SPSS (version 21.0) and Microsoft Excel. Hot spot maps were produced with ESRI's GIS software ArcMap (version 10.4).

3. Results and discussion

3.1. Descriptive statistics for TOC contents and pH values

It is well-known that geochemical data do not belong to the classical Euclidean space and should be considered in their own Euclidean geometry on the simplex (Aitchison, 1986; Filzmoser et al., 2009, 2010, 2014; Egozcue and Pawlowsky-Glahn, 2011; Reimann et al., 2012). However, the Gi* statistic is a parametric spatial statistic that depends on classical statistical parameters (mean and standard deviation). Hence, it is necessary to estimate some parametric descriptive statistics for the raw data of TOC and pH (Table 1). The median values of TOC and pH are 1.80 and 5.77 wt%, respectively. For pH values, the median value (5.77) for the GEMAS agricultural soil samples is very close to that (5.5) of the FOREGS data set (De Vos et al., 2006), indicating that the majority of European agricultural soil samples are acidic or weakly acidic. The GEMAS data median for TOC (1.80%) is also close to the FOREGS data set in topsoil (1.73 wt%).

The large differences among the median, 75th percentile and the maximum value indicate the existence of potential high-value outliers (Zhang et al., 2009). Histograms of TOC contents with the normal distribution curve superimposed are shown in Fig. 2a. The raw data exhibit a long tail towards higher TOC contents, suggesting the existence of high-value outliers. The normal score transformed data of TOC contents display a relatively symmetrical distribution in Fig. 2b (K-S test p value = 0.001 < 0.05). Although it does not perfectly obey the normal distribution, it still greatly reduces the influences of outliers in original data set. Therefore, the normal score transformed data were used for further spatial analysis in this study.

Table 1
Descriptive statistics for TOC and pH in European agricultural soil (TOC in wt%; n = 2095)).

Parameter	Min	25%	Mean	Median	75%	Max	SD
TOC	0.40	1.2	2.55	1.80	2.6	46.0	3.98
pH	3.32	4.96	5.88	5.77	7.03	7.98	1.10

SD: standard deviation.

The pH values show an overall symmetrical statistical distribution, as depicted by the superimposed normal distribution curve (Fig. 2c), because the logarithmic transformation was already performed on concentrations of H^+ to obtain the pH value. However, the pH histogram shows two distinct peaks with a break point at about 6.5, indicating the existence of two 'populations', and the majority of Ap soil samples being overall acidic (<7 pH). The samples in the 'population' with high pH values (e.g., 7.00–7.98) are mostly located in the Mediterranean area (see Fig. 1), where the major soil parent materials are limestone and marble. For consistency and to obtain a symmetrical distribution, the pH values were also subjected to the normal score transformation prior to spatial analysis (Fig. 2d; K-S test p value = 0.2 > 0.05).

3.2. Hot spot analysis of TOC

Hot spot analysis is an important tool in identifying spatial patterns by pinpointing the location and clustering in the TOC data. The hot spot identification results of TOC are affected by various factors, including the logarithmic transformation of the raw data set. To reduce the effects of spatial outliers, the spatial weight relationships between each feature, and the choice of distance band are important factors (Zhang et al., 2008). Two key factors on hot spot analysis are discussed here: data transformation and the distance band.

3.2.1. Effects of data transformation on identification of hot and cold spots

Comparison of the spatial distribution between the raw and normal score transformed data of TOC contents is shown in Fig. 3. For the purposes of an absolute comparison, the input parameters for both maps were kept the same, i.e., a fixed distance band of 100 km. The maps show great differences between the raw (Fig. 3a) and normal score transformed data (Fig. 3b). The map of normal score transformed data identifies a greater number of significant hot spots in northern Europe, the United Kingdom and Ireland. Significant cold spots are shown in central and southern Europe, i.e., Portugal, Spain, south-eastern and north-central France, northern Italy, eastern Sicily, Hellas, Hungary, north-eastern Germany, Poland and Ukraine (Fig. 3b). In contrast, there was no significant cold spot on the raw data map, an expected outcome because of the strongly right-skewed data distribution (Fig. 3a).

After data transformation, the effects of outliers (extremely high values) on the spatial analysis of TOC contents were reduced, as they were 'scaled' closer towards the majority of the data. Before data transformation, the mean value of TOC contents was comparatively high due to the existence of outliers (see Fig. 2a), and the total number of significant hot spots identified by Gi* statistics was relatively small (Fig. 3a). Therefore, the number of TOC hot spots was less than that after data transformation. In addition, a number of cold spots were identified on the map (see Fig. 3b). This is because the normal score transformation makes both high and low values evenly distributed by ranking them in order (Zhang et al., 2008). It is, therefore, demonstrated that data transformation is an important influencing factor in identifying spatial distribution patterns on hot and cold spot analysis maps. Due to the non-normality of geochemical data distributions (Reimann and Filzmoser, 2000) and the existence of outliers, the transformed TOC and pH data were used for hot spot analysis.

3.2.2. Effects of different distance bands on identification of hot and cold spots

There is no widely accepted criteria to select the optimal distance band, because this depends on the sampling density of the geochemical

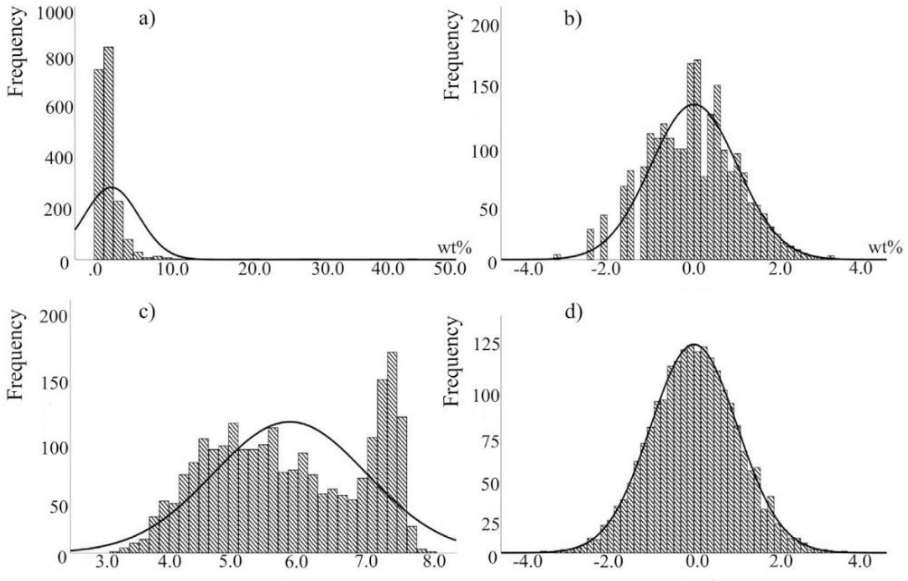


Fig. 2. Histograms and normal distribution curve of TOC contents and pH values (n = 2095): a) raw data of TOC; b) normal score transformed data of TOC; c) raw data of pH; d) normal score transformed data of pH.

survey. In general, the distance band should not be longer than half the maximum distance between all sample pairs and not shorter than the sampling interval (Zhang et al., 2008). To investigate the effects of different distance bands on hot spot analysis of TOC data, four distance bands were applied in this study: 20, 50, 75 and 100 km, and the resulting maps are shown in Fig. 4.

It can be clearly seen that the results are different for the four distance bands. When the shortest distance band (20 km; Fig. 4a) is examined, the hot and cold spot results for the majority of soil samples are insignificant. Only a few hot spots were identified in northern Europe and the United Kingdom, with a few cold spots in Spain, north-east Germany and Poland.

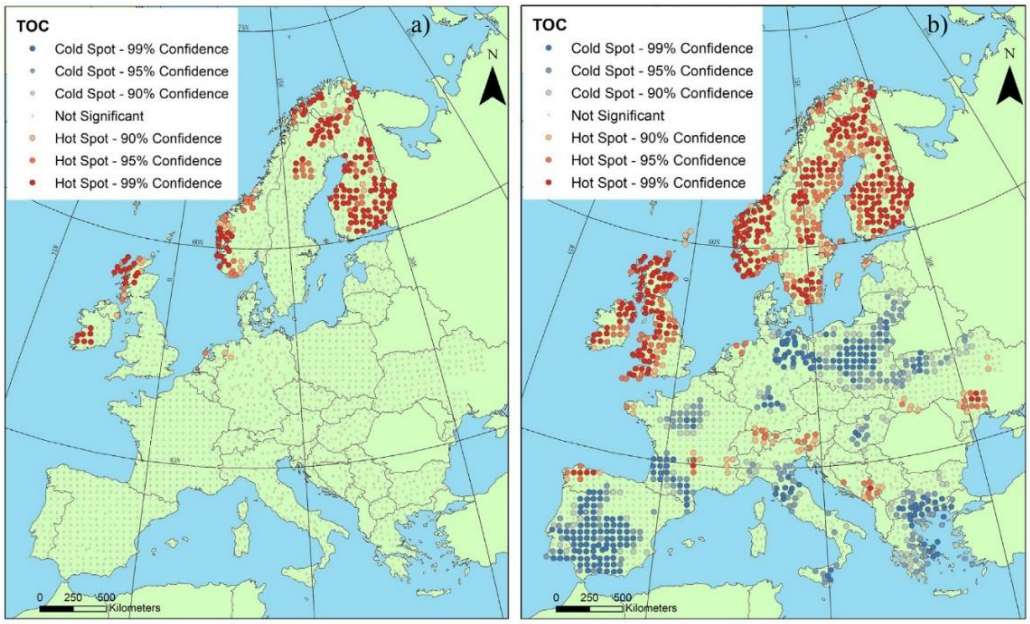


Fig. 3. Spatial distribution map of significant TOC hot and cold spots calculated using a distance band of 100 km: a) raw data; b) normal score transformed data.

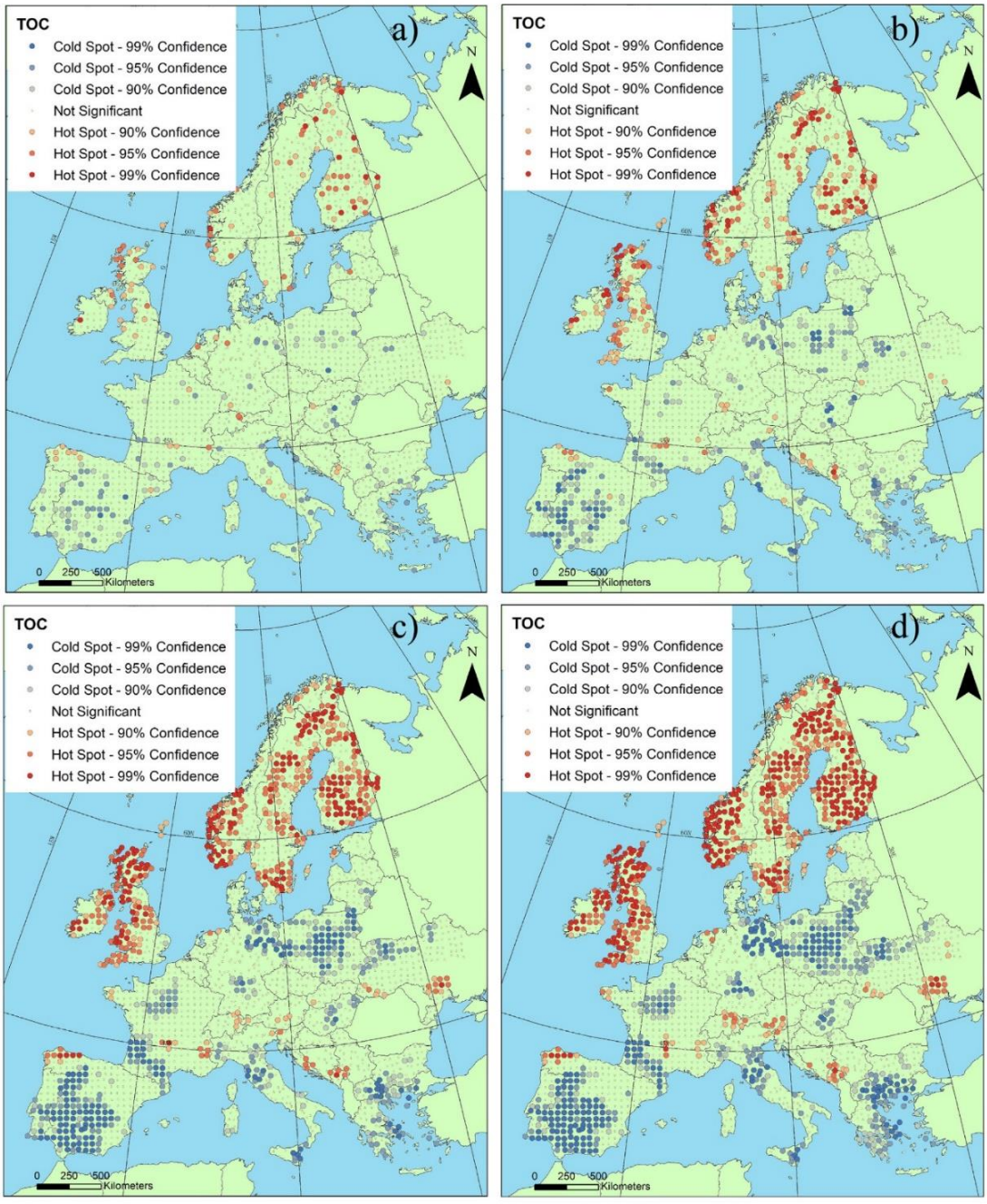


Fig. 4. Spatial distribution maps of significant TOC hot and cold spots calculated by using different distance bands: a) 20 km, b) 50 km, c) 75 km and d) 100 km.

When the distance band is increased to 50 km (Fig. 4b), the number of significant hot and cold spots increases. Hot spots are clustered in northern Europe, mainly in Fennoscandia, United Kingdom and Ireland, while cold spots show more intensive clusters in central and southern Europe.

The total number of significant hot and cold spots increased enormously at the distance band of 75 km, which clearly reveals the spatial distribution patterns of TOC in European agricultural soil (Fig. 4c). Overall, the TOC content of agricultural soil in northern Europe is higher than that in central and southern Europe. In Fig. 4c, the majority of significant

hot spots are located in Fennoscandia, United Kingdom and Ireland, with a small number of hot spots clustered in Switzerland, Austria, southern France, north-west Spain, Croatia, Montenegro, south-western and south-central Ukraine. Cold spots mainly occur in Poland, Germany, France, north-western Ukraine and southern Europe (e.g., Portugal, Spain, Italy and Hellas).

When the distance band is increased to 100 km, there is a greater number of hot spots shown on the map (Fig. 4d), suggesting a strong influence of distance band on hot spot analysis results. As the distance band is increased, the number of significant hot spots increases. However, the numbers of significant hot and cold spots did not change much as the distance band increased further than 100 km. Therefore, taking into consideration the different distance band influence, it seems that the distance band of 100 km is the optimal among the four selected bands for revealing the spatial patterns of TOC between north and south at the European-scale. The range of influence depends, of course, on the distance between samples, and the studied variable, as well as the interpolation algorithms. In another GEMAS spatial analysis study of the Ni distribution in agricultural soil, the triangular irregular network (TIN) raster map with a smoothing window size of 110 × 110 km revealed best the large-scale spatial trends and patterns (Jordan et al., 2018). Therefore, the optimal distance band for the GEMAS sampling design appears to be around 100 km.

3.3. Spatial relationships between TOC and pH in European agricultural soil

3.3.1. Spatial distribution of TOC and pH

There is a generally negative relationship between TOC and pH values, i.e., soil samples with high TOC contents contain more organic matter and more organic acids, resulting in low pH values (Fabian et al., 2014). However, due to complicated influencing factors, such a relationship may be interfered. The spatial patterns of pH hot and cold spots are shown in Fig. 5. When compared with the TOC hot and cold spots spatial patterns in Fig. 4, the overall negative relationship between TOC and pH data in European agricultural soil is observed. Soil TOC in southern Europe is significantly lower than that in northern Europe, while soil pH values in southern Europe are significantly higher than those in northern Europe, showing opposite spatial distribution patterns between TOC and pH on hot and cold spot maps.

The most interesting pattern found in this hot and cold spot analysis study is that both low TOC contents and low pH values are observed in north-central Europe (Poland, Germany and north-western Ukraine).

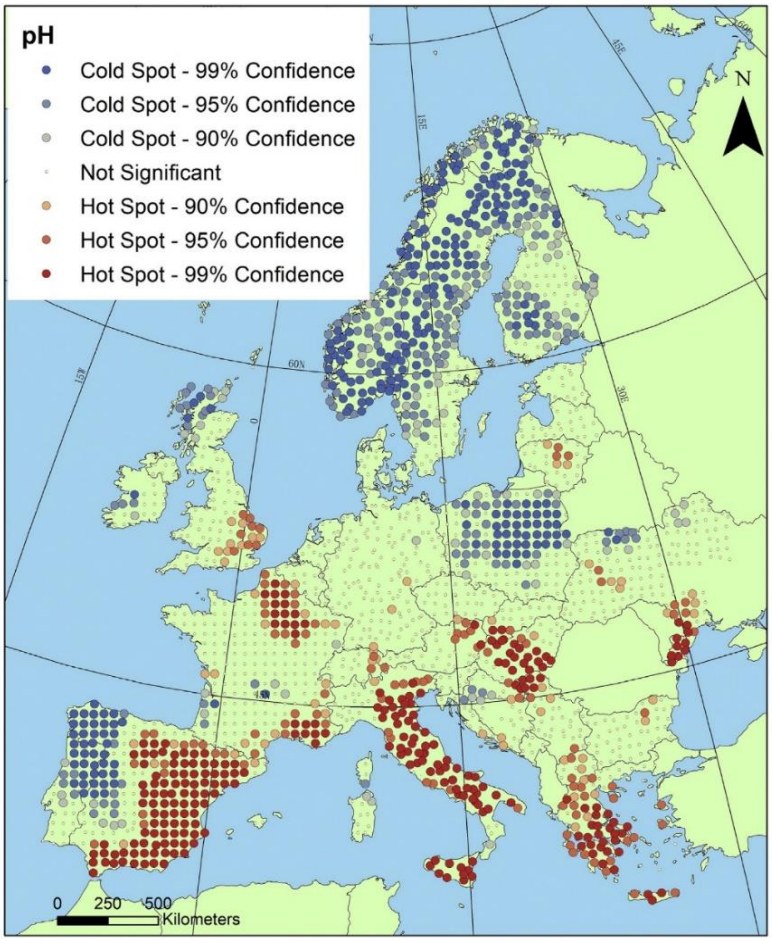


Fig. 5. Hot and cold spot map of pH values calculated at a distance band of 100 km.

This feature does not follow the overall negative correlation between TOC and pH values. The agricultural soil at these locations has a particularly high concentration of SiO_2. Here coarse-grained sediments of the last glaciation occur (Reimann et al., 2014a). A different soil parent material can thus cause special patterns of TOC and pH values in north-central Europe, which will be further explored in the following sections.

3.3.2. Environmental factors associated with negative relationship between TOC and pH

Spearman rank correlation coefficients were calculated to measure the relationship among TOC, pH and other selected environmental parameters (Table 2). The Spearman rank correlation coefficient between two variables is equivalent to the Pearson correlation coefficient between the rank values of the two variables (Caruso and Cliff, 1997), while the Spearman index does not require the normality of the data. Due to sample size effect of the large number of samples (e.g., $n > 100$) used for correlation analysis in this study, the significant levels need to be carefully interpreted (Zhang et al., 2005).

Except altitude, all other parameters (temperature, precipitation, clay+silt and SiO_2) showed significant correlations with TOC and pH. Nevertheless, temperature and clay+silt contents showed weak negative correlation with soil TOC, and weak positive correlation with pH values in European agricultural soil. Interestingly, the concentration of SiO_2 is negatively correlated with both TOC and pH ($r = -0.379$ and $r = -0.322$, respectively), which could be affected by the special relationship between high SiO_2 and both low TOC contents and pH values in central-eastern Europe.

The overall negative correlation between TOC contents and pH values could be related to multiple processes. Organic Carbon is the progenitor of carbonic acid, which contributes to reducing the rate of degradation of organic matter by microorganisms in soil (McGrath and Zhang, 2003). Meanwhile acid deposition influences soil pH values in forests (White et al., 1995; White and Cresser, 1998). Although this process is obvious in forests, it will also affect agricultural soil to a certain degree if there is no other buffer system in soil (Fabian et al., 2014). Thus, due to the limited biomass of plants, comparatively low residue input contributes to a low soil organic matter level. In addition, the leaching of dissolved organic carbon is generally accelerated by relatively high pH values, resulting in a decreased organic carbon content in surface soil (Andersson and Nilsson, 2001).

Climate also plays a key role on soil TOC contents and pH values. High precipitation and low temperature tend to cause a decreased decomposition of organic matter, favouring, thus, the accumulation of humus (Jenny, 1980), and in turn affecting the content of TOC. Furthermore, several studies have reported that climate (characterised by temperature and precipitation) is an important determinant of physicochemical properties in soil (Liu et al., 2013). For instance, an increase in rainfall may contribute to a higher salt leaching rate. Leaching of base cations can result in significant reduction in soil pH values (Darilek et al., 2009). Wet conditions facilitate the formation of stable soil organic carbon (SOC) mineral surfaces by enhanced weathering of soil parent materials (Mikutta and Kaiser, 2011; Doetterl et al., 2015). Temperature also greatly influences the microbial decomposition of organic matter because of its complex molecular properties with the high intrinsic sensitivity to temperature (Davidson and Janssens, 2006; Conant et al., 2011).

Although this relationship is controlled by a variety of constraints, many studies suggest that TOC contents decrease with increasing temperature (Jobággy and Jackson, 2000; Sleutel et al., 2007), and humid and cool conditions favour the stock of soil organic carbon at global scale (Post et al., 1982).

It is clear from the hot and cold spot maps (Figs. 4c and 5) and spatial distribution maps, plotted with a different interpolation method (see Reimann et al., 2014b, p.193), that the TOC contents in Fennoscandia are significantly higher than in other areas. This is due to the natural factor that agricultural soil in northern Europe is generally acidic, resulting from long-term low temperature and high average annual rainfall. In contrast, soil in southern Europe has comparatively high pH values, higher annual temperature and less rainfall resulting in lower TOC contents. Therefore, mapping TOC and pH by hot spot analysis at the European scale can spatially identify their relationships.

3.3.3. Effects of quartz on both low soil TOC contents and pH values in north-central Europe

The co-existence of low TOC contents and low pH values was observed in north-central Europe, both of which are clearly identified as cold spots on the hot and cold spot analysis maps (see Fig. 5). They are further confirmed as low values on the colour surface interpolated maps (see Reimann et al., 2014b, p.132, 193). Due to the high concentration of silica in the coarse-grained sediments from the last glaciation, the soil in this region does not follow the same pattern between TOC and pH at the European scale. These glacial sands consist almost exclusively of quartz (SiO_2) and some feldspar. A hot and cold spot map of SiO_2 concentration was plotted showing that most of the SiO_2 hot spots occur in north-central Europe, covering north-western Ukraine, Poland, Lithuania, Estonia, Denmark, northern Germany, The Netherlands and Belgium (Fig. 6). While the cold spots are mainly concentrated in eastern Spain, southern France, Italy and parts of the Balkan countries. Poland is characterised by a particularly high concentration of SiO_2 (quartz) in agricultural soil, which accounts for 73.6% in the soil parent materials (SPM) based on the GEMAS Ap data set. These coarse-grained, quartz-rich sediments were deposited by the glaciers of the last ice age (Holzhauser et al., 2005; Piotrowski et al., 2006; Woronko and Bujak, 2018).

The importance of SPM on TOC can be attributed to differences in quartz (SiO_2) contents (Badgery et al., 2013). Generally, the occurrence of quartz is directly related to coarse-grained sandy soil. Soil formed on these coarse-grained glacial sediments contains larger particles and a greater proportion of sand. Some researchers have demonstrated a direct effect of soil properties (e.g., particle size, sand proportion) on SOC contents (Kern, 1994; Homann et al., 1998; Percival et al., 2000). Several studies have reported that the sequestration of organic materials in soil is texture-dependent and highly correlated with the percentage of fine particles (Scott and Cole, 1996; Hassink et al., 1997). Soil with high sand content exhibits low OC storage due to its low aggregate stability (Le Bissonnais and Arrouays, 1997), while heavy clay and silty clay show higher OC contents.

However, the low soil pH values can also be related to high concentration of SiO_2 (quartz). These glacial sediments and moraines contain little Ca^{2+} to buffer the soil pH values due to the coarse-grained particles (Fabian et al., 2014), leading to acidification of agricultural soil. In addition, the perennial cold climate in north-central Europe has also a

Table 2
Spearman's rank correlation coefficients of TOC contents and pH values and selected environmental variables (n = 2095).

Parameter	TOC (wt%)	pH	Altitude (m)	Clay + silt (%)	Temperature (°C)	Precipitation (mm/yr)	SiO_2 (%)
TOC	1.000	**−0.162**[a]	−0.003	**−0.082**[a]	**−0.408**[a]	**0.213**[a]	**−0.379**[a]
pH	**−0.162**[a]	1.000	0.002	**0.410**[a]	**0.449**[a]	**−0.107**[a]	**−0.322**[a]

[a] Correlation is significant at the 0.01 level (2-tailed).

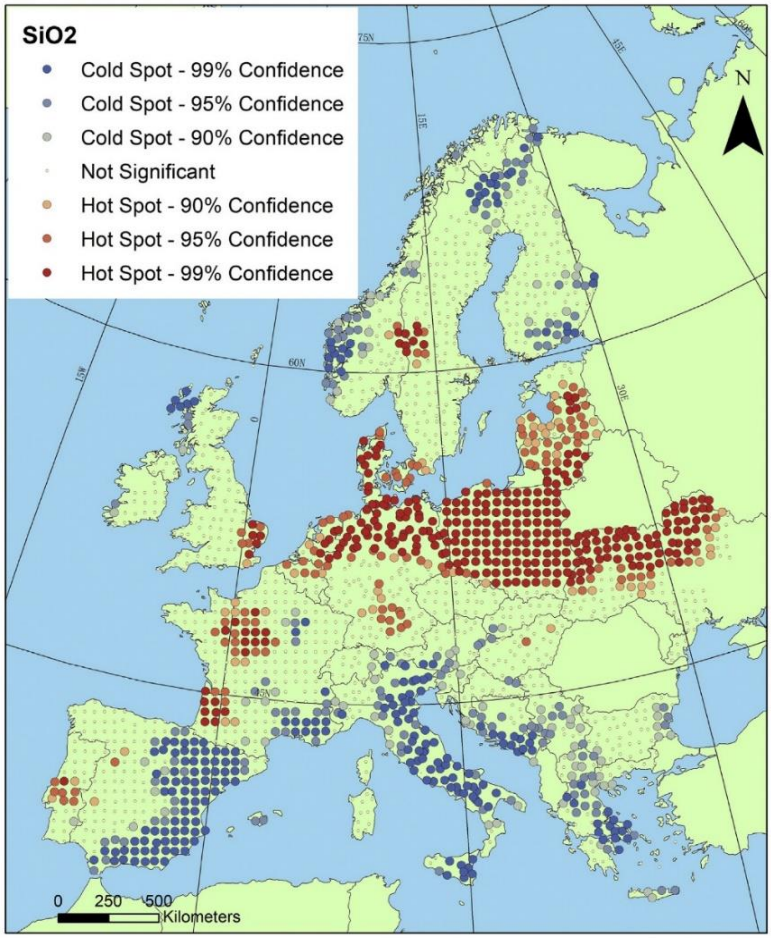


Fig. 6. Hot and cold spot map of SiO_2 concentrations calculated at a distance band of 100 km.

direct impact on acid soil in these countries, which has been confirmed by global temperature models (e.g., Hijmans et al., 2005).

It should be mentioned that although the glacial deposits extend northward to the Baltic States (Estonia, Latvia, see Fig. 6), the soil pH values were not low in these areas. This may be attributed to the occurrence of limestone in these areas resulting in a sufficient amount of Ca^{2+} to buffer the acid in soil, leading to comparatively higher pH values than those in north-central Europe.

4. Conclusions

This study identified the spatial distribution patterns of soil TOC contents in agricultural soil at the European scale based on the GEMAS Ap data set. The comparison of TOC contents and pH values in European agricultural soil shows that they are negatively correlated. However, a significant observation is the co-existence of both low TOC contents and low pH values in agricultural soil in north-central Europe. This 'unusual' spatial pattern is attributed to the high concentration of SiO_2 (quartz) in the coarse-grained glacial sediments, resulting in a high proportion of sand in the coarse-textured soil of this area. Climate (temperature and precipitation), soil parent material and soil texture appear to be the main influencing factors for the observed overall spatial distribution patterns of TOC and pH in Europe. Their special spatial pattern in north-central Europe is mainly related to the special pattern of the last glacier deposition. The observation of such special spatial pattern found in this study was achieved based on hot spot analysis, demonstrating its potential power in identifying hidden spatial patterns in environmental geochemical studies. For the successful application of the method the data needed to be normal-score transformed to reach an as symmetrical distribution as possible. Testing different distance bands, a distance of 100 km provided optimal results for the GEMAS Ap data set.

Conflict of interest

The authors declare that they have no actual or potential financial interest.

Acknowledgements

We acknowledge participants of EU COST Action ES1406 (https://www.cost.eu/actions/ES1406/#tabs|Name:overview) for useful comments and suggestions in an earlier version of this paper. Also, the authors wish to thank all European sampling teams and colleagues

involved in the GEMAS project for planning and sampling, and colleagues at the State Geological Institute of Dionyz Stur in Slovakia for sample preparation.

The GEMAS project is a cooperation project of the EuroGeoSurveys Geochemistry Expert Group with a number of outside organisations (e.g., Alterra, Netherlands; Norwegian Forest and Landscape Institute; Research Group Swiss Soil Monitoring Network, Swiss Research Station Agroscope Reckenholz-Tänikon, several Ministries of the Environment and University Departments of Geosciences Chemistry and Mathematics in a number of European countries and New Zealand; ARCHE Consulting in Belgium; CSIRO Land and Water in Adelaide, Australia) and Eurometaux.

Sampling was carried out by the participating organisations in their own country. The analytical work was co-financed by the following organisations: Eurometaux, Cobalt Development Institute (CDI), European Copper Institute, Nickel Institute, Europe, European Precious Metals Federation, International Antimony Association, International Manganese Institute, International Molybdenum Association, ITRI Ltd. (on behalf of the REACH Tin Metal Consortium), International Zinc Association, International Lead Association-Europe, European Borates Association, the (REACH) Vanadium Consortium and the (REACH) Selenium and Tellurium Consortium.

The GEMAS project was managed by the Geological Survey of Norway with financial support from EuroGeoSurveys. Finally, the Directors of the European Geological Surveys, and the additional participating organisations, are thanked for making the sampling of almost all of Europe in a tight time schedule possible.

References

Alessa, L., Kliskey, A., Brown, G., 2008. Social-ecological hotspots mapping: a spatial approach for identifying coupled social-ecological space. Landscape Urban Plan 85 (1), 27–39.

Andersson, S., Nilsson, S.I., 2001. Influence of pH and temperature on microbial activity, substrate availability of soil-solution bacteria and leaching of dissolved organic carbon in a mor humus. Soil Biol. Biochem. 33, 1181–1191.

Ansong, D., Renwick, C.B., Okumu, M., Ansong, E., Wabwire, C.J., 2018. Gendered geographical inequalities in junior high school enrollment: do infrastructure, human, and financial resources matter? J. Econ. Stud. 45 (2), 411–425.

Aitchison, J., 1986. The statistical analysis of compositional data. Chapman & Hall, London, p. 416.

Badgery, W.B., Simmons, A.T., Murphy, B.M., Rawson, A., Andersson, K.O., Lonergan, V.E., van de Ven, R., 2013. Relationship between environmental and land-use variables on soil carbon levels at the regional scale in central New South Wales, Australia. Soil Res 51, 645–656.

Baritz, R., Ernstsen, V., Zirlewagen, D., 2014. Carbon concentrations in European agricultural and grazing land soil. Chapter 6. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils–Part B: General Background Information and Further Analysis of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102). Schweizerbarth, Hannover, pp. 117–129 eds.

Barro, A.S., Kracalik, I.T., Malania, L., Tsertsvadze, N., Manvelyan, J., Imnadze, P., Blackburn, J.K., 2015. Identifying hotspots of human anthrax transmission using three local clustering techniques. Appl. Geogr. 60, 29–36.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. Eur. J. Soil Sci. 47, 151–163.

Birke, M., Reimann, C., Fabian, K., 2014. Analytical methods used in the GEMAS project. Chapter 5. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils – Part a: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102). Schweizerbarth, Hannover, pp. 41–46.

Bradley, R.I., Milne, R., Bell, J., Lilly, A., Jordan, C., Higgins, A., 2005. A soil carbon and land use database for the United Kingdom. Soil Use Manag. 21, 363–369.

Braithwaite, A., Li, Q., 2007. Transnational terrorism hot spots: identification and impact evaluation. Conflict Manag. Peace 24 (4), 281–296.

Caruso, J.C., Cliff, N., 1997. Empirical size, coverage, and power of confidence intervals for Spearman's Rho. Educ. Psychol. Meas. 57 (4), 637–654.

Conant, R.T., Ryan, M.G., Ågren, G.I., Birge, H.E., Davidson, E.A., Eliasson, P.E., Evans, S.E., Frey, S.D., Giardina, C.P., Hopkins, F.M., Hyvönen, R., Kirschbaum, M.U.F., Lavallee, J.M., Leifeld, J., Parton, W.J., Megan, S.J., Wallenstein, M.D., Martin, W.J.Â., Bradford, M.A., 2011. Temperature and soil organic matter decomposition rates -synthesis of current knowledge and a way forward. Glob. Chang. Biol. 17 (11), 3392–3404.

Darilek, J.L., Huang, B., Wang, Z.G., Qi, Y.B., Zhao, Y.C., Sun, W.X., Gu, Z.Q., Shi, X.Z., 2009. Changes in soil fertility parameters and the environmental effects in a rapidly developing region of China. Agric. Ecosyst. Environ. 129, 286–292. https://doi.org/10.1016/j.agee.2008.10.002.

Davidson, E.A., Janssens, I.A., 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440 (7081), 165–173.

De Vos, W., Tarvainen, T. (chief-editors), Salminen, R., Reeder, S., De Vivo, B., Demetriades, A., Pirc, S., Batista, M.J., Marsina, K., Ottesen, R.T., O'Connor, P.J., Bidovec, M., Lima, A, Siewers, U., Smith, B., Taylor, H., Shaw, R., Salpeteur, I., Gregorauskiene, V., Halamic, J., Slaninka, I., Lax, K., Gravesen, P., Birke, M., Breward, N., Ander, E.L., Jordan, G., Duris, M., Klein, P., Locutura, J., Bel-lan, A., Pasieczna, A., Lis, J., Mazreku, A., Gilucis, A., Heitzmann, P., Klaver, G., Petersell, V., 2006. Geochemical Atlas of Europe. Part 2 - Interpretation of Geochemical Maps, Additional Tables, Figures, Maps, and Related Publications. Geological Survey of Finland, Espoo, 692 pp.; http://weppi.gtk.fi/publ/foregsatlas/.

Demetriades, A., Reimann, C., Filzmoser, P., 2014. Evaluation of GEMAS project quality control results. Chapter 6. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils – Part A: Methodology and Interpretation of the GEMAS Data Set Geologisches Jahrbuch (Reihe B102). Schweizerbarth, Hannover, pp. 47–60.

Di Minin, E., Hunter, L.T.B., Balme, G.A., Smith, R., Goodman, P.S., Slotow, R., 2013. Creating larger and better connected protected areas enhances the persistence of big game species in the Maputaland-Pondoland-Albany biodiversity hotspot. PLoS One 8 (8), e71788.

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Pinto, M.C., Casanova-Katny, A., Munoz, C., Boudin, M., Venegas, Ez., Boeckx, P., 2015. Soil carbon storage controlled by interactions between geochemistry and climate. Nat. Geosci. 8 (10), 780–783.

EC, 2006. Regulation (EC) no 1907/2006 of the European Parliament and of the council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) no 793/93 and Commission Regulation (EC) no 1488/94 as well as council directive 76/769/EEC and commission directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Off. J. Eur. Communities 1–849 30.12.2006, L396.

ECHA, 2012. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.16: Environmental Exposure Assessment. Version 2.1. European Chemicals Agency (147 pp).

EGS, 2008. EuroGeoSurveys geochemistry working group. EuroGeoSurveys geochemical mapping of agricultural and grazing land in Europe (GEMAS) - field manual. Norges Geologiske Undersøkelse Report, 2008.038 46 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2008/2008_038.pdf.

ESRI, 2016. How Hot Spot Analysis (Getis-Ord Gi*) Works. http://desktop.arcgis.com/en/arcmap/10.4/tools/spatial-statistics-toolbox/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm.

Eswaran, H., Van Den Berg, E., Reich, P., 1993. Organic carbon in soils of the world. Soil Sci. Soc. Am. J. 57, 192–194.

Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E., 2014. GEMAS: spatial distribution of the pH of European agricultural and grazing land soil. Appl. Geochem. 48, 207–216.

Fang, X., Zhang, J., Meng, M., Guo, X., Wu, Y., Liu, X., Zhao, K., Ding, L., Shao, Y., Fu, W., 2017. Forest-type shift and subsequent intensive management affected soil organic carbon and microbial community in southeastern China. Eur. J. Forest Res. 136 (4), 689–697.

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data — problems and possibilities. Sci. Total Environ. 407, 6100–6108.

Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Sci. Total Environ. 408, 4230–4238.

Filzmoser, P., Reimann, C., Birke, M., 2014. Univariate Data Analysis and Mapping. Chapter 8. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. (Eds.), Chemistry of Europe's agricultural soils - Part A: Methodology and interpretation of the GEMAS data set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 67–81.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3), 189–206.

Guo, L., Gifford, R.M., 2002. Soil carbon stocks and land use change: a meta analysis. Glob. Change Biol. 8, 345–360.

Hassink, J., Whitmore, A.P., Kubat, J., 1997. Size and density fractionation of soil organic matter and the physical capacity of soils to protect organic matter. Eur. J. Agron. 7, 189–199.

Hati, K.M., Swarup, A., Dwivedi, A.K., Misra, A.K., Bandyopadhyay, K.K., 2007. Changes in soil physical properties and organic carbon status at the topsoil horizon of a vertisol of central India after 28 years of continuous cropping, fertilization and manuring. Agric. Ecosyst. Environ. 119, 127–134.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Holzhauser, H., Magny, M., Zumbuehl, H.J., 2005. Glacier and lake-level variations in west-central Europe over the last 3500 years. Holocene 15 (6), 789–801.

Homann, P.S., Sollins, P., Fiorella, M., Thorson, T., Kern, J.S., 1998. Regional soil organic carbon storage estimates for western Oregon by multiple approaches. Soil Sci. Soc. Am. J. 62, 789–793.

ISO, 1995. ISO 10694:1995. Soil Quality - Determination of Organic and Total Carbon after Dry Combustion (Elementary Analysis). International Organization for Standardization (ISO), Geneva (7 pp).

Jackson, R.B., Banner, J.L., Jobbagy, E.G., Pockman, W.T., Wall, D.H., 2002. Ecosystem carbon loss with woody plant invasion of grasslands. Nature 418, 623–626.

Jandl, R., Lindner, M., Vesterdal, L., Bauwens, B., Baritz, R., Hagedorn, F., Johnson, D.W., Minkkinen, K., Byrne, K.A., 2007. How strongly can forest management influence soil carbon sequestration? A review. Geoderma 137, 253–268.

Jenny, H., 1980. The Soil Resource, Origin and Behavior. Springer-Verlag, New York (392 pp.).

Jobággy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10 (2), 423–436.

Johnston, C.A., Groffman, P., Breshears, D.D., Cardon, Z.G., Currie, W., Emanuel, W., Gaudinski, J., Jackson, R.B., Lajtha, K., Nadelhoffer, K., Nelson, D.J., Post, W.M., Retallack, G., Wielopolski, L., South Dakota State University, 2004. Carbon cycling in soil. Front. Ecol. Environ. 2, 522–528.

Jordan, G., Petrik, A., De Vivo, B., Albanese, S., Demetriades, A., Sadeghi, M., The GEMAS Project Team, 2018. GEMAS: spatial analysis of the Ni distribution on a continental-scale using digital image processing techniques on European agricultural soil data. J. Geochem. Explor. 186, 143–157. https://doi.org/10.1016/j.gexplo.2017.11.011.

Kern, J.S., 1994. Spatial patterns of soil organic carbon in the contiguous United States. Soil Sci. Soc. Am. J. 58, 439–455.

Kumar, D., Singh, A., Jha, R., Sahoo, S., Jha, V., 2018. Using spatial statistics to identify the uranium hotspot in groundwater in the mid-eastern Gangetic plain, India. Environ. Earth Sci. 77 (19), 1–12.

Lal, R., 2005. Forest soils and carbon sequestration. Forest Ecol. Manag. 220, 242–258.

Le Bissonnais, Y., Arrouays, D., 1997. Aggregate stability and assessment of soil crustability and erodibility. II. Application to humic loamy soils with various organic carbon contents. Eur. J. Soil Sci. 48, 39–48.

Lenka, N.K., Lal, R., 2013. Soil aggregation and greenhouse gas flux after 15 years of wheat straw and fertilizer management in a no-till system. Soil Tillage Res. 126, 78–89.

Liu, Y., Lv, J., Zhang, B., Bi, J., 2013. Spatial multi-scale variability of soil nutrients in relation to environmental factors in a typical agricultural region, Eastern China. Sci. Total Environ. 450, 108–119. https://doi.org/10.1016/j.scitotenv.2013.01.083.

Lu, N., Noori, M., Liu, Y., 2017. Fatigue reliability assessment of welded steel bridge decks under stochastic truck loads via machine learning. J. Bridg. Eng. 22, e04016105.

Mackových, D., Lučivjanský, P., 2014. Preparation of GEMAS project samples and standards. Chapter 4. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils. Part a: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 37–40.

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8, 1053–1065.

Mathieu, J.A., Hatté, C., Balesdent, J., Parent, É., 2015. Deep soil carbon dynamics are driven more by soil type than by climate: a worldwide meta-analysis of radiocarbon profiles. Glob. Chang. Biol. 21, 4278–4292.

Matschullat, J., Reimann, C., Birke, M., Dos Santos Carvalho, D., GEMAS Project Team, 2018. GEMAS: CNS concentrations and C/N ratios in European agricultural soil. Sci. Total Environ. 627, 975–984.

McGrath, D., Zhang, C.S., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem. 18, 1629–1639.

Meersmans, J., Van Wesemael, B., Goidts, E., Van Molle, M., De Baets, S., De Ridder, F., 2011. Spatial analysis of soil organic carbon evolution in Belgian croplands and grasslands. Glob. Chang. Biol. 17, 466–479. https://doi.org/10.1111/j.1365-2486.2010.02183.x.

Mikutta, R., Kaiser, K., 2011. Organic matter bound to mineral surfaces: resistance to chemical and biological oxidation. Soil Biol. Biochem. 43 (8), 1738–1741.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27 (4), 286–306.

O'Rourke, S.M., Angers, D.A., Holden, N.M., Mcbratney, A.B., 2015. Soil organic carbon across scales. Glob. Chang. Biol. 21, 3561–3574.

Pan, G., Li, L., Wu, L., Zhang, X., 2003. Storage and sequestration potential of topsoil organic carbon in China's paddy soils. Glob. Chang. Biol. 10, 79–92.

Percival, H.J., Parfitt, R.L., Scott, N.A., 2000. Factors controlling soil organic carbon levels in New Zealand grasslands: is clay content important? Soil Sci. Soc. Am. J. 64, 1623–1630.

Piotrowski, J.A., Larsen, N.K., Menzies, J., Wysota, W., 2006. Formation of subglacial till under transient bed conditions: deposition, deformation, and basal decoupling under a Weichselian ice sheet lobe, central Poland. Sedimentology 53 (1), 83–106.

Post, W.M., Emanuel, W.R., Zinke, P.J., Stangenberger, A.G., 1982. Soil carbon pools and world life zones. Nature 298 (5870), 156–159.

Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol. 39 (9), 1001–1014.

Reimann, C., Siewers, U., Tarvainen, T., Bityukova, L., Eriksson, J., Gilucis, A., Gregorauskiene, V., Lukashev, V.K., Matinian, N.N., Pasieczna, A., 2003. Agricultural Soils in Northern Europe: A Geochemical Atlas. E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart (279 pp).

Reimann, C., Demetriades, A., Eggen, O.A., Filzmoser, P., The EuroGeoSurveys Geochemistry Expert Group, 2009. The EuroGeoSurveys Geochemical Mapping of Agricultural and Grazing Land Soils Project (GEMAS) – Evaluation of Quality Control Results of Aqua Regia Extraction Analysis. Geological Survey of Norway, Trondheim, NGU Report 2009.049. 94 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2009/2009_049.pdf.

Reimann, C., Demetriades, A., Eggen, O.A., Peter Filzmoser, P., the EuroGeoSurveys Geochemistry Expert Group, 2011. The EuroGeoSurveys GEochemical Mapping of Agricultural and Grazing Land Soils Project (GEMAS) – Evaluation of Quality Control Results of Total C and S, Total Organic Carbon (TOC), Cation Exchange Capacity (CEC), XRF, pH, and Particle Size Distribution (PSD) Analysis. Geological Survey of Norway, Trondheim, NGU Report 2011.043. 90 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2011/2011_043.pdf.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., The GEMAS Project Team, 2012. The concept of compositional data analysis in practice — Total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210. https://doi.org/10.1016/j.scitotenv.2012.02.032.

Reimann, C., Demetriades, A., Birke, M., Filzmoser, P., O'Connor, P., Halamić, J., Ladenberger, A., the GEMAS Project Team, 2014a. Distribution of elements/parameters in agricultural and grazing land soil of Europe. Chapter 11. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils – Part a: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 103–474.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014b. Chemistry of Europe's Agricultural Soils, Part a: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover (523 pp).

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P., 2014c. Chemistry of Europe's Agricultural Soils, Part B: General Background Information and Further Analysis of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B103), Schweizerbarth, Hannover (349 pp).

Reisser, M., Purves, R.S., Schmidt, M.W.I., Abiven, S., 2016. Pyrogenic carbon in soils: a literature-based inventory and a global estimation of its content in soil organic carbon and stocks. Front. Earth Sci. 4 (80), 14. https://doi.org/10.3389/feart.2016.00080.

Rodríguez Martín, J.A., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J.J., Grau Corbí, J.M., Boluda, R., 2016. Assessment of the soil organic carbon stock in Spain. Geoderma 264, 117–125.

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. https://doi.org/10.1038/nature10386.

Scott, N.A., Cole, C.V., 1996. Soil textural control on decomposition and soil organic matter dynamics. Soil Sci. Soc. Am. J. 60, 1102–1109.

Sleutel, S., De Neve, S., Hofman, G., 2007. Assessing causes of recent organic carbon losses from cropland soils by means of regional-scaled input balances for the case of Flanders (Belgium). Nutr. Cycl. Agroecosyst. 78 (3), 265–278.

Smith, P., Martino, D., Cai, Z., Gwary, D., Janzen, H., Kumar, P., Mccarl, B., Ogle, S., O' Mara, F., Rice, C., Scholes, B., Sirotenko, O., Howden, M., Mcallister, T., Pan, G., Romanenkov, V., Schneider, U., Towprayoon, S., Wattenbach, M., Smith, J., 2008. Greenhouse gas mitigation in agriculture. Philos. T. R. Soc. 363, 789–813.

Tiessen, H., Cuevas, E., Chacon, P., 1994. The role of soil organic matter in sustaining soil fertility. Nature 371, 783–785.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46 (Supplement), 234–240.

Tran, D.X., Pla, F., Latorre-Carmona, P., Myint, S.W., Caetano, M., Kieu, H.V., 2017. Characterizing the relationship between land use land cover change and land surface temperature. ISPRS J. PHOTOGRAMM. 124, 119–132.

Wang, F., Guo, D., McLafferty, S., 2012. Constructing geographic areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois. Appl. Geogr. 35, 1–11.

Wang, Y., Yang, Y., Shi, X., Mao, S., Shi, N., Hui, X., 2016. The spatial distribution pattern of human immunodeficiency virus/acquired immune deficiency syndrome in China. Geospat. Health 11, 104–109.

White, C., Cresser, M., 1998. Sensitivity of Scottish upland moorland podzols derived from sandstones and quartzites to acidification: the potential importance of the mobile anion effect. Water Air Soil Poll 103 (1), 229–244.

White, C., Dawod, A., Cruickshank, K., Gammack, S., Cresser, M., 1995. Evidence for acidification of sensitive Scottish soils by atmospheric deposition. Water Air Soil Poll 85 (3), 1203–1208.

Woronko, B., Bujak, L., 2018. Quaternary aeolian activity of Eastern Europe (a Poland case study). Quatern. Int. 478, 75–96.

Xu, X., Liu, W., Kiely, G., 2011a. Modeling the change in soil organic carbon of grassland in response to climate change: effects of measured versus modelled carbon pools for initializing the Rothamsted carbon model. Agric. Ecosyst. Environ. 140, 372–381.

Xu, X., Liu, W., Zhang, C.S., Kiely, G., 2011b. Estimation of soil organic carbon stock and its spatial distribution in the Republic of Ireland. Soil Use Manag. 27, 156–162.

Yu, G., Chen, Z., Piao, S., Peng, C., Ciais, P., Wang, Q., Lia, X., Zhua, X., 2014. High carbon dioxide uptake by subtropical forest ecosystems in the East Asian monsoon region. P. Natl. Acad. Sci. USA. 111 (13), 4910–4915.

Zhang, C.S., McGrath, D., 2004. Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of southeastern Ireland from two different periods. Geoderma 119, 261–275.

Zhang, C.S., Manheim, F.T., Hinde, J., Grossman, J.N., 2005. Statistical characterization of a large geochemical database and effect of sample size. Appl. Geochem. 20 (10), 1857–1874.

Zhang, C.S., Luo, L., Xu, W., Ledwith, V., 2008. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci. Total Environ. 398, 212–221.

Zhang, C.S., Tang, Y., Luo, L., Xu, W., 2009. Outlier identification and visualization for Pb concentrations in urban soils and its implications for identification of potential contaminated land. Environ. Pollut. 157, 3083–3090.

Zhang, P., Wong, D.W., So, B.K.L., Lin, H., 2012. An exploratory spatial analysis of western medical services in Republican Beijing. Appl. Geogr. 32, 556–565.

Zhang, H., Wu, P., Fan, M., Zheng, S., Wu, J., Yang, X., Zhang, M., Yin, A., Gao, C., 2018. Dynamics and driving factors of the organic carbon fractions in agricultural land reclaimed from coastal wetlands in eastern China. Ecol. Indic. 89, 639–647.

4.2 Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression

Xu, H.F., Zhang, C.S., 2021. Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression. Sci. Total Environ., 752, 141977.

Summary: This paper investigated the spatial relationships between TOC contents and pH values using GWR based on 2,108 topsoil samples that collected from GEMAS project in European agricultural soil. The existence of the spatially varying relationships between these two variables were revealed in more than 50% of the study area, with negative and positive local coefficients simultaneously observed on the continental level. The novel finding of 'special' positive correlations was observed in central-eastern Europe, while original negative correlations were found mainly clustered in northern Europe. Mixed relationships occurred in southern Europe. A strong association between the positive patterns and specific natural factors, especially the quartz-rich soil was revealed in the central-eastern part of Europe. Also, the mixed relationships in southern European areas indicated the influence from anthropogenic inputs. Our results proved that the GWR is a powerful and effective technique for revealing the spatially varying relationships, and thus provides a new way to further explore the related influencing factors on the spatial distribution of TOC and pH.

My contribution in this paper accounted for ~90% in reviewing literatures, exploring data and writing manuscript.

Research paper

Science of the Total Environment 752 (2021) 141977

Contents lists available at ScienceDirect

Science of the Total Environment

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

Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression

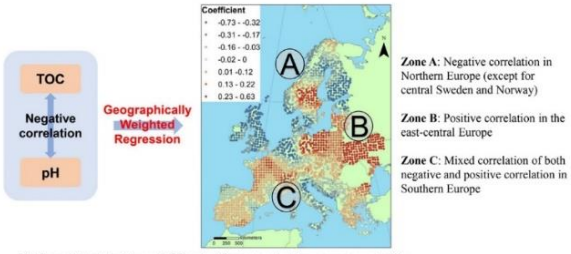
Haofan Xu, Chaosheng Zhang *

International Network for Environment and Health (INEH), School of Geography and Archaeology & Ryan Institute, National University of Ireland, Galway, Ireland

HIGHLIGHTS

- Spatially varying relationships between TOC and pH are revealed by GWR.
- Positive correlations between TOC and pH cluster in central-eastern Europe.
- Negative and mixed correlations are observed in northern and southern Europe.
- Quartz-rich soil is the main contributing factor to the positive relationship.

GRAPHICAL ABSTRACT



Positive relationship between TOC and pH in east-central Europe was revealed by Geograpically Weighted Regression, and the main contributing factor is quartz-rich soil.

ARTICLE INFO

Article history:
Received 26 June 2020
Received in revised form 14 August 2020
Accepted 23 August 2020
Available online 26 August 2020

Editor: Jay Gan

Keywords:
Total organic carbon
pH
Geographically weighted regression (GWR)
European agricultural soil

ABSTRACT

Total organic carbon (TOC) has received increased attention in recent years, not only as an important indicator in soil fertility, but also due to its close relationship with the atmosphere. Generally, soil TOC and pH values follow a negative correlation, which was revealed by traditional statistical methods. However, the conventional global models lack the ability to capture the spatial variation locally. In this study, spatially varying local relationships between TOC and pH values are studied by geographically weighted regression (GWR) on continental-scale data of European agricultural soil from the project 'Geochemical Mapping of Agricultural and Grazing land Soil' (GEMAS). In this study, TOC is the dependent and pH the independent variable. Both negative and positive local correlation coefficients are observed, showing the existence of 'special' spatially varying relationships between TOC and pH values. Original negative relationships change to positive values in more than 50% of the study area. Novel finding of significant positive correlations is observed in central-eastern Europe, while negative correlations are found mainly in northern Europe. Mixed relationships occur in southern Europe. These special patterns are strongly associated with specific natural factors, especially the extensive occurrence of quartz-rich soil in the central-eastern part of Europe. Anthropogenic inputs may have also played a role in the mixed southern European areas. The GWR technique is powerful and effective for revealing spatially varying relationships at the local level. Thus, it provides a new way to further explore the related influencing factors on the TOC and pH spatial distribution.

1. Introduction

In recent years, more researchers are focusing on global climate change, especially global warming caused by the greenhouse effect

* Corresponding author.
E-mail addresses: h.xu2@nuigalway.ie (H. Xu), Chaosheng.Zhang@nuigalway.ie (C. Zhang).

https://doi.org/10.1016/j.scitotenv.2020.141977

(e.g., Tashi et al., 2016). This is an urgent problem that needs to be solved by international collaboration as the concentration of carbon dioxide in the atmosphere continues to increase (O'Rourke et al., 2015). Soil is regarded as the largest organic carbon (OC) pool in the terrestrial ecosystems, and even a slight variation of OC stock in soil can influence the atmospheric CO_2 concentration (Batjes, 1996; Johnston et al., 2004; Martin et al., 2011; Schmidt et al., 2011; Lenka and Lal, 2013). Therefore, conservation (e.g., carbon sequestration) of this large soil OC pool is considered to be one of the most vital solutions to offset CO_2 emissions in controlling the concentration of atmospheric CO_2 (Pan et al., 2003; Conforti et al., 2017). Numerous studies have been conducted in terrestrial ecosystems to investigate the temporal and spatial variation of soil total organic carbon (TOC) storage at different scales (e.g., Pan et al., 2010; Arrouays et al., 2012; Stockmann et al., 2015) and the influencing factors (Jenny, 1980; Jackson et al., 2002; McGrath and Zhang, 2003; Reisser et al., 2016). These studies have demonstrated that TOC stocks and dynamics are related to atmospheric CO_2 concentration, while its soil fertility and quality are also linked to agricultural productivity as other important indicators (Stockmann et al., 2015).These factors can be categorised into natural, such as soil properties (i.e., soil texture, pH), climate (i.e., temperature and precipitation), soil parent materials, topography, and anthropogenic, including land use and management, human input (e.g., fertilisers) (Jenny, 1980; Post et al., 1982; Jackson et al., 2002; Stockmann et al., 2013; Wiesmeier et al., 2015). Although all these factors have effects on soil organic carbon contents in different regions, some of them generally follow similar spatial patterns (Wiesmeier et al., 2019). For example, a cold and wet environment is conducive to the accumulation of soil organic carbon, while high temperature and low rainfall may lead to low soil organic carbon storage at both regional (Baritz et al., 2010; Badgery et al., 2013) and continental scales (Jobággy and Jackson, 2000). As a result, TOC and pH have been found to maintain a generally negative correlation under natural conditions at various scales, a feature attributed to their innate internal relationship (Andersson and Nilsson, 2001; Reisser et al., 2016). Organic matter on decomposition releases organic acids, leading to lower soil pH values. In addition, relatively high pH values accelerate the decomposition of soil organic carbon, resulting in a decrease in TOC storage capacity (Andersson and Nilsson, 2001; McGrath and Zhang, 2003). The ability of soil to maintain and supply nutrients is closely related to its cation exchange capacity (CEC). However, the cation and anion exchange capacity are affected by soil pH values. TOC is an important indicator for evaluating soil fertility and nutrition. Therefore, systematic research on the internal relationship between TOC and pH is an important topic, which is helpful on the agricultural management.

Recently, Xu et al. (2019) identified a 'concealed pattern' with positive correlations between TOC and pH values in central-eastern Europe and related it to the coarse soil materials of last glacier deposit. This indicated the existence of different relationships between these two variables at the regional scales and does not follow the general negative relationship. However, the concept of 'spatially varying relationship' between TOC and pH values was not developed until we used a different method of GWR in this paper. The negative correlation between TOC and pH has been well recognized in the literature. It can be regarded as the original and innate relationship between TOC and pH value so most papers only have a general discussion on it (e.g. Andersson and Nilsson, 2001; Fabian et al., 2014; Reisser et al., 2016). As examples, we have chosen four references with quantitative statistical analysis on the relationship between them (McGrath and Zhang, 2003; Korkanç, 2014; Wang et al., 2016; Gebrehiwot et al., 2018). However, there are contradictory results of positive correlation reported in a limited number of two papers (Wang et al., 2010; Luo et al., 2017). The positive correlation is due to the combination of complex influencing factors, which deserves further investigations Meanwhile, the relationships can be different at different locations of sub-regions (de Moraes Sa et al., 2009; Xu et al., 2019) or different soil layers (Zhang et al., 2018). While the 'location' is considered, the concept of 'spatially varying relationship' has not been well recognized which is the focus of this study. Use of the local statistical methods such as geographically weighted regression (GWR) is helpful to quantify the varying relationship over space in an objective way, while the division of sub-regions in the existing literature was fairly arbitrary (see Table 1).

Based on the literature, the following research challenges and gaps exist:

1. Contradictory results of the relationship between TOC and pH value have been found in the literature, making the relationship an interesting topic for further investigation.
2. These contradictory results were obtained using global parameter of correlation analysis applied in arbitrarily divided sub-regions.
3. The reasons to the positive correlation remain complicated, which warrantees the need for further investigations in different study areas to explore the reasons.
4. No concept of 'spatially varying relationship between TOC and pH value' has been proposed in the literature.

This paper attempts to address these challenges and gaps using the local statistical method of GWR to analyse the TOC contents and pH values in European agricultural soil, extracted from the database of the project 'Geochemical Mapping of Agricultural and grazing land Soil' (GEMAS), which covers 33 European countries and an area of about 5,600,000 km^2 (Reimann et al., 2014a). As explained in Xu et al. (2019), TOC is affected not only by pH but by many other factors such as climate, soil type, and human activities. It needs to be acknowledged that besides pH value, other factors can be included in the GWR model. However, in this study, our focus is not modelling TOC, but revealing the spatially varying relationship between TOC and pH values. Thus, we have only chosen pH value with a focus on the assumption that the relationship between TOC and pH value is not always negative as commonly known, but spatially varying. In addition, it should be emphasised that the concepts of 'spatial heterogeneity' and 'spatially varying relationship' are different. 'Spatial heterogeneity' mostly refers to the variation of concentrations of chemicals over space in this context, while we are focusing on the 'relationship'. To our knowledge, this is the first research focusing on spatially varying relationships between TOC and pH values.

The objectives of this research are: (1) to study the spatially varying relationships between TOC contents and pH values in European agricultural soil; (2) to investigate the effects of different bandwidths in GWR for identifying different patterns of the spatially varying relationships, and (3) to explore the related influencing factors on the spatially varying relationships between TOC and pH values.

2. Methods

2.1. Soil sampling

The GEMAS agricultural soil data were used in this study. GEMAS is one of the largest European projects carried out by the Geochemistry Expert Group of EuroGeoSurveys (EGS) in collaboration with Eurometaux associated companies (Reimann et al., 2014a, 2014b). This project aimed at establishing an internally consistent soil geochemical data set at the European scale according to the specifications of the REACH regulation (EC, 2006). From summer 2008 to early 2009, a total of 2108 agricultural (Ap) soil samples were collected across 5.6 million square kilometres in 33 European countries at a sampling density of 1 sample/2500 km^2 (Reimann et al., 2014d, p.24). The REACH regulation specifies that agricultural soil samples should be taken at a depth of 0–20 cm (ECHA, 2012), which is the ordinary ploughing depth (Ap horizon). Each sample of about 3.0 kg weight was a composite from five sub-sites, taken from the four corners and centre of a 10×10 m square. Duplicate field samples were collected at every 20th site (EGS, 2008).

The maps of TOC contents and pH values in European agricultural soil samples are shown in Fig. 1. There are 13 missing TOC values in

Table 1
Overview of relationships between TOC and pH values in the past literatures.

Relationship	Author(s)	Method	Descriptions	Reasons
Negative correlation	McGrath and Zhang, 2003	Global Pearson (linear) correlation coefficient	Negative correlation between TOC and pH value in Ireland ($r = 0.17$)	Innate relationship: organic acids lead to low pH
	Korkanç, 2014	Global Pearson (linear) correlation coefficient	Negative correlation between TOC and pH value ($r = -0.274$)	Innate relationship: organic acids lead to low pH
	Wang et al., 2016	Global Redundancy analysis (RDA)	pH value negatively correlated with impact on soil organic carbon	Innate relationship: organic acids lead to low pH
	Gebrehiwot et al., 2018	Global Pearson (linear) correlation coefficient	($r = -0.126$) Weak negative correlation was observed between TOC and pH value	Innate relationship: organic acids lead to low pH
Positive correlation	Wang et al., 2010	Global Pearson (linear) correlation coefficient	Positive relationship between TOC and pH value in the study area ($r = 0.549$, $p = 0.000$).	Complex influences of soil bulk density and landscape, while the causative relationship between TOC and pH is complicated.
	Luo et al., 2017	Path model (i.e., structural equation model)	pH significantly and positively associated with TOC	Combination of soil properties (e.g. particle size, CEC, clay and silt) and climate factors
Contradictory relationships	de Moraes Sa et al., 2009	Global Pearson (linear) correlation coefficients	Changes in the relationships between TOC concentration and pH values in the tillage chronosequence	Tillage chronosequence is the key factor to influence the soil pH values, and thus the original negative relationship changes.
	Zhang et al., 2018	Liner regression model	Significant and positive correlation of soil pH value with storage of TOC was found for all soil layers except for 10–20 cm (negative).	Slope, climate, grazing intensity are main reasons for the contradictory relationships.
	Xu et al., 2019	Global Spearman correlation coefficients and Hot spot analysis (Getis-Ord Gi*)	Negative in European continent Positive in north-central Europe	Innate relationship of negative correlation at regional level, while local statistics observed positive correlation due to the influence of quartz in central Europe.

the GEMAS database, so the corresponding samples with pH values were excluded. Therefore, the total number of soil samples used in this study is 2095 (see Fig. 1). Due to the spatial scale issue, the number of samples can be endless. Our purpose is to investigate the spatial patterns at the continental-scale, and to discover if local and regional patterns emerge. While more samples are favourable, the GEMAS project's average sampling density at 1 sample/2500 km^2 is good enough to reveal spatial patterns at the continental-scale.

Total organic carbon contents are discussed in Baritz et al. (2014) and Matschullat et al. (2018), and pH values in Fabian et al. (2014). The detailed description of TOC and pH in agricultural soil samples can be found in the GEMAS project atlas (Reimann et al., 2014c).

2.2. Sample preparation and analysis

In the GEMAS project, to produce comparable data sets across national borders, all soil samples were prepared and analysed in the same laboratory for the same suite of determinants following a strict quality control procedure, which is described in Reimann et al. (2009, 2011) and Demetriades et al. (2014). Sample preparation involving

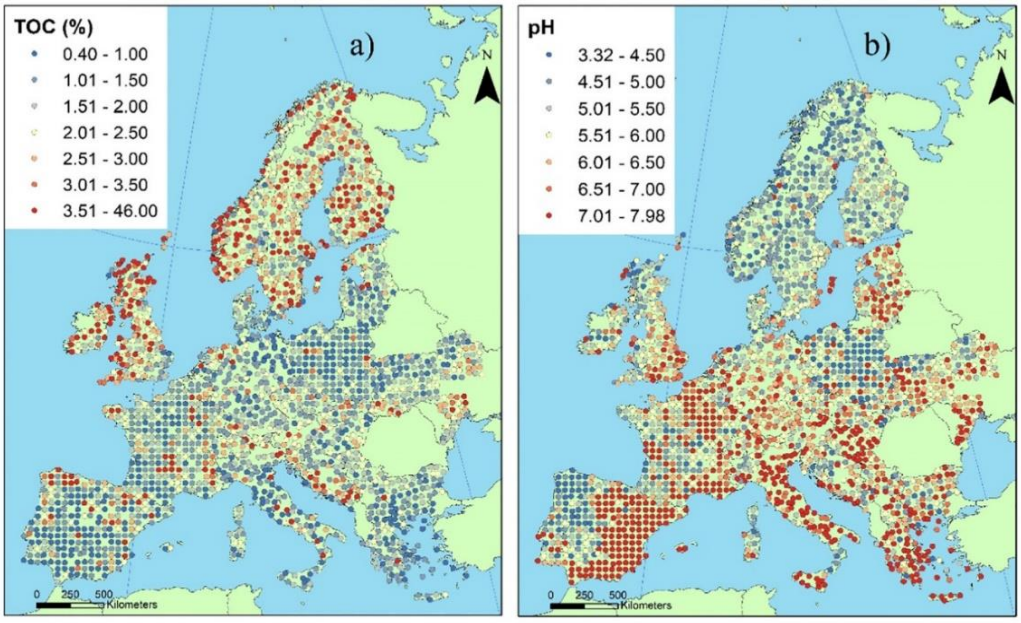


Fig. 1. Coloured symbol maps showing agricultural soil sampling locations, with (a) TOC contents and (b) pH values in European agricultural soil samples based on GEMAS project data ($n = 2095$).

air-drying, sieving through a 2 mm nylon sieve, homogenisation and splitting into 10 aliquots was performed at the Geological Survey of Slovakia (Mackových and Lučivjanský, 2014). Randomisation of samples and insertion of analytical replicates and project reference samples were carried out at the Geological Survey of Norway (NGU).

For the parameters used in this study, TOC was determined by the ISO standard 10,694 method (ISO, 1995) at FUGRO Consult GmbH (now KIWA) in Germany, and pH was measured in 0.01 M $CaCl_2$-solution at the Geological Survey of Norway (Reimann et al., 2011, 2014c).

2.3. Geographically weighted regression

Geographically weighted regression (GWR) is known as a powerful method for capturing the spatially varying relationships and exploring spatial non-stationarity since the 1990s (Brunsdon et al., 1996; Fotheringham et al., 2002). It has been widely used in various fields, including agricultural, urban (e.g., land use), social environmental and health studies (Kumar et al., 2012; Lu and Liu, 2016; Li et al., 2017; Margaritis and Kang, 2017; Feuillet et al., 2018). This approach is an extension of ordinary regression, and is used to reveal spatial relationships between the dependent and independent variable(s) at the local level (Fotheringham et al., 2001). The routine generates a set of regression parameters at the local level that reveal how the relationship between the input variables change over space (Fotheringham et al., 2002). The spatial patterns of the local parameters are used to further investigate the possible factors governing the varying relationships that are not identified by the traditional regression model, such as Ordinary Least Square (OLS). Traditional regression estimates the global statistic that assumes the relationship studied is linear and spatially constant, so the estimated parameters remain the same for the whole study area (Tu and Xia, 2008). It estimates local regression coefficients at each sample site by measuring the 'features' locally, allowing the parameter estimation between the dependent and independent variable(s) to vary concurrently at each location (Fotheringham et al., 2002; Kumar et al., 2012). Therefore, GWR can explore the spatially varying relationships between variables by including the spatial coordinates of each sample site, which are ignored in the traditional linear regression modelling. In this study, TOC is used as the dependent variable and pH as the sole independent variable. The traditional OLS equation is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \quad (1)$$

where y_i is the value of the dependent variable (TOC) at the ith location, x_{i1} is the value of independent variable (pH) at the ith location, β_0 is the intercept on the y-axis, β_1 is the regression coefficient that is estimated for the independent variable at location i, and ε_i is the error term.

Based on Fotheringham et al. (2002), the GWR model can estimate local coefficients rather than global ones by adding the geographical location in the function, which is expressed as:

$$y_i = \beta_0(\mu_i, v_i) + \beta_1(\mu_i, v_i)x_{i1} + \varepsilon_i \quad (2)$$

where (μ_i, v_i) represent the coordinates for sample location i, $\beta_0(\mu_i, v_i)$ is the intercept for location i, and $\beta_1(\mu_i, v_i)$ is the local regression coefficient for the independent variable (pH) at location i.

In contrast with the ordinary regression model, the local regression coefficients in GWR can be estimated by using a weighted function (Fotheringham et al., 2002), as expressed by the following equation:

$$\hat{\beta}(\mu_i, v_i) = \left(X^T W(\mu_i, v_i)X\right)^{-1} X^T W(\mu_i, v_i)Y \quad (3)$$

where X is the matrix formed by the values of the independent variable x (pH); Y is the corresponding matrix generated by the values of the dependent variable y (TOC); $W(\mu_i, v_i)$ represents the weight matrix chosen to ensure that observations closer to the specific location (μ_i, v_i) have greater influence on the final result.

The adaptive kernel type was selected as the weight function because it can reduce the 'border effect' when sample sites are located near to coastal or country border areas (Zhang et al., 2011), which is suitable for this study area. It needs to be mentioned that while the term 'bandwidth' is widely used in GWR, it is more suitable to use the term 'area of influence' in the two-dimensional geographical space. However, for the sake of consistency, the term 'bandwidth' is used in this paper. There are two types of bandwidth in the GWR model, one is the spatial distance and the other is the distance between the neighbours of the sample points. By applying the adaptive kernel function, the latter bandwidth was selected for this study by ascribing a weight of '1' for sample sites within the selected bandwidth and '0' for sample sites outside. The bandwidth was chosen by using the Akaike Information Criterion (AIC), which is effective in finding the 'optimal' distance band in the GWR model (Fotheringham et al., 2002). This method can calculate the most suitable bandwidth for the model by weighing the relationship between the goodness-of-fit and the simplicity of the model (Akaike, 1974), which is widely applied in statistical inference. Regarding technical details, there is no consensus on the choice of the 'optimal' bandwidth to define the 'local' range of influence, which is an important input parameter in the GWR model. This has been extensively debated in the literature (e.g., Farber and Paez, 2007; Guo et al., 2008). The results vary by selecting different bandwidths and spatial weights. The GWR technique captures the spatially varying relationship by fitting a local regression model at each sample location, weighing the values of neighbouring sampling sites by a function of distance band from that particular sample location. The neighbours closer to the sample location have a stronger influence on the model results than the observations that are farther away (Fotheringham et al., 2002). In other words, with larger bandwidth, the GWR approach tends to reach a global regression, and the spatial patterns of the estimated parameters become larger and smoother across the study area. Due to its importance on the estimation of local coefficients, eight different bandwidths (with the number of neighbours being 25; 50; 75; 100; 125; 150; 200; 250) were investigated in this research.

In order to discuss the correlation between the dependent (TOC) and independent (pH) variables, a local correlation coefficient (r) was calculated, using the following equation:

$$Coefficient\ (r) = \sqrt{R^2_{local}} \times {}^{C}/_{|C|} \quad (4)$$

where R^2_{local} is the deterministic coefficient R^2 from the GWR model, indicating how well the variation of TOC contents can be explained by pH values (ranging from 0 to 1), and C is the local regression coefficient which is the same as $\beta_1(\mu_i, v_i)$ in Eq. (3).

The original results of local regression coefficients only represent the slope coefficients (Gao and Li, 2011). The higher values of regression coefficients demonstrate that changes of soil pH values can lead to greater changes of TOC contents. However, by calculating the local correlation coefficient (r), strong and weak correlations can be presented between dependent and independent variables. Eq. (4) is equivalent to Geographically Weighted Pearson Correlation Coefficient (GWPCC), which is a statistic based on geographical location weighting moments, adopting the concept of geographical weights around samples for calculating local statistics (Kalogirou, 2014). The core concept of GWPCC is the same with that of GWR (Fotheringham et al., 2002), and it can provide a significance test of local coefficients. By controlling the same input parameters with the GWR model, the significance test can indicate whether the spatially varying relationships at the local (site-specific) level are significant or not.

2.4. Data transformation and software

It is well-known that geochemical data do not belong to the classical Euclidean space and should be considered in their own Euclidean

geometry on the simplex (Aitchison, 1986; Filzmoser et al., 2009, 2010, 2014; Egozcue and Pawlowsky-Glahn, 2011; Reimann et al., 2012). However, classical statistical methods are still used in the treatment of geochemical data because they provide interpretable patterns. So, in this case, all data sets (TOC and pH) were subjected to a normal score transformation in SPSS (ver. 24) to deal with 'non-normality' of the data and to reduce the effects of potential outliers (Zhang et al., 2008). Geographically weighted regression and local correlation coefficients (r) were estimated in ArcGIS (ver. 10.4), while the significance test was calculated in the R package lctools (ver. 3.56, in http://cran.r-project.org/web/packages/lctools/index.html). All statistics were estimated in SPSS (ver. 24) and Microsoft Excel (ver. 2016), and all maps produced in ArcGIS (ver. 10.4).

3. Results and discussion

3.1. Basic statistics and background of TOC and pH in European agricultural soil

The basic statistics for TOC and pH in European agricultural soil can be found in the Table 1 in Xu et al. (2019). The median values of TOC and pH are 1.80 (wt%) and 5.77, respectively, which are close to those of the FOREGS data set (De Vos et al., 2006). The large difference between the minimum (0.40 wt%) and maximum (46.0 wt%) for TOC contents indicates the strong variation across the study area. For soil pH values, the low median value (5.77) indicates that agricultural soil is generally acidic at the European scale. More details of the two variables across different countries can be found in the boxplot comparison in the GEMAS atlas (see Reimann et al., 2014a, p.131, 190). Generally, higher TOC contents with comparatively lower pH values are mainly concentrated in northern Europe (i.e., Fennoscandia, the United Kingdom and Ireland). The majority of samples with lower TOC contents and relatively higher pH values occur in southern Europe (see Fig. 1).

The Quantile-Quantile (Q-Q) plot (Fig. 2) displays the expected values of normal distribution against the actual values for TOC and pH. The Q-Q plot shows whether the variables follow the normal distribution (Wilk and Gnanadesikan, 1968). If the values are normally distributed, the points should cluster near to a straight line on the plot, which is not the case for the studied variables. The long tail towards higher TOC contents indicates the existence of potential outliers (extremely high values) in the data set (Fig. 2a). Due to the non-normality and existence of potential outliers, data transformation is widely applied prior to use of GWR (e.g. Fotheringham et al., 2002; Zhang et al., 2011; Joseph et al., 2012; Yuan et al., 2020). To limit the influences of potential outliers and to satisfy the normality requirement of GWR, normal score transformation (NST) was applied to the TOC data, as NST is an effective statistical data transformation for such a purpose (Fotheringham et al., 2002; ESRI, 2016). In Fig. 2b, the pH values show that values between 4.0 and 6.0 are close to the normal distribution line. This is because the concentrations of H^+ have already been logarithmically transformed to obtain the pH values. Prior to the analysis of this study, the pH values were also transformed to a normal distribution to maintain consistency and to achieve a near symmetrical distribution.

A Spearman's rank correlation coefficient was calculated between TOC and pH; the result shows a weak negative correlation ($r = -0.19$, $p < 0.01$). In fact, these two variables generally follow a weak negative correlation at the European continental scale. Soil with higher organic carbon (OC) content generally contains higher organic matter (OM), which secretes organic acids, making the soil acidic and, thus, resulting in a lower pH value (Fabian et al., 2014). However, due to the complex factors affecting each sampling site, the relationship may be interfered at the local scale.

3.2. Spatially varying relationships between TOC and pH in European agricultural soil

As motioned earlier, bandwidth is regarded as an important parameter in applying GWR (Guo et al., 2008). In this study, the Akaike Information Criterion (AIC) was applied to find the 'optimal' bandwidth at the European scale. Based on this 'optimal' bandwidth, seven additional bandwidths were used for comparison. The results of local regression coefficients between TOC and pH, using eight different bandwidths in the GWR model ($n = 25$; 50; 75; 100; 125; 150; 200 and 250), are shown in Fig. 3.

Spatial variation between TOC and pH is observed among all bandwidths (see Fig. 3), suggesting that spatially varying relationships exist between the two variables in the study area. When using the optimal distance band calculated by the AIC (n = 75, see Fig. 3c), patterns showing different degrees of spatial variation are observed. The spatial patterns can be divided into three groups: northern, central-eastern and southern Europe. Except for the central part of Sweden and Norway, there appears to be a negative correlation between TOC contents and pH values in agricultural soil in northern Europe, with original negative relationships between the two variables maintained. The central-eastern part of Europe is the most noteworthy, with only positive coefficients being observed. Especially in Ukraine, Poland and eastern Germany, the local coefficients are relatively high, ranging from 0.27 to 0.72, suggesting that changes of pH values are associated with rapid changes of TOC contents. Relatively mixed relationships are found in southern Europe, with the local coefficient showing a significant variation. Except for the positive regression coefficients in central France, north-west Italy, Switzerland and north-eastern Bulgaria, the

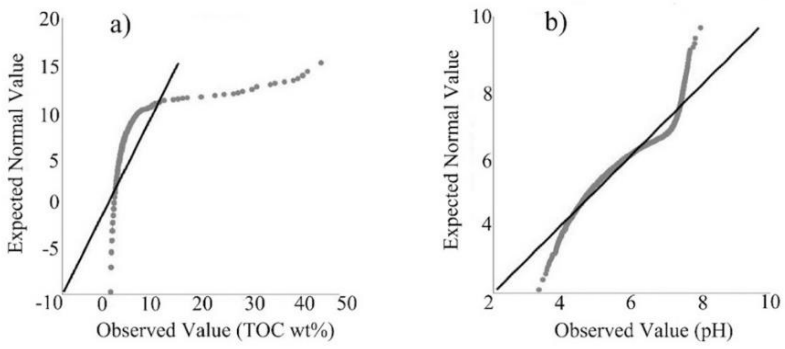


Fig. 2. Q-Q plots of TOC contents and pH values in European agricultural soils: raw data of (a) TOC (%), and (b) pH. Black line shows the expected normal distribution, and the grey dots the actual sample values.

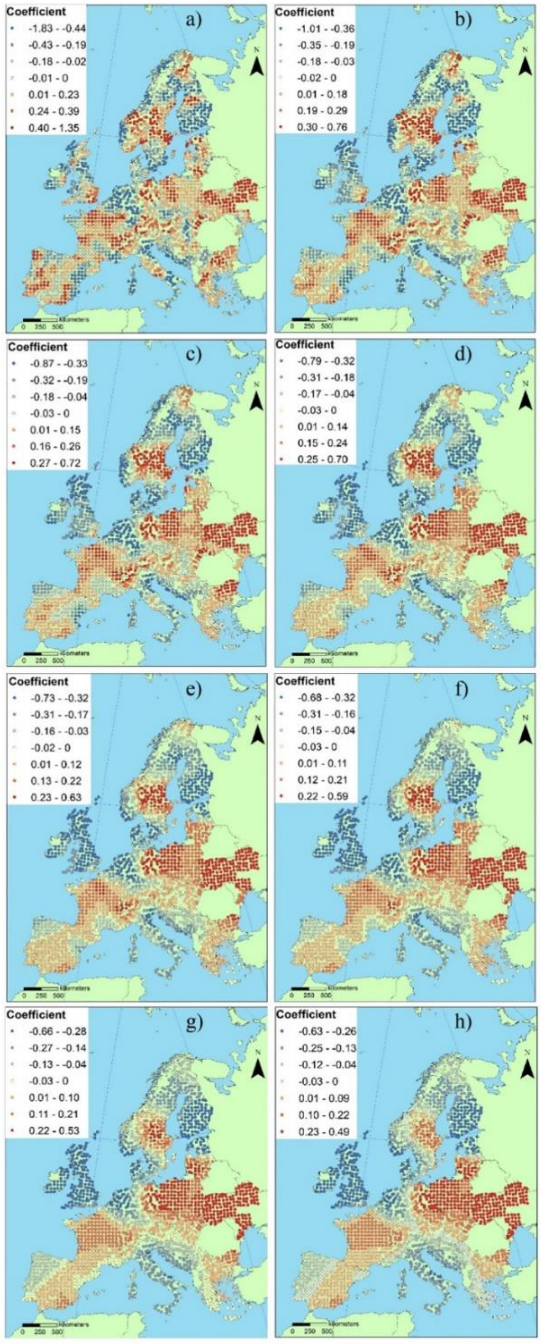


Fig. 3. Local regression coefficient maps between TOC and pH using different number of neighbours (bandwidths): a) n = 25; b) n = 50; c) n = 75; d) n= 100; e) n = 125; f) n = 150; g) n = 200; h) n = 250.

relationship in other regions is generally weak positive to negative. These patterns suggest that soil TOC contents and pH values in agricultural soil in southern Europe may be influenced by other factors, such as climate and, possibly, agricultural activities. This implies that the relationship between TOC and pH is greatly interfered, even with changed relationships from generally negative to positive direction.

The local regression coefficient of the smaller bandwidths (n = 25, 50) is estimated from the values of neighbouring sample sites, which are closer to the sample location for which the local regression coefficient is calculated, leading to results that change rapidly across the whole area (Foody, 2004; Bickford and Laffan, 2006). This is reflected in the increased total number of local regression coefficients that change from negative to positive values (see Table 2). However, comparing these results with those of the optimal bandwidth (AIC, n = 75), the increased number of high positive local regression coefficients only appear in the United Kingdom (UK) and Ireland, as well as in the northernmost parts of Finland and Norway. With the larger bandwidths (n = 100, 125, 150 and 250), the variation of the local regression coefficients becomes smaller and some variation is less visible on the maps (e.g., south-western UK, central part of Italy and Estonia). This is due to the fact that the larger bandwidths tend to make GWR a relatively more global model (Fotheringham et al., 2002; Guo et al., 2008).

There are no standard criteria for selecting the 'best' bandwidth in GWR. By comparing multiple bandwidths, the spatially varying relationships between TOC and pH can be effectively revealed at different scales. The statistics of performance for GWR results are summarised in Table 3, including values for AICc, R^2 and adjust R^2. Comparing the optimal bandwidth with the other distance bands, the model appears to perform better with the 75-neighbour bandwidth because it has relatively higher R^2 and lowest AICc values. Highest R^2 and adjust R^2 were found in 25-neighbour bandwidth. However, AICc value is the preferred measurement to compare GWR models (ESRI, 2016), while lower AICc value means better performance of the GWR model. The second lowest AICc value is found in 125-neighbour bandwidth (see Table 3). In fact, this bandwidth (Fig. 3e) reveals large and continuous spatially varying relationships that divide the whole continent into three zones (i.e., northern, central-eastern and southern Europe). With increasing bandwidth (number of neighbours), large spatial patterns become visible and local variations disappear or are subdued. Results from smaller bandwidths tended to be noisy, thus we focused on the distance band of n = 125 which showed large patterns while revealed sufficient details for capturing the spatially varying relationships between TOC and pH values.

3.3. Spatially varying relationships between TOC and pH

3.3.1. Explanation of the results in GWR

Mapping the local regression coefficient, local R^2, local correlation coefficient (r), significance and standardized residuals, provide an efficient and direct approach to investigate the spatially varying relationships between TOC and pH values. Both positive and negative local regression coefficients are observed (see Figs. 3e and 4a), and the percentage of positive values account for 52.2% of the variation (see Table 2), suggesting that the relationship between TOC and pH is perturbed over more than half of the European continent.

Strong positive correlations are observed in central-eastern Europe (i.e., north-eastern Germany, Poland and Ukraine), central part of Norway and Sweden, central France and north-eastern Bulgaria. Strong negative correlations are mainly found in Ireland, the UK, southern Finland, Estonia, Denmark, The Netherlands, Belgium, Luxembourg, south-western Germany, Croatia, Bosnia and Herzegovina, and Italy (Fig. 4a). Some weak positive correlations are observed in northernmost Finland (except Estonia), Latvia, Mediterranean areas, Czech Republic, Slovakia, Austria, Hungary, Switzerland, Spain and Hellas. Weak negative correlations occur in northern Sweden and Finland, western

Table 2
Summary statistics for GWR output of local regression coefficients at different bandwidths, and the percentage of negative and positive coefficients.

Bandwidth (n)[a]	Min.	Q25	Median	Q75	Max.	% of negative values	% of positive values
25	−1.83	−0.25	0.04	0.28	1.35	44.9	55.1
50	−1.01	−0.23	0.03	0.21	0.76	46.9%	53.1
75	−0.87	−0.22	0.02	0.17	0.72	47.8	52.2
100	−0.79	−0.21	0.01	0.15	0.70	48.3	51.7
125	−0.73	−0.20	0.01	0.15	0.63	47.8	52.2
150	−0.68	−0.19	0	0.14	0.59	48.9	51.1
200	−0.66	−0.18	−0.01	0.12	0.53	52.1	47.9
250	−0.63	−0.17	0.01	0.11	0.49	53.6	46.4

[a] n: number of neighbours.

Norway, north-western Iberia, Serbia, central and southern Italy, and Sicily.

Higher local R^2 and significant ($p < 0.05$) positive correlation is only observed in central Sweden and Norway, central-eastern Europe, north-central France and north-eastern Bulgaria (Figs. 4b, c). This suggests that the spatially varying relationship in these agricultural soil samples is more significant, and the variation of TOC contents can be better explained by pH values than in other areas. In contrast, significant negative correlations with higher local R^2 (Figs. 4a, c) occur in Ireland, UK, northern Sweden, southern Finland, south-western Norway, Denmark, The Netherlands, Belgium, Luxembourg, south-western Germany, north-western Iberia, Croatia, Bosnia and Herzegovina, Italy, Sicily, Sardinia and Corsica. However, the spatial relationship between TOC and pH in the remaining areas cannot be explained well in GWR ($p > 0.05$). Therefore, the spatial distribution of the GWR model results can effectively reveal the spatially varying relationships between TOC and pH in European agricultural soil at the local scale. However, it is acknowledged that the R^2 values are still low, and they are related to multiple influencing factors. Within our expectation, there are large areas of non-significant findings, due to the complicated influencing factors. The non-significant results are also a part of the 'spatially varying relationships' found in this study, which are equally important as the 'significant' results. The map for standardized residuals is illustrated in Fig. 4d, which did not show obvious spatial patterns. Also, the spatial autocorrelation test (Moran's I) on the standardized residuals indicated the significant results ($p < 0.05$, z-score = 2.99). However, it needs to be recognized that the significance level is related to the large sample size, and thus should be carefully interpreted (e.g. Cornfield, 1966; Gingerich, 1995; Zhang et al., 2005). More importantly, the purpose of this study is not to model TOC, but to reveal the relationship between TOC and pH value only.

3.3.2. Potential factors influencing the spatially varying relationships

It is acknowledged that it is challenging to explore the causal effects between TOC contents and pH values in European agricultural soil, which it is the universal issue for all the statistical methods: relationship does not mean causal effects. However, based on the association of the patterns of spatially varying relationships in different areas, the potential influencing factors can be explored. In addition, the Spearman correlation coefficient analysis of relevant environmental factors between TOC and pH was conducted in our previous study (see Table 2 in Xu et al., 2019). Based on this, the spatially varying relationships between TOC and pH can be discussed quantitatively, although not in depth.

Table 3
Summary of the performance statistics of eight bandwidths for GWR results.

Bandwidth (n)	AICc	R^2	Adjust R^2
25	5192.17	0.58	0.43
50	5053.85	0.48	0.39
AIC (75)	5039.47	0.46	0.42
100	5048.57	0.44	0.39
125	5036.14	0.42	0.38
150	5079.80	0.38	0.35
200	5111.82	0.36	0.34
250	5142.41	0.34	0.33

The spatial distribution of TOC is strongly influenced by environmental factors, like pH, potential hazardous and nutrient elements (Khaledian et al., 2017; Wiesmeier et al., 2019). It is reported that the impacts of chemical elements and physicochemical parameters in European agricultural soil are dominated by natural conditions (e.g., Fabian et al., 2014; Matschullat et al., 2018; Négrel et al., 2019). Agricultural soil in northern Europe shows a large area of negative correlation between TOC and pH, except for central Sweden and Norway (Fig. 4a). This can be attributed to the acidic soil under natural conditions in Fennoscandia, with higher TOC contents in the long-term cold and wet environment.

In central-eastern European countries where the positive relationships between TOC and pH values are clearly observed, quartz-rich soil tends to show strong positive correlation across this large and continuous area (see Fig. 1 and Reimann et al., 2014c, p. 98, 193). Similar positive correlations are also found in other areas (i.e., Sweden, Norway and France; see Figs. 3d and 4a). This is due to the high concentration of SiO_2 in the coarse-grained sediments from the last ice age (Piotrowski et al., 2006). To further explore this feature, the spatial relationships between TOC and SiO_2 at the local level were analysed using the GWR model and the results are shown in Fig. 5. Only negative local regression coefficients and correlation are observed (Figs. 5a, b), indicating the spatially 'stationary' relationships that exist between these two variables. It is worth noting that the local coefficients of all sample points are significant ($p < 0.05$), suggesting the strong and relatively spatially stable relationship between SiO_2 and TOC. Relatively stronger correlation and higher R^2 values are clustered in northern and central Europe. However, the overall values of local R^2 (0–0.68) in this model are much higher than those in the model between TOC and pH (see Fig. 5c), suggesting that their relationships are strongly correlated, and the concentration of silica can largely explain the variation of TOC contents. In other areas, lower quartz contents may also play a role in the spatially varying relationships between TOC and pH. Thus, the spatial variations are mixed and complicated (see Fig. 4c).

The glacial sediments in central-eastern Europe are almost entirely composed of quartz (SiO_2) and some feldspars, featuring large particle size and good permeability. The soil parent material (PM) and the concentration of quartz are vital factors associated with soil organic carbon (SOC) contents (Badgery et al., 2013). Soil formed on these coarse-

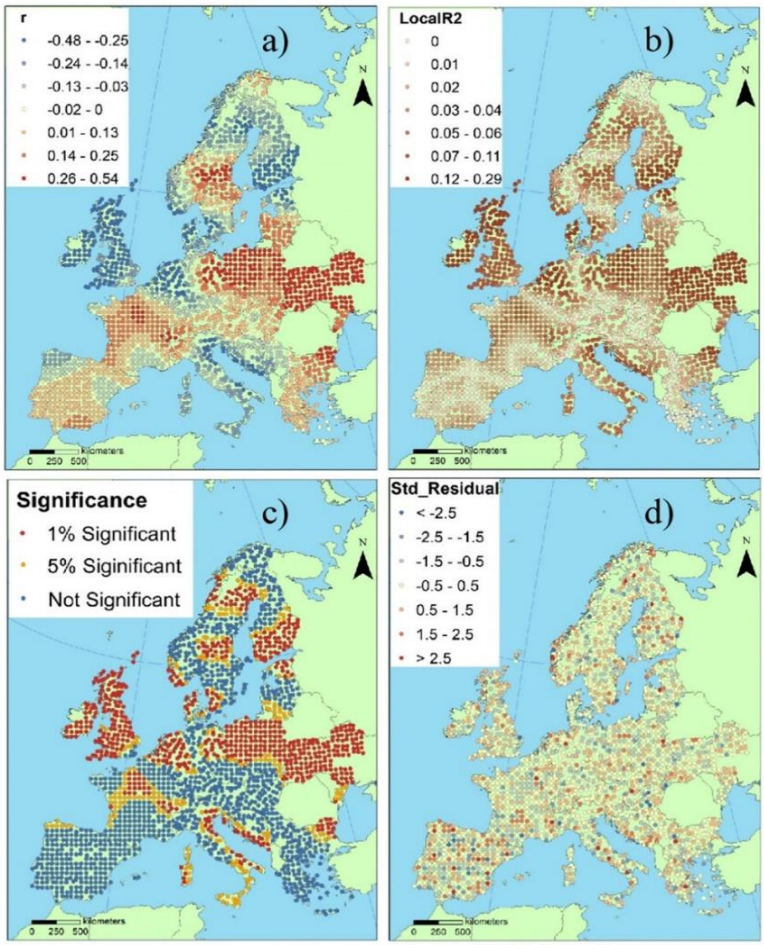


Fig. 4. Spatial variation of GWR regression outputs using 125 number of neighbours: a) local correlation coefficient (r); b) Local R^2; c) Significance, and d) Standardized residuals.

textured glacial deposits contain a larger proportion of quartz with larger particle sizes, which contribute to lower soil organic carbon stock due to the lower aggregation stability in this sandy soil (Kern, 1994; Le Bissonnais and Arrouays, 1997; Homann et al., 1998; Wiesmeier et al., 2019), resulting in a significant decrease in TOC content. In addition, the low soil pH values can also be attributed to these quartz-rich soils. Due to the coarse-grained particle sizes, these soils contain little Ca^{2+} to buffer the soil pH (Fabian et al., 2014), leading to lower pH values of east-central European agricultural soil. Therefore, the overall negative relationship between TOC and pH is locally disturbed. Both quartz and clay are significantly related to TOC, while the absolute value of correlation coefficient with quart ($r = -0.379$) is higher than clay ($r = 0.202$), highlighting the main control of quartz on the positive relationship between TOC and pH in the central-eastern Europe. When the particle size of quartz-rich soil is large, it is not tended to favour the TOC storage.

In southern Europe, the spatial relationship between soil TOC contents and pH values becomes mixed and could not completely correspond to natural factors. On the other hand, anthropogenic inputs, such as lime, fertilisers and tillage management may play some roles in the mixed areas. Due to the special lower TOC contents in southern Europe, crops need to rely on external inputs to sustain growth and nutrition. Such practices, however, change the soil properties, resulting in the overall negative correlation to be interfered at the local level.

4. Conclusions

This study investigated the spatially varying relationships between soil TOC contents and pH values across the European continent by using the GEMAS project agricultural soil data. The results confirmed that the relationships between TOC and pH are spatially varying in European agricultural soil samples at the local level, whereas both negative and positive correlations are identified using the GWR technique. Negative correlations are mainly observed in northern Europe and significant positive correlations are clustered in central-eastern areas, while comparatively mixed relationships between TOC and pH occur in southern Europe. The positive correlation between these two variables in central-eastern Europe is attributed to natural factors, i.e., these areas have low pH values, quartz-rich soil (i.e., high concentration of SiO_2), resulting in low TOC contents. Use of different bandwidths can also affect the GWR spatial statistical results, while the bandwidth of 125 neighbours (based on adaptive kernel type) is apparently the most

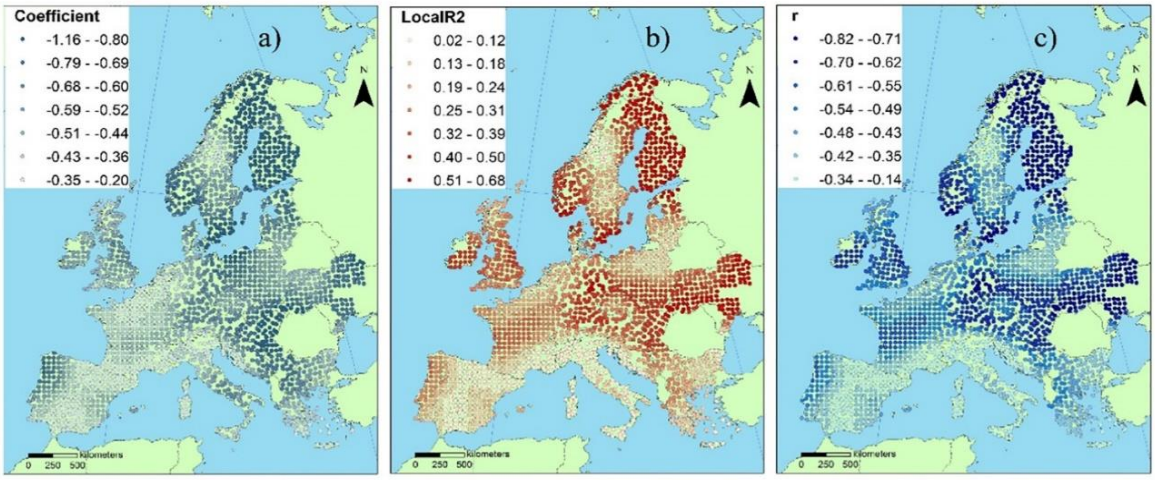


Fig. 5. Spatial relationships between TOC and SiO_2 calculated in GWR using 125 number of neighbours: a) local regression coefficient; b) local R^2 and c) correlation coefficient (r).

suitable, as shown by this study, because it provides a continuous and smooth pattern across the European continent.

The main scientific contributions of this research including: (1) The introduction and proof of 'spatially varying relationship between TOC and pH' provide added value and clarification to the understanding of the controversy of their complicated relationship in the literature; (2) The effective method of using the local statistics GWR has provided a solution to answer the controversy, and such a way of thinking can be expanded to other study areas and other relationships. Based on our results, it can be concluded that the spatially varying relationship between TOC contents and pH values can be mapped by the GWR technique, suggesting that it is an efficient and powerful tool to explore the spatial variations, and provides a new approach to identify potential influencing factors. The local statistical method provides an effective approach to explore the research problem of complex relationship between TOC and pH value, and the novel finding of the 'spatially varying relationship between TOC and pH value' in European agricultural soils enriches the existing literature. While GWR is widely used to explore the spatially varying relationship, it is always a challenge to link it to actual scientific problems and to interpret the results. In this study, we have limited our focus on the relationships between TOC and pH value, not the modelling of TOC which could be further explored in the future by considering more factors.

CRediT authorship contribution statement

Haofan Xu: PhD students, Literature review, Data analysis, Computer software implementation, Data analysis, Production of maps, Figure and Tables, Analysis of results, Writing - Original draft preparation.

Chaosheng Zhang: Supervision, Research design, Data source collection, Overall methodology design, Analysis of results, Writing - Review & Editing.

Declaration of competing interest

The authors declare that there is no actual or potential conflict of interest in this study.

Acknowledgements

The GEMAS project is a collaborative project between the EuroGeoSurveys Geochemical Expert Group and many external organizations (e.g., Alterra, Netherlands; Norwegian Forest and Landscape Institute; Research Group Swiss Soil Monitoring Network, Swiss Research Station Agroscope Reckenholz-Tänikon, several Ministries of the Environment and University Departments of Geosciences Chemistry and Mathematics in a number of European countries and New Zealand; ARCHE Consulting in Belgium; CSIRO Land and Water in Adelaide, Australia) and Eurometaux. The analytical process was co-financed by Eurometaux, Cobalt Development Institute (CDI), European Copper Institute, Nickel Institute, Europe, European Precious Metals Federation, International Antimony Association, International Manganese Institute, International Molybdenum Association, ITRI Ltd. (on behalf of the REACH Tin Metal Consortium), International Zinc Association, International Lead Association-Europe, European Borates Association, the (REACH) Vanadium Consortium and the (REACH) Selenium and Tellurium Consortium. We acknowledge the assistance of the GEMAS Project Team: S. Albanese, M. Andersson, R. Baritz, M.J. Batista, A. Bel-lan, M. Birke, D. Cicchella, A. Demetriades, B. De Vivo, W. De Vos, E. Dinelli, M. Ďuriš, A. Dusza-Dobek, M. Eklund, V. Ernstsen, P. Filzmoser, B. Flem, D.M.A. Flight, S. Forrester, M. Fuchs, U. Fügedi, A. Gilucis, M. Gosar, V. Gregorauskiene, W. De Groot, A. Gulan, J. Halamić, E. Haslinger, P. Hayoz, R. Hoffmann, J. Hoogewerff, H. Hrvatovic, S. Husnjak, L. Janik, G. Jordan, J. Kirby, V. Klos, F. Krone, P. Kwecko, L. Kuti, A. Ladenberger, A. Lima, J. Locutura, P. Lucivjansky, A. Mann, D. Mackovych, M. McLaughlin, B.I. Malyuk, R. Maquil, J. Matschullat, R.G. Meuli, G. Mol, P. Négrel, P. O'Connor, K. Oorts, A. Pasieczna, V. Petersell, S. Pfleiderer, M. Poňavič, C. Prazeres, U. Rauch, S. Radusinović, C. Reimann, M. Sadeghi, I. Salpeteur, R. Scanlon, A. Schedl, A. Scheib, I. Schoeters, E. Sellersjö, I. Slaninka, J.M. Soriano-Disla, A. Šorša, R. Svrkota, T. Stafilov, T. Tarvainen, V. Trendavilov, P. Valera, V. Verougstraete, D. Vidojević, A. Zissimos, Z. Zomeni.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London (416 pp).

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.

Andersson, S., Nilsson, S.I., 2001. Influence of pH and temperature on microbial activity, substrate availability of soil-solution bacteria and leaching of dissolved organic carbon in a mor humus. Soil Biol. Biochem. 33, 1181–1191.

Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Martin, M.P., Bellamy, P.H., Lark, R.M., Kibblewhite, M., 2012. Generic issues on broad-scale soil monitoring schemes: a review. Pedosphere 22, 456–469.

Badgery, W.B., Simmons, A.T., Murphy, B.M., Rawson, A., Andersson, K.O., Lonergan, V.E., van de Ven, R., 2013. Relationship between environmental and land-use variables

on soil carbon levels at the regional scale in central New South Wales, Australia. Soil Res 51 (7–8), 645–656.
Baritz, R., Seufert, G., Montanarella, L., Van Ranst, E., 2010. Carbon concentrations and stocks in forest soils of Europe. For. Ecol. Manag. 260 (3), 262–277.
Baritz, R., Ernstsen, V., Zirlewagen, D., 2014. Carbon concentrations in European agricultural and grazing land soil. Chapter 6. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils–Part B: General Background Information and Further Analysis of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 117–129.
Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. Eur. J. Soil Sci. 47, 151–163.
Bickford, S.A., Laffan, S.W., 2006. Multi-extent analysis of the relationship between pteridophyte species richness and climate. Glob. Ecol. Biogeogr. 15, 588–601. https://doi.org/10.1111/j.1466-8238.2006.00250.x.
Brunsdon, C., Fotheringham, A.S., Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geogr. Anal. 28, 281–298.
Conforti, M., Matteucci, G., Buttafuoco, G., 2017. Organic carbon and total nitrogen topsoil stocks, biogenetic natural reserve 'Marchesale' (Calabria region, southern Italy). J. Maps 13 (2), 91–99.
Cornfield, J., 1966. Sequential trials, sequential analysis and the likelihood principle. Am. Statistic. 20, 18–23.
Taylor, H., 2006. In: De Vos, W., Tarvainen, T., Salminen, R., Reeder, S., De Vivo, B., Demetriades, A., Pirc, S., Batista, M.J., Marsina, K., Ottesen, R.T., O'Connor, P.J., Bidovec, M., Lima, A., Siewers, U., Smith, B., Shaw, R., Salpeteur, I., Gregorauskiene, V., Halamic, J., Slaninka, I., Lax, K., Gravesen, P., Birke, M., Breward, N., Ander, E.L., Jordan, G., Duris, M., Klein, P., Locutura, J., Bel-lan, A., Pasieczna, A., Lis, J., Mazreku, A., Gilucis, A., Heitzmann, P., Klaver, G., Petersell, V. (Eds.), Geochemical Atlas of Europe. Part 2 - Interpretation of Geochemical Maps, Additional Tables, Figures, Maps, and Related Publications. Geological Survey of Finland, Espoo chief-editors. 692 pp. http://weppi.gtk.fi/publ/foregsatlas/.
Demetriades, A., Reimann, C., Filzmoser, P., 2014. Evaluation of GEMAS project quality control results. Chapter 6. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils – Part a: Methodology and Interpretation of the GEMAS Data Set Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 47–60.
EC, 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Off. J. Eur. Communities L396, 1–849 30.12.2006.
ECHA, 2012. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.16: Environmental Exposure Assessment. Version 2.1. European Chemicals Agency (147 pp).
Egozcue, J.J., Pawlowsky-Glahn, V., 2011. Basic concepts and procedures. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. Wiley, Chichester, pp. 12–28.
EGS, 2008. EuroGeoSurveys Geochemistry Working Group. EuroGeoSurveys Geochemical Mapping of Agricultural and Grazing Land in Europe (GEMAS) - Field Manual. Norges Geologiske Undersøkelse Report. 2008.038, 46 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2008/2008_038.pdf.
ESRI, 2016. Normal score transformation.Available at. https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/normal-score-transformation.htm Accessed date: 21/04/2020.
Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E., 2014. GEMAS: spatial distribution of the pH of European agricultural and grazing land soil. Appl. Geochem. 48, 207–216.
Farber, S., Paez, A., 2007. A systematic investigation of crossvalidation in GWR model estimation: empirical analysis and Monte Carlo simulations. J. Geogr. Syst. 9, 371–396. https://doi.org/10.1007/s10109-007-0051-3.
Feuillet, T., Commenges, H., Menai, M., Salze, P., Perchoux, C., Reuillon, R., Kesse-Guyot, E., Enaux, C., Nazare, J.A., Hercberg, S., Simon, C., Charreire, H., Oppert, J.M., 2018. A massive geographically weighted regression model of walking-environment relationships. J. Transp. Geogr. 68, 118–129.
Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data - problems and possibilities. Sci. Total Environ. 407, 6100–6108.
Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Sci. Total Environ. 408, 4230–4238.
Filzmoser, P., Reimann, C., Birke, M., 2014. Univariate data analysis and mapping. Chapter 8. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P. (Eds.), Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 67–81.
Foody, G.M., 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. Glob. Ecol. Biogeogr. 13, 315–320. https://doi.org/10.1111/j.1466-822X.2004.00097.x.
Fotheringham, A.S., Charlton, M.E., Brunsdon, C., 2001. Spatial variations in school performance: a local analysis using geographically weighted regression. Geogr. Environ. Model. 5, 43–66.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons Ltd, Chichester, UK (284 pp).
Gao, J.B., Li, S.C., 2011. Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using Geographically Weighted Regression. Appl. Geogr. 31, 292–302.
Gebrehiwot, K., Desalegn, T., Woldu, Z., Demissew, S., Teferi, E., 2018. Soil organic carbon stock in Abune Yosef afroalpine and sub-afroalpine vegetation, northern Ethiopia. Ecol. Process. 7 (1), 6.
Gingerich, P.D., 1995. Statistical power of EDF tests of normality and the sample size required to distinguish geometric-normal (lognormal) from arithmetic-normal distributions of low variability. J. Theor. Biol. 173, 125–136.
Guo, L., Ma, Z., Zhang, L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can. J. For. Res. 38, 2526–2534.
Homann, P.S., Sollins, P., Fiorella, M., Thorson, T., Kern, J.S., 1998. Regional soil organic carbon storage estimates for western Oregon by multiple approaches. Soil Sci. Soc. Am. J. 62, 789–793.
ISO, 1995. ISO 10694:1995. Soil Quality - Determination of Organic and Total Carbon After Dry Combustion (Elementary Analysis). International Organization for Standardization (ISO), Geneva (7 pp).
Jackson, R.B., Banner, J.L., Jobággy, E.G., Pockman, W.T., Wall, D.H., 2002. Ecosystem carbon loss with woody plant invasion of grasslands. Nature 418, 623–626.
Jenny, H., 1980. The Soil Resource, Origin and Behavior. Springer-Verlag, New York (392 pp).
Jobággy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10 (2), 423–436.
Johnston, C.A., Groffman, P., Breshears, D.D., Cardon, Z.G., Currie, W., Emanuel, W., Gaudinski, J., Jackson, R.B., Lajtha, K., Nadelhoffer, K., Nelson, D.J., Post, W.M., Retallack, G., Wielopolski, L., South Dakota State University, 2004. Carbon cycling in soil. Front. Ecol. Environ. 2, 522–528.
Joseph, M., Wang, L., Wang, F., 2012. Using Landsat imagery and census data for urban population density modelling in Port-au-Prince, Haiti. GISci. Remote Sens. 49 (2), 228–250.
Kalogirou, S., 2014. A spatially varying relationship between the proportion of foreign citizens and income at local authorities in Greece. Proceedings of the 10th International Congress of the Hellenic Geographical Society. vol. 5, pp. 1458–1466.
Kern, J.S., 1994. Spatial patterns of soil organic carbon in the contiguous United States. Soil Sci. Soc. Am. J. 58, 439–455.
Khaledian, Y., Pereira, P., Brevik, E.C., Pundyte, N., Paliulis, D., 2017. The influence of organic carbon and pH on heavy metals, potassium, and magnesium levels in Lithuanian Podzols. Land Degrad. Develop. 28, 345–354.
Korkanç, S.Y., 2014. Effects of afforestation on soil organic carbon and other soil properties. Catena 123, 62–69.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 189, 627–634.
Le Bissonnais, Y., Arrouays, D., 1997. Aggregate stability and assessment of soil crustability and erodibility. II. Application to humic loamy soils with various organic carbon contents. Eur. J. Soil Sci. 48, 39–48.
Lenka, N.K., Lal, R., 2013. Soil aggregation and greenhouse gas flux after 15 years of wheat straw and fertilizer management in a no-till system. Soil Till. Res. 126, 78–89.
Li, H.L., Peng, J., Liu, Y.X., Y.N., H., 2017. Urbanization impact on landscape patterns in Beijing City, China: a spatial heterogeneity perspective. Ecol. Indic. 82, 50–60.
Lu, C., Liu, Y., 2016. Effects of China's urban form on urban air quality. Urban Stud. 53 (12), 2607–2623.
Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Chang. Biol. 23 (10), 4430–4439.
Mackových, D., Lučivjanský, P., 2014. Preparation of GEMAS project samples and standards. Chapter 4. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils. Part a: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, pp. 37–40.
Margaritis, E., Kang, J., 2017. Relationship between green space-related morphology and noise pollution. Ecol. Indic. 72, 921–933.
Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8, 1053–1065.
Matschullat, J., Reimann, C., Birke, M., Dos Santos Carvalho, D., GEMAS Project Team, 2018. GEMAS: CNS concentrations and C/N ratios in European agricultural soil. Sci. Total Environ. 627, 975–984.
McGrath, D., Zhang, C.S., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem. 18, 1629–1639.
de Moraes Sa, J.C., Cerri, C.C., Lal, R., Dick, W.A., de Cassia Piccolo, M., Feigl, B.E., 2009. Soil organic carbon and fertility interactions affected by a tillage chronosequence in a Brazilian Oxisol. Soil Till. Res. 104 (1), 56–64.
Négrel, P., Ladenberger, A., Reimann, C., Birke, M., Demetriades, A., Sadeghi, M., The GEMAS Project Team, 2019. GEMAS: geochemical background and mineral potential of emerging tech-critical elements in Europe revealed from low-sampling density geochemical mapping. Appl. Geochem. 111, 104425 20 pp. https://doi.org/10.1016/j.apgeochem.2019.104425.
O'Rourke, S.M., Angers, D.A., Holden, N.M., McBratney, A.B., 2015. Soil organic carbon across scales. Glob. Chang. Biol. 21, 3561–3574.
Pan, G., Li, L., Wu, L., Zhang, X., 2003. Storage and sequestration potential of topsoil organic carbon in China's paddy soils. Glob. Chang. Biol. 10, 79–92.
Pan, G., Xu, X., Smith, P., Pan, W., Lal, R., 2010. An increase in topsoil SOC stock of China's croplands between 1985 and 2006 revealed by soil monitoring. Agric. Ecosyst. Environ. 136 (1), 133–138.
Piotrowski, J.A., Larsen, N.K., Menzies, J., Wysota, W., 2006. Formation of subglacial till under transient bed conditions: deposition, deformation, and basal decoupling under a Weichselian ice sheet lobe, central Poland. Sedimentology 53 (1), 83–106.
Post, W.M., Emanuel, W.R., Zinke, P.J., Stangenberger, A.G., 1982. Soil carbon pools and world life zones. Nature 298 (5870), 156–159.

Reimann, C., Demetriades, A., Eggen, O.A., Filzmoser, P., The EuroGeoSurveys Geochemistry Expert Group, 2009. The EuroGeoSurveys geochemical mapping of agricultural and grazing land soils project (GEMAS) – evaluation of quality control results of aqua regia extraction analysis. Geological Survey of Norway, Trondheim, NGU report. 2009.049, 94 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2009/2009_049.pdf.

Reimann, C., Demetriades, A., Eggen, O.A., Peter Filzmoser, P., the EuroGeoSurveys Geochemistry Expert Group, 2011. The EuroGeoSurveys GEochemical Mapping of Agricultural and grazing land Soils project (GEMAS) – evaluation of quality control results of total C and S, total organic carbon (TOC), cation exchange capacity (CEC), XRF, pH, and particle size distribution (PSD) analysis. Geological Survey of Norway, Trondheim, NGU Report 2011.043. 90 pp. http://www.ngu.no/upload/Publikasjoner/Rapporter/2011/2011_043.pdf.

Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., The Gemas Project Team, 2012. The concept of compositional data analysis in practice—total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O' Connor, P., 2014a. Chemistry of Europe's Agricultural Soils, Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover (528 pp).

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O' Connor, P., 2014b. Chemistry of Europe's Agricultural Soils, Part B: General Background Information and Further Analysis of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B103), Schweizerbarth, Hannover (352 pp).

Reimann, C., Demetriades, A., Birke, M., Filzmoser P., O'Connor, P., Halamić, J., Ladenberger, A., the GEMAS Project Team, 2014c. Distribution of elements/ parameters in agricultural and grazing land soil of Europe. Chapter 11 In: C. Reimann, M. Birke, A. Demetriades, P. Filzmoser, P. O'Connor (Editors), Chemistry of Europe's Agricultural Soils – Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, 103–474.

Reimann, C., Demetriades, A., Birke, M., Schoeters, I., 2014d. The GEMAS project - concept and background. Chapter 1. In: Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O'Connor, P. (Eds.), Chemistry of Europe's Agricultural Soils – Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B 102), Schweizerbarth, Hannover, pp. 23–25.

Reisser, M., Purves, R.S., Schmidt, M.W.I., Abiven, S., 2016. Pyrogenic carbon in soils: a literature-based inventory and a global estimation of its content in soil organic carbon and stocks. Front. Earth Sci. 4, 80. https://doi.org/10.3389/feart.2016.00080.

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. https://doi.org/10.1038/nature10386.

Stockmann, U., Adams, M.A., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A.B., de Courcelles, V.D.R., Singh, K., Wheeler, I., Abbott, L., Angers, D.A., Baldock, J., Bird, M., Brookes, P.C., Chenu, C., Jastrow, J.D., Lal, R., Lehmann, J., O'Donnellk, A.G., Parton, W.J., Whitehead, D., Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric. Ecosyst. Environ. 164, 80–99.

Stockmann, U., Padariana, J., McBratney, A.B., Minasny, B., de Brogniez, D., Montanarella, L., Hong, S.Y., Rawlins, B.G., Field, D.J., 2015. Global soil organic carbon assessment. Glob. Food Sec. 6, 9–16.

Tashi, S., Singh, B., Keitel, C., Adams, M., 2016. Soil carbon and nitrogen stocks in forests along an altitudinal gradient in the eastern Himalayas and a meta-analysis of global data. Glob. Chang. Biol. 22, 2255–2268.

Tu, J., Xia, Z.G., 2008. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: model design and evaluation. Sci. Total Environ. 407 (1), 358–378.

Wang, Z.M., Zhang, B., Song, K.S., Liu, D.W., Ren, C.Y., 2010. Spatial variability of soil organic carbon under maize monoculture in the Song-Nen Plain, Northeast China. Pedosphere 20 (1), 80–89.

Wang, T., Kang, F., Cheng, X., Han, H., Ji, W., 2016. Soil organic carbon and total nitrogen stocks under different land uses in a hilly ecological restoration area of North China. Soil Tillage Res. 163, 176–184.

Wiesmeier, M., Lützow, M.V., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., Kögel-Knabner, I., 2015. Land use effects on organic carbon storage in soils of Bavaria: the importance of soil types. Soil Tillage Res. 146, 296–302.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., Von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - a review of drivers and indicators at various scales. Geoderma 333, 149–162.

Wilk, M.B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika 55 (1), 1–17.

Xu, H.F., Demetriades, A., Reimann, C., Jiménez, J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94–104.

Yuan, Y.M., Cave, M., Xu, H.F., Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater. 393, 122377.

Zhang, C.S., Manheim, F.T., Hinde, J., Grossman, J.N., 2005. Statistical characterization of a large geochemical database and effect of sample size. Appl. Geochem. 20 (10), 1857–1874.

Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146, 378–390.

Zhang, C.S., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. 26, 1239–1248.

Zhang, X., Liu, M., Zhao, X., Li, Y., Zhao, W., Li, A., Cheng, S., Cheng, S., Han, X., Huang, J., 2018. Topography and grazing effects on storage of soil organic carbon and nitrogen in the northern China grasslands. Ecol. Indic. 93, 45–53.

4.3 Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis

Xu, H.F., Croot, P., Zhang, C.S., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environ. Int. 151, 106456.

Summary: This study investigated the spatial clustering patterns of 15 PTEs and 6,862 topsoil samples that collected from Tellus project of NI by two SML techniques of hot spot analysis and K-means clustering analysis. The spatial clusters of hot and cold spots for the 15 PTEs were revealed, showing clear associations with different geological features, especially peat and basalt. Peat was associated with high concentrations of Bi, Pb, Sb and Sn, while basalt was associated with high concentrations of Co, Cr, Cu, Mn, Ni, V and Zn. The high concentrations of As, Ba, Mo and U were associated with mixture of various lithologies, indicating the complicated influences on them. Moreover, three hidden patterns in soil samples were also identified by K-means clustering analysis. These hidden patterns of soil samples were consistent with the spatial clustering patterns for PTEs, highlighting the dominant control of peat and basalt in the topsoil of NI. Our results demonstrated that these two SML techniques are powerful and effective in identifying hidden spatial patterns, providing evidences to extract geochemical knowledge in environmental studies.

My contribution in this paper accounted for ~90% in reviewing literatures, exploring data and writing manuscript.

Research paper

Environment International 151 (2021) 106456

Contents lists available at ScienceDirect

Environment International

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

Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis

Haofan Xu [a], Peter Croot [b], Chaosheng Zhang [a,*]

[a] International Network for Environment and Health (INEH), School of Geography, Archaeology & Irish Studies, National University of Ireland, Galway, Ireland
[b] iCRAG (Irish Centre for Research in Applied Geoscience), Earth and Ocean Sciences, School of Natural Sciences and the Ryan Institute, National University of Ireland Galway, Galway, Ireland

ARTICLE INFO

Handling Editor: Frederic Coulon

Keywords:
Potentially toxic elements
Hot spot analysis
K-means clustering analysis
Hidden spatial patterns
Geochemical association

ABSTRACT

The understanding of sources and controlling factors of potentially toxic elements (PTEs) in soils plays an important role in the improvement of environmental management. With the rapid growth of data volume, effective methods are required for data analytics for the large geochemical data sets. In recent years, spatial machine learning technologies have been proven to have the potential to reveal hidden spatial patterns in order to extract geochemical information. In this study, two spatial clustering techniques of Getis-Ord G_i^* statistic and K-means clustering analysis were performed on 15 PTEs in 6,862 topsoil samples from the Tellus datasets of Northern Ireland to investigate the hidden spatial patterns and association with their controlling factors. The spatial clustering patterns of hot spots (high values) and cold spots (low values) for the 15 PTEs were revealed, showing clear association with geological features, especially peat and basalt. Peat was associated with high concentrations of Bi, Pb, Sb and Sn, while basalt was associated with high concentrations of Co, Cr, Cu, Mn, Ni, V and Zn. The high concentrations of As, Ba, Mo and U were associated with mixture of various lithologies, indicating the complicated influences on them. In addition, three hidden patterns in the 6,862 soil samples were revealed by K-means clustering analysis. The soil samples in the first and second clusters were overlaid on the peatland and basalt formation, respectively, while the samples in the third cluster were overlaid on the mixture of the other lithologies. These hidden patterns of soil samples were consistent with the spatial clustering patterns for PTEs, highlighting the dominant control of peat and basalt in the topsoil of Northern Ireland. This study demonstrates the power of spatial machine learning techniques in identifying hidden spatial patterns, providing evidences to extract geochemical knowledge in environmental studies.

1. Introduction

Understanding the controlling factors of the potentially toxic elements (PTEs) is crucial for environmental management due to their toxic effects on organisms. Sources of PTEs in urban and rural areas include both natural and anthropogenic factors (Rodrigues et al., 2009; Argyraki and Kelepertzis, 2014). Natural factors are mainly associated with geogenic occurrences, as well as soil formation and parent materials (Tipping et al., 2006; Jordan et al., 2007; Zhang et al., 2008a; Reimann et al., 2014; Birke et al., 2017). Anthropogenic sources are attributed to human activities, including industrial, waste, traffic (vehicle emissions, fuel), agricultural inputs and atmospheric deposition (e.g. Cloquet et al., 2006; Ettler et al., 2008; Aelion et al., 2009; Davis et al., 2009; Okorie et al., 2011; Dao et al., 2014). The application of pesticides and fertilisers can cause pollution of agricultural soils in rural areas (Faria et al., 2012). Generally, anthropogenic pollution is characterized by points and dispersion patterns on the spatial distribution maps of PTE concentrations, which can be observed around urban areas (Zhang, 2006; Marchant et al., 2011; Delbecque and Verdoodt, 2016). On the other hand, the elevated concentrations of PTEs caused by geogenic sources are usually reflected in large and continuous geochemical patterns.

To assess the controlling factors of PTEs, statistical and geostatistical methods have widely been applied, including neighbourhood analysis, multivariate statistical analysis, spatial autocorrelation analysis and geographically weighted regression (e.g. Zhang, 2006; Zhang et al., 2007; Zhang et al., 2008b; Bhowmik et al., 2015; Buccianti et al., 2015;

* Corresponding author.
E-mail addresses: h.xu2@nuigalway.ie (H. Xu), peter.croot@nuigalway.ie (P. Croot), Chaosheng.Zhang@nuigalway.ie (C. Zhang).

https://doi.org/10.1016/j.envint.2021.106456
Received 11 December 2020; Received in revised form 13 January 2021; Accepted 7 February 2021
Available online 1 March 2021

Petrik et al., 2018; Thiombane et al., 2018; Fei et al., 2019; Meng et al., 2020; Yuan et al., 2020). The geochemical data are collected through sampling at a specific location, which play an important role in environmental management and assessment (Li and Heap, 2011). However, with the increasing volume of geochemical data and the complexity of soil chemical elements, an effective way is required to investigate the controlling factors on PTEs in multivariate datasets. This paper attempts to reveal the spatial patterns of PTEs using the spatial machine learning (SML) technique. The SML techniques are the application of machine learning algorithms into spatial data (Kanevskij et al., 2009; Li et al., 2011). As an emerging concept, it has been gradually applied in various fields, including environmental science (Meyer et al., 2018), agriculture (Ludwig et al., 2019) and social economics (Fan et al., 2018), etc. The typical machine learning algorithms are mainly divided into three categories: supervised learning, unsupervised learning and reinforcement learning (Bishop, 2006). Supervised learning is training from labelled data, and each sample contains an input object and the corresponding expected output value. In contrast to the supervised learning, unsupervised learning algorithms aim to identify hidden patterns from unlabelled data (Jordan and Mitchell, 2015). They are able to learn by themselves without being explicitly told whether what they have done is correct (LeCun et al., 2015). Thus, these techniques have the potential to reveal hidden spatial patterns which are helpful to extract useful geochemical knowledge and association (Xie et al., 2004; Meshkani et al., 2011; Sergeev et al., 2019; Rahmati et al., 2020). In recent years, they have also received an increasing attention in identification of pollution sources and environmental monitoring (e.g. Boente et al., 2018; Kelepertzis et al., 2019; Klapstein et al., 2020). In this study, two of the spatial machine learning approaches: hot spot analysis (Getis-Ord G_i^* statistic) and K-means clustering analysis were explored to identify spatial clustering and hidden patterns of topsoil samples based on 15 PTEs in Northern Ireland, and to reveal their association with controlling factors. The Getis-Ord G_i^* statistic is a parameter spatial statistic which are popularly in applied for the identification of the spatial clustering patterns of environmental variables (Bagstad et al., 2017; Xu et al., 2019). The K-means clustering analysis is an unsupervised algorithm which is used to cluster the samples in data mining (Hartigan and Wong, 1979). As a simple and powerful algorithm, it has been widely applied to discover the hidden information and structure of unlabelled samples (Zuo, 2017). Both methods can be regarded as the currently emerging concept of SML techniques. The spatial patterns of soil geochemical features are complicated and hard to identify due to the strong influences of both natural and human factors. Based on the different spatial patterns revealed by these two techniques, the association between the patterns and influencing factors can be obtained, providing an effective way to understand the sources of PTE enrichment in the topsoil.

Reimann et al. (2018) defined the background knowledge and thresholds of PTEs in detail based on GEMAS data in European agricultural soil. A total of 15 PTEs, including arsenic (As), barium (Ba), bismuth (Bi), chromium (Cr), cobalt (Co), copper (Cu), nickel (Ni), manganese (Mn), molybdenum (Mo), lead (Pb), antimony (Sb), tin (Sn), uranium (U), vanadium (V) and zinc (Zn) in the topsoil of Northern Ireland were selected from Tellus project database. The factors that are related with some individual PTEs (e.g. As, Cr, Co, Cu, Ni, Pb and Zn) have been investigated in previous researches (e.g. Kelepertsis et al. 2006; Zhang et al., 2007; Ajmone-Marsan et al. 2008; Palmer et al. 2013; McIlwaine et al., 2014; Meng et al., 2020). A detailed introduction to these PTEs was reported in the guide book of Tellus data in Young and Donald (2013). As a step forward, this study attempts to investigate the spatial patterns for 15 PTEs, and then to associate these patterns with their controlling factors from the spatial perspective.

The objectives of this study were: (1) to identify the spatial clustering patterns of high and low values for 15 PTEs using hot spot analysis; (2) to reveal the hidden spatial patterns in the 6,862 soil samples using K-means clustering analysis; and (3) to further explore the geochemical association between these spatial patterns and the controlling factors on PTEs.

2. Materials and methods

2.1. Study area

Despite that the total area is only 14,120 km^2 (13,480 km^2 of land area and 640 km^2 of inland water area), Northern Ireland is a microcosm of the earth's geology (Zhang et al., 2007). Bedrock history includes almost every period from Mesoproterozoic to Palaeogene, and almost every known type of rocks can be found there. A simplified geological map is shown in Fig. 1, with the locations of the peat overlaid. The history of Northern Ireland involves the development of ice sheets and meltwater from the last 100,000 years, which caused more than 80% of the bedrock to be covered by various superficial deposits (e.g. alluvium, peat). It is reported that the peatland accounted for over 12% of the total land area (Davies and Walker, 2013), which is a major soil subgroup in Northern Ireland. The northeast part is composed of a large area of extrusive Palaeogene basalt, and the northwest is dominated by psammites (schist) that are mainly Neoproterozoic in age. The southwestern terrain is a mixture of sandstone, mudstone and limestone, mainly Carboniferous in age. While southeast is controlled by greywacke shales, significant granite intrusions were found in this area. The diverse types of soil and lithology provide unique opportunities for investigating the spatial distribution and classification of PTEs, which can be beneficial to environmental research and assessment in Northern Ireland and elsewhere.

Northern Ireland is rich in minerals. Historically, the major minerals mined in Northern Ireland include iron ore, lead, coal and salt. Nowadays, there are more than 2,000 abandoned mines, most of which worked between the 18th and early 20th centuries. In recent years, gold, lignite and industrial minerals are dominated in commercial mining exploration activities in Northern Ireland. For example, it is reported that County Tyrone holds "one of the most promising undeveloped gold deposits" in the world (Dalradian, 2019). A detailed description about the extent and spatial distribution of mineralization can be found in Mitchell (2004). There are two main urban areas in Northern Ireland: the Belfast Metropolitan Area and Londonderry.

2.2. Soil sampling and analyses

The Tellus project was a national collaborative project designed to collect geophysical and geochemical data across the island of Ireland. In the part of Northern Ireland, it was managed and undertaken by Geological Survey of Northern Ireland (GSNI). During 2004 to 2006, nearly 30,000 samples of soils, stream sediments and stream water samples were collected in the geochemical survey. A total of 6,862 regional topsoil samples (5–20 cm depth) were used in this study, with sample locations displayed in Fig. 2. Samples were taken as composite samples of five auger flights (approx. 750 g), with two composite samples at one site. The sampling density is on an average of 1 per 2 km^2. After the collection process, the soil samples were shipped to the storage for drying in oven at 30 °C for 2 to 3 days. Then, samples were sieved with the <2 mm pore size nylon mesh, while repetition was prepared by shallow-splitting of each duplicate sample. Further information of soil sampling and preparation is provided by Young and Donald (2013).

After soil preparation, X-ray fluorescence (XRF) analysis was performed for total concentrations of trace elements and major oxides in the British Geological Survey (BGS) laboratory. A series of quality control was conducted during the analytical process. The detailed description of methodology and quality control procedures of Tellus program can be found in Smyth (2007) and Jordan et al. (2007). Meanwhile, the spatial distribution maps showing the concentrations of PTEs can be found in Young and Donald (2013).

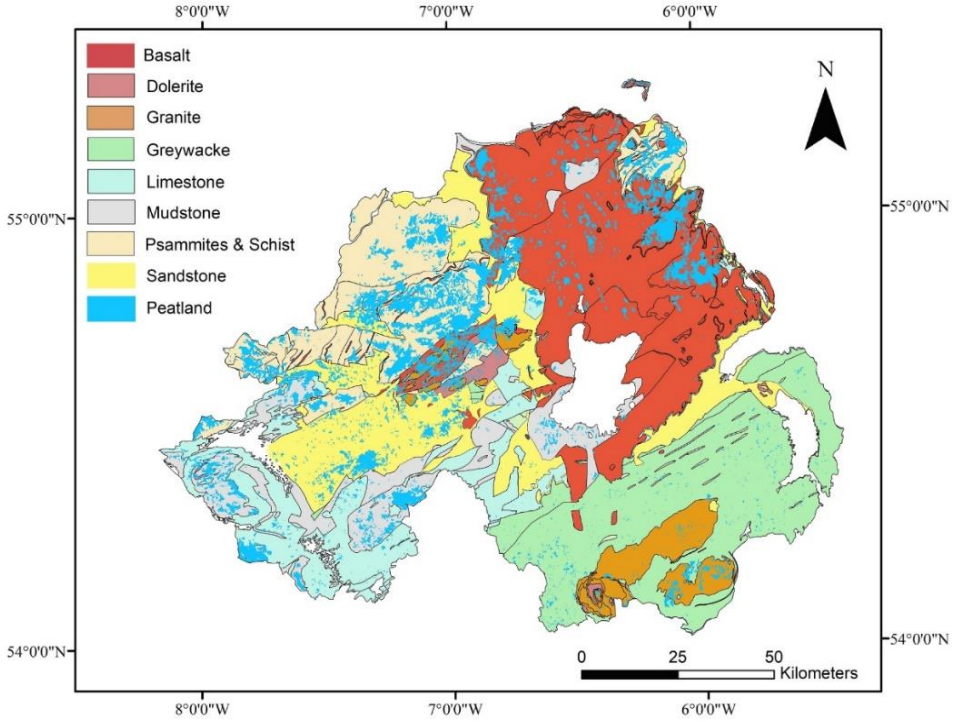


Fig. 1. Simplified bedrock geology maps of Northern Ireland and areas of peatland (original GIS shapefiles from GSNI, 1998).

2.3. Hot spot analysis (Getis-Ord Gi* statistic)

Hot spot analysis is a mapping technology that can reveal the hidden spatial clusters based on the distance between samples, which can identify locations with statistically significant high and low values in a certain geographic area. Getis-Ord G_i^* statistic is a measure of spatial autocorrelation at the local scales (Ord and Getis, 1995), indicating high and low values that are associated with the hot spot and cold spot cluster patterns, respectively. The local G_i^* statistic returns the z-scores and p-values for all features in the datasets by calculating each feature and its neighbours. The statistically significant hot spot is returned with high z-scores and small p-values. In contrast, the high negative z-scores and small p-values indicate the significant cold spots. The function of Getis-Ord G_i^* statistic is showing as follows (Getis and Ord, 1992):

$$G_i^* = \frac{\sum_{j=1}^{n}\omega_{i,j}x_j - \overline{X}\sum_{j=1}^{n}\omega_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n}\omega_{i,j}^2 - \left(\sum_{j=1}^{n}\omega_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where i is the centre of the local neighbourhood; x_j is the value of the variable in the sample at location j; $\omega_{i,j}$ is the spatial weight between sample locations i and j; n is the total number of samples.

The following equation calculates the mean of the whole datasets:

$$\overline{X} = \frac{\sum_{j=1}^{n}x_j}{n} \quad (2)$$

and the standard deviation of the whole datasets is calculated by the following equation:

$$S = \sqrt{\frac{\sum_{j=1}^{n}x_j^2}{n} - \left(\overline{X}\right)^2} \quad (3)$$

In this study, Getis-Ord G_i^* statistic was applied on the 15 PTEs to identify their spatial clustering patterns in the topsoil of Northern Ireland based on the Tellus datasets.

2.4. K-means Clustering analysis

Cluster analysis is the method of classifying samples into different groups or subsets, with all samples in the same group having relatively similar properties. Various clustering methods exist, such as hierarchical clusters, partitioning clusters, fuzzy methods and model-based methods (Reimann et al., 2008). K-means clustering algorithm is a typical partitioning method, which is adopted as the most widely used cluster methods in machine learning and data mining due to its simplicity and efficiency (Han and Kamber, 2006). It is usually performed as the initial step of data analysis, which has been proved to be powerful for capturing the hidden information of geochemical patterns (e.g. Bengio, 2013; LeCun et al., 2015; Zuo, 2017). It aims to partition the space into k non-overlapping clusters, and classify each observation to the nearest centre in order to maximize the between-cluster variance as well as minimize the within-cluster variance (Hartigan, 1975; Alizadeh et al., 2017). K-means is a distance-based clustering algorithm, and thus the variance here is calculated based on the distance. It is worth noting that the distance in K-means clustering analysis does not refer to the actual distance between two observations or samples. It is used to measure the similarity between two observations or samples in the algorithm. Based on the selection of different distance, the measurement of similarity is also different. In this study, Euclidean distance was selected for the K-

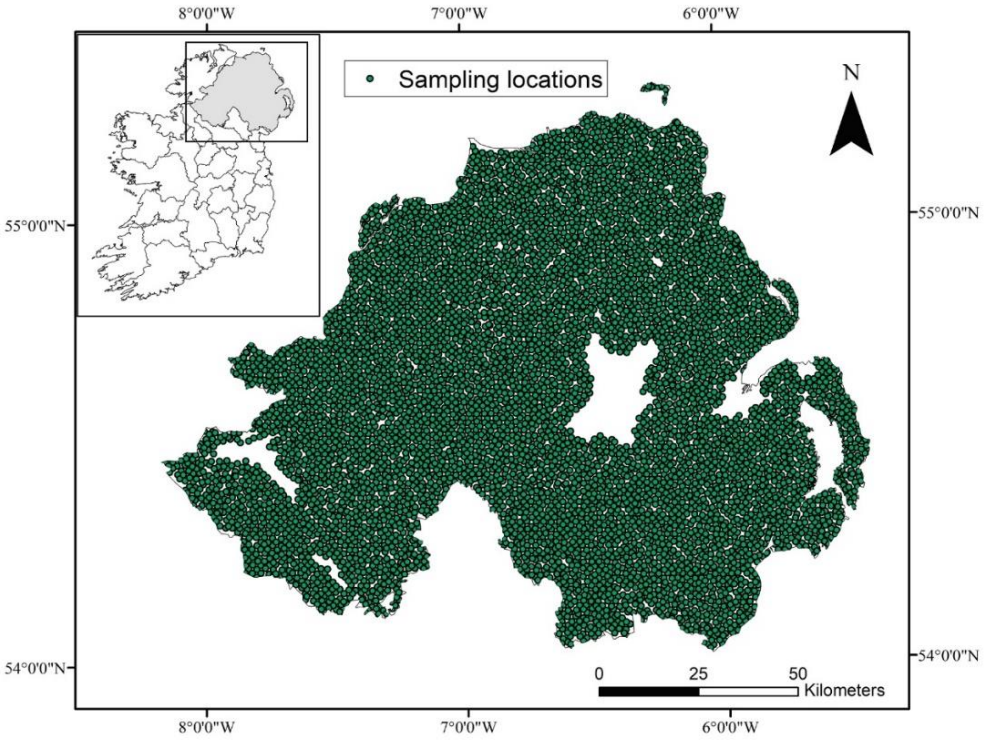


Fig. 2. Spatial distribution map with sample locations of Tellus topsoil data sets in the Northern Ireland.

means algorithm. The function can be represented as follow (MacQueen, 1967; Hartigan and Wong, 1979):

$$J = \sum_{i=1}^{k} \sum_{j \in C_i}^{n_i} \|x_j - \mu_i\|^2 \quad (4)$$

where J is the objective function, C_i is the i^{th} cluster, n_i is the number of samples in i^{th} cluster, distance function $d_{ji} = ||x_j - \mu_i||^2$ represents the calculation of the distance between each sample point x_j and centroid μ_i in the i^{th} cluster. The centroid μ_i can be calculated based on the function as below:

$$\mu_i = \frac{1}{|C_i|} \sum_{j \in C_i} x_j \quad (5)$$

The implementation of K-means clustering algorithm can be summarized in the following steps (Zagouras et al., 2013):

(1) Randomly initializing the cluster centroid $\mu_1, \mu_2, \ldots, \mu_k$;
(2) Calculating the distance function d_{ji} between each sample point x_j and centroid μ_i in the i^{th} cluster. The distance function d_{ji} was based on the Euclidean distance in this study.
(3) Moving each sample point x_j to the cluster of its nearest centroid $\mu_{nearest}$, and update cluster centroids from which sample points have been disjointed or reassigned.
(4) Computing the objective function J, as given above in formula (1). If function J converges, the centroids do not change from the previous iterations, and the K-means clustering algorithm derives the final centroids of cluster. Otherwise, the step 2 and 3 are repeated until the objective function J converges.

Another parameter needed to be considered in partition clustering is the optimal cluster number (Weatherill and Burton, 2008). This can be achieved by various methods and the prior knowledge including Davies-Bouldin Index (DBI) (Davies and Bouldin, 1979), Silhouette method (Rousseeuw, 1987), elbow method (Ketchen and Shook, 1996), information criterion approach (Goutte et al., 2001). In this study, Silhouette method was applied to choose the appropriate cluster number. It can provide succinct graphics to display the quality of classification, as well as silhouette values to interpret and validate the consistency of clusters within samples (Rousseeuw, 1987). The silhouette values can represent how similar an observation belongs to its cluster compared to others, where a high value implies the good cohesion of one object to its own cluster and poor match with adjacent clusters. This principle corresponds well to the classification criteria of cluster analysis.

2.5. Principle component analysis

Principal component analysis (PCA) is one of the most popular methods in multivariate statistics. It combines multiple correlated variables into fewer principle components based on correlation or covariance matrix. These components are not correlated to each other, which can represent the interrelationships between the multi-variables in the original data set (Jolliffe, 2002). In geochemical studies, PCA is widely used to extract useful geochemical information and has become a standard approach. The advantage of adopting extract components is that the input datasets can be replaced by fewer comprehensive indicators with as little loss of information as possible (Jolliffe, 2002), which is called dimension reduction. The appropriate number of components can be determined by a significant inflection point on the output scree plot (Cattell, 1966). In addition, PCA can enhance the interpretability among multiple variables with appropriate rotation methods

(Cheng et al., 2006). Varimax (Kaiser, 1958) is the most common rotation option in PCA. Besides, there are some other methods such as Promax, Oblimin and Quartimin (Carroll, 1953; Hendrickson and White, 1964; Harman, 1976). The detailed comparison of rotation methods was discussed in Reimann et al. (2002). More detailed information on the principles and application of PCA can be found in Davis (2002) and Cheng et al. (2006).

The PCA was performed using Varimax rotation in this study and the new variables were saved as component scores. Then, the K-means clustering analysis were conducted using the derived component scores as input variables instead of the 15 PTEs for reducing the dimension.

2.6. Data preparation and software

Due to the strong complexity of geochemical data, the results of spatial clustering techniques are usually affected by the data preparation and the choice of clustering analysis (Templ et al., 2008). Reviewing past literature, the two main problems before applying spatial clustering techniques on geochemical compositional data are (e.g. Yeung and Ruzzo, 2001; Templ et al., 2008; Zuo, 2017; Zuo et al., 2019; Tepanosyan et al., 2020): (1) data transformation; (2) whether to reduce the high dimension of the raw datasets.

The first problem is the method of data transformation. Data without transformation can lead to relatively unreliable results of Hot spot analysis and cluster analysis. For hot spot analysis, transformed data can lead to clear clustering patterns of hot spots and cold spots (Zhang et al., 2008b; Xu et al., 2019). For K-means clustering analysis, performing centred log-ratio (clr) transformation and isometric log-ratio (ilr) transformation seems have better performance to capture spatial hidden patterns in the geochemical datasets (Templ et al., 2008). Geochemical data is usually taken as compositional data, which is considered as being 'closed' (Aitchison, 1986; Buccianti et al., 2006; Filzmoser et al., 2010). Before performing analysis on the compositional data, it is suggested to use data transformation to open the data in order to destroy the closure effects (Aitchison, 1986; Egozcue et al., 2003). Thus, a clr-transformation was conducted to the raw data based on 15 variables of PTEs. In addition, data transformation on the geochemical data is able to reduce the influences of outliers and to obtain a relatively symmetrical distribution (Zhang et al., 2008b). The transformed data were used for the following hot spot analysis, PCA and K-means analysis.

A further challenge is dimension reduction. In fact, K-means clustering analysis, as an effective machine learning algorithm (Kanungo et al., 2002), can be used to process high-dimensional datasets. The advantage of using dimension reduction is that tools such as PCA can project the original data from the high-dimensional space onto the low-dimensional space and preserve the useful information from the original datasets (Saaltink et al., 2014), which is more efficient during the algorithm computation. However, for example, Yeung and Ruzzo (2001) proposed that clustering based on high-dimensional variables cannot be replaced by fewer number of orthogonal components or factors. In this paper, our research only focused on the spatial overlap association between hidden patterns of samples and the controlling factors after clustering. Therefore, the mainstream method was adopted to perform PCA on the 15 PTEs in Tellus database for dimension reduction before K-means clustering analysis.

The raw Tellus soil data was stored in Microsoft Excel (ver. 2016), and data transformation was computed in R project (ver. 3.56). Hot spot analysis (Getis-Ord G_i^* statistic) was performed using ArcGIS (ver. 10.4), while K-means clustering analysis was performed using 'cluster' package (ver. 2.10) in R project (Maechler et al., 2019; https://cran.r-project.org/web/packages/cluster/cluster.pdf). Principle component analysis was conducted in SPSS (ver. 24). Data statistics were compiled in Microsoft Excel (ver. 2016) and SPSS (ver. 24), and all the spatial distribution maps were produced using ArcGIS (ver. 10.4).

3. Results and discussion

3.1. Basic statistics for 15 PTEs in topsoil in Northern Ireland

Table 1 summarises the basic statistics for 15 PTEs in the topsoil of Northern Ireland. The values below the detection limits (DLs) were replaced by half of the DL values of their corresponding elements for further statistical analysis. The significant differences between the maximum and minimum values indicated strong variation in all the 15 PTEs. In addition, the large coefficients of variation and differences between 95% percentiles and maximum values in the raw datasets (e.g. Mn, Ni, Pb, Sb, and U) suggested that potential outliers existed within the datasets. Thus, performing appropriate data pre-processing is necessary to reduce the impact of outliers on clustering results of geochemical elements (Templ et al. 2018).

Histograms with the normal distribution curve for the raw data and transformed data are shown in Fig. 3, using Ni as an example. The raw dataset displayed a long tail towards higher values (see Fig. 3a), implying the existence of high-value outliers. The significance ($p < 0.05$) of Kolmogorov-Smirnov normality test (K-S test) also suggested the non-normality of raw data. The 'non-normality' feature of soil geochemical elements has been widely reported (e.g. Reimann and Filzmoser, 2000; Zhang et al., 2005). Thus, data transformation was necessary for the raw data to reduce the effects of outliers and to obtain a symmetrical distribution. Although the result still did not pass the K-S test after clr-transformation ($p < 0.05$), a comparatively symmetrical distribution of Ni concentrations was displayed in Fig. 3b.

3.2. Identification of hidden spatial patterns of soil samples based on 15 PTEs

3.2.1. Spatial clustering patterns for 15 PTEs

The Getis-Ord G_i^* statistic was performed to identify the spatial clustering patterns of hot and cold spots for the 15 PTEs in the topsoil of Northern Ireland. The hot and cold spots reflected spatial clusters of high and low values for each PTE, showing the hidden spatial patterns. Different clustering patterns are shown in the hot spot maps for 15 PTEs (Fig. 4), indicating the complexity of geological processes in Northern Ireland.

The hot spots of As were mainly concentrated in the western and south-eastern areas which overlaid on the schist, mudstone and greywacke shale, while large and continuous pattern of cold spots was observed on the eastern areas that overlaid on the basalt formation. In addition to As, the similar clustering patterns were also observed on Ba and U, with the same controlling factors on their low concentrations. However, the hot spots of Ba were mainly observed overlaid on the schist, sandstone and greywacke shale, while most hot spots of U were found on the schist and granite. This indicated that the spatial association with the controlling factors of basalt on the low concentrations, while the high values of them are controlled by a mixture of various geological processes.

Only a few hot spots and cold spots of Bi were identified on the maps, with hotspot patterns clustered in the western and northern areas, showing a clear spatial association with peat on the high concentrations of Bi. The cold spots were scattered on other lithologies, mainly including schist, sandstone and limestone.

The large and continuous pattern for hot spots of Co were observed in the north-eastern areas overlaid on the basalt formation, while cold spots were mainly clustered in the western and southern areas overlaid on the other geological features (e.g. peat, schist, sandstone). Interestingly, the very similar clustering patterns were also observed on Cr, Cu, Mn, Ni and V, suggesting the same association with controlling factors of basalt on the high concentrations of these PTEs. In addition, the spatial clustering pattern of Zn was also very similar, with the only difference was that more hot spots were observed overlaid on greywacke shale.

The hot spots of Mo were mainly scattered in the southern areas

Table 1
Statistical parameters for 15 potentially toxic elements in topsoil in Northern Ireland.

Parameter	Min	Q25	Median	Mean	Q75	Q95	Max	DL	CV (%)
As	<0.90	6.5	8.7	10.5	11.7	20.5	271.2	0.90	95
Ba	112	250	347	341	423	550	2361	1.0	39
Bi	<0.30	0.3	0.5	0.5	0.7	0.8	24.1	0.30	83
Co	<1.50	6.3	10.9	15.2	18.7	44.4	205.1	1.50	88
Cr	4.1	56.7	94.1	131.0	161	375.5	1228.8	3.0	92
Cu	<1.30	18.6	31.6	39.7	49.5	105.2	1510.1	1.30	96
Mn	<0.01	0.029	0.058	0.084	0.101	0.21	14.99	0.01	310
Mo	<0.20	0.6	0.8	0.8	1	1.5	5.4	0.20	49
Ni	1.40	14	29.1	46.2	56.1	155	333.6	1.40	105
Pb	2.20	22.2	28.8	41.7	41	92.3	18756.8	1.30	562
Sb	0.70	1	1	1.1	1.2	1.5	156.9	0.50	177
Sn	1.7	2.3	2.4	2.6	2.7	3.6	37.5	0.50	38
U	<0.50	1.7	2.3	2.5	2.8	4	142.9	0.50	115
V	5.90	56.5	85	99.7	121.3	234	401.6	2.90	65
Zn	2.80	47	71.8	78.4	101.9	149.9	2460.5	1.20	69

*Units are reported in mg/kg except for Mn; Mn is in %; Min: minimum; Max: maximum; Q25 - Q95: quantiles; DL: detection limit; CV: coefficient of variation.

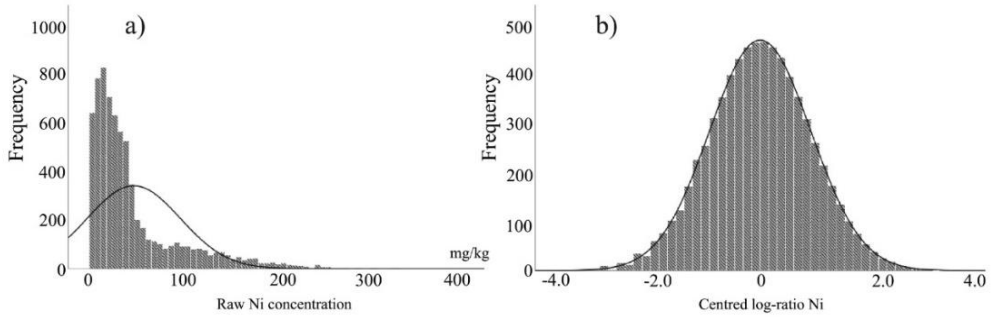


Fig. 3. Histograms and normal distribution curves of Ni concentrations: a) raw data of Ni; b) centred log-ratio transformed data of Ni.

overlaid on multiple lithologies, including limestone, sandstone and granite. The cold spots were clustered in the western and northern areas, showing no spatial association with local lithology, but with peat. This reflected the complicated relationships with mixture of different controlling factors on its concentrations.

The spatial clustering patterns for Pb, Sb and Sn were similar, with hot spots mainly clustered in the western, south-eastern and a small part of north-eastern areas. The clustering patterns of the hot spots of Pb, Sb and Sn were noisy and does not show a clear association with local lithology. However, these patterns were able to associate with peatland at spatial level, suggesting the controlling effects of peat on the high values of Pb, Sb and Sn. This can be attribute to the atmospheric decomposition associated with human activities (Coggins et al., 2006; De Vleeschouwer et al., 2007). In addition, the clustering pattern of Pb hot spots surrounded the Belfast city also suggested the association with anthropogenic influences.

3.2.2. *Spatial patterns of soil samples*

K-means clustering analysis was performed to reveal the hidden spatial patterns on the 6,862 soils samples based on the 15 PTEs. As mentioned earlier, principal component analysis (PCA) was performed first following the clr-transformation for the 15 PTEs in order to reduce the dimension of input variables. The appropriate number of components were chosen as three based on the scree plot of eigenvalues. The results of PCA are presented in Table 2, showing the loading coefficients as well as percentage of variance and cumulative proportion. First component (PC1) accounted for approximately 42% of the variance, with significant positive loadings of Mn, V, Cr, Co, Ni, Cu and Zn. All of these elements are 1st row transition elements and their distribution likely controlled by geogenic sources. These PTEs were shown similar spatial clustering patterns. The second component (PC2) which was characterised by high positive loadings of ore-forming elements explained nearly 22% of variance, including Sn, Sb, Pb and Bi. Generally, these elements are associated with relict hydrothermal processes in the study areas (Wang et al., 2017). They have low temperature volatility and can be found enriched in sulphides from hydrothermal activity. The variance explained by the third component (PC3) was 7.6%, with elements Mo and As showing noticeable positive loadings in the structure. The cumulative percentage of these three factors was around 71%, indicating the majority of input variation were explained. After obtaining a simplified component structure on the 15 PTEs, derived component scores on all sample points were generated as well.

Maps of spatial distribution of the derived component scores for the three indices are shown in Fig. 5. For PC1, high scores were mainly clustered in the northern and eastern parts of Northern Ireland, while low values were mainly concentrated in the northwest and southeast areas. For PC2, the majority of high values were observed in the northeast, central part of northwest and southwest areas, with some scattered patterns observed in the southeast areas. The high scores for the PC3 were mainly concentrated in the southeast and southwest areas, with some small patterns in the northwest areas. Most of the low values were observed in the northwest areas.

Then, K-means clustering analysis was performed on the component scores derived from PCA. Before performing the cluster analysis, the appropriate cluster number was determined by the average silhouette method (Fig. 6). The silhouette map was produced in R project using '*factoextra*' package (Kassambara and Mundt, 2017). The maximum value appeared at cluster number three, suggesting that three clusters were the most appropriate classification in this study. However, considering the potential instability of a specific criterion and the

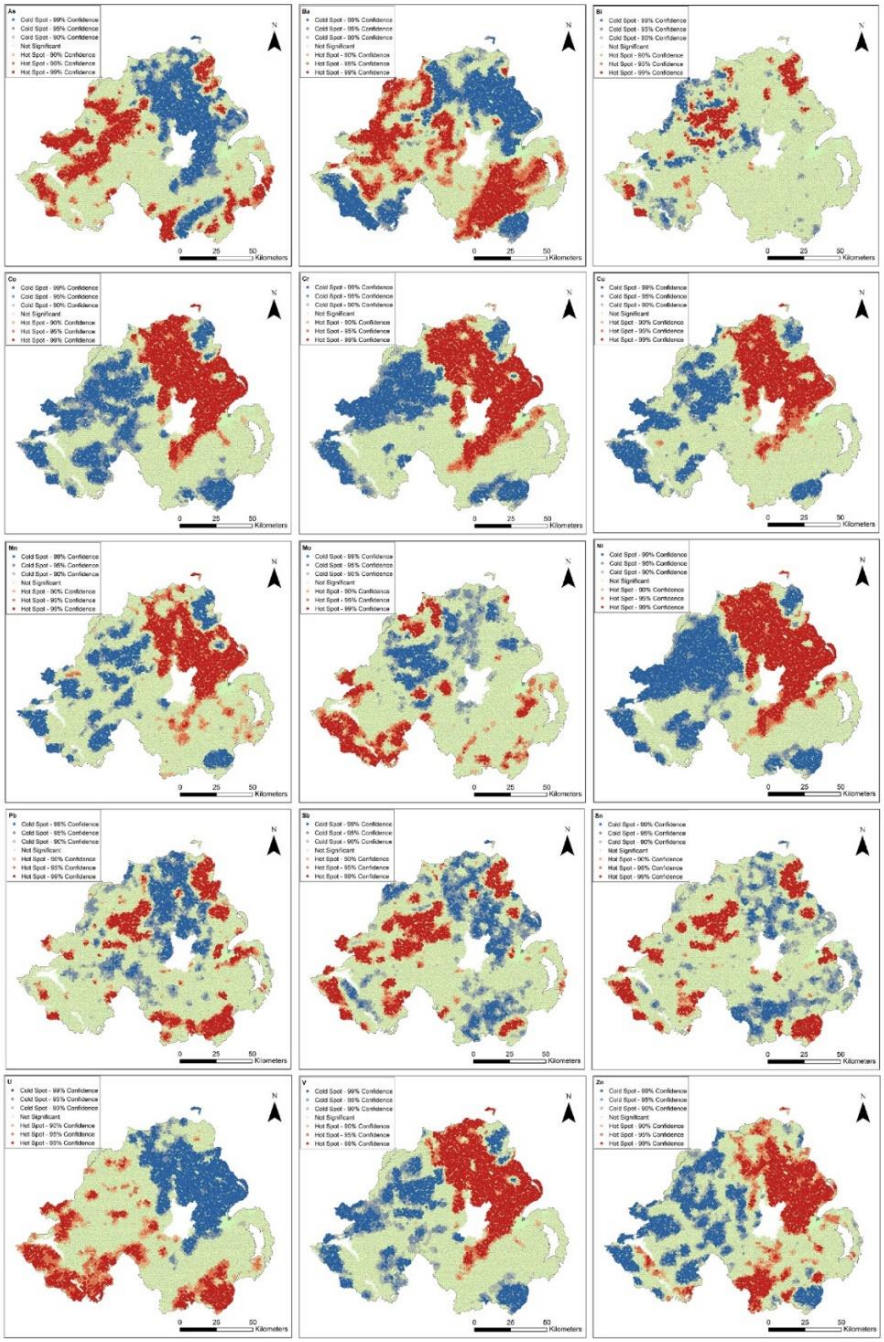


Fig. 4. Hot spot maps showing of spatial clustering patterns for 15 PTEs in the topsoil of Northern Ireland: red dots are hot spots; blue dots are cold spots. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

relatively similar Silhouette values, three adjacent cluster numbers (i.e. 2, 3 and 4) were selected for comparison in order to determine the optimal number of clusters in this study (Fig. 7).

Among the three selected cluster numbers, three and four clusters revealed more details of geochemical information, whereas the results were uninterpretable when the samples were divided into only two clusters (see Fig. 7a). When the number of clusters was three, the cluster patterns showed clear association with peat, basalt and other lithologies,

Table 2
Results showing loading coefficients, raw and cumulative percentage for PCA on 15 PTEs.

	F1	F2	F3
As	−0.507	0.389	0.324
Ba	−0.352	−0.677	−0.015
Bi	−0.188	0.518	−0.409
Co	0.907	−0.287	−0.008
Cr	0.838	−0.389	−0.076
Cu	0.796	−0.092	−0.006
Mn	0.714	−0.436	0.167
Mo	−0.101	−0.028	0.857
Ni	0.943	−0.189	−0.050
Pb	−0.226	0.818	0.119
Sb	−0.477	0.692	−0.105
Sn	−0.413	0.697	−0.078
U	−0.754	0.079	0.227
V	0.837	−0.432	−0.047
Zn	0.730	0.279	0.136
Variance (%)	41.560	21.681	7.640
Cumulative (%)	41.560	63.241	70.881

respectively (see Fig. 1). However, when the samples were grouped into four clusters, the number of samples in the first cluster associated with peat decreased, while the third and fourth clusters seemed to be associated with multiple lithologies (i.e. greywacke shale, limestone and schist), and the results became fairly unclear, making it complicated to interpret the clustering patterns. Therefore, combining Silhouette value and prior knowledge of lithology and PTEs in Northern Ireland (e.g. Zhang et al., 2007; McKinley et al., 2018), three clusters were selected as the most suitable number of clusters in this study.

The spatial patterns for K-means clustering results using three clusters of soil samples overlaid on the simplified geology map are shown in Fig. 8. A total of 6,862 samples were identified into three hidden clustering patterns, with the number of samples for the three clusters of 673, 1,772 and 4,417, respectively. The samples in the first cluster overlaid perfectly on the peatland, showing clear association with peat. For the second cluster, the majority of the samples were overlaid on the basalt formation (1,684, nearly 95%), while only 5% of them were observed on other geological features. Comparing the first two clusters, the samples in the third cluster overlaid on more complicated geological features. The clustering results are impressive due to the clear spatial patterns overlaid on different geological features were revealed, showing spatial associations with peat, basalt, and other lithologies. It is worth noting that the hidden spatial patterns for the soil samples are consistent with the spatial clustering patterns for the 15 PTEs, highlighting the dominant controlling effects of peat and basalt in the Northern Ireland. Combination of these two spatial machine learning techniques effectively revealed hidden spatial and clustering patterns, thereby extracted clear geochemical associations with geological features.

In addition, it should be noted that some sample points belonging to the second cluster were found overlaid on the geological features other than basalt, especially in the southern and eastern areas, respectively. It is reported that the soils in these areas are rich in lead–zinc deposits (Young and Donald, 2013; McIlwaine et al., 2014). Therefore, these samples can be regarded as outliers due to the mineralisation, which were identified after clustering by K-means clustering algorithm. As an effective unsupervised learning algorithm, K-means clustering analysis can be used to identify ore-related anomalies in the datasets (Zuo, 2017; Zhou et al., 2018; Ghezelbash et al., 2020). The anomalies of PTEs in the topsoil of Northern Ireland deserve more in-depth investigations.

3.3. Exploring the spatial association with controlling factors for 15 PTEs

After identifying the clustering patterns of both PTEs and soil samples, in order to better explore the spatial association and extract geochemical knowledge among the 15 PTEs based on the three clusters (defined as 'peat', 'basalt' and 'others'), boxplots for the concentrations of 15 PTEs are shown in Fig. 9. For the concentrations of most PTEs, lower median values can be clearly observed in the peat group, while the samples in the basalt formation exhibited higher values. In the previous studies in Northern Ireland, the peat and basalt have been proposed to be key factors in controlling the distribution of multiple elements (Palmer et al., 2013; Zhang et al., 2007). Furthermore, the concentrations of these 15 PTEs in different bedrock and geological features are summarised based on the existing literatures (Table 3). In general, low PTE concentrations have been reported in peatland (e.g. Joint Nature Conservation Committee, 2011; Young and Donald, 2013), whereas higher PTEs concentrations were observed in basalt formation (Barrat and Nesbitt 1996; Zhang et al., 2007; McIlwaine et al., 2014). This is due to the enrichment of organic matter contents in the peatland, resulting in

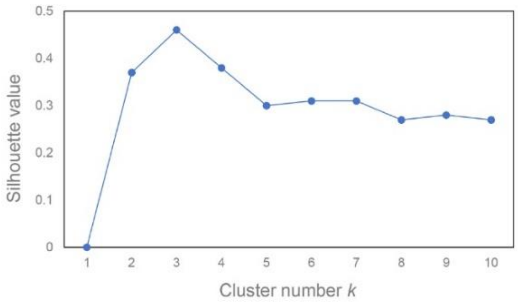


Fig. 6. Results of the silhouette value for K-means cluster analysis under different cluster number.

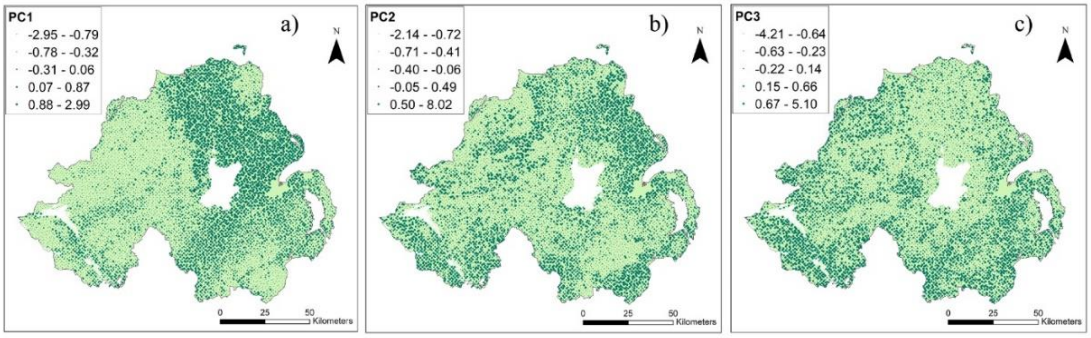


Fig. 5. Spatial distribution maps showing component scores for (a) PC 1; (b) PC 2 and (c) PC 3.

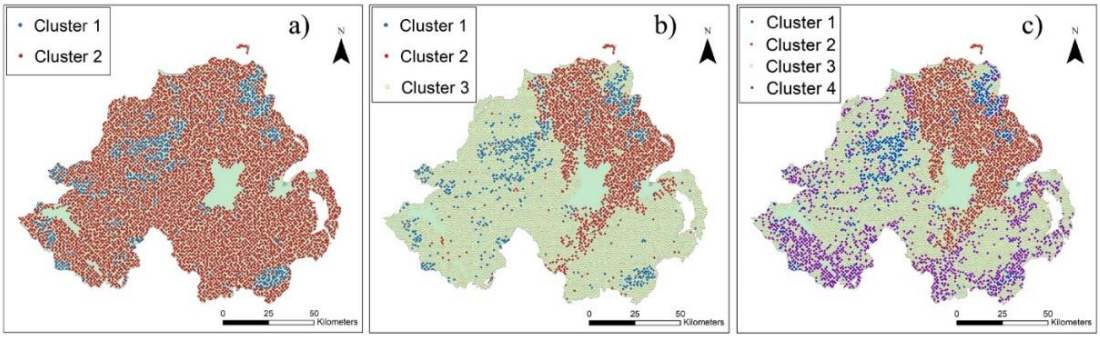


Fig. 7. Visualisation of the results with different numbers of clusters: a) 2 clusters; b) 3 clusters; c) 4 clusters.

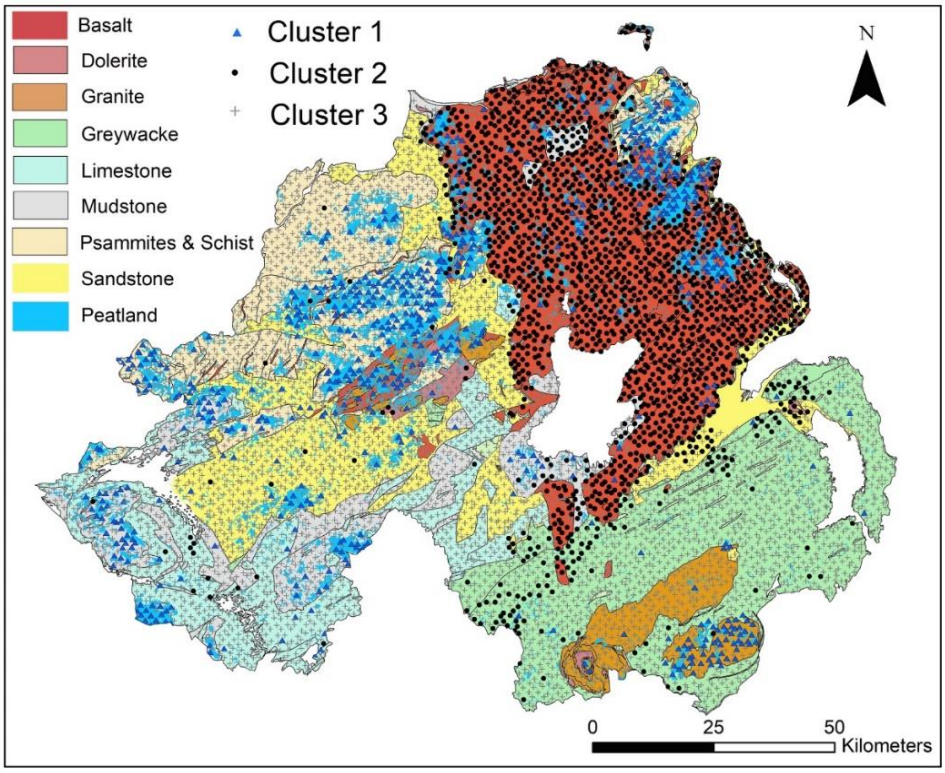


Fig. 8. K-means clustering analysis results overlaid on the simplified geology map.

lower PTE contents in topsoil. However, the elevated concentrations of Pb, Sb and Sn in the peatland reflected the association with the control of organic matter (e.g. peat, coal; Reimann et al., 2014; McIlwaine et al., 2015; Palmer et al., 2015). These three elements are reported as lower concentrations in basic rocks (e.g. basalt, limestone) in other research areas, indicating that they are more susceptible to the influence of superficial deposits as natural sources. The iron family elements Co, Ni and other metal elements Cr, Cu, Mn, V and Zn were mainly associated with the geogenic control of basalt, with elevated median concentrations comparing to other two clusters. The other elements As, Ba, Mo, and U seem to be controlled by various geological processes, with the highest median values in the third cluster. The similar conclusions in other study areas were also proposed (see Table 3).

The spatial association that extracted from the hidden spatial patterns in our study were consistent with the current knowledge, highlighting the dominant control of peat and basalt on the concentrations of topsoil PTEs in Northern Ireland. Moreover, according to the clustering results, a clear understanding of the classification for the 15 PTEs can be obtained into three groups: (a) Bi, Pb, Sn and Sb associated with peat; (b) Co, Cr, Cu, Ni, Mn, V and Zn associated with basalt; and (c) As, Ba, Mo, and U associated with other lithologies.

The PTE concentrations in soils, as well as other geochemical

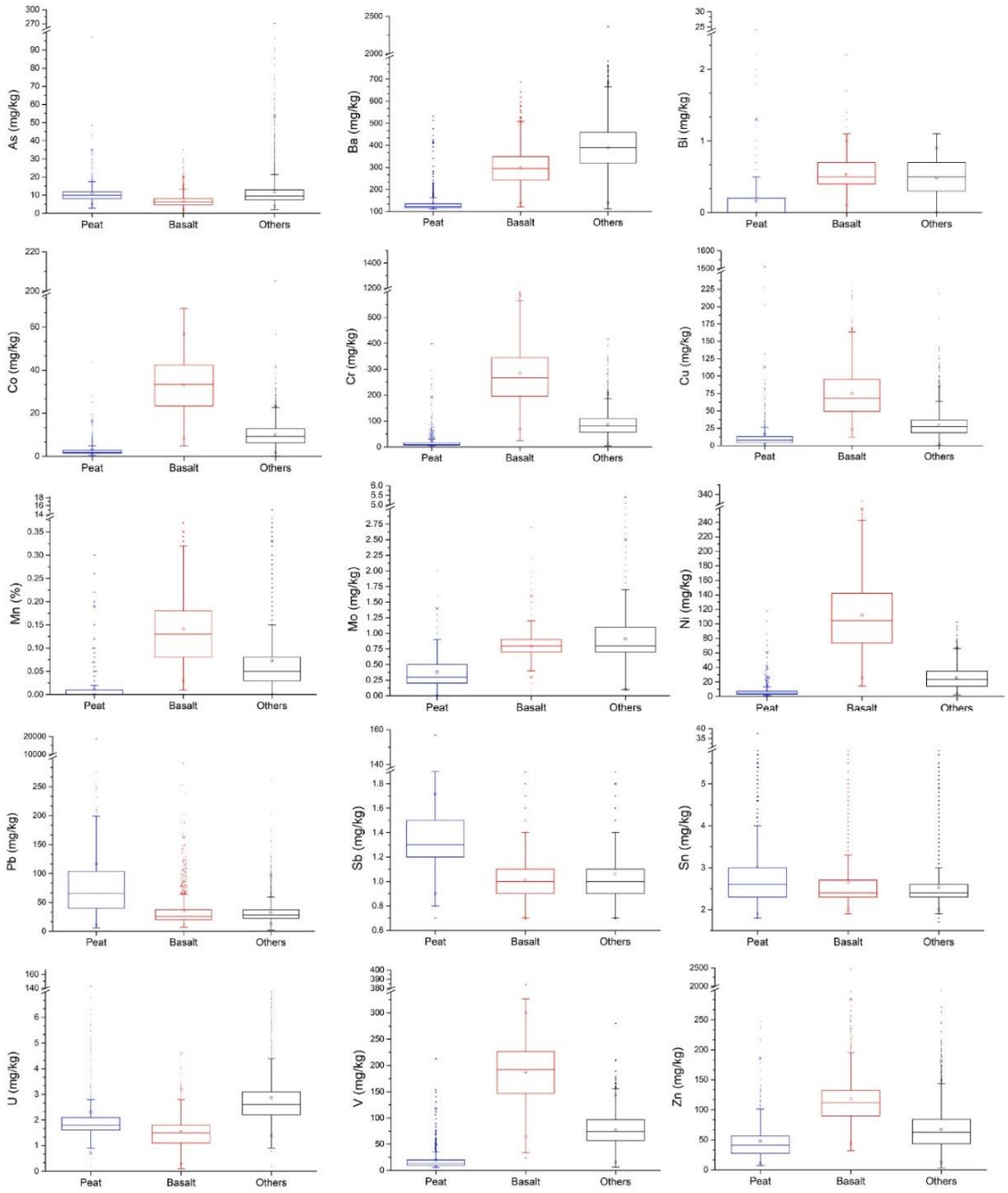


Fig. 9. Boxplots showed comparison of 15 PTEs concentrations in the three clusters.

features, are spatially continuously distributed in soils. However, the data have to be collected based on discrete sampling locations. To produce the spatially continuously distributed GIS maps for PTEs, spatial interpolation is needed based on the discrete sampling data. Therefore, the spatial distribution patterns of typical elements in each group of PTEs including Pb, Ni and Ba are displayed in Fig. 10. The concentrations of these three elements corresponded well to the boxplot results. Higher values of Pb were found in the peatland areas of the Sperrin Mountain and north-eastern parts, while lower values were found in northern areas overlaid on the basalt formation. This suggested that Pb is mainly controlled by superficial deposits rather than bedrocks. The increase of Pb concentration in the mountainous area is worth

Table 3
Summary of the existing literatures on concentrations for 15 PTEs in different bedrocks and geological features.

Element	High concentration	Low concentration	Reference
As	Greywacke (shale), mudstone, schist,	Basalt, quartzite, sandstone	Smedley and Kinniburgh, 2002; Tarvainen et al., 2013; McIlwaine et al., 2017
Ba	Carbonate, granite	Limestone, mafic (basalt)	Reimann et al., 2007; Reimann et al., 2014
Bi	Granite, shale	Sandstone	Reimann et al., 2014
Co	Greenstone, basalt,	Limestone, sandstone	Farmer, 2014; McIlwaine et al., 2014; Albanese et al., 2015
Cr	Basalt	limestone, granite	Farmer, 2014; McIlwaine et al., 2014; Albanese et al., 2015
Cu	Basalt, shale	Granite, organic matter	Wedepohl, 1978; Reimann et al., 2014; Albanese et al., 2015
Mn	Basalt	Granite, quartzite, schist	Reimann et al., 2007; Reimann et al., 2014
Mo	Granite, greywacke (shale), schist	Basalt	McIlwaine et al., 2017; Reimann et al., 2018
Ni	Basalt, shale	Granite, limestone, quartzite, sandstone	Farmer, 2014; Reimann et al., 2014; Albanese et al., 2015; Jordan et al., 2018
Pb	Granite, peat, shale	Basalt, limestone	McIlwaine et al., 2014; Reimann et al., 2014; Palmer et al., 2015
Sb	Coal, peat	Basalt, sandstone	Reimann et al., 2014; McIlwaine et al., 2015
Sn	Granite, peat, shale	Basalt, limestone	Reimann et al., 2014; McIlwaine et al., 2015
U	Granite, shale	Basalt, sandstone	Alloway, 2013; McKinley et al., 2013; Négrel et al., 2018
V	Basalt, shale	Limestone	Barsby et al., 2012; Reimann et al., 2014
Zn	Alluvium, basalt, shale	Granite	Reimann et al., 2014; McIlwaine et al., 2017

noting, which reflected the stronger control of peat in the highland rather than lowland (Young and Donald, 2013, p.27). However, higher values were also observed in eastern areas of Antrim Plateau, southern areas and Belfast metropolitan areas. This could be attributed to the influence of local mineralisation and urbanisation. Pb is well-known as an ore-forming element which is susceptible to human activities. It is also reported that petrol and vehicle emissions, as well as coal burning are the main anthropogenic factors for resulting elevated Pb values in urban and rural soil in Northern Ireland (Young and Donald, 2013). Moreover, lead pollution in urban areas is able to enrich on highland peat by atmospheric deposition (e.g. De Vleeschouwer et al., 2007), which explains the correlation between highland peat and elevated Pb concentration in Northern Ireland.

Higher concentrations of Ni were found in the basalt formation, whereas lower values were shown in other geological features. As mentioned earlier, there was strong association between Ni and basalt, reflecting the geogenic control of bedrock in Northern Ireland. Relatively homogeneous contents of geochemical elements were reported in the basalt formation (Zhang et al., 2007). The elevated concentration areas especially on the Antrim Plateau were attributed to the influence of iron-rich soil. In addition, comparatively higher concentrations were also observed in greywacke shale areas. It has been reported that greywacke shale is another dominant factor controlling the concentrations of iron family elements in Northern Ireland (Zhang et al., 2007). The concentration of Ni exhibited large and continuous patterns at the regional scale, indicating the controlling effect of basic rocks (basaltic rock and shale) on its distribution. Such geogenic control is also found on the contents of elements Cr and V (Albanese et al., 2015).

The overall concentration of Ba was low in peat and basalt areas. Elevated values were mainly observed on other geological features, especially in the western and south-eastern areas, mainly including granite, schist and greywacke shale. These patterns reflected that Ba was controlled by mixture of various geological features in Northern Ireland. The source of the anomalies in the western and southern areas was mainly related to hydrothermal fluids (Young and Donald, 2013), due to its strong correlation of gold deposit and other ore-forming elements (Lusty et al. 2009; Wang et al., 2017).

Visualising the spatial distribution for the concentrations of typical elements in the three groups can reflect the corresponding controlling factors on its entire group of PTEs, proving that the results of hot spot analysis and K-means clustering analysis were reasonable. The spatial association extracted from the clustering patterns of PTEs and soil samples strengthened existing findings in Northern Ireland (e.g. Hill et al., 2001; Zhang et al., 2007; McIlwaine et al., 2014; Albanese et al., 2015; McIlwaine et al., 2017), which provided a better understanding of sources and classification for PTEs at regional scale. Furthermore, combing other detailed supplementary data, the hidden spatial patterns can be used to explore more geochemical information (e.g. geochemical anomalies). This could be a future study direction of data analysis for the Tellus project. Future studies may also find it worthwhile to explore the relative performances of different clustering techniques such as K-means clustering, fuzzy clustering and density-based clustering algorithms into the spatial data sets, however this is outside the scope of the present work.

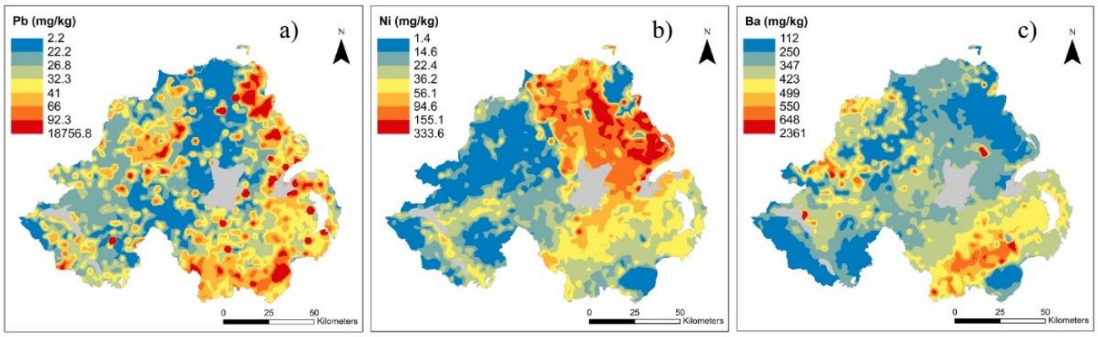


Fig. 10. Spatial distribution maps of typical elements of three groups: a) Pb; b) Ni and c) Ba.

4. Conclusion

This study investigated the hidden spatial patterns for 15 PTEs and topsoil samples in Northern Ireland using Hot spot analysis and K-means clustering analysis. The hot spot analysis results revealed different spatial clustering patterns for the 15 PTEs, showing clear association with different geological features, especially peat and basalt. Peat is associated with high concentrations of Bi, Pb, Sb and Sn, while basalt is associated with high concentrations of Co, Cr, Cu, Mn, Ni, V and Zn. The high concentrations of As, Ba, Bi, Mo and U are associated with mixture of various lithologies (e.g. schist, greywacke shale, sandstone, limestone and granite), indicating the complicated controlling effects on them. In addition, K-means clustering results revealed three hidden patterns in 6,862 soil samples, showing spatial overlay relationships with controlling factors on the simplified geology map. The samples in the first, second and third clusters were overlaid on peatland, basalt formation and other lithologies, respectively. The results of two spatial clustering techniques were consistent with each other, highlighting the major controlling effects of peat and basalt for both PTEs and soil samples. Furthermore, the boxplot results indicated the differences among the three clusters were significant for all 15 PTEs, which confirmed the accuracy and rationality of our clustering results in this study.

Our results revealed hidden spatial patterns in a study area that have been widely studied. These spatial patterns and association enhanced the current knowledge of the controlling factors on the selected 15 PTEs in the topsoil of Northern Ireland. Moreover, this study provides a clear demonstration on the efficiency of spatial machine learning techniques in discovering hidden spatial patterns and extract geochemical association in the multivariate datasets, which can be applied for environmental study in other unexplored areas.

CRediT authorship contribution statement

Haofan Xu: Conceptualization, Formal analysis, Data curation, Methodology, Software, Validation, Visualization, Writing - original draft. **Peter Croot:** Writing - review & editing. **Chaosheng Zhang:** Data curation, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

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

References

Aelion, C.M., Davis, H.T., McDermott, S., Lawson, A.B., 2009. Soil metal concentrations and toxicity: associations with distances to industrial facilities and implications for human health. Sci. Total Environ. 407, 2216–2223.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Wiley, New York.

Ajmone-Marsan, F., Biasioli, M., Kralj, T., Grcman, H., Davidson, C.M., Hursthouse, A.S., Madrid, L., Rodrigues, S., 2008. Metals in particle-size fractions of the soils of five European cities. Environ. Pollut. 152, 73–81.

Albanese, S., Sadeghi, M., Lima, A., Cicchella, D., Dinelli, E., Valera, P., Falconi, M., Demetriades, A., De Vivo, B., The GEMAS Project Team, 2015. GEMAS: cobalt, Cr, Cu and Ni distribution in agricultural and grazing land soil of Europe. J. Geochem. Explor. 154, 81–93.

Alizadeh, M.J., Shahheydari, H., Kavianpour, M.R., Shamloo, H., Barati, R., 2017. Prediction of longitudinal dispersion coefficient in natural rivers using a cluster-based Bayesian network. Environ. Earth Sci. 76 (2), 86.

Alloway, B.J., 2013. Bioavailability of Elements in Soil. In: Selinus, O. (Ed.), Essentials of Medical Geology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4375-5_15.

Argyraki, A., Kelepertzis, E., 2014. Urban soil geochemistry in Athens, Greece: the importance of local geology in controlling the distribution of potentially harmful trace elements. Sci. Total Environ. 482–483, 366–377.

Bagstad, K.J., Semmens, D.J., Ancona, Z.H., Sherrouse, B.C., 2017. Evaluating alternative methods for biophysical and cultural ecosystem services hotspot mapping in natural resource planning. Landsc. Ecol. 32 (1), 77–97.

Barrat, J.A., Nesbitt, R.W., 1996. Geochemistry of the tertiary volcanism of Northern Ireland. Chem. Geol. 129, 15–38.

Barsby, A., McKinleya, J.M., Ofterdinger, U., Young, M., Cave, M., Wraggd, J., 2012. Bioaccessibility of trace elements in soils in Northern Ireland. Sci. Total Environ. 433, 398–417.

Bengio, Y., 2013. Deep learning of representations: looking forward. In: International Conference on Statistical Language and Speech Processing. Springer, Berlin, Heidelberg, pp. 1–37.

Bhowmik, A.K., Alamdar, A., Katsoyiannis, I., Shen, H., Ali, N., Ali, S.M., Bokhari, H., Schäfer, R.B., Eqani, S., 2015. Mapping human health risks from exposure to trace metal contamination of drinking water sources in Pakistan. Sci. Total Environ. 538, 306–316.

Birke, M., Reimann, C., Rauch, U., Ladenberger, A., Demetriades, A., Jähne-Klingberg, F., Oorts, K., Gosar, M., Dinelli, E., Halamić, J., 2017. GEMAS: Cadmium distribution and its sources in agricultural and grazing land soil of Europe — Original data versus clr-transformed data. J. Geochem. Explor. 173, 13–30.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer. ISBN 978-0-387-31073-2.

Boente, C., Albuquerque, M.T.D., Fernández-Braña, A., Gerassis, S., Sierra, C., Gallego, J. R., 2018. Combining raw and compositional data to determine the spatial patterns of Potentially Toxic Elements in soils. Sci. Total Environ. 631–632, 1117–1126. https://doi.org/10.1016/j.scitotenv.2018.03.048.

Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V., (Eds.) 2006. Compositional data analysis in the geosciences – from theory to practice. Geological Society of London, Special Publication 264.

Buccianti, A., Lima, A., Albanese, S., Cannatelli, C., Esposito, R., Vivo, B., 2015. Exploring topsoil geochemistry from the CoDA (Compositional Data Analysis) perspective: The multi-element data archive of the Campania Region (Southern Italy). J. Geochem. Explor. 159, 302–316.

Carroll, J.B., 1953. An analytic solution for approximating simple structure in factor analysis. Psychometrika 18, 23–38.

Cattell, R.B., 1966. The scree test for the number of factors. Multivar. Behav. Res. 1 (2), 245–276.

Cheng, Q., Jing, L., Panahi, A., 2006. Principal component analysis with optimum order sample correlation coefficient for image enhancement. Int. J. Remote Sens 27 (16), 3387–3401.

Cloquet, C., Carignan, J., Libourel, G., 2006. Isotopic composition of Zn and Pb atmospheric depositions in an urban/periurban area of northeastern France. Environ. Sci. Technol. 40, 6594–6600.

Coggins, A.M., Jennings, S.G., Ebinghaus, R., 2006. Accumulation rates of the heavy metals lead, mercury and cadmium in ombrotrophic peatlands in the west of Ireland. Atmos. Environ. 40, 260–278.

Dalradian, 2019. Making the most of County Tyrone's gold deposits. Available at: https://www.newsletter.co.uk/business/making-the-most-of-county-tyrone-s-gold-deposits-1-9081043.

Dao, L.G., Morrison, L., Zhang, H., Zhang, C., 2014. Influences of traffic on Pb, Cu and Zn concentrations in roadside soils of an urban park in Dublin, Ireland. Environ. Geochem. Health 36, 333–343.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1 (4), 224–227.

Davies, H., Walker, S., 2013. Strategic planning policy statement (SPPS) for Northern Ireland: Strategic Environmental Assessment (SEA) Scoping report. Leeds.

Davis, J.C., 2002. Statistics and Data Analysis in Geology, 3rd ed. John Wiley & Sons Inc., New York.

Davis, H.T., Marjorie Aelion, C., McDermott, S., Lawson, A.B., 2009. Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. Environ. Pollut. 157, 2378–2385.

Delbecque, N., Verdoodt, A., 2016. Spatial patterns of heavy metal contamination by urbanization. J. Environ. Qual. 45, 9–17.

De Vleeschouwer, F., Gérard, L., Goormaghtigh, C., Mattielli, N., Le Roux, G., Fagel, N., 2007. Atmospheric lead and heavy metal pollution records from a Belgian peat bog spanning the last two Millenia: Human impact on a regional to global scale. Sci. Total Environ. 377, 282–295.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

Ettler, V., Sebek, O., Grygar, T., Klementova, M., Bezdicka, P., Slavikova, H., 2008. Controls on metal leaching from secondary Pb smelter air-pollution-control residues. Environ. Sci. Technol. 42, 7878–7884.

Fan, C., Cui, Z., Zhong, X., 2018. House prices prediction with machine learning algorithms. In: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 6–10.

Faria, P.B.F., He, Z.L., Stoffella, P.J., Montes, C.R., Melfi, A.J., Baligar, V.C., 2012. Nutrients and nonessential elements in soil after 11 years of wastewater irrigation. J. Environ. Qual. 41, 920–927.

Farmer, G.L., 2014. Continental basaltic rocks. In: Chapter 4.3 in R.L. Rudnick, H. Holland, K. Turekian (Eds.), The Crust, 2nd ed., Treatise on Geochemistry, no. 4, pp. 75–100.

Fei, X., Christakos, G., Xiao, R., Ren, Z., Liu, Y., Lv, X., 2019. Improved heavy metal mapping and pollution source apportionment in Shanghai City soils using auxiliary information. Sci. Total Environ. 661, 168–177.

Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Sci. Total Environ. 408, 4230–4238.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3), 189–206.

Ghezelbash, R., Maghsoudi, A., Carranza, E.J.M., 2020. Optimization of geochemical anomaly detection using a novel genetic K-means clustering (GKMC) algorithm. Comput. Geosci. 134, 104335.

Goutte, C., Hansen, L.K., Liptrot, M.G., Rostrup, E., 2001. Feature-space clustering for fMRI meta-analysis. Hum. Brain Mapp. 13 (3), 165–183.
GSNI, 1998. The solid geology of Northern Ireland: a vector map at 1:250,000 scale. Geological Survey of Northern Ireland, Belfast.
Han, J., Kamber, M., 2006. Data Mining, Concepts and Techniques. Morgan Kaufman Publishers, San Francisco, USA.
Harman, H.H., 1976. Modern Factor Analysis, 3rd ed. University of Chicago Press, Chicago.
Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. J. Roy. Stati. Soci. Ser. C (Appl. Stat.) 28, 100–108.
Hendrickson, A.E., White, P.O., 1964. PROMAX: a quick method for rotation to oblique simple structure. Brit. J. Stat. Psychol. 17, 65–70.
Hill, I.G., Worden, R.H., Meighan, I.G., 2001. Formation of interbasaltic laterite horizons in NE Ireland by early tertiary weathering processes. Proc. Geol. Assoc. 112 (4), 339–348.
Joint Nature Conservation Committee, 2011. Towards an assessment of the state of UK Peatlands, JNCC report No. 445.
Jolliffe, I.T., 2002. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed. Springer, New York, p. 487.
Jordan, C., Zhang, C.S., Higgins, A., 2007. Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland. J. Geochem. Explor. 93, 135–152.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science 349 (6245), 255–260.
Jordan, G., Petrik, A., De Vivo, B., Albanese, S., Demetriades, A., Sadeghi, M., Team, T.G. P., 2018. GEMAS: spatial analysis of the Ni distribution on a continental-scale using digital image processing techniques on European agricultural soil data. J. Geochem. Explor. 186, 143–157. https://doi.org/10.1016/j.gexplo.2017.11.011.
Kaiser, H.F., 1958. The Varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200.
Kanevskij, M., Pozdnoukhov, A., Timonin., V., 2009. Machine Learning for Spatial Environmental Data: Theory, Applications and Software. Epfl Press, Lausanne.
Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7), 881–892.
Kassambara, A., Mundt, F., 2017. Package 'factoextra'. Extract and visualize the results of multivariate data analyses.
Kelepertsis, A., Argyraki, A., Alexakis, D., 2006. Multivariate statistics and spatial interpretation of geochemical data for assessing soil contamination by potentially toxic elements in the mining area of Stratoni, North Greece. Geochem-Explor. Env. A. 6, 349–355.
Kelepertzis, E., Argyraki, A., Botsou, F., Aidona, E., Szabo, A., Szabo, C., 2019. Tracking the occurrence of anthropogenic magnetic particles and potentially toxic elements (PTEs) in house dust using magnetic and geochemical analyses. Environ. Pollut. 245, 909–920.
Ketchen Jr., D.J., Shook, C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manag. J. 17 (6), 441–458.
Klapstein, S.J., Walker, A.K., Saunders, C.H., Cameron, R.P., Murimboh, J.D., O'Driscoll, N.J., 2020. Spatial distribution of mercury and other potentially toxic elements using epiphytic lichens in Nova Scotia. Chemosphere 241, 125064.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Li, J., Heap, A., 2011. A review of comparative studies of spatial interpolation methods: performance and impact factors. Ecol. Inform. 3–4, 228–241.
Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26 (12), 1647–1659.
Ludwig, M., Morgenthal, T., Detsch, F., Higginbottom, T.P., Lezama Valdes, M., Nauß, T., Meyer, H., 2019. Machine learning and multi-sensor based modelling of woody vegetation in the Molopo Area, South Africa. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2018.12.019.
Lusty, P.A.J., McDonnell, P.M., Gunn, A.G., Chacksfield, B.C., Cooper, M., 2009. Gold potential of the dalradian rocks of north-west Northern Ireland: Prospectivity analysis using tellus data. British Geological Survey Internal Report OR/08/39:74 pp.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, CA, pp. 281–297.
McKinley, J.M., Grunsky, E., Mueller, U., 2018. Environmental Monitoring and Peat Assessment Using Multivariate Analysis of Regional-Scale Geochemical Data. Math. Geosci. 50, 235–246. https://doi.org/10.1007/s11004-017-9686-x.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0. Available at: https://cran.r-project.org/web/packages/cluster/cluster.pdf. [Accessed date: 08/12/2020].
Marchant, B.P., Tye, A.M., Rawlins, B.G., 2011. The assessment of point-source and diffuse soil metal pollution using robust geostatistical methods: a case study in Swansea (Wales, UK). Eur. J. Soil Sci. 62, 346–358.
McIlwaine, R., Cox, S., Doherty, R., Palmer, S., Ofterdinger, U., McKinley, J., 2014. Comparison of methods used to calculate typical threshold values for potentially toxic elements in soil. Environ. Geochem. Health 36, 953–971.
McIlwaine, R., Cox, S.F., Doherty, R., 2015. When are total concentrations not total? Factors affecting geochemical analytical techniques for measuring element concentrations in soil. Environ. Sci. Pollut. Res. 22, 6364–6371. https://doi.org/10.1007/s11356-015-4204-5.
McIlwaine, R., Doherty, R., Cox, S.F., Cave, M., 2017. The relationship between historical development and potentially toxic element concentrations in urban soils. Environ. Pollut. 220, 1036–1049.
McKinley, J.M., Ofterdinger, U., Young, M., Barsby, A., Gavin, A., 2013. Investigating local relationships between trace elements in soils and cancer data. Spat. Stat. 5, 25–41.
Meng, Y., Cave, M., Zhang, C., 2020. Identifying geogenic and anthropogenic controls on different spatial distribution patterns of aluminium, calcium and lead in urban topsoil of Greater London Authority area. Chemosphere 238, 124541.
Meshkani, S.A., Mehrabi, B., Yaghubpur, A., Alghalandis, Y.F., 2011. The application of geochemical pattern recognition to regional prospecting: A case study of the Sanandaj-Sirjan metallogenic zone, Iran. J. Geochem. Explor. 108 (3), 183–195.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw. 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.
Mitchell, W.I. (Ed.), 2004. The Geology of Northern Ireland: Our Natural Foundation, 2nd ed. Geological Survey of Northern Ireland, Belfast.
Négrel, P., De Vivo, B., Reimann, C., Ladenberger, A., Cicchella, D., Albanese, S., Birke, M., De Vos, W., Dinelli, E., Lima, A., O'Connor, P.J., Salpeteur, I., Tarvainen, T., the GEMAS Project Team., 2018. U-Th signatures of agricultural soil at the European continental scale (GEMAS): distribution, weathering patterns and processes controlling their concentrations. Sci. Total Environ. 622–623, 1277–1293.
Okorie, A., Entwistle, J., Dean, J.R., 2011. The application of in vitro gastrointestinal extraction to assess oral bioaccessibility of potentially toxic elements from an urban recreational site. Appl. Geochem. 26, 789–796.
Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27 (4), 286–306.
Palmer, S., Ofterdinger, U., McKinley, J.M., Cox, S., Barsby, A., 2013. Correlation analysis as a tool to investigate the bioaccessibility of Nickel, Vanadium and Zinc in Northern Ireland Soils. Environ. Geochem. Health 35, 569–584.
Palmer, S., McIlwaine, R., Ofterdinger, U., Cox, S.F., Mckinley, J.M., Doherty, R., Wragg, J., Cave, M., 2015. The effects of lead sources on oral bioaccessibility in soil and implications for contaminated land risk management. Environ. Pollut. 198, 161–171.
Petrik, A., Albanese, S., Lima, A., De Vivo, B., 2018. The spatial pattern of beryllium and its possible origin using compositional data analysis on a high-density topsoil data set from the Campania Region (Italy). Appl. Geochem. 91, 162–173. https://doi.org/10.1016/j.apgeochem.2018.02.008.
Rahmati, O., Falah, F., Dayal, K.S., Deo, R.C., Mohammadi, F., Biggs, T., Moghaddam, D. D., Naghibi, S.A., Bui, D.T., 2020. Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. Sci. Total Environ. 699, 134230.
Reimann, C., Filzmoser, P., 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ. Geol. 39 (9), 1001–1014.
Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 17, 185–206.
Reimann, C., Arnoldussen, A., Englmaier, P., Filzmoser, P., Finne, T.E., Garrett, R.G., Koller, F., Nordgulen, O., 2007. Element concentrations and variations along a 120-km transect in southern Norway – anthropogenic vs. geogenic vs. biogenic element sources and cycles. Appl. Geochem. 22, 851–871.
Reimann, C., Filzmoser, P., Garrett, G.R., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons Ltd., p. 359
Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., Connor, P.O', 2014. Chemistry of Europe's Agricultural Soils, Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover.
Reimann, C., Fabian, K., Birke, M., Filzmoser, P., Demetriades, A., Négrel, P., Oorts, K., Matschullat, J., de Caritat, P., The GEMAS Project Team, 2018. GEMAS: Establishing geochemical background and threshold for 53 chemical elements in European agricultural soil. Appl. Geochem. 88, 302–318.
Rodrigues, S., Urquhart, G., Hossack, I., Pereira, M.E., Duarte, A.C., Davidson, C., Hursthouse, A., Tucker, P., Roberston, D., 2009. The influence of anthropogenic and natural geochemical factors on urban soil quality variability: a comparison between Glasgow, UK and Aveiro, Portugal. Environ. Chem. Lett. 7, 141–148.
Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Saaltink, R., Griffioen, J., Mol, G., Birke, M., 2014. Geogenic and agricultural controls on the geochemical composition of European agricultural soils. J. Soil. Sediment. 14 (1), 121–137.
Sergeev, A.P., Buevich, A.G., Baglaeva, E.M., Shichkin, A.V., 2019. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena 174, 425–435.
Smedley, P.L., Kinniburgh, D.G., 2002. A review of the source, behaviour and distribution of arsenic in natural waters. Appl. Geochem. 17, 517–568.
Smyth, D., 2007. Methods used in the Tellus geochemical mapping of Northern Ireland. British geological survey open report, 89 pp.
Tarvainen, T., Albanese, S., Birke, M., Poňavič, M., Reimann, C., 2013. Arsenic in agricultural and grazing land soils of Europe. Appl. Geochem. 28, 2–10.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. Appl. Geochem. 23 (8), 2198–2213.
Tepanosyan, G., Lilit, S., Nairuhi, M., Armen, S., 2020. Combination of Compositional Data Analysis and Machine Learning Approaches to Identify Sources and Geochemical Associations of Potentially Toxic Elements in Soil and Assess the Associated Human Health Risk in a Mining City. Environ. Pollut. 261, 114210.

Thiombane, M., Martin-Fernández, J.A., Albanese, S., Lima, A., Doherty, A., De Vivo, B., 2018. Exploratory analysis of multi-element geochemical patterns in soil from the Sarno River Basin (Campania region, southern Italy) through compositional data analysis (CODA). J. Geochem. Explor. 195, 110–120. https://doi.org/10.1016/j.gexplo.2018.03.010.
Tipping, E., Lawlor, A., Lofts, S., Shotbolt, L., 2006. Simulating the long-term chemistry of an upland UK catchment: heavy metals. Environ. Pollut. 141, 139–150.
Wang, J., Zuo, R., Caers, J., 2017. Discovering geochemical patterns by factor-based cluster analysis. J. Geochem. Explor. 181, 106–115.
Weatherill, G., Burton, P.W., 2008. Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. Geophys. J. Int. 176 (2), 565–588.
Wedepohl, K.H., 1978. Handbook of Geochemistry. Springer-Verlag, Berlin-Heidelberg.
Xie, X., Liu, D., Xiang, Y., Yan, G., Lian, C., 2004. Geochemical blocks for predicting large ore deposits — concept and methodology. J. Geochem. Explor. 84, 77–91.
Xu, H.F., Demetriades, A., Reimann, C., Jiménez, J.J., Filser, J., Zhang, C.S., 2019. Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data. Sci. Total Environ. 678, 94–104.
Yeung, K., Ruzzo, W., 2001. An empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774.
Young, M.E., Donald, A.W. (Eds.), 2013. A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast.
Yuan, Y.M., Cave, M., Xu, H.F., Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater. 393 (5), 122377.
Zagouras, A., Kazantzidis, A., Nikitidou, E., Argiriou, A.A., 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. Sol. Energy 97 (5), 1–11.
Zhang, C.S., Manheim, F.T., Hinde, J., Grossman, J.N., 2005. Statistical characterization of a large geochemical database and effect of sample size. Appl. Geochem. 20, 1857–1874.
Zhang, C.S., 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ. Pollut. 142, 501–511.
Zhang, C.S., Jordan, C., Higgins, A., 2007. Using neighbourhood statistics and GIS to quantify and visualize spatial variation in geochemical variables: An example using Ni concentrations in the topsoils of Northern Ireland. Geoderma 137, 466–476.
Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008a. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146, 378–390.
Zhang, C.S., Luo, L., Xu, W., Ledwith, V., 2008b. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci. Total Environ. 398 (1–3), 212–221.
Zhou, S.G., Zhou, K.F., Wang, J.L., Wang, S.S., 2018. Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. Front. Earth Sci. 12, 491–505.
Zuo, R.G., 2017. Machine Learning of Mineralization-Related Geochemical Anomalies: A Review of Potential Methods. Nat. Resour. Res. 26, 457–464.
Zuo, R.G., Xiong, Y., Wang, J., Carranza, E.J.M., 2019. Deep learning and its application in geochemical mapping. Earth Sci. Rev. 192, 1–14.

4.4 Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient

Xu, H.F., Croot, P., Zhang, C.S., 2021. Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient. (Under review).

Summary: This paper investigated the spatial relationships between Pb and Al in the topsoil samples that collected from currently available Tellus data set in the northern half of Ireland using GWPCC. Both positive and negative correlation coefficients were observed, suggesting the existence of spatially varying relationships between Pb and Al concentrations. The 'special' negative correlations were observed in more than 35% of the whole study area, mainly clustered in the northern-western and north-eastern of Ireland. The positive correlations were observed in the central-western and midlands. Mixed relationships of both negative and positive correlations occurred in the eastern coastal areas. The majority of negative correlation patterns showed clear association with blanket peat, which can be attribute to long-distance transportation of Pb from atmospheric deposition. Moreover, the weakened the relationships in the eastern coastal areas indicated the influences related to anthropogenic activities. Our results demonstrated the efficiency of GWPCC in exploring the spatially varying relationships between environmental variables and identifying association with influencing factors, which could be hardly achieved by traditional techniques.

My contribution in this paper accounted for ~90% in reviewing literatures, exploring data and writing manuscript.

1. Introduction

Potentially toxic elements (PTEs) are usually found at trace levels. Although some PTEs (e.g., Cu and Zn) are regarded as essential elements, the presence of other PTEs can be highly toxic to plants and organisms, such as Cd, Cr and Pb (Kabata-Pendias, 2004; Hooda, 2010; Zeng et al., 2011). Under natural conditions, the concentrations of PTEs are influenced by soil-forming parent material and processes (Alloway, 1995; Aelion et al., 2009). However, since the industrial revolution, human activities have greatly increased the input of PTEs (Biney et al., 1994). The sources of anthropogenic pollution include traffic emission, fossil fuel burning, metalliferous industries, construction, medical and electronic waste (Nriagu and Pacyna, 1988; Zhang, 2006; Wu et al., 2019; Zhao et al., 2019), which are directly discharged into soil, water and air. Soil is regarded as the most important sink of PTEs (Wong et al., 2006), receiving pollution from both surface disposal and atmospheric deposition. Excessive levels of PTEs in the soil will increase the risk of ingesting toxic metals into human body through food chain or soil dust, especially for children (Odukoya et al., 2000). Among all the PTEs in the soil, the spatial distribution and variation of Pb are of particular concern in environmental studies (Nriagu, 1983). This is not only because of the adverse effects of Pb on human health, but also that its concentration in the soil is strongly interfered by anthropogenic factors. Previous studies have widely reported the abnormally elevated concentrations of Pb in soil caused by human activities in urban and industrial areas (e.g., McGrath et al., 2004; Appleton et al., 2013; Li et al., 2014; Liu et al., 2015; Marrugo-Negrete et al., 2017). In addition, the previous use of leaded gasoline and traffic emission in urban areas (e.g., phased out in Germany and the USA in 1996) caused significant Pb pollution in the air, which in turn polluted rural soil through atmospheric deposition (Shotyk, 2002; Novák et al., 2003). Therefore, it is challenging to identify the sources of Pb to prevent the spread of soil contamination and maintain the sustainable development at the regional level.

Conventional statistical analysis and multivariate analysis have been widely applied to identify the potential sources of Pb contamination in the topsoil of urban and rural areas

(e.g., Madrid et al., 2002; Zhang et al., 2007; Acosta et al., 2011; Bhowmik et al., 2015; Meng et al., 2020). The problem, however, is that these traditional techniques (e.g., ordinary linear regression) assume the studied relationship between variables is linear and spatially constant across the space, so the parameter estimation remains the same for the whole study area (Guo et al., 2008). Due to the complexity of soil properties and disturbances of intensive anthropogenic activities, the pollution sources on the spatial variation of Pb are controlled by a mixture of multiple factors (Franco-Uria et al., 2009; Martín et al., 2013). Thus, the relationships between Pb and other geochemical elements or environmental variables may be spatially varying at different locations. The traditional statistics, which should be regarded as global techniques, are likely to mask the spatially varying relationships due to the neglect of spatial heterogeneity (Su et al., 2012). In recent years, the concept of spatially varying relationships was proposed to study the concealed patterns between geochemical elements and environmental variables (Tu and Xia, 2008; Li et al., 2017; Yuan et al., 2020; Yang et al., 2020; Ballard and Bone, 2021; Xu and Zhang, 2021), which has been proved as an effective way to identify the association with related influencing factors from the spatial perspective.

In light of this, the use of advanced spatial techniques that consider local statistics such as geographically weighted regression (GWR) is more appropriate to investigate the soil Pb pollution by capturing the spatially varying relationships (Brunsdon et al., 1996; Fotheringham et al., 1998). This technique is an extension of traditional ordinary linear regression, which can generate local regression coefficients at each sample point (Fotheringham et al., 2001; Fotheringham et al., 2002). However, it has been proposed that the local coefficients of GWR can only represent the ‘slope’ coefficients between the dependent and independent variables rather than correlation (Gao and Li, 2011; Xu and Zhang, 2021). Therefore, an improved method called Geographically Weighted Pearson Correlation Coefficients (GWPCC) was performed in this study to investigate the spatial correlations between Pb and aluminium (Al) concentration in the topsoil of Ireland. This technique is a combination of traditional Pearson correlation coefficient and geographically

weighted (GW) framework (Kalogirou, 2014), which can identify the strong and weak correlations between input parameters at each sample point.

The reason to choose Al for comparison is that these two elements are generally reported to maintain a positive correlation under most natural conditions (Schropp and Windom, 1988), which could be expected for soils derived from continental crust (Walsh and Barry, 1957). In addition, the element Al is a basic constituent of silicate clays and Pb can not only be adsorbed to clay but it is also present in primary silicates as K-feldspar and mica (Spark, 2010). It is a conservative lithogenic element and often used as reference element (Shotyk et al., 2002; Sezgin et al., 2003; Le Roux et al., 2004), which is chemically stable and its fate in the environment media is not easily affected by human activities. Relevant references with statistical analyses were summarised as examples in Supplemental Table S1 (Zhang et al., 2008a; Shaheen, 2009; Vasić et al., 2012; Guo et al., 2019; Zhang et al., 2019). However, in some limited cases, the contradictory result of negative correlation was also recorded (El Bilali et al., 2002; Zhang et al., 2014), while this 'special' relationship was reported to be an implication with different pollution sources in their studies. Considering the existing literature has not deeply explored the contradictory relationships between these two variables, not only in Ireland but also in other study areas over the world. Therefore, the exploration of the spatially varying relationships between Pb and Al concentration based on the current available Tellus data sets seems to be an interesting and important topic in environmental studies. As it is a promising way to identify the potential influencing factors on Pb in the topsoil, which can provide enhanced understanding of spatial variation of PTEs for current literature. In this case, it needs to be acknowledged that we do not attempt to explore all the influencing factors of Pb in this study and the conventional exploration of multivariate relationships is out of the focus of this study.

The objectives of this study are: (1) to investigate the spatial relationships between Pb and Al concentrations using GWPCC based on the currently available Tellus data set in the topsoil of northern half of Ireland; (2) to identify the spatial associations with different

influencing factors from the local correlation patterns; (3) to further explore the underlying mechanisms between the 'special' negative correlation and potential pollution sources on the Pb distribution.

2. Materials and methods

2.1 Soil sampling and analyses

The Tellus project is a collaboration of national project to collect geophysical and geochemical data across Ireland. It is undertaken by Geological Survey Ireland (GSI) and Geological Survey of Northern Ireland (GSNI) in the part of Republic of Ireland and Northern Ireland, respectively. A total of 17,798 topsoil samples (surface to 20 cm depth) were used in this study, covering the northern half of Ireland (Fig. 1). Each sample was taken as composite sample from five sub-sites. The sampling density averaged one site per 4 km^2 in the Republic of Ireland and per 2 km^2 in Northern Ireland, respectively, and was increased to one site per 2 km^2 in the urban areas of Galway and Dublin in the Republic of Ireland. Samples were collected in paper bags and air-dried initially before further preparation process.

Then, all samples were sieved through a 2 mm pore size nylon mesh to remove stones and plant roots, and the repetition was prepared by shallow-splitting of duplicate samples to create the quality control (QC) samples. After sample preparation, the geochemical composition was analysed in the laboratory by Inductively Coupled Plasma (ICP-OES/-MS) method following aqua regia digestion with a series of strict QCs during the analytical process (e.g., randomisation of sample IDs; blind insertion of internal or secondary reference materials). Detailed QC process can be found in Knights (2013) and Young and Donald (2013). More specific details on the sampling program, including protocols, and all data are publicly available from the Geologic Survey of Ireland (https://www.gsi.ie/tellus).

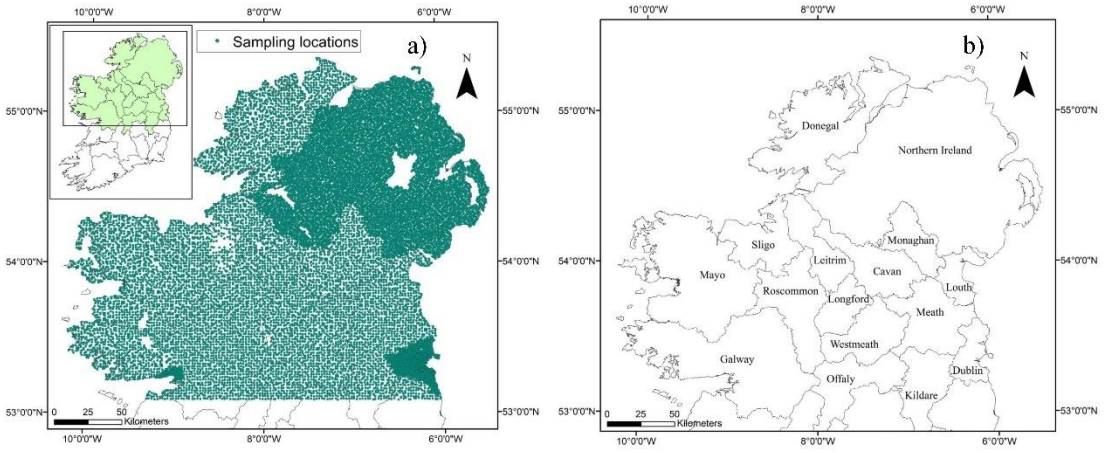


Figure 1. Spatial distribution maps showing: a) sampling locations (n = 17,798) and b) county names in the study area.

2.2. Geological background of study area

The simplified bedrock map of the study area was classified based on the bedrock unit map from GSI (McConnell and Gately, 2006), mainly consisting of basalt, clay, granite, greywacke shale, limestone, sandstone and schist (Fig. 2a). In addition, peat is reported to be regarded as the major soil subgroup covering the bedrock and is related to the elevated concentrations of Pb in the topsoil of Ireland (Davies and Walker, 2013; Xu et al., 2021), which should be considered separately from other types of soil. There are two major types of peat including blanket peat and basin peat in Ireland (Rosca et al., 2018). The blanket peat is mostly concentrated in the mountains of the north-eastern and western coastal of Ireland, while basin peat is mainly distributed in the central part of midlands of Ireland.

The other controlling factors on Pb distribution need to be noted are the existence of mineralised areas and urban areas. It was reported that Ireland hosts the one of the world's major Pb-Zn deposits (Banks et al. 2002; Lusty et al., 2012), and has a long mining history can be traced back to Bronze Age with approximately 450 mining locations recorded (Stanley et al., 2009). There are two major urban areas, including Greater Dublin area in the central-east and Belfast Metropolitan area in the north-east. Considering the potential

influencing factors except bedrocks on the Pb concentration, the spatial locations for peat, urban areas and lead deposits are displayed in Fig. 2b.

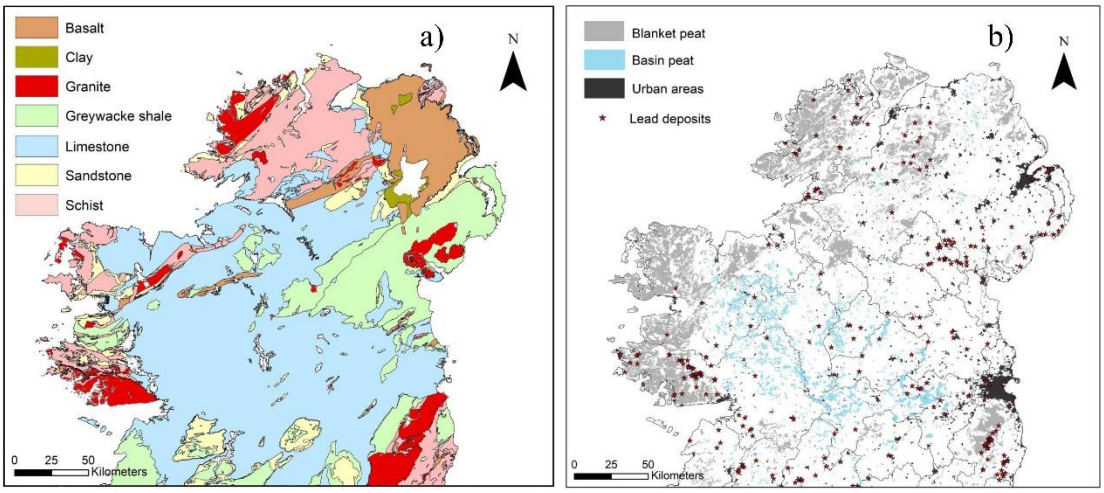


Figure 2. Maps showing background of study area: a) simplified bedrock map (original 1:500,000 shapefile from GSI, 2006); b) spatial distribution of locations for peatland, urban areas and recorded lead deposits.

2.3. Geographically Weighted Pearson Correlation Coefficient (GWPCC)

Since the first proposed in 1990s, the GWR has received extensive attention from environmental studies due to its powerful function of exploring spatial non-stationarity and spatially varying relationships (Brunsdon et al., 1996; Fotheringham et al., 2001; Páez et al., 2011; Oshan and Fotheringham, 2018). In addition, a variety of global statistical models are extended or improved to explore local parameters based on the GW framework such as Geographically Weighted Logistic Regression (GWLR); Geographically Weighted Lasso (GWL), Geographically Weighted Principal Component Analysis (GWPCA), Geographically Weighted Pearson Correlation Coefficients (GWPCC) and Multiscale Geographically Weighted Regression (MGWR), etc. (Atkinson et al., 2003; Wheeler, 2009; Harris et al., 2011; Kalogirou, 2014; Fotheringham et al., 2017).

The GWPCC is an extension of traditional Pearson correlation coefficient (PCC) which adopts the concept of geographical weights around observations for calculating local statistics (Fotheringham et al., 2002; Kalogirou, 2012). The traditional statistical analysis, including PCC, is regarded as a global statistic that assumes the correlation between two variables are spatially constant and remain the same in the whole study area (Tu and Xia, 2008), and thus cannot explore how the relationship changes over space. The GWPCC estimates the local correlation coefficients at each sample point by measuring the parameters of relationship locally, allowing the estimation of parameters (i.e., correlation coefficients) at each location simultaneously (Fotheringham et al., 2002). Therefore, it has the potential to capture the spatially varying relationships between input variables by including the information of spatial locations for each sample site, which are usually ignored by the traditional PCC. Moreover, the local coefficients of GWPCC can represent strong or weak correlation between variables, instead of the 'slope' coefficients in the GWR model (Xu and Zhang, 2021). Moreover, a series of significance tests are provided by GWPCC, which can identify the spatial variations at different significance levels (Kalogirou, 2014). The formula of traditional PCC is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where x_i is the value of Al at the i^{th} location, y_i is the value of Pb at the i^{th} location, $\bar{x}$ is mean value of Al which is calculated by $\sum_{i=1}^{n} x_i/n$, $\bar{y}$ is the mean value of Pb which calculated by $\sum_{i=1}^{n} y_i/n$, n is the total number of samples.

The GWPCC can estimate local correlation coefficients (r_i) at a location i by adding geographical weighting w_{ij} in the equation, which is expressed as (Kalogirou, 2014):

$$gwpcc_i = \frac{\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} w_{ij}(y_i - \bar{y})^2}} \quad (2)$$

where $\bar{x}$ is the geographically weighted mean value of Al which calculated by $\sum_{i=1}^{n} w_{ij}x_i \Big/ \sum_{i=1}^{n} w_{ij}$, $\bar{y}$ is the geographically weighted mean value of Pb which calculated by $\sum_{i=1}^{n} w_{ij}y_i \Big/ \sum_{i=1}^{n} w_{ij}$.

The weights are calculated by a bi-square function expressed as:

$$w_{ij} = \begin{cases} [1 - \left(\frac{d_{ij}}{h_i}\right)^2]^2 & \text{if } \mathrm{d_{ij}} < h_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where d_{ij} is the distance between location i and j, h_i is the selected bandwidth (nearest neighbours) using adaptive kernel type function of location i.

The bandwidth is an important parameter in the GWPCC and also other GW models, which has been extensively debated in the current literature (e.g., Farber and Páez, 2007; Guo et al., 2008; Gao and Li, 2011). However, there is no consensus on the 'best' bandwidth, while it depends on different research purposes. The results may vary using different bandwidths and spatial weights. With a larger bandwidth, the GWPCC tends to reach a global statistic to reveal larger patterns by including more samples in local coefficient estimation, and vice versa (Song et al., 2016). From technical perspective, the 'optimal' bandwidth can be determined by minimising model fit diagnostic such as Akaike Information Criterion (AIC) (Akaike, 1974) or cross validation (CV) (Bowman, 1984). In this study, the bandwidth is the number of nearest neighbours instead of spatial distance due to the adaptive kernel type as weight function in the GWPCC (Fotheringham et al., 2002). This is useful to reduce the 'border effects' when samples were located in the costal and border areas (Zhang et al., 2011). Before exploring the spatially varying relationships between Pb and Al, the initial bandwidth was chosen using the AIC, which is effective to calculate the most suitable bandwidth by evaluating how well the model fit the data and compared with different possible models (Akaike, 1974). Then, due to the effects of

different bandwidths on the results, a total of 6 bandwidths including the AIC and 5 larger ones (with number of nearest neighbours being 43; 100; 150; 200; 250; 300) were investigated to achieve the purpose of focusing on large and smooth patterns of spatially varying relationships between Pb and Al from the local perspective.

2.5 Data preparation and software

As a parametric spatial statistic that depends on classic statistical parameters (i.e., mean values), it requires the normality of distribution for the dataset prior to the implementation of GWPCC. However, it is well known that geochemical data do not follow a normal or log-normal distribution (e.g., Reimann and Filzmoser, 2000; Zhang et al., 2005), thus appropriate data transformation process is necessary. The normal score transformation (NST) was performed on the Pb and Al concentrations in Tellus dataset in order to meet the normality requirement of GW model (Fotheringham et al., 2002). The NST and statistical analyses were conducted in SPSS (ver. 24), and the GWPCC, including its local correlation coefficients and significance tests were calculated in the R package '*lctools*' (ver. 3.56, in http://cran.r-project.org/web/packages/lctools/index.html). All the spatial distribution maps were produced using inverse distance weighted (IDW) interpolation in ArcGIS (ver. 10.4).

3. Results and discussion

3.1 Descriptive statistics for Pb and Al concentrations

The basic statistics for raw data of Pb and Al concentrations in the topsoil of Ireland are presented in Table 1. The mean values of these two elements were higher than the median values, suggesting the right skewed distribution in the raw data set. The large difference between minimum, median, 95% and maximum values for Pb indicated the strong variation and potential outliers (extremely high values) across the study area (Zhang et al., 2009). Thus, it is necessary to perform data transformation to reduce the effects of potential outliers on the spatial statistics from the raw data sets. The NST was applied to Pb and Al

data, and the transformed data were used for further analysis in this study. In addition, the detection limits (DLs) for two elements are also provided in Table 1, and the values below the DLs were replaced as half of the DL values for further statistical analysis as well.

Table 1. Basic statistics of Pb and Al concentrations in the topsoil of Ireland.

Element	Min.	Q25	Median	Mean	Q75	Q95	Max.	Std. Dev	DL
Pb (mg/kg)	<0.2	17.5	24.4	36.5	36.2	93.6	3120	68.24	0.2
Al (%)	<0.01	0.65	1.24	1.43	1.92	3.7	9.27	1.13	0.01

Units: Pb (mg/kg); Al (%); DL: detection limit

3.2 Spatial distribution of Pb and Al in the topsoil of Ireland

The spatial distribution maps based on the IDW interpolation for Pb and Al concentrations are shown in Fig. 3, displaying relatively similar distribution patterns with high values in the east and low values in the west of Ireland. The high concentration of Pb in the eastern areas are not only related to natural background source (i.e., greywacke shale), but also affected obviously by anthropogenic influences, especially near the urban areas of Belfast and Dublin. The eastern parts of the study area are featured extensive traffic and metals industry comparing with other areas. In addition, the high-value patterns of Pb are also observed in some areas in the western regions (i.e., Donegal, Galway and Mayo), showing good spatial associations with peatland (see Fig. 2b). It has been reported that peat is a major controlling factor on Pb concentration in the topsoil in Ireland (Palmer et al., 2013; McIlwaine et al., 2014), which is related to the spread of pollution from urban areas through atmospheric deposition to rural areas. For Al, the high concentrations in the east are mainly controlled by geogenic factors include basalt and greywacke shale, displaying large-scale patterns from north-eastern to central midlands. The spatial patterns of Pb and Al are consistent with the previous studies (e.g., Zhang et al., 2008a; Young and Donald, 2013). Furthermore, the results of global PCC ($r = 0.027$, $p < 0.01$) also indicated the existence of overall positive correlation between these two variables. However, this natural positive

relationship may be altered on the local scale due to the spatial variability of Pb influenced by anthropogenic factors, thus the varying relationships can be associated with the potential sources of Pb.

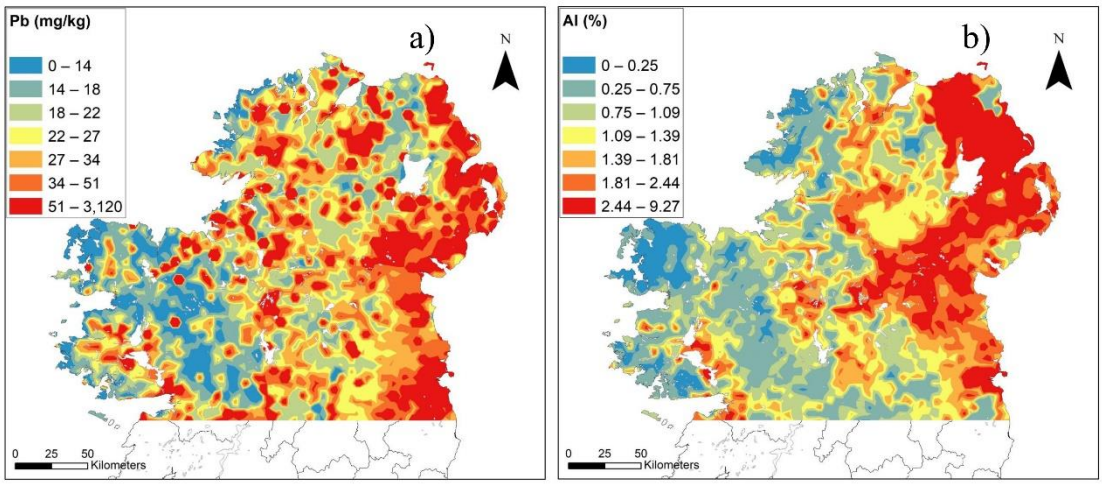


Figure 3. Spatial distribution maps of Pb and Al concentrations in the topsoil of Ireland: a) Pb (mg/kg); b) Al (%).

3.3 Effects of bandwidths on the results of spatially varying relationships between Pb and Al

Fig. 4 exhibited the results of local correlation coefficients of GWPCC using selected six bandwidths, and the corresponding statistics of different bandwidths are summarised in Table S2. The spatial variation patterns of both positive and negative correlation coefficients between Pb and Al concentrations were observed among all the six bandwidths. More than 35% of the sampling locations were found to have negative correlations at all distance bands (see Table S2). When choosing the bandwidth calculated by the AIC (n = 43), complex patterns of spatial variation were observed in the whole study area (Fig. 4a). The positive correlation was clustered across the whole study area, especially in the midlands of Ireland. The negative correlation was mainly clustered in northern, western and central-eastern parts of Ireland. Compared with other bandwidths, these patterns

showed scattered distributions with more details of spatially varying relationships between Pb and Al, while it is not conducive to identifying the smooth patterns from regional perspective.

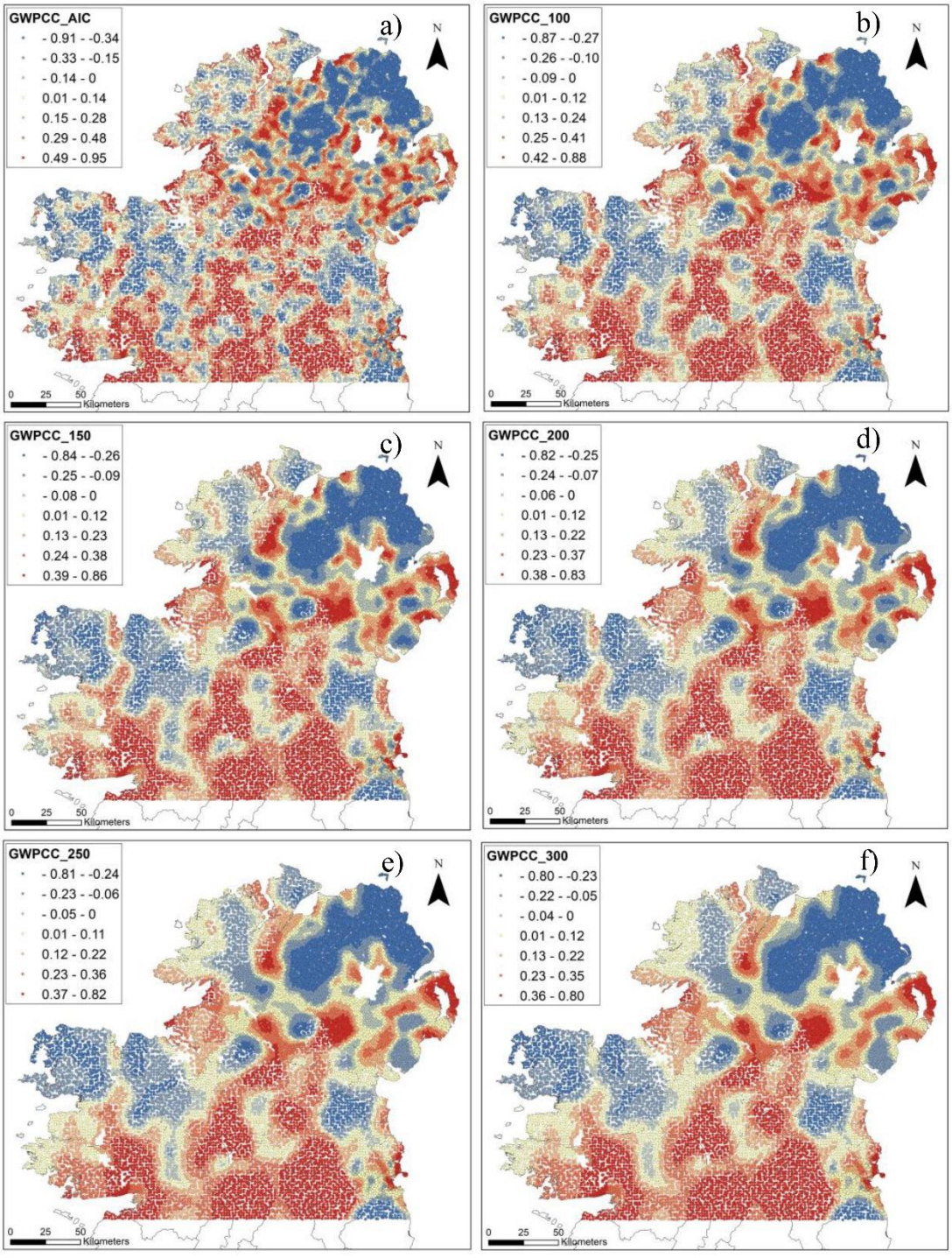


Figure 4. Local correlation coefficients showing spatially varying relationships between Pb and Al at six different bandwidths (number of nearest neighbouring samples): a) 43 (AIC); b) 100; c) 150; d) 200; e) 250; f) 300.

As the distance band gradually increased (from 100 to 300 samples), the range of the coefficients became smaller, and some scattered patterns of spatial variation between Pb and Al disappeared. This is because the local correlation coefficients are calculated based on the selected bandwidth using the nearest neighbouring samples. The smoother and continuous patterns for the spatially varying relationships were revealed when using larger bandwidths. It is worth noting that when the bandwidth increased to a certain extent (i.e., n = 200), both of the local correlation coefficients and patterns did not change significantly, which demonstrated that the selected bandwidths were good enough to reveal the spatially varying relationships between Pb and Al in the study area. Considering that there is no standard criterion for the 'best' bandwidth, it seems that the 200 nearest neighbouring samples was the most suitable choice to meet our research objectives. It reveals comparatively large and smooth patterns at the local level, while maintaining much detailed variation simultaneously. Therefore, this bandwidth was selected for further study.

3.4 Exploration of associations between the potential influencing factors and the spatially varying relationships

As mentioned early, the spatial association of potential influencing factors can be investigated through the patterns of different correlations that revealed by the GWPCC. The patterns of positive correlation suggested that the original relationship between Pb and Al concentrations reserved, however, the 'special' negative and mixed correlations implied the interferences of external factors, which can be used to explore the potential influencing factors at the local level. In order to obtain interpretive results of the influences on the spatial relationships between Pb and Al concentrations, the maps showing significance levels of the local correlation coefficients in the major areas are produced in Fig. 5. The possible explanation to the different relationships in major areas are also summarised in Table 2. For clarity, the spatial relationships between Pb and Al concentrations were reclassified as five classes based on the significance test, including 1% negative significant, 5% negative significant, 1% positive significant, 5% positive significant and not significant.

1) North-western and north-eastern areas (A1, A2 and A3): The majority of significant negative correlation were observed in these areas, extending from western to north-eastern Ireland. These patterns showed a clear association with the locations of blanket peat (Fig. 2b), especially in the northern and western parts (county Mayo, Donegal and Londonderry). The comparatively higher concentration of Pb was distributed, while the Al concentration was generally low (see Fig. 3). As well known that the rich organic matter content in blanket peat can bind Pb in the surface soil from atmospheric deposition (e.g., Cheng et al., 2015; Shotyk et al., 2016), which played an important role on controlling the elevated Pb concentrations in these areas. Although Al would be accumulated via atmospheric dust in some cases, however, its concentration in peat bogs may be also controlled by soil pH and precipitation conditions (Takahashi and Dahlgren, 2016). It was reported that humic substances have the biding capability of Al^{3+} to form multi-dentate complexes under a rainy environment (Rouff et al., 2012), thus resulted decreased Al concentration in the topsoil and further confirmed in the spatial interpolated maps. Therefore, the 'special' negative correlation was established and then captured by GWPCC approach. In the north-eastern region (area A3; county Londonderry and Antrim), the negative correlation patterns displayed the overlay associations with not only blanket peat, but also the basalt formation (Fig. 2b). This indicated more complicated influential mechanisms, as the Pb distribution was controlled totally different in these two geological features (Jordan et al., 2007; Xu et al., 2021). From the spatial distribution map (see Fig. 3), it can be clearly observed that high value of Pb gathered on the eastern and western sides of this part (area A3), overlaying on the peatland and showing a decreasing trend on the basalt formation. In contrast, high value of Al was found overlaid on the basaltic rocks, with a decreasing trend in the peatlands on both sides. The spatial variation and distribution between these two elements presented a completely opposite trend, which leads to a negative correlation between them. However, it should be noted that the north-western and north-eastern areas belong to remote rural soils. Therefore, these negative correlation patterns have a strong association with atmospheric pollution, and the fates of Pb may

related to long-distance transportation from human activities in urban areas, such as traffic emissions and mining.

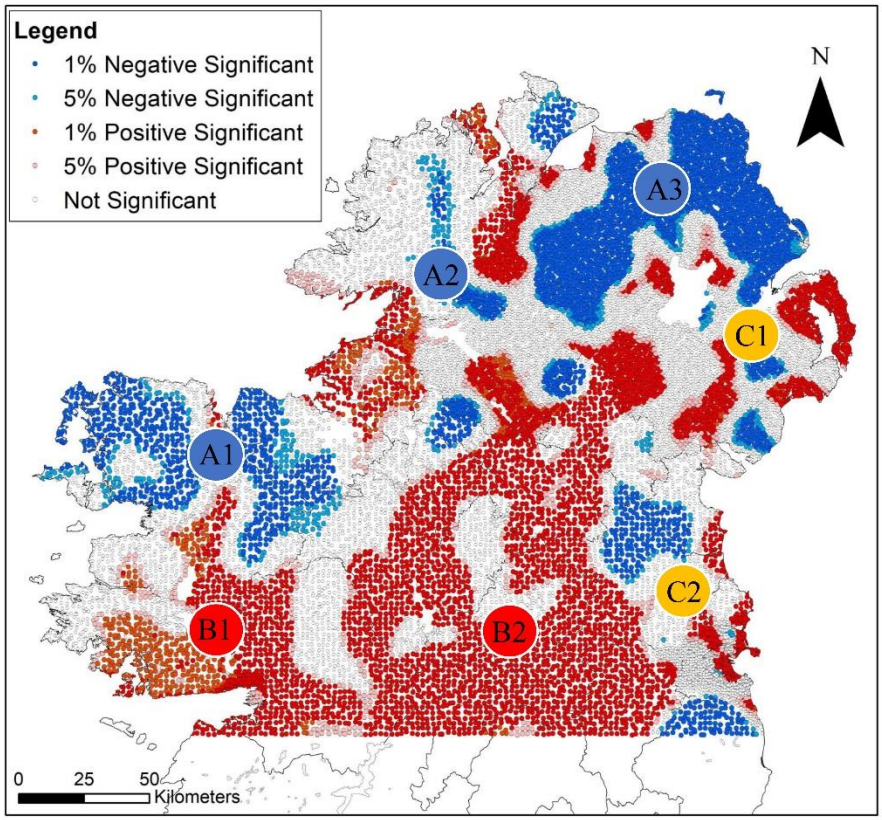


Figure 5. Spatial distribution map showing the significance levels of local correlations in the topsoil of northern half of Ireland.

Table 2. Summarised explanation for different relationships in the major areas

Area	Relationship	Association	Possible explanation
Western (A1)	Negative	Blanket peat	Atmospheric deposition
North-western (A2)	Negative	Blanket peat	Atmospheric deposition
North-eastern (A3)	Negative	Basalt; blanket peat	Geogenic control; atmospheric deposition

Central-western (B1)	Positive	Pb ore deposits	Large-scale Pb-Zn mineralisation
Midlands (B2)	Positive	Basin peat	Natural mineral soil
Eastern (C1)	Mixed	Greywacke shale; urbanisation	Mining; traffic emission
Central-eastern (C2)	Mixed	Blanket peat; urbanisation	Mining; traffic emission

2) Central western and midlands (B1 and B2): Most samples of significant positive correlation were maintained and found across the central midlands of Ireland, extending from the north-central to western areas. The consistent patterns between Pb and Al can be observed from Fig. 3, that is, the concentrations of these two elements are both relatively low. The large-scale limestone in the central areas as well as granite and schist in the central-western regions are not conducive to their accumulation in the topsoil, as Pb and Al are reported with low values in the soils formed on these rocks (Reimann et al., 2014; McIlwaine et al., 2015). In addition, the basin peat in the central region is formed on the natural mineral soil in low-land areas and contain less organic matter content comparing with blanket peat bogs (Fay et al., 2007). Thus, the influence is not as strong as that in the blanket peat, whereas the positive correlation reserved as normal. However, it is worth noting that some significant patterns of positive correlation even extend to the central-western (county Galway), overlaying on the blanket peat. From the perspective of spatial analysis, the reason for the existence of the positive correlation is not yet clear, however, it is likely to be closely associated with Pb-Zn ore deposits (see Fig. 2) and deserves further investigation.

3) Eastern coastal areas (C1 and C2): Mixed and weakened relationships were observed (Fig. 4d), with both significant positive and negative correlation scattered in these two regions. Although both elevated values of Pb and Al clustered in these areas (see Fig. 3), the spatial pattens are not continuous, implying more complicated factors with a

mixture of both natural (i.e., greywacke shale) and anthropogenic influences. The spatial relationships between these two variables were interfered more by anthropogenic inputs in the eastern urban areas (i.e., Belfast and Dublin) due to the urbanisation level of Ireland. Extensive research has been conducted on the influence of urban development on the elevated Pb concentrations (e.g., Zhang et al., 2008b; Laidlaw et al., 2012; Appleton and Cave, 2018). Except for urbanisation, the distribution of Pb in the topsoil of urban areas may also be related to traffic emission (Johnson et al., 2017). The combustion of leaded gasoline can pollute soils and air through vehicle emissions (Xu and Liao, 2004), and even pollute suburban soils through long-distance transportation, which is the case of the blanket peatland in the western areas of this study. Given that leaded petrol has been eliminated since 2000, the enrichment of Pb may be attribute to the historical factors. Due to the large bandwidth selected in this study, the weakened relationships by anthropogenic interference cannot pass the significant test in the GWPCC, and thus displayed scattered patterns. Moreover, the spatial patterns of large regional data sets are mainly associated with the influences of natural factors, which is in line with the conclusions of recent studies (e.g., Matschullat et al., 2018; Négrel et al., 2018; Xu et al., 2019).

3.5 Limitations and future work

It should be acknowledged that it is hard to determine the specific pollution sources from the spatial perspective, while the association with atmospheric deposition and anthropogenic influence can be identified instead. For example, we have identified the clear association between negative patterns and peat in the northern and western of Ireland, which is the novelty point of this study by using GWPCC. However, this approach cannot quantify the influencing factors such as atmospheric pollution in the surface peat. Most studies used Pb isotopes to fingerprint the timeline of atmospheric pollution in the peatland (e.g., Coggins et al., 2006; Allan et al., 2013; Rosca et al., 2018). But this deserves more physico-chemical data and improvement of this approach in our future research.

Additionally, some results cannot be fully interpreted and further investigation is needed. For example, the positive patterns in the central western (county Galway) remain unclear based on the current dataset. It should be noted that although much more spatial variation can be revealed and be related to potential pollution sources at the smaller bandwidths (see Fig. 4a, b), the explanation of these patterns may require a very large amount of work. Therefore, we decided to focus on large and smooth patterns that identified using larger bandwidth in this study. This is also one of the reasonable goals for using large-scale regional datasets in environmental studies.

In this study, we only selected two elements for comparison, because Pb and Al have their own representativeness. We acknowledge that investigations of all the influencing factors and conventional multivariate relationships, in combination with the concept of spatially varying relationships can be considered in future studies. Our results provided a demonstration on application of spatial machine learning approach to explore the relationships between environmental variables, and can be regarded as an effective method to identify the spatial association between potential influencing factors and pollution. In the big data era, it is promising and efficient to extract hidden geochemical information from the spatial patterns, contributing to the improvement of environmental assessment sustainable policy-making at the regional or even global scale.

4. Conclusion

This study investigated the spatially varying relationships between Pb and Al concentrations in the soil of Ireland based on the current Tellus data sets. The results of GWPCC technique found that the relationships between these two variables are spatially varying, with both positive and negative correlations identified at the local level. Original positive correlation was observed in central-western and midlands, while the ‘special’ negative correlation was clustered in the north-western and north-eastern of Ireland. The comparatively mixed correlations were found in the eastern coastal areas. Clear association between the significant negative correlations and blanket peat in the topsoil was identified,

which can be attribute to atmospheric deposition of Pb from long-distance transportation. Moreover, anthropogenic activities weakened the relationships in the eastern coastal areas. Our findings highlighted the efficiency of spatial machine learning technique in identifying the association between PTE distribution and potential pollution sources. Such new findings are revealed by GWPCC technique, which is proved as a powerful approach in investigating the spatially varying relationships and exploring the potential influencing factors, and can be applied into environmental studies elsewhere.

4.5 Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR)

Yuan, Y.M., Cave, M., **Xu, H.F.**, Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater., 393, 122377.

Summary: This paper investigated the spatial relationships between Pb and Al concentration in urban soils of London using GWR based on 6,467 samples collected by British Geological Survey. The local regression coefficients of GWR revealed that the relationships between Pb and Al were spatially varying, with different relationships in different areas. The strong negative relationships were found in north-eastern and northern areas, while weak negative relationships were clustered in central areas. The association between positive patterns and large parklands and greenspaces were found in the south-eastern and south-western areas, where the natural geochemical signatures were reserved due to less influences from anthropogenic activities. On the contrary, the 'special' negative relationship indicated the association with the impact of anthropogenic activities on Pb concentration, such as road traffic, industry activities and construction. Such results demonstrated the efficiency of GWR in revealing the spatially varying relationships between environmental variables, which provides better understanding of the complicated influencing factors in environmental studies.

My contribution in this paper accounted for ~20% in exploring data and writing-reviewing manuscript.

Research paper

Journal of Hazardous Materials 393 (2020) 122377

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

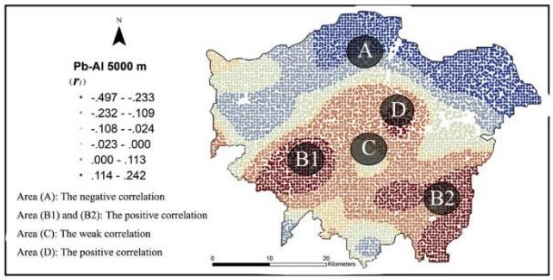
Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR)

Yumin Yuan[a], Mark Cave[b], Haofan Xu[a], Chaosheng Zhang[a,*]

[a] International Network for Environment and Health (INEH), School of Geography, Archaeology & Irish Studies & Ryan Institute, National University of Ireland, Galway, Ireland

[b] British Geological Survey, Environmental Science Centre, Nottingham, United Kingdom

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Deyi Hou

Keywords:
Geographically weighted regression (GWR)
Spatially varying relationships
Urban soil
Lead
Aluminium

ABSTRACT

In this study, geographically weighted regression (GWR) was applied to reveal the spatially varying relationships between Pb and Al in urban soils of London based on 6467 samples collected by British Geological Survey. Results showed that the relationships between Pb and Al were spatially varying in urban soils of London, with different relationships in different areas. The strong negative relationships between Pb and Al were found in the northeast and north areas and weak relationships were located in central areas, implying the links with the impact of anthropogenic activities on Pb concentration, while road traffic, industry activities and construction in centre of London may be linked to the weakened or changed direction of the relationship. However, positive relationships between Pb and Al were found in large parklands and greenspaces in the southeast and southwest as well as a small area in central London, due to less influences from human activities where the natural geochemical signatures were preserved. This study suggests that GWR is an effective tool to reveal spatially varying relationships in environmental variables, providing improved understanding of the complicated relationships in environmental parameters from the spatial aspect, which could be hardly achieved using conventional statistical analysis.

* Corresponding author.
E-mail address: Chaosheng.Zhang@nuigalway.ie (C. Zhang).

https://doi.org/10.1016/j.jhazmat.2020.122377
Received 4 October 2019; Received in revised form 20 February 2020; Accepted 21 February 2020
Available online 22 February 2020
0304-3894/

1. Introduction

Heavy metals are important chemical components in soils. For the health of humans and plants, some heavy metals such as Pb, Cd and Hg belong to the non-essential elements (De Miguel et al., 1998; Kabata-Pendias, 2004), which are considered as contaminants because of their toxicity and difficulty in degradation (Zhang et al., 2012). In natural conditions, concentrations of geochemical elements in soils are affected by parent materials (Alloway, 1995; De Temmerman et al., 2003; Aelion et al., 2009; Ballesta et al., 2010; Cai et al., 2010), as a result of geological and pedologic processes that rule soil formation (A. Castrignanò et al., 2000). In urban areas, additional sources of heterogeneity are caused by anthropogenic activities. Moreover, due to the complex and heterogeneous nature of urban soils, the spatial distribution of geochemical elements is affected by multiple factors (Franco-Uria et al., 2009; Qishlaqi et al., 2009; Yang et al., 2009; Martín et al., 2013), resulting in different relationships among them in different locations (Zhang et al., 2007; Lv et al., 2013). Therefore, it is necessary to find an efficient way to reveal such spatially varying relationships between geochemical elements in urban soils across local areas, which would be helpful to better understand the complicated relationships in urban geochemistry as well as to reveal their association with influencing factors.

The traditional statistical methods, such as correlation analysis and ordinary least square (OLS) regression, produce 'average' or 'global' parameters to estimate the spatial relationships (Ali et al., 2007), which are reflected equally over the whole study area. Therefore, the impacts of local variations could be hidden (Bacha, 2003; Batisani and Yarnal, 2009; Geri et al., 2010). In light of this, an important contribution of geographically weighted regression (GWR) is to build regression models to explore how one dependent variable changes in response to one or more independent variables at the local scale (McMillen, 1996; Fotheringham et al., 1998; Leung et al., 2000a; Yu and Wu, 2004; Deller and Lledo, 2007; Waller et al., 2007). The GWR model takes the samples within a defined neighbourhood into calculation by giving more weights to nearby samples than those further away (Wheeler and Calder, 2007; Zhang et al., 2011). The GWR results depend on the observations that are in close proximity to the subject point, so they reveal the relationships within the neighbourhood (Fotheringham et al., 2002; Foody, 2004; Bickford and Laffan, 2006). Thus, the GWR can be used to explore the spatially varying relationships between variables (Tu, 2011).

Among heavy metals, of particular interest is the spatial variation of Pb in urban soils, not only because it contains toxicity, but also it can be strongly influenced by human activities. For example, industrial discharges (Mattuck and Nikolaidis, 1996; Aelion et al., 2009), vehicle emissions (Sansalone and Buchberger, 1997) and construction are considered as the major influencing factors of Pb in urban soils (De Temmerman et al., 2003; Zhang, 2006; Delbecque and Verdoodt, 2016; Wu et al., 2019). In addition, previous studies widely reported the elevated concentrations of Pb in soils influenced by human activities including traffic, especially in the urbanized and industrialized areas (Zheng et al., 2002; Madrid et al., 2002; Qin, 2008; Zhang et al., 2011; Sayyed and Sayadi, 2011; Raju et al., 2013; Su, 2014). Zhang et al. (2008) indicated that the spatial distribution of Pb in the urban soil of Galway was related to traffic pollution (Zhang, 2006) and historical rubbish dumping (Carr et al., 2008). To date, the use of Pb gasoline for vehicles has declined in many countries, while the concentrations of Pb in urban soils remain a concern because the anthropogenic sources of Pb are particularly dense in the urban environment (Clark et al., 2006; Rawlins et al., 2012). Despite other potential toxic elements could also be interesting for investigation, this study focuses on Pb, which could be helpful for us to seek the links with the influencing factors.

In this study, the GWR was applied to explore the spatially varying relationships between Pb and Al in urban soils of London. The chemical element Al was chosen as the independent variable for the dependent variable of Pb. The reason why Al was chosen was because not only it has been commonly used as a reference element of lithogenic origin in multivariate statistical analyses, but also it shares similar features as Pb under the natural environment where they are strongly bound by fine-particles of clay minerals (Spark, 2010). However, this positive correlation may be disturbed or changed by anthropogenic influences, as Pb is strongly influenced by human activities. Another element Ti was chosen as the dependent variable of Al for comparison. The element Ti and Al have been frequently used as reference elements, as they are components of minerals resistant to chemical weathering, and they are less affected by anthropogenic factors (Sezgin et al., 2003; Tylmann, 2004). The relationship between these conservative elements is expected to be less spatially variable, providing a good comparison with that between Pb and Al.

The objectives of this study were: (1) to reveal the spatially varying relationships between Pb and Al in London soils at the regional scale; (2) to investigate the effects of different bandwidths on the results of GWR for the purpose of identifying spatial patterns of the spatially varying relationships; and (3) to explore the associations between the spatially varying relationships and the related influencing factors.

2. Methods

2.1. Geology and soil geochemistry data

The bedrock geology of the Greater London Authority (GLA) showed a wide range of Cretaceous and Palaeogene bedrocks in the north and south areas (Miles and Appleton, 2005; British Geological Survey, 2011) (Fig. 1). Palaeogene bedrocks in the north area were composed of the Bagshot Formation, Thames Group (clay), Lambeth and Thanet sand Formation. The White chalk, Grey Chalk and Thanet Group (sand) belonging to Cretaceous deposits were found in the south area. Quaternary superficial deposits occurred in the central area with the most extensive alluvium, river terrace deposits and brickearth. Relatively small patches of Clay-with-flints and Head (clay-silt) were found in the south area.

In the GLA study area, a total of 6467 topsoil samples were collected on a grid system by British Geology Survey at a density of 4 samples per km^2. Each composite sample was obtained by collecting 5 subsamples from the centre and corners of a 20 m square at each sampling site. Topsoil samples were collected at a depth of $0-20$ cm, after removal of surface vegetation, litter and rootlet zone (usually < 5 cm). The nominal sampling depth is therefore $5-20$ cm (Johnson, 2005). Analyses were performed for total concentrations of 48 elements by X-ray fluorescence spectrometry (XRFS) and for loss on ignition (LOI at 450 °C) and pH. Detailed information for sampling and quality control is available in Allen et al. (2011) and Johnson (2011).

2.2. Land use data

The land use data used in this study is the GLUD 2005 obtained from the website of the London Data store (http://data.london.gov.uk/). Five simplified land use classes: farm-land, industry, greenspace, built-up and others were generated using ArcGIS® software (Fig. S1).

2.3. Geographically weighted regression

The GWR reveals the spatially varying relationships between the dependent and independent variables, and a set of location specific parameter estimates. Based on Fotheringham et al. (2002), the GWR model with one independent variable can be expressed as:

$$y_i = \beta_0 (u_i,v_i) + \beta_1 (u_i,v_i)x_i + \varepsilon_i$$

Where u_i and v_i are the coordinates of the i^{th} location, and $\beta_0 (u_i,v_i)$ is the local intercept for i^{th} location, $\beta_1 (u_i,v_i)$ is the estimated local regression

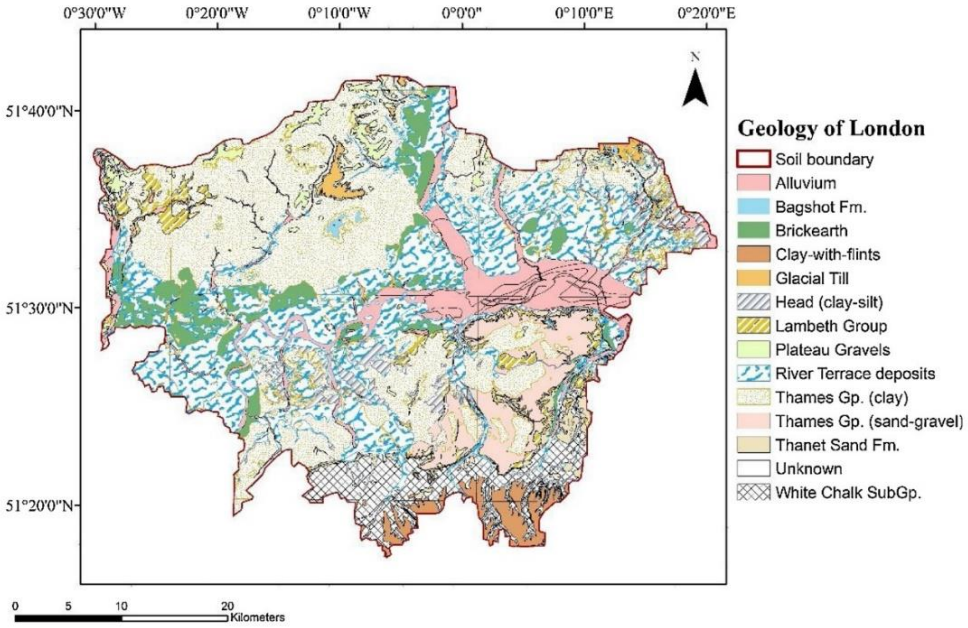


Fig. 1. Geology map of the London region.

coefficient for the i^{th} location and ε_i is the random error at the i^{th} location. As data included in the calculation are geographically weighted, the local intercepts and local regression coefficients are different at different locations, which is different from OLS where these parameters remain the same for the whole dataset. The parameters are estimated from:

$$\beta\,(u_i,v_i) = (X^T W\,(u_i,v_i)\,X)^{-1}\,X^T W\,(u_i,v_i)\,Y,$$

Where $\beta\,(u_i,v_i)$ represents the unbiased estimate of β, $W\,(u_i,v_i)$ is the weighting matrix which acts to ensure that observations near the specific point have larger weighting values.

Before using GWR technology, some issues need to be considered. For example, Mei et al. (2016) pointed out that the excessive flexibility of GWR's calibration method may lead to spurious spatially varying relationships or even reverse the correlation between variables. In addition, multiple collinearity of independent variables may lead to strong correlations (Páez et al., 2011). However, in this paper, using only Pb as the dependent variable and Al as the independent variable can effectively avoid the above problems. One issue to be discussed in this paper is the choice of bandwidth, which is the key controlling parameter for GWR results (Guo et al., 2008; Gao and Li, 2011). In practice, the bandwidth is the key controlling parameter for GWR results (Guo et al., 2008; Gao and Li, 2011). The process of choosing the weighting matrix is important to predetermine an optimum bandwidth. The optimal bandwidth for GWR was determined by minimizing some model fit diagnostic, such as cross-validation (CV) score (Bowman, 1984) or the Akaike Information Criterion (AIC) (Akaike, 1973. Considering the soil sampling sites on a generally regular grid, in order to calibrate the spatial weighting function and determine the optimal bandwidth for the models used in this study, both AIC and fixed distance bands ranging from 1000 to 50,000 m were applied for comparison. Considering that the results of local regression coefficients only represent the slop coefficients Gao and Li, 2011), a local correlation coefficient (r_i) was calculated to reveal the correlation between the dependent variable (Pb) and independent variable (Al). The formula can be expressed as:

$$r_i = \sqrt{R^2_i} \times \beta 1\,(ui, vi) \Big/ |\beta 1\,(ui, vi)|$$

Where R^2_i is one of the local deterministic coefficient R^2 from GWR model, indicating how strong the two variables correlate with each other linearly, and $\beta 1\,(ui, vi)$ is the local regression coefficient (Yu, 2006; Clement et al., 2009). The local correlation coefficient is equivalent to Geographically Weighted Pearson Correlation (GWPCC), which is based on the concept of local statistics (Kalogirou, 2014).

2.4. Data transformation and computer software

A normal score transformation was applied to the data for GWR analyses due to the non-normality and skewness problem of the raw data (Zhang et al., 2008a). All maps were produced using ArcGIS (ver.10.4) software. The conventional statistical analyses were carried out using SPSS (ver. 21.0).

Table 1
Basic statistics of Pb and Al concentrations in urban soils of London (Al: in %, Pb: in mg/kg).

Element	Min	10 %	25 %	Median	75 %	90 %	95 %	Max	Avg	StdDev
Pb	10.8	60.9	97.3	180.1	340	606.2	857.1	10000	295.6	430.4
Al	0.4	2.6	3.2	4.0	5.1	6.3	6.9	11	4.2	1.5

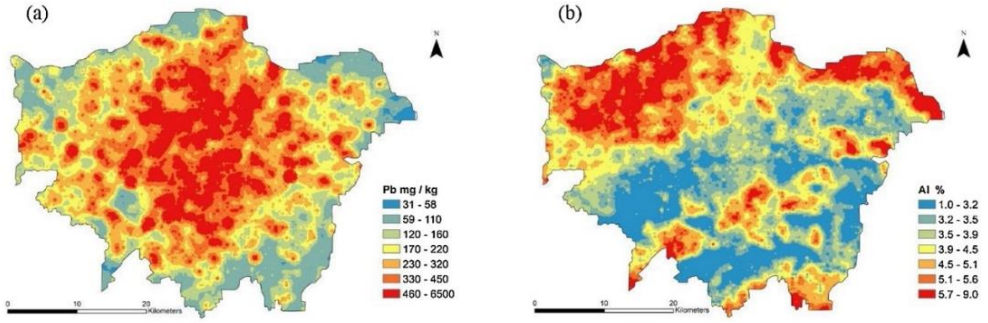


Fig. 2. Spatial distribution maps in London soils: a) Pb; b) Al.

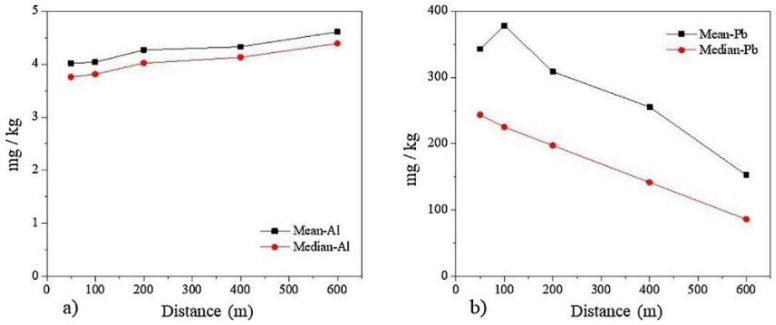


Fig. 3. The median and mean of topsoil a) Al, and b) Pb with distance from the nearest road. Data grouped into 0-50 m, 50-100 m, 200-400 m, 400-600 m.

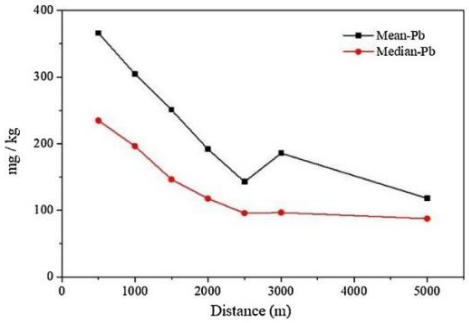


Fig. 4. The median and mean of topsoil Pb with distance from the nearest industry site. Data grouped into 0-500 m, 500-1000 m, 1000-1500 m, 1500-2000 m, 2000-2500 m, 2500-3000 m, 3000-5000 m.

3. Results and discussions

3.1. Basic statistics for Pb and Al concentrations in London soils

The basic statistics for Pb and Al concentrations in urban soils of GLA are listed in Table 1. The 90th percentile of Pb 606 mg/kg had exceeded the provisional Category 4 Screening Levels for lead (pC4SL, 130 − 330 mg/kg for residential area SP1010, 2014. The large difference between the percentiles implied that there were strong variation and heterogeneity of Pb concentrations in soils over the study area, with the range from 10.8 mg/kg to 10,000 mg/kg. Meanwhile, the mean value for Pb was significantly much higher than the median, implying that there were high value outliers or extreme values which skewed the distribution of the data set. The variation of Al was also strong, ranging between 0.4 % and 11 %.

3.2. Spatial distribution of Pb and Al

The spatial distribution maps for Pb and Al concentrations in London soils based on inverse distance weighted (IDW) interpolation are illustrated in Fig. 2. The relatively low values of Pb concentrations were located in the north part and some small areas in the southeast, where rock types Thames Gp, White Chalk and Clay-with-flints were located (Fig. 1). Elevated values of Pb concentrations were observed in the central part of GLA, indicating that the topsoil Pb was dominantly influenced by anthropogenic factors which are also spatially variable. The high values of Pb in the central areas may be associated with the high traffic volume, especially in the central area where is highly urbanized with an extensive road network. In order to investigate the influences of traffic on Pb concentration, the road buffer zones (0 − 50 m, 50 − 100 m, 200 − 400 m, 400 − 600 m) were generated using the buffer tool in ArcGIS, according to the major roads of London downloaded from Geofabrik (2016) (Fig. 3). Sites closer to major roads had higher Pb concentrations, and the reduction is up to 65 % if the distance is 500 m away from the road (Fig. 3). These results are in agreement with previous studies showing that the concentrations of Pb in soils increased with traffic volume while decreased with the distance from the roads (Thorpe and Harrison, 2008; Pant and Harrison, 2013; Wang et al., 2017). Moreover, Appleton and Cave (2018) reported that the bombing of UK increased Pb and other heavy metals spread on the soils and deposition of airborne particulates during the period 1940-41. In residential areas across the UK, Pb concentrations in domestic garden soils were consistently higher than those in soils of public parks

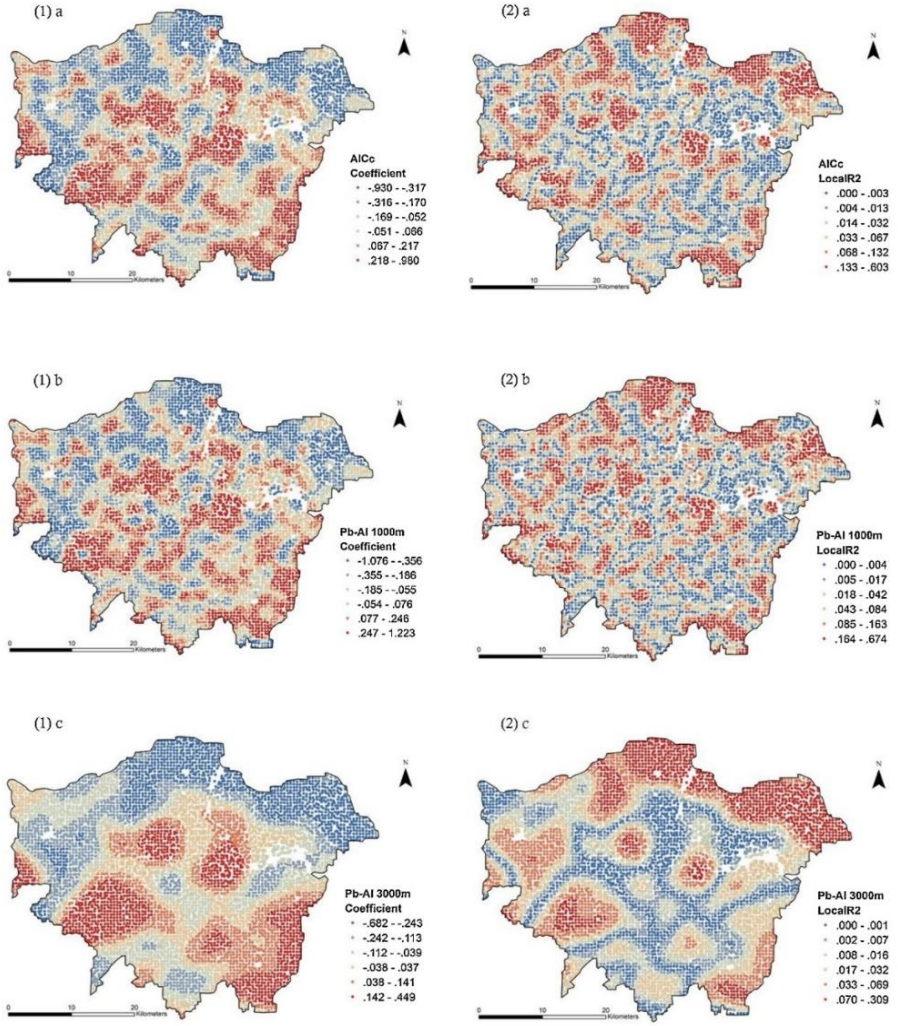


Fig. 5. Spatial patterns of (1) local regression coefficient and (2) Local R^2 for different bandwidths: a AIC, b 1000 m, c 3000 m, d 5000 m, e 10,000 m, f 20,000 m and g 50,000 m, respectively.

(Thornton et al., 1990), indicating that the historical construction as well as lead-based paints were also likely to be contributory factors (Milke et al., 2001).

Another important influencing factor is industry. Relatively high Pb concentrations were also associated with the impact of industrial development especially in the Thames and Lee valleys. There was a decreasing trend in Pb concentration with the increase of distance away from the nearest industrial sites within 2000 m. Beyond this distance, Pb concentration remained stable (Fig. 4). Elevated Pb concentrations were also associated with other industrial activities such as landfill, metal recycling and transport functions. On the other hand, relatively low values were found over the major parks, with little impact from the significant urban development throughout the 200–300 years history of London (Knights and Scheib, 2011; Scheib et al., 2011; Appleton and Cave, 2018). High concentrations of Al were observed in the topsoil of north London where London clay was the dominant soil parent material, while low concentration was found in the central and south areas, which are associated with the Alluvium, River terrace deposits, Plateau Gravels and Chalk (Fig. 1).

3.3. Effects of bandwidth

To investigate the effects of different bandwidths on the GWR results, the AIC and six bandwidths ranging from 1000 m to 50,000 m were considered Fig. 5). The AIC was first selected because it can effectively find the "optimal" bandwidth in GWR model. A number of scattered patterns were observed in the whole study area when the AIC method was chosen (Fig. 5a). Compared with the results from AIC, when the shortest bandwidth of 1000 m was used (Fig. 5b, the spatial patterns of GWR results were discovered at local level with a large number of small scattered patterns, showing more details of spatial variation. When the bandwidth increased to 3000 m, some small

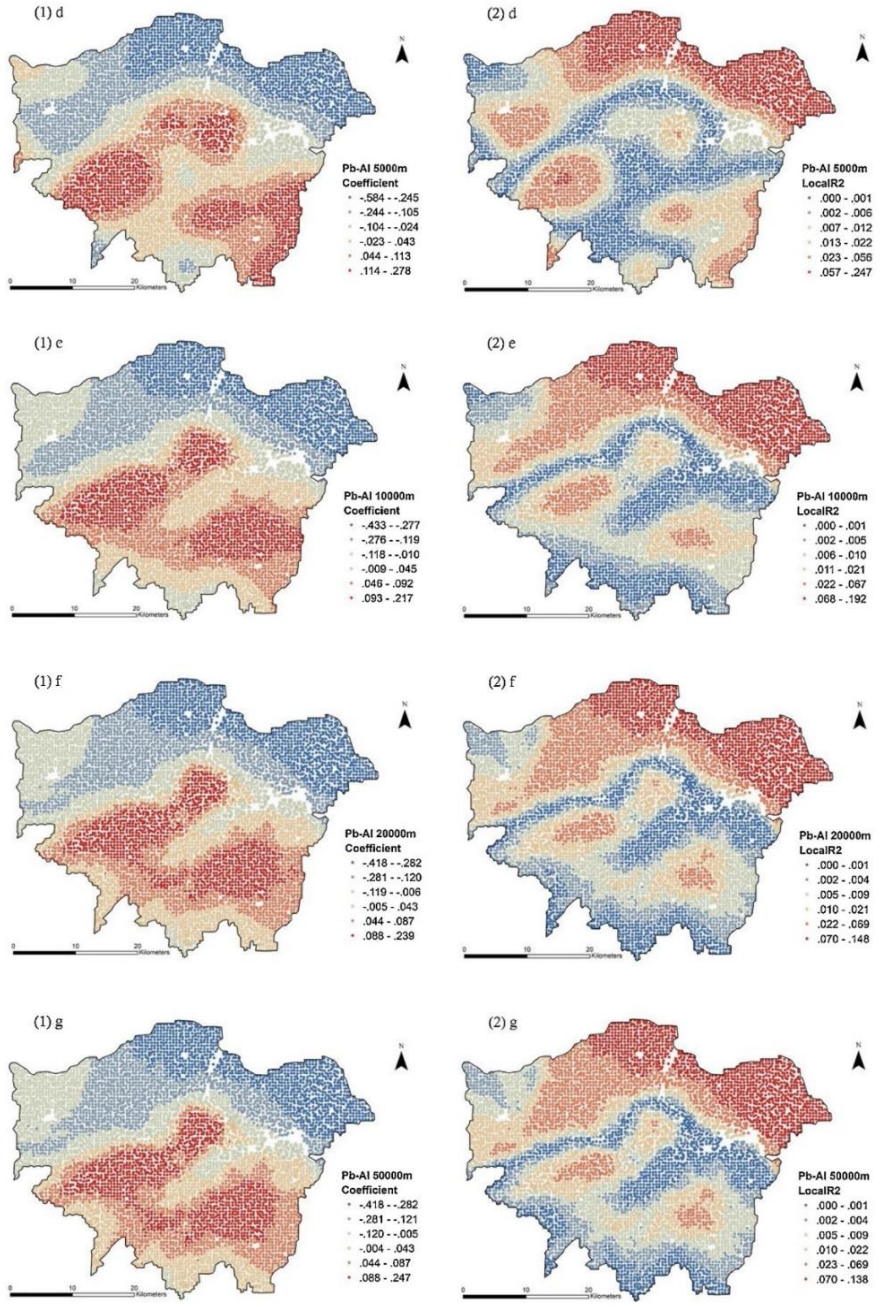


Fig. 5. (*continued*)

scattered patterns in the north area disappeared. These small scattered patterns were merged in several large patterns near the city centre. The patterns became much simpler and clearer when the bandwidth increased from 5000 m to 50,000 m, showing similar and consistent spatial patterns for the 4 bandwidths Fig. 5c, d, e, f and g. With the longest bandwidth of 50,000 m used in this study, most details of local

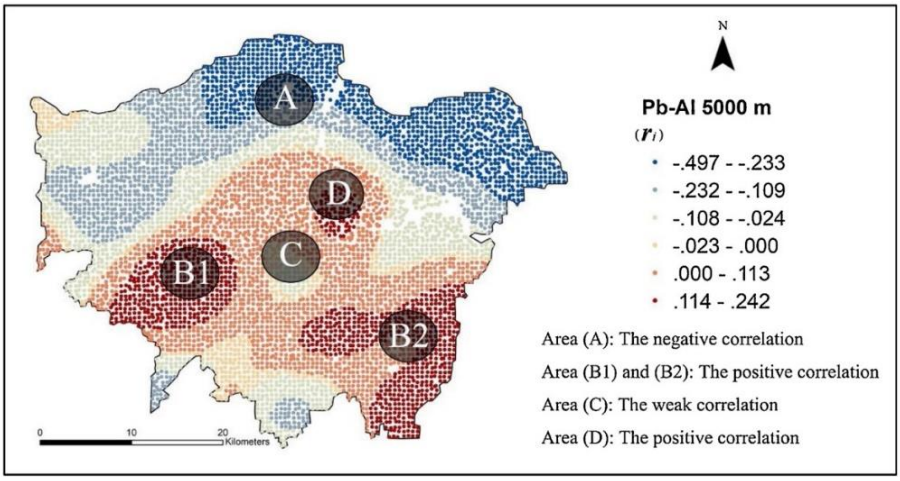


Fig. 6. The spatial distribution of local correlation coefficients (r_i) between Pb and Al.

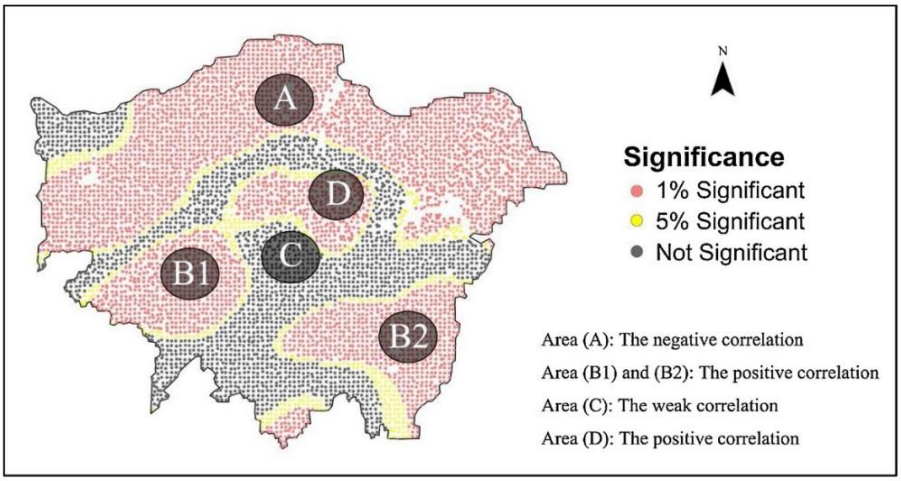


Fig. 7. Spatial distribution of local significance values of correlation between Pb and Al.

spatial variation disappeared. Comparisons of R^2 and AIC_C between the AIC and six bandwidths are displayed in Table S1. In order to further test whether the reasonable bandwidth have better performance than the others, the assessment was performed by comparing the R^2 and the AICc values from both AIC and six bandwidths, together with consideration of the spatial patterns of the GWR results. The R^2 values of the GWR models with the AIC, 1000 m, 3000 m and 5000 m bandwidths were all higher than those with longer bandwidths, with their AICc values showing an opposite trend, demonstrating that the GWR models with short bandwidths performed better in modelling the relationships between Pb and Al (Huang et al., 2017). However, the short bandwidths revealed more details of spatial variation making it hard to identify the large-scale patterns. The moderate bandwidth of 5000 m appears to be a compromised choice among all the bandwidth parameters tested, with relatively high R^2, low AICc while clearly depicting the large-scale patterns of the underlying spatially varying relationships in the study area, with fewer small scattered patterns. Our interpretation to the results will focus on this bandwidth.

3.4. Spatially varying relationships between Pb and Al revealed by GWR

3.4.1. Local regression coefficients and local R^2

The variation of local regression coefficients for the independent variable Al and the spatial patterns of local R^2 are already shown in Fig. 5. The local regression coefficients are the slope values showing how strongly the changes of the explanatory variable Al impacts the value of the dependent variable Pb locally. Both positive and negative local regression coefficients were found, demonstrating the existence of both positive and negative correlations between Pb and Al in London soils at the local scale.

Negative regression coefficients were located in the north part of London, varying from -0.05 to -0.6. A clear directional feature was observed extending in the west-east direction. The high absolute values of regression coefficients indicated that Pb concentrations changed

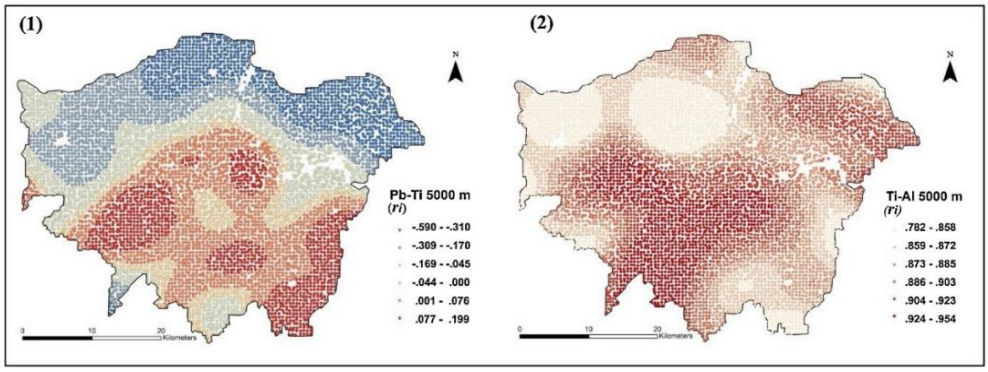


Fig. 8. The spatial distribution of local correlation coefficients (r_i) between (1) Pb and Ti; (2) Ti and Al.

more rapidly with the change of the Al in the north edge area. The decreasing trend of the absolute values of regression coefficients towards central London showed the slower changes of Pb with the change of Al. Some interesting positive local relationships with Al were identified in southeast and southwest part of the London. Relatively high regression coefficients (0.1-0.3) were found in large greenspace areas and low regression coefficients (0-0.1) were shown around in some built-up areas (see Fig. S1), indicating the relatively more rapid changes of the values Pb in large greenspace areas than those in the built-up areas.

The relatively high values of local R^2 were located in the north area, indicating that the local linear regression model performed well in this area. In central London and the south area, the local R^2 were generally low, implying the linear relationships between Pb and Al were weaker in these areas, and thus the variation in Al values explained a smaller percentage of the variation in Pb. These areas were more affected by human activities. There were three areas with relatively higher R^2 values in the southeast, southwest and the centre, respectively. These areas are line with the locations of large parks and greenspace, which are relatively less interfered by anthropogenic activities which are highly spatially variable.

3.4.2. Local correlation coefficient (r_i)

For the convenience of exploration of the spatially varying relationships between Pb and Al in this study, the local results from GWR for the dependent variable Pb were used to calculate local correlation coefficients (r_i). The spatially varying relationships between Pb and Al in London soils were clearly revealed by the local correlation coefficients and its significance levels (Figs. 6 and 7), showing strong negative correlations in the northeast and north area, strong positive correlations in two areas in southwest and southeast, while relatively weak correlations in central London, except for a small area in the city centre with strong positive correlations.

1) Northeast and north area (Area A): The strong negative correlations ($p < 0.01$) in the northeast area extended from northeast to the north (Figs. 6 and 7). This is in line with the spatial distribution patterns of Pb and Al (Fig. 2): high values of Al were located in the north side, with a decreasing trend towards central London, while high values of Pb were distributed in central London, with a decreasing trend towards the north (including northeast and northwest). The spatial variations of the two elements were in clearly opposite directions, resulting in their strong negative correlations in this area. The high values of Pb in central London could be related to human activities including industries, traffic and construction. On the other hand, under natural conditions, Al concentrations were controlled by geology, with high values in line with the distribution of Thames Gp (clay) in the north area. This result was supported by the strongly positive correlations between Ti and Al in the whole study (see Fig. 8).
2) Two areas in southeast and southwest (Areas B1 and B2): Strong positive correlations ($p < 0.01$) were found in two areas of southeast and southwest London, respectively (Figs. 6 and 7). Clear spatial patterns of both low value of Pb and Al could be identified in these two areas from the Fig. 2, with a generally increasing trend in their surrounding areas. The consistent spatial distribution patterns for the two elements resulted in their positive correlations in these two areas. This result was related to the distributions of parent materials (PMs) and parklands. Both low value of Pb and Al was presented in top-soil samples overlying Alluvium, River Terrace Deposits, Thanet Sand Formation as well as White Chalk containing a large number of quartzite clasts and gravelly-sand PMs with larger particle sizes (Figs. 1 and 2). It is well known that the adsorption properties of sand soils to metals are weak due to the presence of high silicate in quartzite clasts and gravelly-sand parent materials (Kern, 1994; Homann et al., 1998), which contribute to both low Pb and low Al values. Moreover, the large parks and greenspace areas may also play a role in relatively low concentrations of both Pb and Al, especially the sandy Richmond Park in the south-west as well as Greenwich Park and Blackheath in the southeast of London. These areas were less influenced by anthropogenic activities which had been historically protected for the past 200–300 years.
3) Central London (Area C): The weak correlations ($p > 0.05$) between Pb and Al were observed in soils of central London, with obviously high and strongly variable Pb values and low and strongly variable Al values (Fig. 2). As mentioned before, Al concentrations tended to be high in clay-rich PMs while low in Alluvium, River Terrace Deposits, which were associated with gravels and sandy PMs. Except for the north London with high Al values, the spatial distribution of Al in central London were generally low and variable. On the other hand, Pb concentrations in central London were strongly affected by the anthropogenic activities with elevated values. Compared with geology, the elevated Pb values were spatially random, mixed and complicated. Therefore, the correlations between Pb and Al were interfered and weakened in central London by anthropogenic activities in general.
4) A small area in City Centre (Area D): Strong positive correlations ($p < 0.01$) were observed in a small area in central London (Figs. 6 and 7). The spatial pattern of both relatively low Pb and Al values can be identified in this area, with the increasing values in their surrounding areas. The spatial variations of the two elements were in the same direction, resulting in their strong positive correlations

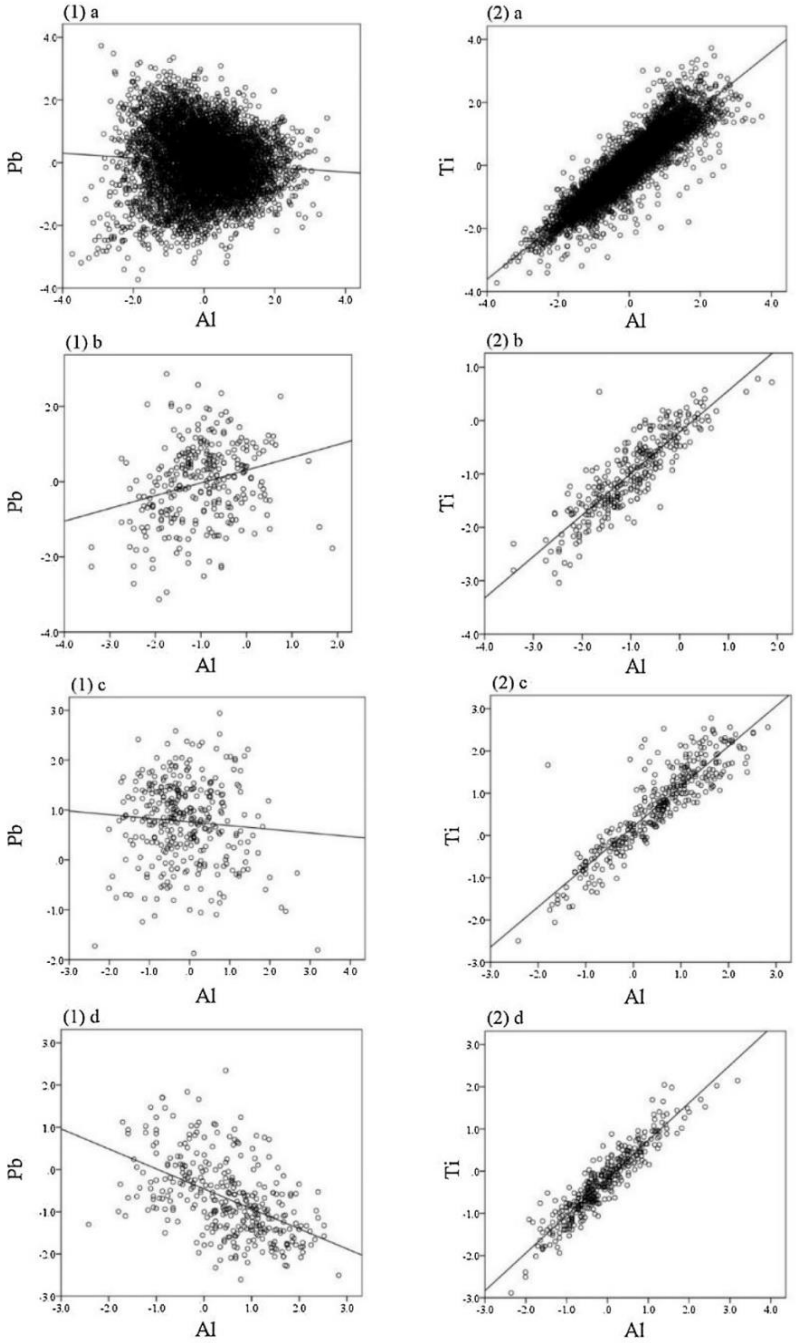


Fig. 9. Scatter plots between 1) Pb and Al, and 2) Ti and Al in a) all samples, b) positive correlation in southwest of London, c) weak correlation areas in the centre of London, and d) negative correlation areas in the northeast of London.

in this area. As mentioned before, the low value of Al in central London could be related to the Alluvium and River Terrace Deposits, while the relatively low values of Pb could be related to small parklands (i.e. The Vitoria park and Queen Elizabeth Olympic park) as the soils of these parklands had retained a more natural geochemical signature than the surrounding built-up areas over the same PMs.

3.4.3. Comparison between spatial relationships of Pb-Al with those of Pb-Ti and Ti-Al

Our hypothesis is that if Pb were not affected by human activities, the relationships between Pb and Al should be generally similar to those between Ti and Al. Therefore, we used the spatial patterns of correlation between Ti and Al to "mimic" the natural relationships between Pb and Al assuming Pb concentrations were not affected by human activities. This approach will be helpful for us to establish the links between the spatially varying correlations between Pb and Al and the influencing factors. In addition to the spatial distribution map of local correlation coefficients between Pb and Al (Fig. 6), such a map for Ti and Al was also produced, together with the map for Pb and Ti for further comparison (Fig. 8). As expected, the correlations between Ti and Al were much less spatially variable, with all the local correlation coefficients being high and positive. Such a result conforms the strong anthropogenic influences on Pb, making the correlations between Pb and Al spatially variable. Furthermore, the spatially varying relationships between Pb and Ti demonstrated the same spatial patterns as those between Pb and Al, e.g., Pb still exhibited the negative correlation with Ti in the northeast and north areas and weak correlations in centre of London, and relatively strong positive correlations were still observed in southeast and southwest areas. Such results further conformed the strong anthropogenic influences on Pb.

In order to further explore the spatially varying relationships, scatter plots between Pb and Al, and Ti and Al were produced (Fig. 9). All samples and three groups of soil samples were arbitrarily selected based on the location from positive correlation, weak correlation and negative correlation areas between Pb and Al (Fig. S2). As expected, strong positive correlations between Ti and Al existed in all samples and all the three groups, implying that their relationships were generally spatially "invariable". In contrast, Pb exhibited the generally positive, negative and weak correlations with Al in the positive, negative, and weak correlation groups, respectively. Due to the complicated factors, the positive and negative relationships between Pb and Al on the scatter plots were still quite weak. This is in contrast to very good linear correlations between Ti and Al for all the samples and groups. The overall results confirm that anthropogenic factors had great impact on the concentration of Pb in the highly urbanized central areas, causing varied correlations between Pb and the conservative element Al.

GWR provides a useful way to explore the spatially varying relationships among environmental parameters, showing a promising approach to investigate the complex relationships at the local level which are useful for improved soil management. It needs to be noted that the concept of "spatially varying relationship" is different from scale effect and not caused by scale effect. It is the local variation related to varying influences of factors at the local level: At different locations, the relationship is different. Specifically, in this paper, the generally positive relationship between Pb and Al under natural conditions can be interfered and even changed to the negative relationship which is related to the influences of human factors. Such relationships remain similar with the changes of scales, e.g., the changes of distance band tested in this study. The similar patterns of local coefficients for different distance band were observed in Fig. 5. Therefore, the varying relationship is indeed different from scale effect. The scale effort mainly affects the details of the patterns, not the overall patterns of the result of spatially varying relationship.

It is also necessary to clarify that the edge effect for such a large number of samples in this study is minimal. The high local R^2 values with negative correlation between Pb and Al are located in the northern border area, which are caused by the elevated values of Pb closer to the city centre side. Such a result is obviously affected by human factors. The higher local R^2 exists throughout the study area, not only in the border areas. These results are expected to remain when the study area is expanded further north, outside the Greater London Area. This could be tested in our future studies when a larger study area is considered.

4. Conclusions

The relationships between Pb and Al in urban London soils were spatially varying, which have not been revealed before. The strong negative relationships between Pb and Al were found in the northeast and north areas and weak relationships were located in central areas, associated with the impact of strong anthropogenic activities on Pb concentration. Road traffic, industry activities and construction in centre of London may be linked to the weakened or changed direction of relationship from positive to negative correlations. The positive relationships between Pb and Al were found in two areas of southeast and southwest and a small area in central London, which were associated with large parklands and greenspaces areas with less influences by anthropogenic activities and natural geochemical signature retained. Such new findings are important for a better understanding of the complicated relationships in urban geochemistry, especially with strong human activities which are strongly spatially variable. The new finding of "spatially varying relationship" between Pb and Al in urban soils of London is important to achieve a better understanding of the complicated relationships in urban geochemistry, which is useful to seek the links with the influencing factors. The newly revealed negative correlation between Pb and Al provides new insight to the existing knowledge. It was achieved through the GWR technology, proving that GWR is an effective tool. in identifying the spatially varying relationships of the research objects, thereby revealing the hidden spatial distribution patterns. It should be noted that the GWR method has only revealed the spatially varying patterns of geochemistry elements in the urban area, implying the spatial associations with influences factors. The causal effects of heavy metals contamination still require more detailed investigations.

Acknowledgement

This study is supported by the Royal Society International Exchange Scheme of the UK (IE141447) (2015-2017). The soil data used in this study were provided by the British Geological Survey from the London Earth Project.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jhazmat.2020.122377.

References

Aelion, C.M., Davis, H.T., McDermott, S., Lawson, A.B., 2009. Soil metal concentrations and toxicity: associations with distances to industrial facilities and implications for human health. Sci. Total Environ. 407, 2216–2223.
Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. Selected papers of Hirotugu Akaike, pp. 199–213.
Ali, K., Partridge, M.D., Olfert, M.R., 2007. Can geographically weighted regressions improve regional analysis and policy making? Int. Reg. Sci. Rev. 30 (3), 300–311.
Allen, M.A., Cave, M.R., Chenery, S.R.N., Gowing, C.J.B., Reeder, S., 2011. Sample Preparation and Inorganic Analysis for Urban Geochemical Survey Soil and Sediment Samples. Mapping the Chemical Environment of Urban Areas. Wiley, pp. 28–46.
Alloway, B.J., 1995. Heavy Metals in Soils: Trace Metals and Metalloids in Soils and Their Bioavailability. Academic & Professional, Blackie.
Appleton, J.D., Cave, M.R., 2018. Variation in soil chemistry related to different classes and eras of urbanisation in the London area. Appl. Geochem. 90, 13–24.
Bacha, C.J.C., 2003. The determinants of reforestation in Brazil. Appl. Econ. 35, 631–639.
Ballesta, R.J., Bueno, P.C., Rubí, J.A.M., Giménez, R.G., 2010. Pedo-geochemical baseline

content levels and soil quality reference values of trace elements in soils from the Mediterranean (Castilla La Mancha, Spain). Cent. Eur. J. Geosci. 2, 441–454.
Batisani, N., Yarnal, B., 2009. Urban expansion in Centre County, Pennsylvania: spatial dynamics and landscape transformations. Appl. Geogr. 29, 235–249.
Bickford, S.A., Laffan, S.W., 2006. Multi-extent analysis of the relationship between pteridophyte species richness and climate. Glob. Ecol. Biogeogr. 15, 588–601.
Bowman, A., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71, 353–360.
British Geological Survey, 2011. British Geological Survey Official Website. Available online at:. http://www.bgs.ac.uk/gbase/londonearth.html/.
Cai, L., Huang, L., Zhou, Y., Xu, Z., Peng, X., Yao, L., Peng, P., 2010. Heavy metal concentrations of agricultural soils and vegetables from Dongguan, Guangdong. J. GEOGR. SCI. 20, 121–134.
Carr, R., Zhang, C.S., Moles, N., Harder, M., 2008. Identification and mapping of heavy metal pollution in soils of a sports ground in Galway City, Ireland, using a portable XRF analyser and GIS. Environ. Geochem. Health 30 (1), 45–52.
Castrignanò, A., Giugliarini, L., Risaliti, R., Martinelli, N., 2000. Study of spatial relationships among some soil physico-chemical properties of a field in central Italy using multivariate geostatistics. Geoderma 97, 39–60.
Clark, H.F., Brabander, D.J., Erdil, R.M., 2006. Sources, sinks, and exposure pathways of lead in urban garden soil. J. Environ. Qual. 35, 2066–2074.
Clement, F., Orange, D., Williams, M., Mulley, C., Epprecht, M., 2009. Drivers of afforestation in Northern Vietnam: assessing local variations using geographically weighted regression. Appl. Geogr. 29, 561–576.
De Miguel, E., De Grado, M.J., Llamas, J.F., Martin-Dorado, A., Mazadiego, L.F., 1998. The overlooked contribution of compost application to the trace element load in the urban soil of Madrid (Spain). Sci. Total Environ. 215, 113–122.
De Temmerman, L., Vanongeval, L., Boon, W., Hoenig, M., Geypens, M., 2003. Heavy metal content of arable soils in northern Belgium. Water Air Soil Pollut. 148, 61–76.
Delbecque, N., Verdoodt, A., 2016. Spatial patterns of heavy metal contamination by urbanization. J. Environ. Qual. 45, 9–17.
Deller, S.C., Lledo, V., 2007. Amenities and rural Appalachian growth. Agric. Resour. Econ. Rev. 36, 107–132.
Foody, G.M., 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. Glob. Ecol. Biogeogr. 13, 315–320.
Fotheringham, A., Brunsdon, C., Charlton, M., 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ. Plan. A 30, 1905–1927.
Fotheringham, A.S., Brunsdon, C.A., Charlton, M.E., 2002. Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. John Wiley & Sons, New York.
Franco-Uria, A., Lopez-Mateo, C., Roca, E., Fernandez-Marcos, M.L., 2009. Source identification of heavy metals in pastureland by multivariate analysis in NW Spain. J. Hazard. Mater. 165, 1008–1015.
Gao, J., Li, S., 2011. Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using geographically weighted regression. Appl. Geogr. 31, 292–302.
Geofabrik, 2016. Geofabrik Official Website. Available online at: (Accessed 15 May.18). http://download.geofabrik.de/europe/great-britain/england/greaterlondon.html.
Geri, F., Amici, V., Rocchini, D., 2010. Human activity impact on the heterogeneity of a Mediterranean landscape. Appl. Geogr. 30, 370–379.
Guo, L., Ma, Z., Zhang, L., 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can. J. For. Res. 38, 2526–2534.
Huang, Y.P., Yuan, M., Lu, Y.P., 2017. Spatially varying relationships between surface urban heat islands and driving factors across cities in China. Urban Analytics and City Science. 46 (2), 377–394.
Johnson, C.C., 2005. 2005 G-BASE Field Procedures Manual. British Geological Survey. Internal Report No. IR/05/097, Keyworth, UK.
Johnson, C.C., 2011. Understanding the Quality of Chemical Data From the Urban Environment E Part 1: Quality Control Procedures. Mapping the Chemical Environment of Urban Areas. Wiley, pp. 61–76.
Kabata-Pendias, A., 2004. Soil–plant transfer of trace elements—an environmental issue. Geoderma 122, 143–149.
Kalogirou, S., 2014. A spatially varying relationship between the proportion of foreign citizens and income at local authorities in Greece. Proceedings of the 10th International Congress of the Hellenic Geographical Society 1458–1466 5.
Knights, K., Scheib, C., 2011. Examining the Soil Chemistry of London's Parklands. Applied Geoscience for Decision-making in London and the Thames Basin, London, UK.
Leung, Y., Mei, C., Zhang, W., 2000a. Statistical tests for spatial nonstationary based on the geographically weighted regression model. Environ. Plan. A 32, 9–32.
Lv, J., Liu, Y., Zhang, Z., Dai, J., 2013. Factorial kriging and stepwise regression approach to identify environmental factors influencing spatial multi-scale variability of heavy metals in soils. J. Hazard. Mater. 261, 387–397.
Madrid, L., Díaz-Barrientos, E., Madrid, F., 2002. Distribution of heavy metal contents of urban soils in parks of Seville. Chemosphere. 49, 1301–1308.
Martín, J.A.R., Ramos-Miras, J.J., Boluda, R., Gil, C., 2013. Spatial relations of heavy metals in arable and greenhouse soils of a Mediterranean environment region (Spain). Geoderma 200, 180–188.
Mattuck, R., Nikolaidis, N.P., 1996. Chromium mobility in freshwater wetlands. J. Contam. Hydrol. 23, 213–232.
McMillen, D.P., 1996. One hundred fifty years of land values in Chicago: a nonparametric approach. J. Urban Econ. 40, 100–124.
Mei, C.L., Xu, M., Wang, N., 2016. A bootstrap test for constant coefficients in geographically weighted regression models. Int. J. Geogr. Inf. Sci. 30 (8), 1622–1643.
Miles, J.C.H., Appleton, J.D., 2005. Mapping variation in radon potential both between and within geological units. J. Radiol. Prot. 25, 257–276.
Milke, R., Wiedenbeck, M., Heinrich, W., 2001. Grain boundary diffusion of Si, Mg, and O in enstatite reaction rims: a SIMS study using isotopically doped reactands. Contrib. Mineral. Petrol. 142, 15–26.
Páez, A., Farber, S., Wheeler, D.C., 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environ. Plann. A. 43 (12), 2992–3010.
Pant, P., Harrison, R.M., 2013. Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: a review. Atmos. Environ. 77, 78–97.
Qin, Y.S., 2008. Study on the influences of combined pollution of heavy metals Cu and Pb on soil respiration. Journal of Anhui Agricultural Sciences. 36 (3), 1117.
Qishlaqi, A., Moore, F., Forghani, G., 2009. Characterization of metal pollution in soils under two land use patterns in the Angouran region, NW Iran; a study based on multivariate data analysis. J. Hazard. Mater. 172, 374–384.
Raju, K.V., Somashekar, R.K., Prakash, K.L., 2013. Spatio-temporal variation of heavy metals in Cauvery River basin. Proc. Int. Acad. Ecol. Environ. Sci. 3 (1), 59–75.
Rawlins, B.G., McGrath, S.P., Scheib, A.J., Breward, N., Cave, M., Lister, T.R., Ingham, M., Gowing, C., Carter, S., 2012. The Advanced Soil Geochemical Atlas of England and Wales. British Geological Survey, Keyworth, Nottingham.
Sansalone, J.J., Buchberger, S.G., 1997. Partitioning and first flush of metals in urban roadway storm water. J. Environ. Eng. 123, 134–143.
Sayyed, M.R.G., Sayadi, M.H., 2011. Variations in the heavy metal accumulations within the surface soils from the Chitgar industrial area of Tehran. Proc. Int. Acad. Ecol. Environ. Sci. 1 (1), 36.
Scheib, A., Flight, D., Lister, B., Scheib, C., 2011. London earth: anthropogenic and geological controls on the soil chemistry of the UK's largest city. In: 25th International Applied Geochemistry Symposium. Rovaniemi, Finland. pp. 22–26.
Sezgin, N., Ozcan, H.K., Demir, G., Nemlioglu, S., Bayat, C., 2003. Determination of heavy metal concentrations in street dusts in Istanbul E-5 highway. Environ. Int. 29, 979–985.
Spark, D.L., 2010. Environmental surfaces and interfaces from the nanoscale to the global scale. J. Environ. Qual.
Su, C., 2014. A review on heavy metal contamination in the soil worldwide: situation, impact and remediation techniques. Environmental Skeptics and Critics. 3 (2), 24.
Thornton, I., Davies, D.J.A., Watt, J.M., Quinn, M.J., 1990. Lead exposure in young children from dust and soil in the United Kingdom. Environ. Health Perspect. 89, 55–60.
Thorpe, A., Harrison, R.M., 2008. Sources and properties of non-exhaust particulate matter from road traffic: a review. Sci. Total Environ. 400, 270–282.
Tu, J., 2011. Spatially varying relationships between land use and water quality across an urbanization gradient explored by geographically weighted regression. Appl. Geogr. 31, 376–392.
Tylmann, W., 2004. Heavy metals in recent lake sediments as an indicator of 20th century pollution: case study on lake Jesien. Limnol. Rev. 4, 261–268.
Waller, L., Zhu, L., Gotway, C., Gorman, D., Gruenewald, P., 2007. Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stoch. Environ. Res. Risk Assess. 21, 573–588.
Wang, G., Zeng, C., Zhang, F., Zhang, Y., Scott, C.A., Yan, X., 2017. Traffic-related trace elements in soils along six highway segments on the Tibetan Plateau: influence factors and spatial variation. Sci. Total Environ. 811–821.
Wheeler, D.C., Calder, C.A., 2007. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. J. Geogr. Syst. 9, 145–166.
Wu, S., Zhou, S., Bao, H., Chen, D., Wang, C., Li, B., Tong, G., Yuan, Y., Xu, B., 2019. Improving risk management by using the spatial interaction relationship of heavy metals and PAHs in urban soil. J. Hazard. Mater. 364, 108–116.
Yang, P.G., Mao, R.Z., Shao, H.B., Gao, Y.F., 2009. An investigation on the distribution of eight hazardous heavy metals in the suburban farmland of China. J. Hazard. Mater. 167, 1246–1251.
Yu, D., 2006. Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation. Ann. Region. Sci. 40, 173–190.
Yu, D., Wu, C., 2004. Understanding population segregation from Landsat ETM+ imagery: a geographically weighted regression approach. Glsci. Remote. Sens. 41, 145–164.
Zhang, C.S., 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ. Pollut. 142, 501–511.
Zhang, C.S., Jordan, C., Higgins, A., 2007. Using neighbourhood statistics and GIS to quantify and visualize spatial variation in geochemical variables: an example using Ni concentrations in the topsoils of Northern Ireland. Geoderma 137, 466–476.
Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008a. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146, 78–390.
Zhang, C.S., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. 26, 1239–1248.
Zhang, Z.Y., Abuduwaili, J., Jiang, F.Q., Tud, M., Wang, S.P., 2012. Contents and sources of heavy metals in surface water in the Tianshan Mountain. China Environ. Sci. 32, 1799–1806.
Zheng, Y.M., Yu, K., Wu, H.T., Huang, Z.C., Chen, H., Wu, X., Tian, Q.Z., Fan, K.K., Chen, T.B., 2002. Lead concentrations of soils in Beijing urban parks and their pollution assessment. Geogr. Res. 21, 418–424 (in Chinese).

Chapter 5
Discussion

5.1 Overview of the Research Process

With the development of GIS and spatial technology, an increasing number of researches in environmental geochemistry are carried out based on spatial analysis and geostatistics (Fotheringham and Rogerson, 2013; Hou et al., 2017). There are two major research directions of soil geochemistry, including the assessment of soil nutrition and contamination (e.g., Schwarzenbach et al., 1993; Macalady, 1998; Eganhouse, 1997). The former is mainly to study the spatial distribution and variation of the soil TOC contents, while the latter is more focused on the factors that condition the fate and transport of PTEs in different environment media. In the era of big data, it is necessary to efficiently conduct data mining on identification of spatial patterns of these variables and to extract hidden environmental information based on large-scale datasets. It is on the basis of these datasets established on a regional scale that we can fully understand how the soil quality and contamination vary geographically (Argyraki and Kelepertzis, 2014; Reimann et al., 2014a; Matschullat et al., 2018). In this thesis, the four papers based on the SML have been applied on soil nutrients (i.e., TOC) and PTEs for providing latest understanding between the spatial variation and potential influencing factors at different scales (i.e., Ireland, Northern Ireland and European continent). To achieve these objectives, on the one hand, the hot spot analysis (Getis-Ord G_i^* statistic) was used to investigate the spatial distribution and relationship between TOC contents and pH values in the European agricultural soil (section 4.1). This technique identified a 'special' co-existence of both positive relationship between these two variables in the north-central Europe, which is a novel finding in Europe that differs from the negative relationship in the previous studies. Then, the GWR model was performed to further explore the spatial relationships between TOC and pH across the whole European continent, and successfully revealed the spatial variation and firstly proposed the concept of 'spatially varying relationships' between these two variables from the local perspective (section 4.2). On the other hand, considering the complex geological processes in NI, the hot spot analysis (Getis-Ord G_i^* statistic) and K-means clustering analysis were applied on identification of hidden spatial patterns and the association with different controlling factors for the 15 PTEs (section 4.3). Subsequently, the GWPCC was

used to investigate the spatial correlations between Pb and Al concentrations in the northern half of Ireland, and successfully revealed the spatially varying relationships that associated with atmospheric pollution in the blanket peat (section 4.4). Additionally, a co-authorship study on the spatial relationships between Pb and Al in urban soil of London was also conducted, and successfully identified the anthropogenic influences on the distribution of Pb in the topsoil. Overall, the integration of these five articles provides useful and clear demonstrations for advanced spatial analysis in the big data era of environmental geochemistry, and enriches the latest understanding of distribution patterns and sources for the soil TOC and PTEs in current knowledge.

5.2 Contributions of Research

The overall contributions of these studies demonstrate the power of GIS-based advanced spatial analysis techniques in identifying spatial patterns and the spatially varying relationships of environmental variables, providing practical examples to efficiently extract geochemical knowledge based on large-scale regional data sets in different study areas. In addition, these studies highlighted the local influencing factors from both natural and anthropogenic factors on soil nutrients and pollution, which can be also applied to environmental studies elsewhere. In the absence of prior knowledge, these examples can provide valuable guidance and assistance for soil management and risk assessment in the broader international research community.

In the European agricultural soil, the spatial distribution patterns which revealed by hot spot analysis (Getis-Ord G_i^* statistic) showing a general negative relationship between TOC and pH at the continental level, while a feature of positive relationship was also observed in the north-central Europe. This hidden pattern indicated the existence of 'special' positive correlation between these two variables that do not follow the normal relationship from the local perspective. The novel findings in the north-central Europe against the conclusions from previous studies (e.g., McGrath and Zhang, 2003; Korkanç, 2014), which

provides new insight and information of soil management and agricultural practice for TOC and pH, such as fertilise and liming.

The GWR model successfully proved and proposed the concept of 'spatially varying relationships' between TOC and pH, which provided a solution for dealing the contradictory results of both negative and positive relationships that we found in the previous literature (e.g., de Moraes Sa et al., 2009; Wang et al., 2010; Wang et al., 2016; Luo et al., 2017; Gebrehiwot et al., 2018; Zhang et al., 2018). For example, these contradictory relationships were obtained using global statistics (i.e., PCC, linear regression model), and the divisions of spatial patterns for sub-regions were fairly arbitrary. The application of GWR illustrated an effective way to obtain the objective division of different relationship patterns, and also reveal the clear associations with related influencing factors at the local scale. Based on our results, such a way of spatial thinking can be expanded to other relationships between environmental variables elsewhere.

The combination of two spatial clustering techniques of hot spot analysis and K-means clustering analysis are proved as a useful way for quantitatively evaluating the controlling factors for 15 PTEs. Due to the complex geochemical processes in the soil, it is challenging to identify the fate and sources of PTEs (Manta et al. 2002; Wong et al. 2006; Xia et al. 2011). The hidden patterns that revealed by these two techniques highlighted the spatial overlay association with different geological features from the local perspective, especially peat and basalt. These results not only enhanced the latest understanding of the controlling factors on the selected 15 PTEs for current literature in the topsoil of an area that have been extensively studied (i.e., NI), but also provides a clear demonstration on the efficiency of SML techniques in discovering hidden spatial patterns and extract geochemical association in the multivariate datasets, which can be applied for environmental study in other unexplored areas.

The hypothesis of the spatially varying relationships between Pb and Al was successfully tested by the GWPCC in the topsoil of northern half of Ireland. The investigation of the local correlations provides a novel and effective way to identify the influencing factors for PTE pollution (Yuan et al., 2020), especially for Pb, as one of the most widely concerned trace element (Nriagu, 1983; Nriagu, 1996). The special negative correlation showed a clear overlay association with blanket peat, highlighting the regional influence of atmospheric pollution on the elevated Pb concentration from anthropogenic factors, such as mining and traffic emission. Moreover, anthropogenic activities weakened the relationships in the eastern coastal areas. Our results provide an effective way to explore the spatially varying relationships and influencing factors of Pb, providing new understanding of controlling factors for PTEs by atmospheric deposition in the topsoil from the spatial perspective, which can be also applied on studying other PTEs elsewhere.

5.3 Advancement

In the field of soil fertility and nutrition, there are three main advancements in this thesis. First of all, the spatial distribution and relationships between TOC contents and pH values was initially revealed in the European agricultural soil. Secondly, a special positive relationship between TOC and pH that associated with coarse-textured glacial sediments (quartz) was identified in central-eastern Europe, which is not usual in other study areas. Thirdly, the topic of ‘spatially varying relationships’ was proposed and proved to provide a new understanding between these two variables which against the contradictory relationships in the previous literature. These findings highlighted the dominant influences of quartz on the not only nutrients (e.g., TOC, S, P), but also the major and trace elements (e.g., Cd, Pb, Zn) in the European agricultural soil, which provided valuable information for soil management and agricultural practice at the continental level.

In the field of soil contamination, there are three main advancements in this thesis. Firstly, the geological controls on the selected 15 PTEs were quantitatively analysed from the

spatial perspective in NI, and the dominant factors including peat and basalt were highlighted by the clustering patterns. Secondly, the hypothesis of spatially varying relationships between Pb and Al was successfully tested, with special negative correlations were found in western and north-eastern Ireland. Thirdly, atmospheric pollution in blanket peat and anthropogenic factors contributed to the negative correlations in Ireland was identified. These findings enhanced the current knowledge of soil contamination from both geogenic sources and anthropogenic inputs against the existing literature and can be applied into environmental studies elsewhere.

5.4 Research limitations in this thesis

It needs to be acknowledged that the research in this dissertation has some unavoidable limitations. The key point is that like correlation analysis, these advanced spatial analysis techniques cannot be used for investigation of causal effects. Instead, the hidden associations can be identified through those spatial patterns, supporting our new findings in the topsoil of Europe and the island of Ireland. In addition, the abrupt geological boundaries may cause discontinuous patterns or difficulty in choosing the scales (e.g., bandwidths) for these advanced spatial analysis techniques, and barriers could be introduced if there are strong evidences of abrupt changes.

Another unavoidable point in geochemical research is the analytical uncertainty, which includes the uncertainty caused by soil sampling and the use of interpolation methods (i.e., IDW). These uncertainties are not specifically measured or evaluated in this dissertation, which may be a direction in future geochemical research.

Reference

Argyraki, A., Kelepertzis, E., 2014. Urban soil geochemistry in Athens, Greece: the importance of local geology in controlling the distribution of potentially harmful trace elements. Sci. Total Environ., 482–483, 366-377, 10.1016/j.scitotenv.2014.02.133.

Coggins, A.M., Jennings, S.G., Ebinghaus, R., 2006. Accumulation rates of the heavy metals lead, mercury and cadmium in ombrotrophic peatlands in the west of Ireland. Atmos. Environ., 40 (2), 260-278.

de Moraes Sa, J.C., Cerri, C.C., Lal, R., Dick, W.A., de Cassia Piccolo, M., Feigl, B.E., 2009. Soil organic carbon and fertility interactions affected by a tillage chronosequence in a Brazilian Oxisol. Soil Till. Res., 104 (1), 56-64.

De Vleeschouwer, F., Gérard, L., Goormaghtigh, C., Mattielli, N., Le Roux, G., Fagel, N., 2007. Atmospheric lead and heavy metal pollution records from a Belgian peat bog spanning the last two Millenia: Human impact on a regional to global scale. Sci. Total Environ., 377, 282-295.

Eganhouse, R.P., 1997. Molecular markers in environmental geochemistry, ACS Symposium Series, American Chemical Society, Washington, DC.

Fotheringham, S., Rogerson, P. (Eds.)., 2013. Spatial analysis and GIS. CRC Press.

Gebrehiwot, K., Desalegn, T., Woldu, Z., Demissew, S., Teferi, E., 2018. Soil organic carbon stock in Abune Yosef afroalpine and sub-afroalpine vegetation, northern Ethiopia. Ecol. Process., 7 (1), p. 6.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environ. Pollut., 231, 1188-1200.

Korkanç, S.Y., 2014. Effects of afforestation on soil organic carbon and other soil properties. Catena, 123, 62-69.

Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Chang. Biol., 23 (10), 4430-4439.

Macalady, D.L., (ed.), 1998. Perspectives in Environmental Chemistry. Oxford University Press, New York, pp. 138-166.

Manta, D. S., Angelone, M., Bellanca, A., Neri, R., & Sprovieri, M. (2002). Heavy metals in urban soils: a case study from the city of Palermo (Sicily), Italy. Science of the Total Environment, 300, 229–243.

Matschullat, J., Reimann, C., Birke, M., dos Santos Carvalho, D., Albanese, S., Anderson, M., ... & Zomeni, Z., 2018. GEMAS: CNS concentrations and C/N ratios in European agricultural soil. Science of the Total Environment, 627, 975-984.

McGrath, D., Zhang, C.S., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem. 18, 1629-1639.

Nriagu, J.O., 1983. Lead and Lead Poisoning in Antiquity Wiley, New York.

Nriagu, J.O., 1996. A history of global metal pollution. Science, 272, 223-224.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., O' Connor, P., 2014a. Chemistry of Europe's Agricultural Soils, Part A: Methodology and Interpretation of the GEMAS Data Set. Geologisches Jahrbuch (Reihe B102), Schweizerbarth, Hannover, 523 pp.

Schwarzenbach, R.P., Gschwent, P.M., Imboden, D.M., 1993. Environmental Organic Chemistry. Wiley, New York.

Wang, T., Kang, F., Cheng, X., Han, H., Ji, W., 2016. Soil organic carbon and total nitrogen stocks under different land uses in a hilly ecological restoration area of North China. Soil Tillage Res., 163, 176-184.

Wang, Z.M., Zhang, B., Song, K.S., Liu, D.W., Ren, C.Y., 2010. Spatial variability of soil organic carbon under maize monoculture in the Song-Nen Plain, Northeast China. Pedosphere, 20 (1), 80-89.

Wong, C. S., Li, X., & Thornton, I. (2006). Urban environmental geochemistry of trace metals. Environmental Pollution, 142, 1–16.

Xia, X., Chen, X., Liu, R., & Liu, H. (2011). Heavy metals in urban soils with various types of land use in Beijing, China. J. Hazard. Mater., 186, 2043–2050.

Yuan, Y.M., Cave, M., Xu, H.F., Zhang, C.S., 2020. Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). J. Hazard. Mater., 393, 122377.

Zhang, X., Liu, M., Zhao, X., Li, Y., Zhao, W., Li, A., Cheng, S. Cheng, S., Han, X., Huang, J., 2018. Topography and grazing effects on storage of soil organic carbon and nitrogen in the northern China grasslands. Ecol. Indic., 93, 45-53.

Chapter 6
Conclusion

6.1 Overview

This Ph.D. thesis summarised the existing problems and possible solutions in environmental geochemistry from the spatial perspective, and provided demonstration of applying four SML techniques on identification of hidden spatial patterns and geochemical association for environmental variables in regional large-scale datasets. Specifically, the spatial distribution of TOC and the special pattern of positive relationship were identified by hot spot analysis in European agricultural soil. The spatially varying relationships between TOC and pH were investigated by GWR at the European continent level. The spatial clustering patterns of 15 PTEs and their controlling factors were quantified by the combination of hot spot analysis and K-means clustering analysis in NI. The spatially varying relationships and special negative correlations were explored by GWPCC in the northern half of Ireland.

6.2 Main conclusions

At present, studies on the soil nutrients and contamination are still regarded as important issues of environmental geochemistry and health. In the era of big data, the results of this thesis have proved that the GIS-based spatial analysis and SML technologies are effective and useful tools to extract geochemical information from the large-scale datasets. The special or interesting patterns of environmental variables can be associated with specific influencing factors, which provides valuable information to stakeholders for soil management and monitoring form the spatial perspective.

6.2.1 Identification of the co-existence of low total organic carbon contents and low pH values in agricultural soil in north-central Europe using hot spot analysis based on GEMAS project data

(1) Spatial distribution patterns of TOC contents and pH values were revealed by hot spot analysis in European agricultural soil.

(2) Soil TOC contents and pH values were negatively correlated at the European scale.

(3) A co-existence of special positive correlation was identified in the north-central Europe.

(4) Both low TOC contents and low pH values were related to coarse-textured glacial sediments.

(5) The special patterns provided important information for agricultural soil management.

6.2.2 Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression

(1) Spatially varying relationships between TOC contents and pH values were revealed by GWR.

(2) The positive correlations between TOC and pH clustered in central-eastern Europe.

(3) Negative and mixed correlations were observed in northern and southern Europe, respectively.

(4) The quartz-rich soil is the main contributing factor to the positive relationship in central-eastern Europe.

(5) Climate and anthropogenic factors weakened the general negative relationship at the continental level.

6.2.3 Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis

(1) The spatial clustering patterns for PTEs in the topsoil were identified by Hot spot analysis.

(2) The hidden patterns of soil samples were revealed by K-means clustering analysis.

(3) The consistent spatial patterns were observed between PTEs and soil samples.

(4) Peat was associated with high concentrations of Bi, Pb, Sb and Sn.

(5) Basalt was associated with high concentrations of Cr, Co, Cu, Mn, Ni, V and Zn.

(6) As, Ba, Mn and U were associated with other lithologies such as granite, schists and greywacke shale.

6.2.4 Exploration of the spatially varying relationships between lead and aluminium concentrations in the topsoil of northern half of Ireland using Geographically Weighted Pearson Correlation Coefficient

(1) Spatially varying relationships between Pb and Al concentrations in soils were revealed by GWPCC.

(2) Special negative correlations occurred in north-western and north-eastern areas.

(3) Original positive correlations clustered in central-western and midland areas.

(4) Atmospheric pollution contributed to negative correlations overlaid on blanket peat.

(5) Anthropogenic factors weakened the relationships in eastern coastal regions.

6.2.5 Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR)

(1) The relationships between Pb and Al concentrations in urban soils of London was spatially varying.

(2) The positive and negative relationships were found in southern and northern areas, respectively.

(3) Anthropogenic factors weakened the original positive relationships between Pb and Al in the central London, while large parks and greenspaces reserved the positive relationships.

(4) GWR is effective in revealing spatially varying relationships in urban soils.

6.3 Recommendations

The results of this thesis would provide several suggestions for carrying out soil management and environmental monitoring, as well as sampling design for future geochemical studies at regional to continental scales.

(1) The agricultural practices such as fertilising and liming in the central-eastern Europe (i.e., Poland, Germany, Ukraine) should be carefully conducted, as the large particle size in these quartz-rich soil could cause excessive human inputs and negative effects on the ecological environment.
(2) Northern Ireland Authority should pay attention to the agricultural activities in the rural soils which covered with basalt and peat, as the relatively high level of PTE concentrations are observed in the topsoil of these areas, which may be accumulated through food chain to humans.
(3) Many leading gardeners still use peat-based compost to some extent, while the blanket peat in the western and north-eastern of Ireland maintains elevated levels of PTEs due to atmospheric pollution (although not as high as in urban areas), and thus alternatives are recommended. In addition, reducing the use of peat in gardening industry contributes to carbon sequestration in order to reduce the CO_2 concentration in the atmosphere.
(4) For geochemical surveys at regional and continental level, the sampling design should consider different geological units and the prior knowledge on the interesting geological feature (e.g., peat and basalt in Ireland). More dense sampling locations could be planned in these soils than the other areas, which is able to discover interesting and novel findings from local to regional or even continental level, maximizing the available budget.

6.4 Future research

Overall, based on the research results obtained in this thesis, several recommendations for future research directions are made as the following:

6.4.1 Predicting and mapping of soil organic carbon contents using machine learning algorithms in the topsoil of Ireland

The research on the distribution and prediction of the concentration of SOC in the topsoil of Ireland were based on previous data sets (McGrath and Zhang, 2003; Zhang and McGrath, 2004; Zhang et al., 2011), and thus the current knowledge of soil properties and quality cannot be updated well. Mapping and prediction of SOC content based on the newly released Tellus database through SML technology can capture its temporal and spatial variation at the regional level, which is able to provide the latest understanding for soil management and agricultural activities.

6.4.2 Investigation of spatially varying relationships between soil total organic carbon content and climate factors in European agricultural soil

Based on our results, climate is the secondary natural factor that affects the variation of TOC content at the continental scale. Huang et al. (2018) used wavelet analysis to identify the latitude-related changes in SOC on a global scale. Considering there is a strong correlation between latitude and climate, therefore, exploring the spatial relationships between TOC and climatic factors can provide better understanding on the changing laws and trends of TOC under natural conditions.

6.4.3 Discovering hidden spatial patterns of selected potentially toxic elements using hot spot analysis and K-means clustering analysis in stream sediments in Northern Ireland

The potential of identification hidden spatial patterns by these two techniques have been proved based on the topsoil data set (Xu et al., 2021). However, stream sediments may be favoured for early-stage mineral exploration over soil sampling, or even done in conjunction with soils to better develop an understanding of the geology. The investigation of the patterns and

associations between PTEs and stream sediment samples can enhance the current knowledge of geological processes in NI.

6.4.4 Exploring spatially varying relationships between Cd, Ni, Zn and Al in the topsoil of Ireland

In addition to Pb, the element Cd, Ni and Zn have also been proposed to be enriched in the blanket peat bogs (Krachler et al., 2003; Rausch et al., 2005). Exploring the patterns of the spatially varying relationships is able to discover novel correlation between these PTEs and reference element, and thus highlight the soil contamination from both anthropogenic and natural sources at the local scale.

Reference

Huang, J., Minasny, B., McBratney, A.B., Padarian, J., Triantafilis, J., 2018. The location- and scale- specific correlation between temperature and soil carbon sequestration across the globe. Sci. Total Environ., 615, 540-548.

Krachler, M., Mohl, C., Emons, H., Shotyk, W., 2003. Two thousand years of atmospheric rare earth element (REE) deposition as revealed by an ombrotrophic peat bog profile, Jura Mountains, Switzerland. J. Environ. Monit., 5 (1), 111-121.

McGrath, D., Zhang, C.S., 2003. Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl. Geochem. 18, 1629-1639.

Rausch, N., Nieminen, T., Ukonmaanaho, L., Le Roux, G., Krachler, M., Cheburkin, A.K., Bonani, G., Shotyk, W., 2005. Comparison of atmospheric deposition of copper, nickel, cobalt, zinc, and cadmium recorded by Finnish peat cores with monitoring data and emission records. Environ. Sci. Technol., 39 (16): 5989-5998.

Xu, H.F., Croot, P., Zhang, C.S., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environ. Int. 151, 106456.

Zhang, C.S., McGrath, D., 2004. Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of southeastern Ireland from two different periods. Geoderma 119, 261-275.

Zhang, C.S., Tang, Y., Xu, X., Kiely, G., 2011. Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Appl. Geochem. 26, 1239-1248.

Appendix A
R program code

A.1 Investigation of spatially varying relationship between TOC and pH values in European agricultural soil

Load package *lctools*

- library(lctools)

Read GEMAS data

- data<- read_excel("F:/NUI Galway/GEMAs Data/GEMAS.xls")

Set coordinates of GWR model

- Coords<-cbind(data$XCOO, data$YCOO)

Complete GWR model using n = 125 bandwidth

- GWR_125<-gwr(TOC ~ pH, data, 125, kernel = 'adaptive', Coords)

Complete GWPCC and significance test using n=125 bandwidth

- GWPCC_125<-lcorrel(data[26:25], 0.06, Coords)

A.2 Clustering 15 topsoil PTEs in Northern Ireland by K-means clustering algorithm

Clustering algorithms and visualisation

- library(cluster)
- library(factoextra)

read Tellus data

- df<- read_excel("F:/NUI Galway/project3/Tellus_NI.xls")

Data scaling and clr-transformation

- library(compositions)
- scale_df<-scale(df)
- clr_df<clr(scllae_df)

Find optimal cluster number using silhouette function

- cluster_number(df, kmeans, method = "silhouette")

Conduct K-means using number of three clusters

- k3 <- kmeans(df, centers = 3, nstart = 25)

Comparison of K-means results using number of two and four clusters

- k2 <- kmeans(df, centers = 2, nstart = 25)
- k4 <- kmeans(df, centers = 4, nstart = 25)

A.3 Investigation of spatially varying relationship between Pb and Al concentrations in the topsoil Ireland

Load package *lctools*

- library(lctools)

Read Tellus data

- data<- read_excel("F:/NUI Galway/project4/Tellus.xls")

Complete GWPCC between Pb and Al and significance test using n=200 bandwidth (change % of bandwidth when using others)

GWPCC_Pb-Al_200<-lcorrel(data[5:4], 0.012, cbind(data$ Easting_ING, data$ Northing_ING))

Complete GWPCC between Pb and Ti and significance test using n=200 bandwidth

GWPCC_Pb-Ti_200<-lcorrel(data[5:6], 0.012, cbind(data$ Easting_ING, data$ Northing_ING))

Complete GWPCC between Al and Ti and significance test using n=200 bandwidth

GWPCC_Al-Ti_200<-lcorrel(data[4:6], 0.012, cbind(data$ Easting_ING, data$ Northing_ING))